

Master Thesis

January 2023

A Methodology to Build Interpretable Machine Learning Models in Organizations

Shruthi Sajid

Master of Science Business Information Technology

Committee:

Dr. F.A. Bukhsh

Faculty of Electrical Engineering, Mathematics Computer Science (EEMCS),
University of Twente

Dr. A. Abhishta

Faculty of Behavioral, Management and Social Sciences (BMS),
University of Twente

Drs. Bart Zeeman

Analytics in Strategy, CTO Office
NXP Semiconductors, Eindhoven



**UNIVERSITY
OF TWENTE.**

Acknowledgements

'Part of the journey is the end.'

Early 2021, I made the tough choice of leaving a massive piece of my heart in India and moving to the Netherlands. Like every other journey we take on, this too has been tumultuous with its many highs and lows. While I have conditioned myself to seize every moment that comes my way, the last two years have tested me in ways I can't begin to imagine, both professionally and personally. Safe to say, I have been able to unpack, unlearn and grow as a person despite the many roadblocks throughout this journey. In retrospect, I am more than grateful for the moments that have challenged my capacity to be resilient and continue to persevere. The thesis marks the final leg of the last two years in the Netherlands to obtain my Master's degree in Business Information Technology with a focus in Data Science and Business at the University of Twente.

I would like to start by first expressing my sincere gratitude to my professors Faiza and Abhishta for their able guidance and confidence in my capabilities throughout my master's thesis. Faiza, I have now worked with you for over a year starting from my internship before the thesis and I must say, you've always been a source of support, motivated me every time we got on a progress call, and pushed me to give my best. Abhishta, the same goes to you, for stepping in as my secondary supervisor right in time, offering constructive feedback whenever required, recognizing my potential in research, and being supportive through it all. Thank you both, once again.

I would be failing in my duties if I don't thank my amazing team at NXP starting with my supervisor, Bart, for trusting me with a portion of one of the most sought-after projects in the organization. You've been incredibly supportive of my work, and never stopped to bounce off ideas together or offer feedback wherever required. I would also like to thank the rest of my team - Cornee, Ben, and Souad who have been absolutely nothing short of a delight to work with. Ben and Souad, what we share is more of a friendship and I'm constantly in awe of how you both carry this project forward. Thank you for letting me pick your brains with questions and doubts all the time. A special mention to some other folks at NXP I've known for a longer time because of my internship before the thesis - Eric, Gabi, and Judith. I can't think of you guys as colleagues anymore, but rather good friends I've made along the way, and I hope we continue to catch up for fun times over a cup of coffee, just as always. To the rest of the folks I've crossed paths with at NXP, I wish you all the best of luck in future endeavors.

I would also like to thank my friends at the University and flatmates in Nijmegen for being supportive throughout my thesis. And last but not the least, I'd like to thank my family and friends back home and the biggest cheerleaders - my parents and siblings. It is true when people say distance means so little when someone means so much to you. I don't think I've gotten past a single day without feeling thankful for everything you have done for me. Despite the distance, you have always stood by me like a rock through the good, and bad and cheered me up on the days I questioned my capabilities. Thank you, once again. Home is indeed where the heart is.

Shruthi Sajid
Nijmegen, January 16, 2023

Contents

	Page
Executive Summary	5
List of Figures	7
List of Tables	8
1 Introduction	9
1.1 Background Information	9
1.2 Problem Context	10
1.3 Company Information	10
1.4 Scope of the Thesis	11
1.5 Research Design	12
1.5.1 Research Objective	12
1.5.2 Research Questions	12
1.5.3 Research Methodology	13
1.6 Thesis Outline	14
2 Literature Review	16
2.1 Systematic Literature Review	16
2.1.1 Search Strategy	16
2.1.2 Inclusion and Exclusion Criteria	17
2.1.3 Data Extraction Strategy	18
2.2 Analytical Methods for Decision Making	19
2.2.1 Descriptive Analytics	19
2.2.2 Diagnostic Analytics	19
2.2.3 Predictive Analytics	20
2.2.4 Prescriptive Analytics	20
2.3 Machine Learning in Business Context	20
2.3.1 Challenges in ML-based decision making for businesses	22
2.4 Explainable AI/ML	22
2.4.1 Need for Explainability & Interpretability	22
2.4.2 Stakeholders and Target Audience	25
2.4.3 Black-box and White-box ML Models	26
2.4.4 Accuracy-Interpretability Trade-off	27
2.4.5 Dimensions of Explainability	28
2.5 Research Gaps	29
2.6 Supporting Background Information	30
2.6.1 Machine Learning	30
2.6.2 Supervised Learning	32
2.6.3 Unsupervised Learning	32
2.6.4 Applications of Machine Learning	33
2.7 Summary	35

3	Data Collection	36
3.1	Need for Qualitative Method	36
3.1.1	Qualitative Research Goal	36
3.2	Qualitative Interviews	37
3.2.1	Interview Design	38
3.2.2	Stakeholder Analysis	39
3.2.3	Synthesis of Results	40
3.3	Summary	43
4	Design and Development	44
4.1	Methodology Design	44
4.1.1	Existing Methodology	44
4.1.2	Observations	45
4.2	Inclusive & Interpretable Cross-Industry Standard Process for ML in Business (iCRISP-ML)	46
4.2.1	Objectives and Goals	47
4.3	Phases of the Methodology	48
4.3.1	Business Understanding	48
4.3.2	Data Understanding	52
4.3.3	Data Preparation	54
4.3.4	Modeling and Evaluation	55
4.3.5	Present to Stakeholders	59
4.3.6	Reporting and Usage	61
4.4	Overview of the Methodology	61
4.5	Summary	62
5	Case Study Evaluation	63
5.1	Case Study Description	63
5.1.1	Problem Context	64
5.1.2	Team Structure and Overview	65
5.1.3	Product Hierarchy within NXP	65
5.1.4	Work Environment - Tools and Technologies	66
5.2	Case Study Execution	67
5.2.1	Business Understanding	67
5.2.2	Data Understanding	70
5.2.3	Data Preparation	73
5.2.4	Model Building and Evaluation	78
5.2.5	Present to Stakeholders	84
5.3	Outcome of Case Study	86
5.3.1	Limitations	87
5.4	Summary	87
6	Expert Interview and Validation	88
6.1	Overview of Validation Process	88
6.1.1	Participant Selection	88
6.1.2	Methodology Presentation	89
6.1.3	Interview Format	90

6.2	Get acquainted with Participants	90
6.3	Design and Development	92
6.3.1	Adequacy of the methodology design	92
6.3.2	Comparison with existing methodology	93
6.3.3	Critical interventions	93
6.4	System Usability Scale (SUS)	95
6.4.1	Results of SUS	96
6.4.2	Summary of SUS Findings	100
6.5	Synthesis of Results	101
6.6	Summary	103
7	Conclusion and Recommendations	104
7.1	Key Takeaways and Contributions	104
7.1.1	Contributions to Science	104
7.1.2	Contributions to Practice	105
7.2	Research Questions	105
7.3	Recommendations	108
7.4	Threats to Validity	109
7.5	Limitations and Future Work	110
	Bibliography	111
	Appendices	116
A	List of Acronyms	116
B	First-level Stakeholder Interview Questions	117
C	Expert Interview: Questions	118
C.1	Agenda	118
C.2	Getting acquainted	118
C.3	Design and Development	118
C.4	Synthesis of Results	119
C.5	Outro	119
D	System Usability Scale - Discussion	120

Executive Summary

Data in an organization is table stakes. But, knowing how to manage and effectively optimize the data to make better business decisions is what makes an organization set itself apart from the competition. One of the most compelling reasons for an organization to manage its data effectively and efficiently is to make better and more informed business decisions. Data-driven insights provide organizations and stakeholders with the required capabilities to generate real-time insights and predictions to optimize and drive better performance and decisions. These insights not only uncover opportunities to expand the organization's growth in the industry, but also provides visibility into internal processes, core activities, and critical business operations within the organization to make more concrete decisions. However, data in an organization is far-reaching because it ranges from simple visualizations driven by analytics to in-depth insights into organizational performance. Moreover, with advances in technology and the many capabilities, there is a paradigm shift towards leveraging advanced analytics and Machine Learning (ML) related suite of techniques to make accurate and informative decisions. Despite the advantages and predictive insights that come with advanced analytic-based decision-making, there is also a certain degree of accountability that needs to be maintained i.e. ensuring the decisions driven by analytics are interpretable to all stakeholders who are within the realm of decision-making. Among these stakeholders, those that are inclined towards business, more often than not, lack the technical knowledge to grasp the logic underpinning ML and analytical models.

The probable solution may not necessarily be to simply find better ways to convey how an analytical system works; rather, it's about creating a transparent, standardized, and interpretable suite of processes that can assist even a data-related stakeholder to build and understand the outcome at ease and then convey it to others. To effectively address these concerns, the objective of the master thesis is to design and develop a methodology that can support the capabilities of ML from an organizational standpoint.

The study first provides a comprehensive systematic literature review to understand the state-of-the-art and best practices to introduce interpretability for a business context. The results of this phase include an overview of the analytical methods for decision-making in organizations, the need for interpretability in ML models, relevant stakeholders and target audiences, and the existence of black-box and white-box ML models. The results of the literature review are then complemented with a qualitative data collection process using interviews with relevant stakeholders. Four stakeholders volunteered to participate in the stakeholder analysis. The results are consolidated to establish the need to design a standardized methodology to achieve the research objective. The qualitative data analysis is followed by the design and development of a business-friendly methodology that bridges knowledge gaps to build better and interpretable ML models in a business setting. The methodology spans six phases such as i.) Business Understanding ii.) Data Understanding iii.) Data Preparation iv.) Model Building and Evaluation v.) Present to Stakeholders and vi.) Reporting and Usage. Further, the practicality of the proposed methodology is also demonstrated by the researcher in a real-world case study. The researcher has also validated the perceived usability with a series of experts within the domain to understand if the methodology would prove to be effective and efficient in forthcoming use cases.

The key takeaway of the findings presented in this study is the introduction of interpretability as a core concept throughout the lifecycle of a project. The methodology aims to be inclusive and interpretable for all stakeholders to be able to easily understand the outcomes of the ML model in a way that

business decisions can be made effectively. The study also presents contributions to science and practice and future work that can be included to improve the current limitations of the methodology.

List of Figures

1	Design Science Research Methodology	14
2	Outline of the Thesis Structure	15
3	Study Selection Process	17
4	Analytical Methods in Business [1]	20
5	Decision-making using ML [2]	21
6	Explainability and Interpretability - A Representation [3]	23
7	Conceptual Model of FAT	24
8	Objectives of Stakeholders [4]	26
9	Black-box vs White-box ML models	27
10	Accuracy-Interpretability Trade-off [5]	28
11	Evolution of Machine Learning	30
12	General structure of an ML model	31
13	Supervised Learning	32
14	Unsupervised Learning	33
15	Stakeholders interacting with the Methodology(SUD)	40
16	Conceptual Model of CRISP DM [6]	45
17	Proposed Methodology: iCRISP-ML	47
18	Impact-Effort Matrix Template	52
19	Steps in Stage 2: Data Understanding	53
20	Test for Interpretability	57
21	Interpretability Checker	59
22	An Overview of iCRISP-ML and its goals	62
23	Overview of the plan for the project <i>[internal]</i>	64
24	Classification of Products within NXP <i>[internal]</i>	65
25	Product Hierarchy <i>[internal]</i>	66
26	Work Environment <i>[internal]</i>	67
27	End-market to Market Segment	71
28	Newly proposed mapping between MAG code and Market segment	72
29	First-level feature ranking using Random Forest Regressor <i>[anonymized]</i>	74
30	Second-level feature ranking using Random Forest Regressor <i>[anonymized]</i>	75
31	Cumulative Variance explained by PCA	77
32	Neural Field for Product Lifecycle <i>[Internal]</i>	78
33	Interpretability Checker for the Case Study	79
34	Test for Interpretability	80
35	Feature Ranking using Random Forest Regressor <i>[anonymized]</i>	81
36	Summary Plot using ShAP for LGBM Model <i>[anonymized]</i>	83
37	Permutation Feature Importance using Random Forest <i>[anonymized]</i>	83
38	Permutation Feature Importance using LGBM <i>[anonymized]</i>	84
39	ML Model Results - Comparison of Model with and without Features	86
40	Overview of Validation Process	88
41	Odd and Even Numbered Statements - Average Scores	97
42	No. of participants vs the overall grade	100
43	Miro Board - Representation of Findings	102

List of Tables

1	Search Queries mapped to Research Questions	17
2	Supervised vs Unsupervised Learning: A comparison	33
3	Summary of Data Collection Methods	37
4	List of Stakeholders	39
5	Observations of the Researcher from a Stakeholder Standpoint	41
6	Questions to identify stakeholders	49
7	RACI Matrix to assign stakeholder responsibilities	51
8	Stakeholder Roles using RACI Matrix	69
9	Benefits and Measure	70
10	Data-related considerations	71
11	Questions for stakeholders based on metrics	85
12	Participant Profile	89
13	Overview gggestions from experts for improving the methodology	94
14	Intention and Points associated with SUS	95
15	Percentile and Grade associated with the SUS score	97
16	Participants and SUS Results	98

1 Introduction

This chapter introduces information on the background of the thesis to understand the problem context and motivation of the study. The chapter also includes a section with information on the company that the researcher has collaborated with for this thesis. It presents the objectives and relevant research questions that have been formulated to guide the research and the research methodology. The remainder of the chapter discusses the outline of the thesis, the mapping of research questions to the relevant chapters, and in turn the research methodology itself.

1.1 Background Information

It is an inherent nature in humans to seek explanation and to an extent, some degree of validation, when concepts are unclear at a first impression. More often than not, explanations are sought for many reasons, such as predicting future events, diagnosing problems, resolving cognitive dissonance, assigning blame, and rationalizing one's action. The same applies to interactions with systems and computing technologies. An effective user experience is shaped by how well they understand how the system operates and their experience while interacting with the system [7]. This is why it is of prime importance for users on the other side of the system to understand how the system works - often referred to as the user mental model. It serves as the foundation for users to anticipate system behaviors, interact effectively and enable ease of use.

Over the years, there have been massive breakthroughs in data processing and advances in data-driven decisions. Organizations and business leaders across various sectors are moving towards leveraging advanced techniques and tools at every step of their operations to optimize business value, improve performance and drive efficiency. In a simple real-world scenario, when tasked with an action item that's waiting on approval, these data-driven approaches based on ML algorithms can handhold stakeholders and decision-makers to base their decisions on historical data, patterns, or evidence. While this not only reduces the practice of guesstimate, it also increases operational efficiency and accelerates informed decision-making within organizations. With such technologies in place, organizations can pull all the stops to extract value, deliver business insights, automate tasks, and make data-informed business decisions. While the umbrella term Artificial Intelligence (AI) can include Expert Systems, Machine Vision, Natural Language Processing (NLP), and Speech Recognition among others, Machine Learning (ML) is indeed the core of it. Owing to the increasing pervasiveness and widespread adoption of ML and its applications, organizations are leaning towards adopting automated decision-making in various facets of their business. From automated cars to product recommendations and making investment decisions to automating employee recruitment, ML has found its way into making data-informed decisions for organizations across different sectors. ML algorithms have proven to be advantageous in areas such as medical diagnosis, financial risk prediction, safety evaluation of robots, aircraft manufacturing, and so on. These algorithms are designed to acquire knowledge from real-time historic data and generate actionable insights to facilitate decision-making. The insights gained from the algorithms not only help with automating business operations but also help identify key drivers for growth and ways to increase business revenue.

As organizations are actively working towards building modernized and optimal processes, ML supports them by handling more complex processes and learning over time. McKinsey [8] conducted a global survey that shows only about 15 percent of respondents have successfully scaled automation across multiple parts of the business and close to 36 percent of respondents said that ML algorithms

had been deployed beyond the pilot stage. The reasons could be many - lack of transparency in ML algorithms, decisions not being easily distilled into digestible formats, sources of information being considered as too high-level or technical, and so on. The caveat to be highlighted is that successful automation simply doesn't end with predictions and having an exhaustive set of insights. The key to driving business efficiency is understanding the insights, trusting the predictions and making informed decisions as next steps.

Accordingly, this thesis dives deep into understanding how organizations can make use of machine learning algorithms in the most optimal way while aligning with the overall business strategy. The remainder of the thesis aims to address the gaps that ML algorithms have brought about in a business setting with respect to a lack of interpretability in model outcomes.

1.2 Problem Context

With rapid advances in analytics and cognitive capabilities in IT, algorithmic-based decision-making plays a critical role in influencing business decisions and deciding the granularity of information people are exposed to. While it is indeed a revolution in the way data is processed and made use of in organizations, it also dramatically increases the complexity of breaking down and understanding the insights and outcomes of ML model predictions. Despite the many benefits that ML algorithms can bring about, stakeholders, especially those that belong to the business and managerial cadre, typically have zero to little insight concerning the knowledge of how these systems make any decisions or predictions due to their lack of transparency or in other words - the "black-box" effect. One possible reason can be that the training data the model learns from could be skewed and riddled with errors, introducing components of bias and unfairness in results. Since these black-boxes suffer from a certain degree of opacity while explaining themselves, it often raises the question of how to 'trust' the predictions that they make. This adds to the growing concern about handing-over the task of critical decision-making to algorithms that suffer from interpretability and explainability. An example can be seen in the field of healthcare where the adoption of an AI-based medical concerning [9] hasn't gained traction among healthcare professionals due to the lack of interpretability and explainability in model predictions. It could also attribute to a lack of trust in how the model makes certain predictions.

Developing a model outcome for business stakeholders to interpret is not intrinsic to the model but lies in the perception and reception of the person receiving the explanation. What makes an explanation provide appropriate information that can be understood and utilized by the business is contingent on the receiver's current knowledge and their goal for receiving the explanation, among other human factors. Therefore, the problem lies in the lack of human-centered approaches that support the technical capabilities of building interpretable models as well as the human experience for the business context.

1.3 Company Information

This research has been conducted in collaboration with NXP Semiconductors. Built on more than 60 years of combined experience and expertise, NXP Semiconductors enables secure communication for a smarter world, advancing solutions to make lives easier, better, and safe. As the world leader in secure connectivity solutions for embedded applications, NXP drives innovation in the Automotive, Industrial and IoT, Mobile, and Communication Infrastructure markets. With approximately 30,000 employees spread across 30+ countries, NXP is ranked first in the Identification industry for bank

cards, e-government, mobile NFC, transportation and access management cards, etc. NXP is also ranked number one in the Automotive industry for Auto Micro Controllers, Auto Non-Power Analog, Car Infotainment, In-Vehicle Networking, etc. The organization is spread across the globe with most of the R&D activities and Manufacturing facilities positioned in Asia, Europe and the United States. The Wafer Fabs, which are semiconductor processing facilities responsible for converting wafers into integrated circuits are located in the United States (Austin, Chandler), Netherlands (Nijmegen), and Singapore. There are also assembling and test facilities located in Bangkok, Kaoshiung, Kuala Lumpur, and Tianjin.

NXP, on a high-level, has business units, core processes, and supporting organizations. Business units are further divided into business lines and product lines that cater to specific areas within the organization. The business units are solely responsible for creating new products with assistance from a centralized R&D Department. This research and thesis have been conducted as part of a collaborative effort between the CTO Office and the R&D IT department within NXP.

1.4 Scope of the Thesis

The advancement of ML techniques in predictions and high-stakes decision-making has seen great success as organizations across the globe are slowly adopting them as the norm. That being said, the adoption of these advanced analytics and algorithms has also faced skepticism from stakeholders for its impediments. In a business context, when investors, consumers, business stakeholders and end-users come together to make informed decisions based on a model, the first question that comes up is 'How can model outcomes be trusted by stakeholders in an organization?'

While the scope of the literature review performed as part of Research Topics prior to the thesis spans various topics like predictive analytics, explainability, interpretability, and trust in model predictions, there are still limitations that need to be bridged. Most of the literature identified offers significant opportunities for organizations by providing means to analyze, diagnose, fine-tune and improve models and the predictions that come with them. However, one of the key weaknesses identified in the existing literature is that the interpretations are not extracted in a way that will be meaningful in a business context i.e. there is little effort on understanding model outcomes in a way that will drive efficiency for businesses and their revenue. Although most literature focuses on forecasting predictions and understanding the nuances of black-box algorithms, they don't necessarily discuss this in light of how the business can derive insights in an actionable manner. The aim is to identify key drivers as a result of insights generated so that business stakeholders can easily make sense of them, look into financing options, and demand planning for operations. In addition, there is also a lack of discovery in formal applications of industry-standard frameworks in business or information systems research. Most of the existing literature highly focuses on the need for a governance framework without going into the specifics of how these algorithms can actually conform to them.

The scope of the thesis revolves around identifying the potential in developing a business-friendly methodology, inclusive to all stakeholders that can be adopted to interpret ML model outcomes and get business stakeholders to trust the final outcomes. To concise the scope of the thesis and specialize in one focus area, the thesis focuses on interpretability and its many dimensions as a core concept to develop the methodology.

1.5 Research Design

This section discusses the objective of the research, the methodology used to reach research objectives and answer the research questions.

1.5.1 Research Objective

The research objective of the study is to develop a methodology that doubles as a comprehensive cookbook for ML and a guideline to help organizations build models that are more focused towards aligning with the overall business strategy. It should strengthen the collaboration between stakeholders across different business lines, managerial levels and drive efficiency in decision-making.

The research process conducted to meet the aforementioned research objective is two-fold. First, it aims to understand the landscape of analytical methods used for decision-making in businesses and how ML algorithms support them. This will be done by analyzing how analytics advances decision-making and the role of ML algorithms in a business context. Secondly, it focuses on understanding how the predictions of ML models can be presented to business stakeholders in a way they can easily interpret and trust them to make actionable business decisions.

1.5.2 Research Questions

This section presents the main research question for the thesis, the sub-research questions that have been framed for background knowledge, and the reasoning behind choosing them. Based on the aforementioned problem statement and research objective, the main research question is formulated as follows:

RQ: What is an appropriate methodology to build interpretable ML models for business stakeholders to trust and make data-informed decisions?

Here, the researcher defines a methodology to be '*appropriate*' in terms of its usability. According to ISO 9241-11 in [10], usability is a more comprehensive concept than is commonly understood by "ease-of-use" or "user friendliness". It enables users to achieve goals effectively, efficiently, and with satisfaction.

The following sub-research questions have been formulated to provide background knowledge for the main research question and guide the remainder of the research. SQ1, SQ2, SQ3, and SQ4 are answered by conducting a Systematic Literature Review. SQ5 and SQ6 are answered using Design, Development, and Evaluation.

SQ1: What are the analytical methods in practice for business decision-making?

This sub-research question serves as the foundation for the overall theme of the research and discusses the most prevalent analytical methods that are in use. It gives a comprehensive overview of how businesses make use of their data to drive decisions.

SQ2: Why do business stakeholders find it difficult to understand the outcomes that ML models generate?

The second sub-research question dives deep into understanding the perspective of business stakeholders while trying to understand how and why an ML model makes a certain decision for the

business. This research question gives a leeway to understand the importance of explainability and interpretability in ML models.

SQ3: What is the difference between explainability and interpretability according to the literature? There have been many attempts within the ML community to distinguish the two terms since they're used interchangeably in different scenarios. The third sub-research question aims to distinguish 'interpretability' and 'explainability' in ML and also discuss the importance of each.

SQ4: Why is it important to introduce a degree of interpretability in ML model outcomes? As a follow-up to the second and third sub-research questions, the fourth question helps identify the importance of building models that have interpretable outcomes for businesses.

SQ5: How can a methodology be designed to build ML models that are interpretable? This sub-research question aims to answer how the author develops a methodology for interpretable and explainable ML models and eventually answers the main research question.

SQ6: How can the developed methodology be evaluated in the organization? This sub-research question evaluates the efficacy of the developed methodology and the benefits of using it in practice.

1.5.3 Research Methodology

To answer the research questions, a methodological approach is required as the foundation. For this research, Design Science Research Methodology (DSRM) [11], a methodology in the field of Information Systems (IS) to design IS artifacts has been adopted. Adopting such a method using the scope of any research would be advantageous as it provides a road map for researchers to use design as a research mechanism. DSRM is a 6-step process model aligned in a nominal sequence. It is structured as follows:

1. **Problem Identification and Motivation:** To define the specific research problem and justify the value of a solution.
2. **Define the objective for a solution:** To define objectives of the solution based on the problem identified and understanding the feasibility of seeing it through.
3. **Design and Development:** To create an artifact i.e. constructs, models, methods, or instantiations.
4. **Demonstration:** To demonstrate the use of the created artifact instances of the problem. This could be in the form of experiments, simulations, case studies, and more.
5. **Evaluation:** To observe and measure how well the artifact supports a solution to the problem using relevant metrics and analysis techniques.
6. **Communication:** To communicate the problem and its importance, the artifact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences.

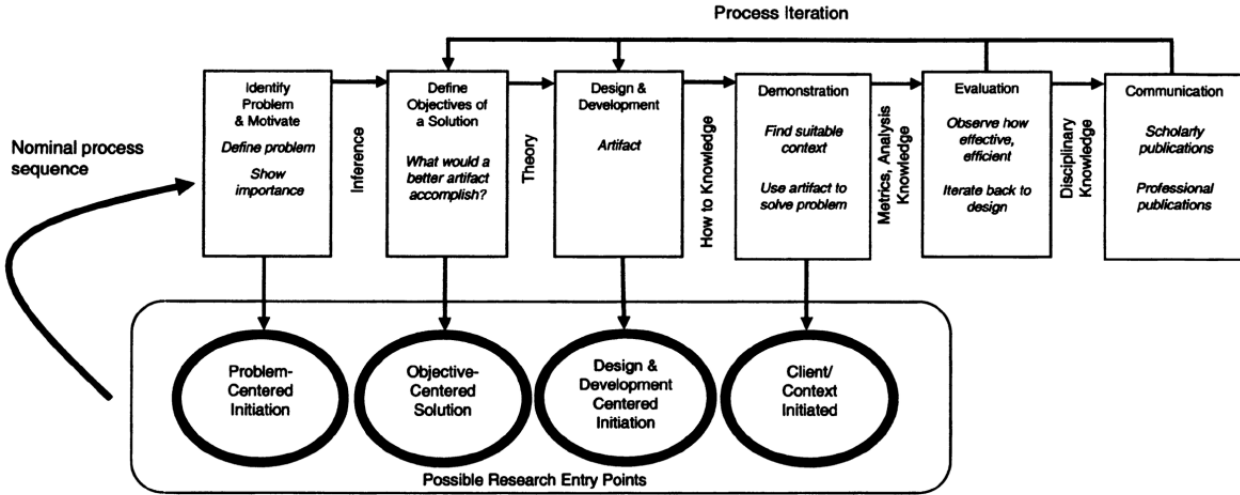


Figure 1: Design Science Research Methodology

Rationale

DSRM has been chosen as the overall research methodology after careful consideration of observations. The primary reason is, it aligns with the scope of the research objective i.e. to create a methodology (artifact). Moreover, the process model includes a demonstration of the designed artifact using an actual implementation after which it will be evaluated on its efficiency and efficacy. This helps with holding the researcher accountable for the developed artifact and its performance in a real-world scenario. Added to this, another observation that is in favor of the process model is the cycle, denoted by arrows, that goes from the evaluation and communication phases back to defining objectives phase. Using this, any gaps that were overlooked in the objectives phase can be re-assessed and added to the process iteratively.

The intent of the box titled 'Possible Research Points' lacked clarity at first for the researcher. On reading the literature further, it is understood that despite the process model being organized in a sequential manner, researchers are not bound to start in that order, rather they can start from any of the entry points. According to the authors, the process model is designed in a way to incorporate the existence of an artifact that has not yet been formally thought through as a solution for the explicit problem domain in which it will be used. This sort of approach gives the researchers flexibility and liberty to choose between many research entry points without having to be bound to one specific entry point. For the scope of this thesis, the author has decided to start with the third entry point i.e. design and development-centered approach. This is because the thesis focuses on building a new methodology by observing the existing methodology in practice and encapsulating the gap and limitations observed as part of it.

1.6 Thesis Outline

The remainder of the thesis is structured as follows. First, Chapter 2 will explain the extensive literature review conducted to answer the main research and sub-research questions SQ1, SQ2, SQ3, and SQ4 that have been formulated. Next, the process of data collection and qualitative approach are discussed in Chapter 3. This is then followed by Chapter 4 that discusses in detail the approach followed to design and develop the methodology to answer sub-research question SQ5. Chapter 5 and Chapter

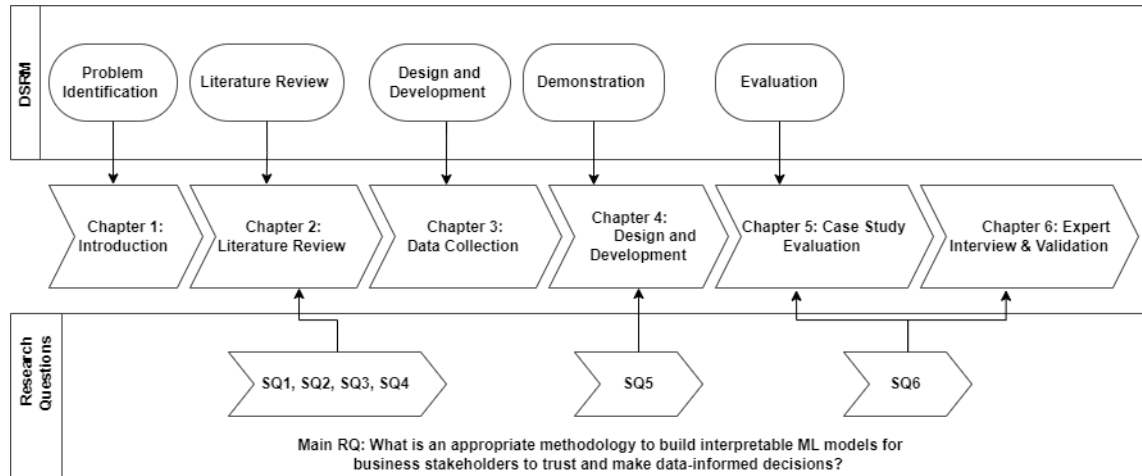


Figure 2: Outline of the Thesis Structure

6 demonstrate the evaluation and validation of the research by means of a case study and expert interviews respectively conducted within the company to answer sub-research question SQ6. Chapter 7 discusses the conclusions by answering the research questions, the contributions towards science and practice, limitations, and recommendations for future research. An overview of the chapters, the related research question and sub-research questions can be seen in Figure 2.

2 Literature Review

This chapter discusses the literature review performed for the scope of the thesis following the guidelines of the Systematic Literature Review. It explains the search strategy followed, the queries and databases used, and the relevant data extraction strategy. The chapter also details supporting literature background to establish the landscape of Machine Learning (ML) and its applications to the audience.

2.1 Systematic Literature Review

The literature review applied for the study follows the guidelines of Systematic Literature Review (SLR) based on Kitchenham et al [12]. This research methodology aims to summarize existing evidence, identify gaps in existing research and provide a framework or approach to include new research possibilities. It also focuses on a search strategy that is less likely to be biased with the search results. This section discusses the research method and search strategy conducted to identify and analyze relevant studies. It also explains the research questions identified along with the search queries used in the databases, inclusion and exclusion criteria, and quality assessment of included studies. The SLR is focused on answering sub-research questions SQ1, SQ2, SQ3, and SQ4 discussed in Chapter 1.

2.1.1 Search Strategy

The search strategy employed in this literature review focuses on gathering literature from the databases to which the University of Twente grants access. Of the 106 databases available for students to access, IEEE Xplore, Scopus, and Web of Science are chosen as the most relevant to the study at hand. The others, despite being exhaustive, don't include relevant studies that complement the areas of interest for the study.

Once the relevant databases and digital libraries are selected for the study, the next step is to create a search string with relevant keywords. To make sure the results are inclusive and accurate, two search strings have been formulated on experimenting with multiple combinations of keywords and synonyms relevant to the scope of the thesis. The first one encapsulates literature on analytic models and business decision-making. The second one dives one step further into the business context while introducing interpretability and explainability in ML models. The queries are as follows:

('Advanced analytics' OR 'Analytical methods' OR 'Analytics') AND ('Business decisions' OR 'Actionable insights' OR 'Data-informed decisions')

('Explainability') AND ('Interpretability') AND ('Business decisions' OR 'Actionable insights' OR 'Data-informed decisions')

Using the strings mentioned above, the search has been performed on the title, abstract, and keywords in the digital libraries. The search strings combined returned a total of 870 articles from the three databases. Zotero [13], an open-source easy-to-use tool to help with the collection, organization, annotation, and citation of research was used for this purpose. As shown in Table 1, the search queries have been mapped to the respect sub-research questions discussed in Chapter 1. The collected articles have been exported to Zotero for the purpose of removing duplicates and screening. It is implied that the search results have to be carefully filtered again based on specific search boundaries to get the

Search Query	Databases	Research Questions
'Advanced analytics' OR 'Analytical methods' OR 'Analytics' AND 'Business decisions' OR 'Actionable insights' OR 'Data-informed decisions'	IEEE, Scopus, Web of Science	SQ1: What are the analytical methods in practice for business decision-making?
'Explainability' AND 'Interpretability' AND 'Business decisions' OR 'Actionable insights' OR 'Data-informed decisions'	IEEE, Scopus, Web of Science	SQ2: Why do business stakeholders find it difficult to understand the outcomes that ML models generate? SQ3: What is the difference between explainability and interpretability according to the literature? SQ4: Why is it important to introduce a degree of interpretability in ML models?

Table 1: Search Queries mapped to Research Questions

right results and take the study forward. To narrow down the search results, a selection criteria has been carefully formulated. Fig 3 shows the overall process of study selection conducted for the thesis.

2.1.2 Inclusion and Exclusion Criteria

The studies selected thus far are on the basis of a simple keyword-search on abstract, title and keywords which returned a large number of results. Next, an attempt has been made to concise the search results further with the following boundaries:

- Range of Publication Years: 2010 to 2022
- Subject Area: Business and Computer Science

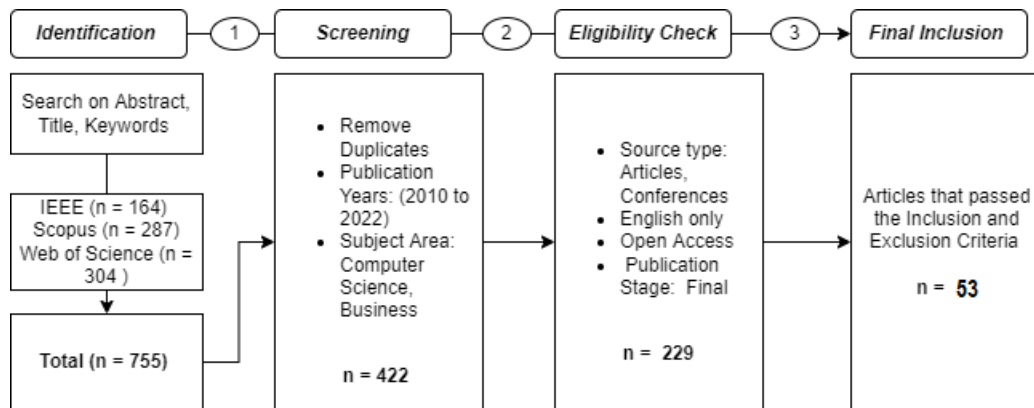


Figure 3: Study Selection Process

The range of publication years for the search strategy has been chosen by keeping in the mind the inception of the Explainable Artificial Intelligence (XAI) program in 2015 by DARPA [14]. The subject area for the search has been limited to business and computer science by keeping in mind the scope and assumption of the thesis: the other areas are decision science, energy, material science, earth and planetary sciences, etc. that are irrelevant to the scope of the thesis. The search string

returned approximately 450 articles from the three selected databases. To further scope it down, a comprehensive set of Inclusion Criteria (ICs) and Exclusion Criteria (ECs) has been formulated. The Inclusion Criteria (ICs) is as follows:

- The paper is in English.
- Source type of material should be conference papers and journal articles.
- The paper is available for download.
- The paper should complement the scope of study.
- The paper should include modifications to an existing approach or introduce new means to approach the scope of study.
- The paper should address one or more research questions.

The Exclusion Criteria (ECs) is as follows:

- Titles with little or no similarity to the scope of study.
- Papers with limited access or only abstract available.
- 'Articles in Press' status of Publication Stage.
- Papers with abstracts that have little or no similarity to scope of study.

The criteria mentioned above has been carefully applied to screen the remaining papers, while also reading their abstracts. This resulted in approximately 55 articles to review and analyze further.

2.1.3 Data Extraction Strategy

In order to answer the research questions mentioned in Chapter 1, it has been decided to formulate an inclusive data extraction strategy that helps with extracting relevant information from the literature.

1. For **SQ1**:, the papers that explain in detail the landscape of analytical methods that are currently in use have been extracted.
2. For **SQ2**:, literature specific to business stakeholders perspective on the difficulty in understanding ML models and their outcomes have been extracted to answer the research question.
3. For **SQ3**:, literature that break-down the concepts of explainable AI and interpretable ML have been extracted to identify the differences between the concepts.
4. For **SQ4**:, as a follow-up to the previous questions, literature that explicitly discuss interpretability as a concept have been analyzed.

The next section discusses in detail the outcome of the literature review performed to answer sub-research questions SQ1, SQ2, SQ3 and SQ4.

2.2 Analytical Methods for Decision Making

The term 'Analytics' refers to a process of discovering, analyzing, interpreting and explaining significant data trends and patterns. With organizations scaling up with data collection and management, what they use it for and how they analyze and interpret that data becomes more nuanced. Data without analytics doesn't make much sense, but analytics is a broad term that can mean a lot of different things depending on where the organization is positioned on the data analytics maturity model. Modern analytics tend to fall in four distinct categories: [15] and [16] classify them as Descriptive analytics, Diagnostic analytics, Predictive analytics and Prescriptive analytics. This makes it easy for key stakeholders and decision-makers to consume the data in a digestible way and make better business decisions. In Fig 4, Gartner categorizes analytical methods in decision-making on the basis of hindsight (to know what has happened), insight (to understand patterns in performance) and foresight (to predict what is most likely to happen).

2.2.1 Descriptive Analytics

Descriptive analytics answers the question of "What happened?" regarding an organization's process. It is typically the starting point in business intelligence and works on historic data and events of the past to help businesses draw comparisons. The reason being, it is of paramount importance for businesses to get an accurate view of what has happened so far and how that differs from other comparable periods. The insights obtained as a result of descriptive analytics can be used to flag areas of strength and weakness to keep the upper-management and business leaders in the know of strategies to implement going forward. Historical data is aggregated and mined to produce visualizations such as line graphs, bar charts, pie charts. Organizations first need to aggregate raw data from various sources and translate it to a digestible format. It presents a clear picture of what has happened in the past but does not make interpretations or advise on future actions.

2.2.2 Diagnostic Analytics

Diagnostic analytics answers the question of 'Why did a certain instance happen?' to examine causes of trends and variations in performance of business processes. Unlike descriptive analytics that gives a surface-level understanding, diagnostic analytics reveals the full spectrum of causes and drills down the data to identify root causes, outliers and causal relationships between data points. Diagnostic analytics also helps with data discovery that can provide valuable insights to explain anomalies identified by descriptive analytics. For example, analyzing external data might reveal change in supply chains, new regulatory requirements, a shift in competitive landscape, and so on. Diagnostic analytics can provide insights into causal relationships and associations in data points. This can be done in several ways:

1. Data drilling - Drilling one level down into a dataset reveals a granular level of detail and informs stakeholders on data aspects that are driving trends and patterns in the organization.
2. Correlation analysis - To examine how strongly different variables are related to each other and find patterns as a result of these relationships.

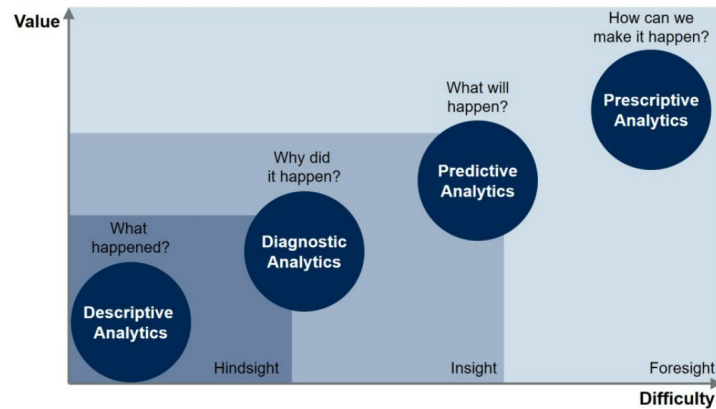


Figure 4: Analytical Methods in Business [1]

2.2.3 Predictive Analytics

Predictive analytics underpinned by algorithm and statistical techniques answers the question 'What is likely to happen?' to determine future outcomes. As compared to the previous two analytical methods, predictive analytics takes the investigation a step further, using statistics, computational modeling, and machine learning to determine the probability of various outcomes. For a typical business context, predictive analytics helps to find what would be the expected sale in the next month, quarter, or year, etc. The goal is to determine a trend, correlation, causation, or probability for the next purchase. With these insights, businesses can anticipate future performance and efficiency of business processes. One of the advantages of predictive analytics is that it can forecast possible future outcomes and identify the likelihood of those events happening as well. This helps organizations plan better and set realistic goals to avoid crossing paths with business risks. Business leaders can take proactive and data-driven decisions by observing how trends will unfold in the future and the potential impact they can drive for the organization.

2.2.4 Prescriptive Analytics

Prescriptive analytics answers the question 'What is going to happen?' and uses the results of descriptive, diagnostic and predictive analytics to suggest actionable decisions and improve business outcomes. With prescriptive analytics, businesses are prescribed best practices and actions to eliminate the likelihood of a roadblock occurring in the future. It leverages advanced tools and technologies, like machine learning, business rules, and algorithm to make timely business decisions.

2.3 Machine Learning in Business Context

Informed decision-making in any organization is table-stakes. Machine learning enables organizations to harvest a high volume of insights from business data and urge them to make decisions that are informed. The myriad uses of machine learning in organizations such as rapid decision-making and accurate demand forecasting capabilities, businesses can uphold their reputations and avoid costly corrective measures. Organizations rely on ML models to learn from historic data and predict future outcomes such as demand and supply, resource allocation, investment opportunities, etc. Business leaders hold a strategic outlook to support long-term opportunities and challenges of the organization based on the wavering unpredictability of the future. They need to keep a tab on the health of the business while continually developing strategies to envision opportunities, contemplate potential

challenges, anticipate the future by creating a series “what if” scenarios to steer the business in the correct direction. It is implied that decision making of a good organization lies at the intersections of Artificial Intelligence, Machine Learning, Deep Learning, and other emerging data sources with a substantial amount of human intervention [17]. For example, as shown in Fig 5, data is generated from the relevant population to train a predefined ML model. Once this is done, the model in turn classifies observations or optimizes a predefined outcome, and its predictions then trigger marketing decisions and actions [2]. Theoretically, decision-making refers to the process of making a choice from a group of alternatives. However, in a broader context, decisions ideally result from being aware of the environment in which a given choice takes place. In this regard, decisions can be categorized as follows:

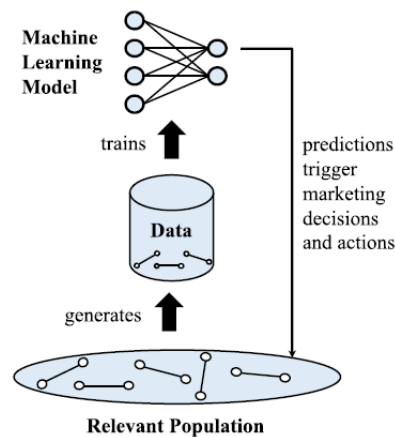


Figure 5: Decision-making using ML [2]

1. **Decisions in conditions of certainty:** These decisions are often made when there is a degree of guarantee and assurance. In these scenarios, the decision-makers are in the know of the possible outcomes and benefits that they can arrive at once the decision is made.
2. **Decisions in conditions of risk:** These decisions are often made when the decision maker is uncertain about the consequences but is aware of the probability of possible outcomes. In business processes, making a decision in this context puts the business at a certain level of risk.
3. **Decisions made under uncertainty:** These decisions are made in uncertain situations and when the probability of consequences is not clear. These decisions are further divided into two. The first type includes decisions made in conditions of ambiguity, where all the effects of the decision are known but the probability of their occurrence is not clear. The second type are decisions made in conditions of ignorance, when even not all the consequences are known to the decision maker [18].

Business data and analytics in combination with expert experience, subject to the caveat of quality and accuracy is the right way to fuel any leader’s decision-making engine [19]. For instance, when organizations make use of predictive analytics, it is possible to make strategic decisions by detecting projects in the lifecycle that are affected by misappropriate budget execution. While this requires identifying the most effective ML algorithms and insightful financial predictors to make use of, business stakeholders can make critical decisions around resource allocation, project schedule, budget expectations, etc. It can also help with minimising negative consequences associated with miscalculated budget allocated for the project such as excessive liabilities or lost opportunities [18].

2.3.1 Challenges in ML-based decision making for businesses

While it is important to transition to automated decision-making, there is also an imperative need to understand how such decisions are made, establish a synergic and collaborative environment between humans and machines. However, these techniques currently lack transparency which makes it difficult for business stakeholders to trust model outcomes easily. Transparency in ML algorithms should be able to reveal how data is integrated into an algorithm, processed by it and the knowledge that is gained there after. One of the main challenges businesses are tasked with algorithms and decisions is that business managers do not know how ML algorithms generate the outputs by processing the input data because the algorithm is either proprietary or the mathematical computational models used in the algorithm are very complex to understand [20]. A collaborative effort for business decisions between machines, algorithms and humans warrants a need for transparency so that stakeholders can trust the final outcomes. Moreover, data scientists and experts may also have troubles explaining why their algorithm made a certain decision. This makes the end-user, on the receiving end of the explanation, more often than not a business stakeholder, hesitant to trust the outcome without sufficient contextual reasoning. Research exploring the models' rationale behind the predictions and how the insights must be provided to the user as an explanation is known as Explainable Artificial Intelligence [21].

2.4 Explainable AI/ML

Explainable AI/ML (XAI) came into existence as a terminology in 2004 by Van Lent et al [22], [23] to describe the ability of their system to explain the behavior of AI-controlled entities in simulation games application. Although the term has recently stirred a series of discussion from academia and practitioners, the problem of explainability has existed since the mid-1970s when researchers studied explanation for expert systems. The motivation to reach a consensus on the problem of explainability and interpretability took a back seat owing to the many advances in ML. Since then the focus of research has shifted towards implementing models and algorithms that have the learning capabilities to generate models with high predictive performance and accuracy. According to [24], ML explanations enable users to understand how the data is processed. They aim to bring awareness to possible bias and system malfunctions.

2.4.1 Need for Explainability & Interpretability

The authors in [25] and [26] identified that the need for interpretability in any algorithm primarily stems from a misconstrued understanding of the formal definition of a machine learning model, its output and finally the real-world impact. Lipton further expanded on interpretability as a concept to build trust in the model, infer causal relationships between the input and produced output, provide introspection and facilitate fair and ethical decision-making. Since then, other researchers have also contributed to this by detailing how cognitive biases can be mitigated using relevant interpretability methods and techniques, proposing taxonomies to arrive at an appropriate interpretability method, design decisions can be carried out with different levels of automation ,etc[27].

Within the research community, despite the many attempts to distinguish 'interpretability' and 'explainability', the two terms are still used interchangeably. Fig 6 shows the relationship between Explainability and Interpretability [3]. This is because earlier work failed to draw a clear line of distinction between interpretability and explainability as these terms are subjective to the stakeholders who need to understand the model and its outcomes. To simplify the two terminologies, they're both

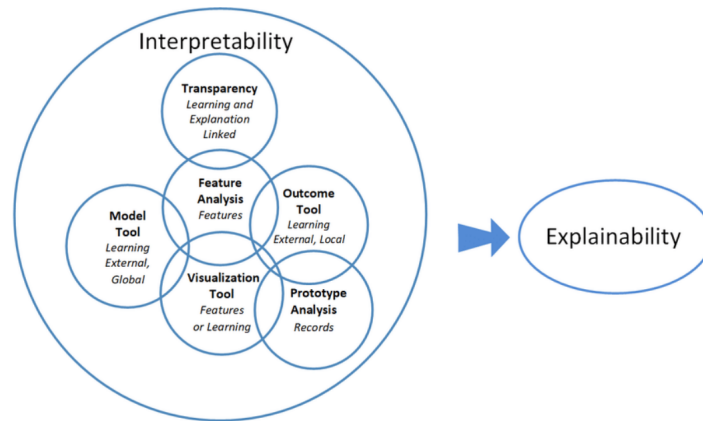


Figure 6: Explainability and Interpretability - A Representation [3]

equally important to nurture a sense of human–machine trust and help the stakeholders on the receiving end of the model outcomes understand how certain decisions are made. It is also important to note that is not limited to statistical metric-based approaches such as accuracy or precision [28]. the authors in [29] say that the goals to be attained through the interpretability of models are trust, reliability, robustness, fairness, privacy, and causality. Explainability can be formulated as the explanation of the decisions made internally by the model that in turn generates the conclusions arrived at by the model. This promotes human understanding of the model’s internal mechanisms and rationale, which in turn helps build user trust in the model. The evaluation criteria for model explainability include comprehensibility by humans, fidelity and scalability of the extracted representations, and scalability and generality of the method. This thesis and research assume the following:

1. Explainability - The ability to discover meaning between input data and model outputs i.e. to take an ML model and explain the behavior in human understandable terms.
2. Interpretability - To be able to understand the inner workings of a model i.e. to understand exactly why and how the model is making predictions.

With a rise in speculations across different sectors on the nuances of entrusting expert systems and algorithms on making key decisions, there has also been an increase in governance frameworks and processes to address these concerns. To this end, the Defense Advanced Research Projects Agency (DARPA), responsible for the development of emerging technologies for use by the military of United States curated an Explainable Artificial Intelligence (XAI) program [14]. The Explainable AI (XAI) program aims to create a suite of machine learning techniques that can help with producing more explainable models, while maintaining a level of performance and prediction accuracy.

Of the many programs that DARPA is currently engaged in, XAI is one of the handful to enable “Third-wave AI Systems” where AI systems should be able to acquire human-like communication and reasoning capabilities to recognize, classify and adapt to new situations autonomously. Over time, machines will be able to understand the context in which they operate and build underlying explanatory models to simulate real-world phenomena. The target audience of XAI is typically an end user who relies on decisions and recommendations produced by automated systems and algorithms. For example, an intelligence analyst who receives recommendations from a big data analytics system needs to understand why it recommended certain activity for further investigation. The program should eventually be able to help users to understand the rationale behind a certain decision and how

it will impact circumstances in the future.

Likewise, Europe has adopted the GDPR, an ambitious set of comprehensive regulations for the collection, storage, and use of personal information with the hopes of making the European Union ‘fit for the digital age’. The law spans across many provisions and among them are some related to automated decision-making. Under the GDPR, Article 22, the Right to Explanation says that ‘*The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*’ [30]. This means that subjects impacted by algorithmic decisions have the right to request to be informed by organizations on how algorithms have made automated decisions [31]. It gives every individual the choice to engage or disengage from something if they find the decisions are biased or unfair. While the law aims to clear the disparity in decision making, it also puts organizations and their business at stake in a plethora of ways. To expect the company and their IT team to explain why their algorithm has made a certain decision seems close to impossible. This is because an AI or ML algorithm trains itself on massive amount of data, while performing multiple bouts of micro-decisions and complex mathematical equations. A human expert, even programmers, would find it extremely difficult to explain an accurate reasoning behind an algorithm’s exact functioning. This warrants the need to establish a strong and well-structured governance framework that organizations can adopt to interpret and explain the decisions made by the algorithms. While there exists a wealth of literature in the quantification of interpretability and explainability for models, the quantification of trust as a component is fairly new and unexplored.

When developing an ML model, the consideration of interpretability as an additional design driver can improve its implementation for 3 reasons — veracity, trustworthiness and impartiality. In respect to this, the FAT framework is responsible for ensuring the algorithm backing a decision stays Fair, Accountable and Transparent. This framework primarily came into existence to tackle ethical challenges and biases in automated decisions. [32] presents a conceptual model of FAT as an antecedent variable affecting satisfaction while trust is a moderator influencing this relation. This can be seen in Fig 7. For example, in the case of expert and recommendation systems, transparency and fairness are key indicators in algorithms to build trust amongst users.

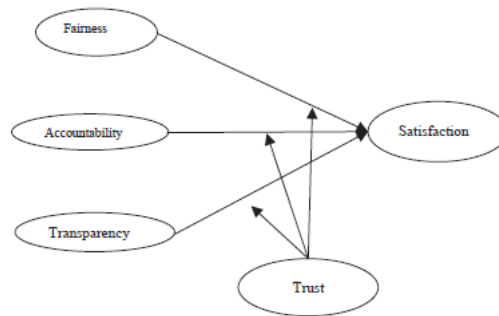


Figure 7: Conceptual Model of FAT

According to [32], the three components of FAT are defined as follows:

1. **Fairness:** To ensure lack of bias in model prediction. Bias can occur in two ways- data bias and model bias. Data bias refers to the dataset serving as input and model bias is a result of unfair influence of model fitting.

2. **Accountability:** To ensure prediction accuracy and coverage of data points with reliable predictions. Accountability bridges the gap between an algorithm generating a high number of unreliable predictions or a lower number of high quality predictions.
3. **Transparency:** Relates to the function of the modelling method. This focuses specifically on the black box models that can generate accurate predictions but cannot explain the logic underlying the predictions.

While prediction accuracy is important for decision-making, it also requires an increase in computational effort and complexity. It boils down to the business leaders having to arrive at a trade-off and decide on the level of accuracy that is satisfactory for business requirements. [32] also suggests an iterative approach to creating fair, accountable and transparent models for business.

1. To tackle transparency, select a general modelling method such as linear models and decision trees that are more transparent.
2. Formulate a set of predictors with a workable solution.
3. Set of checks based on accountability conditions such as prediction accuracy, coverage, model stability, replicability, etc.
4. Next, include a fairness criteria to check for bias. Suggested approach is to we remove outliers and partition the dataset into relevant subsets for calculation. Re-processing the dataset and repeating computation of predictors with different approach could also be another way to deal with model bias.

2.4.2 Stakeholders and Target Audience

The need for explainability and interpretability came into existence due to the fact that stakeholders are often reluctant to trust ML projects because they don't understand what they do. The reason being, it's difficult to garner support that a model can be trusted to make decisions especially when there is no explanation of how that decision was made. Moreover, business stakeholders often lack the ability to grasp the technical logic underpinning ML models to make decisions. Some of the pressing questions at the top of a business leader's head could be:

- How did the model predict or make a decision?
- How are we sure there is no bias creeping into algorithms?
- Is there enough transparency and interpretability to trust the model's decision?

In order to accelerate adoption of automated decision making, enable accountability and provide strategic insights, the ML models and the outcomes presented to stakeholders need to be explainable and interpretable. This helps build trust and confidence of stakeholders in the process of ML and promotes executive buy-in and sponsorship for ML-related projects. Moreover, these models can also provide valuable insights like such as sales in dollar value, employee turnover rate, customer churn, etc. that the business could possibly be interested in.

It is important to note that the target audience for explainable models are not limited to business stakeholders. As shown in Fig 8, the stakeholders that could benefit from understanding the various aspects of a model for decision-making are as follows:

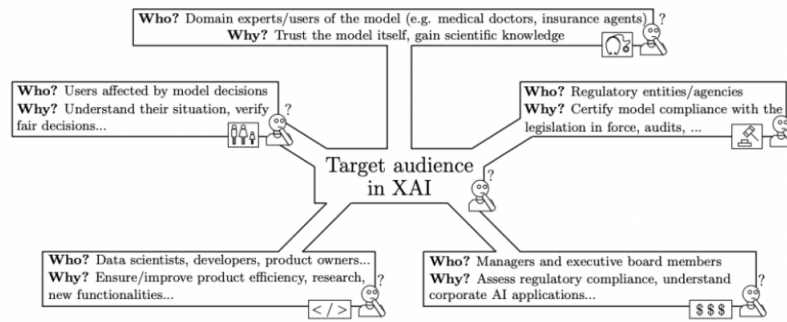


Figure 8: Objectives of Stakeholders [4]

1. **Domain experts & End users:** The stakeholders who are directly involved with the outcomes of the ML model. They need to be able to expand their scientific knowledge, understand how the model makes a certain prediction and trust the results.
2. **Consumers:** The consumers require transparency about how decisions are taken, and how they could potentially affect them.
3. **Data Scientists:** Data scientists, responsible for building the model themselves require a concrete explanation on the outcome to solve impending bugs, if there are any, and improve performance.
4. **Business Owners:** Business owners, leaders and stakeholders require understanding of the model outcome to make informed decisions for the business and ensure they are in line with the overall strategy.
5. **Regulators:** As part of the AI/ML governance team, regulators need to be able to inspect the reliability of a model, as well as measure the impact of its decisions on customers.

2.4.3 Black-box and White-box ML Models

The empirical success of ML over the last couple of years attributes to the combination of efficient learning algorithms and the multiple parameters used to achieve prediction and performance accuracy. However, the scientific community has also sparked discussions on the existence of algorithms that are difficult to break down in terms of their explainability and interpretability. This has led to ML algorithms being labelled and classified as black-box and white-box approaches based on their transparency and opaqueness in inferring the rationale behind their internal working. As the name suggests, black-box models are deemed opaque and contain complex mathematical functions, require a deeper understanding of distance functions and the representation space. These functions are very hard to explain and to be understood by experts, specifically the ones who lack an understanding of the technical logic underpinning models. Due to the black box nature of the models, data science experts, engineers and developers who build the models find it increasingly hard to explain the results to stakeholders who need to make decisions for the business [33]. On the other hand, ML models based on patterns, if-then rules and decision trees are labelled as white-box models as they can be easily interpreted by a simple visualization [34].

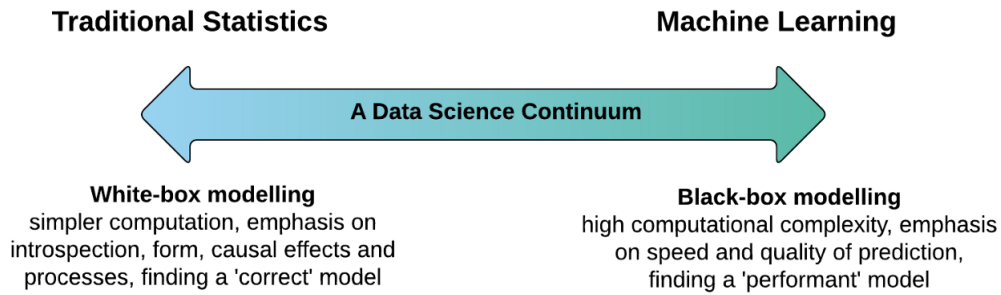


Figure 9: Black-box vs White-box ML models

Black-box Models

While the applications of ML models are advancing further, the understanding of how they work and how decisions are made is not advancing at the same pace [35]. Black-box models are often built directly from the data by an algorithm, so much so that even the data scientists responsible for them cannot understand how variables and parameters work in collaboration to make prediction. These predictive models can be complicated functions such that no human can understand how they are related to each other to create a final outcome. Despite the fact that black-box models such as neural networks, gradient boosting models and complicated ensembles provide great accuracy, the inner workings of them are harder to understand. Moreover, there is no clear interpretation or an estimate of the importance of each feature on the model predictions, nor is it easy to understand how the different features interact.

White-box Models

White-box models whose inner logic, workings and programming steps are transparent to the audience at the receiving end of the explanation makes it easier for decision-making. Simple decision trees are the most common example of white-box models while other examples include linear regression models and bayesian networks. However, the downside of using white-box models is the fact that they suffer from a poor predictive performance and can not be always relied upon to model a complex dataset.

2.4.4 Accuracy-Interpretability Trade-off

The accuracy-interpretability [5] trade-off is based on the assumption that ‘explainability is an inherent property of the model’. Inherently interpretable models such as white-box models provide a better understanding of how predictions and model outcomes are generated. It can be made very clear how variables in the model are jointly related to form the final prediction making it easier for data scientists to explain their models. On the other hand, flexible methods that are capable of estimating more complex functions and mathematical calculations way less interpretable making them black-box. As show in Fig 10, the models at the top of the graph in the shade of dark blue can guarantee accurate predictions and high performance while suffering from interpretability. The models towards the bottom of the graph on the other hand are easy to understand and interpret but do not produce accurate results.

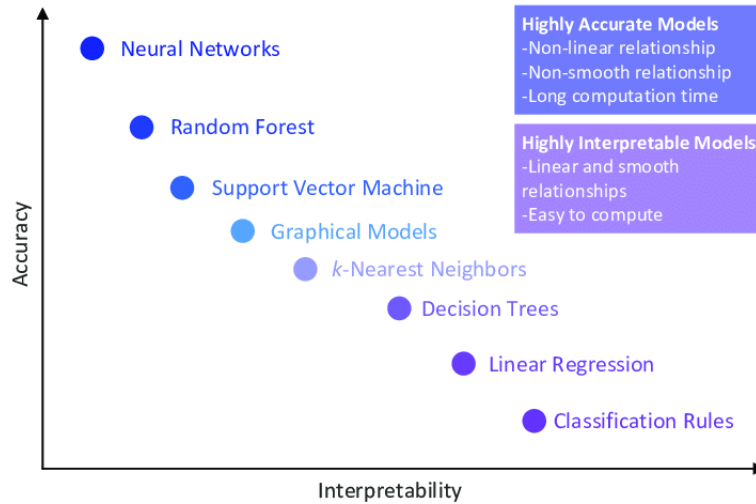


Figure 10: Accuracy-Interpretability Trade-off [5]

While estimating an unknown function 'f', data scientists are often focused on two things for the model - prediction accuracy and inference capability. In the case of the former scenario, where in the focus is on creating a model with high performance and prediction accuracy, data scientists need to train a model can give an accurate prediction of the target variable. Here, the model will be treated as a black-box in the sense that the inference capability or shape of the estimated 'f' is not important as long as it can lead to accurate predictions. On the other hand, when it comes to the latter, the models are often treated as white-box since the aim is to understand the relationship between the input 'X' and output variable 'Y'. From this, we can infer that, the more complicated the shape of the function, the more difficult it will be for data scientists and other stakeholders to understand the relationship between the predictor variables 'X' and the target variable 'Y'.

While there needs to be a trade-off, the belief that accuracy must suffer at the cost of interpretability is inaccurate and vice-versa. In [7], research has shown that in many contexts, especially with well-structured datasets and meaningful features, directly interpretable models can reach comparable performance to opaque models. With the advances being made in methodologies and techniques to combat the problem of arriving at accurate yet interpretable methods, it entirely boils down to making the right choice of ML algorithms and techniques that are deployed to see this through.

2.4.5 Dimensions of Explainability

Recent studies [36], [37] have identified two types of explainability - local and global explainability. In local explainability, an explanation behind an individual decision or prediction is provided, while in global explainability only a single explanation is given for the whole dataset. Added to this, explanation techniques can also be classified as ante-hoc and post-hoc. Ante-hoc methods such as linear regression, decision trees and random forest have a degree of explainability incorporated in the model itself while post-hoc techniques rely on other models to be trained and provide explanations. In addition to this, there are questions [38] that can be used to validate the explanations the model predicts. They are as follows:

1. How to explain?

Relates to how a predictive model derives predictions based on the inputs. It can be done in the form of a proxy model, feature importance or visualization.

2. How much to explain?

Relates to the level of granularity to which models can be explained i.e. local explanations and global explanations.

3. How to present?

Relates to choosing ways to present an explanation the stakeholders based on the characteristics of the end user, their level of expertise, scope of explanation and the purpose it brings forth. It can be done by visual explanations, verbal explanations or analytical explanations.

4. When to explain?

Relates to the point in time an explanation should be provided to stakeholders. This can be intrinsic, where explainability is introduced during the process of building the model, or post-hoc where explainability is established on the basis of model outcomes.

Human-centered Outcomes

In [39], the author steers the research towards the direction of philosophy, psychology and cognitive science with respect to interpretability. The author goes on to discuss if the goal is to build a successful model that can convince stakeholders to trust the outcomes eventually, then there needs to be an understanding that:

- Explanations are contrastive. It is important to note that stakeholders more often than not do not ask why an event 'E' happened, rather why event 'E' happened instead of an another event 'F'.
- Explanations need to be considered as a social conversation and interaction for knowledge transfer. To make it simpler, the explainer must be able to leverage the mental model (existing knowledge) of the explainee while engaging in the explanation process.

On a an overall note, the ways in which models should be built can be classified into two categories.

1. Choosing a directly interpretable or white-box model such as decision tree, rule-based model, linear regression, etc.
2. Choosing a complex, opaque, black-box model such as deep-learning, neural networks, ensemble ML models followed by techniques to generated further explanations.

2.5 Research Gaps

While the literature review performed as part of the thesis present extensive information on the concept of explainability and interpretability required to convince business stakeholders to make informed business decisions, the remainder of the thesis will focus on how the information collated can help with the design and development of a busienss-friendly framework that can be adopted by organizations. To summarize the learnings:

- There is no one-size-fits-all solution in the domain of explainability and interpretability. The technical choices should be driven by business objectives, stakeholder needs, costs and resources allocated, etc. *Chapter 4 will discuss the ways in which a methodology can be designed to address these concerns.*

- For any organization involved in ML-based decision making, explainability and interpretability of their ML models should not be an afterthought that's done at the end of the ML workflow. Instead, these components should be integrated and applied every step of the way—from mapping out business objectives, collecting data, processing to model training and finally, model evaluation. *While chapter 4 will discuss the development of the methodology, Chapter 5 will demonstrate the proposed methodology as a working model using a case study.*
- Human cognition and behaviors play a significant role in understanding the outcome of a model. While evaluating the interpretability of ML models with stakeholders, these parameters need to be taken into consideration. In the existing literature, there are many available human-grounded metrics to validate the stakeholder reception of an ML outcome, however, there is no formal application of them in a real-world context.

2.6 Supporting Background Information

This section establishes context for the researcher and audience to understand the landscape of Machine Learning (ML) in a general context and the various application domains.

2.6.1 Machine Learning

The architecture of the earliest computers were designed in a way to perform complex calculations and allowed for the storage of data and instructions to manipulate that data. It was attributed by data processing using mathematical terms where in the computer would follow the instructions in-built in them. It slowly evolved to a stage where a set of instructions were carefully created for computers to learn from experience rather than merely following them. By means of this, computers were able to make sense of the instructions independently, extract rules from large amount of data and use them for prediction. This set things in motion for Machine Learning (ML) to come into existence. The evolution of ML can be seen in Fig 2. ML has developed distinct wide theoretical and practical tracks that revolve around analytics and improving performance with experience. Machine learning has incorporated different methods from different origins, some of them have their origins from artificial intelligence whereas others are coming from applied statistics. It was in the 1990s that scientists made a shift from a knowledge-driven approach to a data-driven approach. They began creating programs for computers to analyze large amounts of data and draw conclusions — or “learn” — from the results.

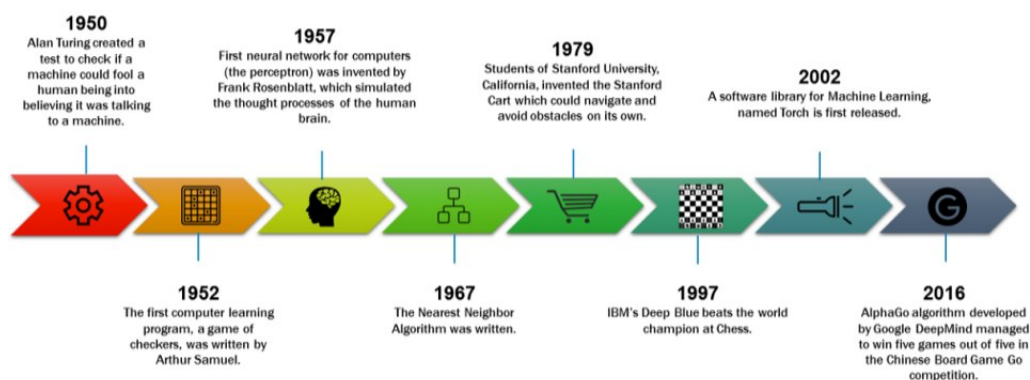


Figure 11: Evolution of Machine Learning

Alan Turing's seminal paper [40] introduced a benchmark standard for demonstrating machine intelligence, such that a machine has to be intelligent and responsive in a manner that cannot be differentiated from that of a human being. Machine Learning (ML) is a subdivision of computational science that deals with training computers to learn automatically through programmed inputs and the incorporation of prior knowledge. It leverages different models and algorithms to approximate complex theories which are difficult to be represented with mathematical and statistical tools and techniques. ML has found its way in different applications such as virtual personal assistants, traffic predictions using GPS navigation, surveillance of cameras to detect crime, face recognition, recommendation systems, email spam filtering, and so on [41]. The premise of ML is about learning from data, making sense of data and discovering patterns and relationships that can describe what will happen in unseen new situations [42]. ML provides systems with the capability to learn from experience without explicitly programming them and for this purpose, ML algorithms are imperative. The learning algorithms can be categorized into Supervised, Unsupervised, Semi-supervised and Reinforcement learning [43]. *The scope of this thesis will focus on supervised and unsupervised learning algorithms.*

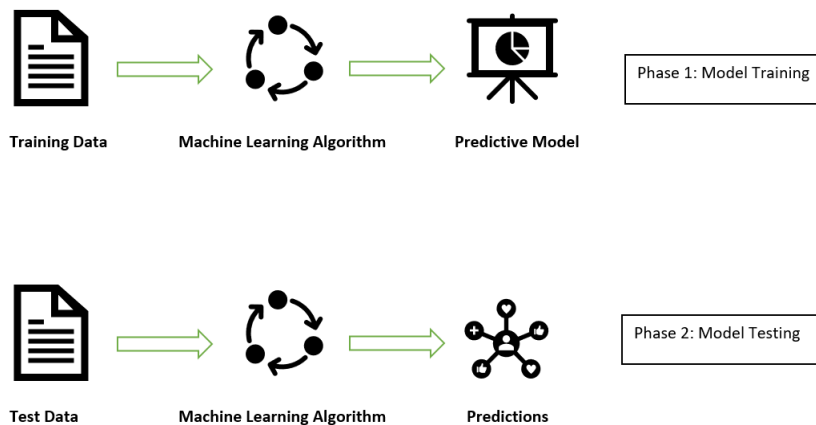


Figure 12: General structure of an ML model

As show in Fig 2, a typical ML model consists of two phases: Train and Test. This is done by slicing a single data set into a training set and test set. The training data is what the data scientist repeatedly feeds the algorithm with as input data. The model in turn evaluates the data repeatedly to learn more about the data and its behavior. Test data, on the other hand, helps with confirm that the ML algorithm was trained effectively by checking its performance on unseen data.

In addition to this, UC Berkeley claims that a typical machine learning algorithm contains three main components:

1. **Decision Process:** ML algorithms are typically deployed for classifications and predictions. Given some input data, be it labelled or unlabelled, the algorithm will look for insights and patterns in the data to help make decisions.
2. **Error Function:** The error function helps with evaluating the prediction and accuracy of the ML model itself.
3. **Model Optimization:** ML algorithms follow the process of updating weights until an accuracy threshold is met. If the model fits well with the data points in a training data set, the weights are adjusted accordingly to reduce discrepancies, if any, between the known data and model estimate.

2.6.2 Supervised Learning

Supervised learning teaches the machine by example and uses labeled datasets to train algorithms that can classify data and predict outcomes accurately. As the model is fed with input data, the algorithm is trained to identify patterns, learn from observations and finally make insightful predictions. The process is repeated until the algorithm reaches a high level of accuracy and performance. More often than not, supervised learning is carried out when certain goals are identified to be accomplished from a certain set of input - a task-driven approach. The most common supervised tasks are “classification” that separates the data, and “regression” that fits the data. An example of supervised learning is the prediction of a class label or sentiment of a piece of text, like a tweet or a product review [44]. Fig 13 shows a general idea of how supervised learning separates classes in a dataset.

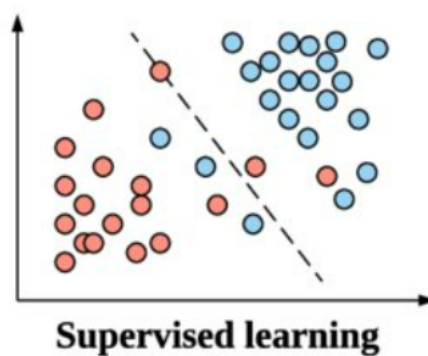


Figure 13: Supervised Learning

1. **Classification:** Classification tasks require ML algorithms to draw conclusions from a set of observed values so that they can be fit into a specific category. For example, in a real-world case of filtering emails as ‘spam’ or ‘not spam’, the algorithm must look at existing observational data and classify the emails accordingly. Linear classifiers, support vector machines, decision trees and random forest are common types of classification algorithms.
2. **Regression:** Regression tasks require machine learning program to estimate and understand the relationships among variables. Regression algorithms are useful for prediction and forecasting as they focus on one dependent variable and a series of other changing variables. Some popular regression algorithms are linear regression, logistic regression and polynomial regression.

2.6.3 Unsupervised Learning

Unsupervised machine learning uses algorithms to identify patterns in unlabeled datasets. This is a data-driven process. Unlike supervised machine learning, these algorithms don’t require human intervention to provide instructions. Instead, the algorithms have the inherent ability to determine correlations and relations by analyzing available data. It has the ability to use clustering techniques to discover similarities and differences in data making it the ideal solution for clustering, density estimation, anomaly detection, exploratory data analysis, customer segmentation, image and pattern recognition. Some of the most commonly used algorithms in unsupervised learning include neural networks, k-means clustering, probabilistic clustering methods, and more. Fig 14 shows a general

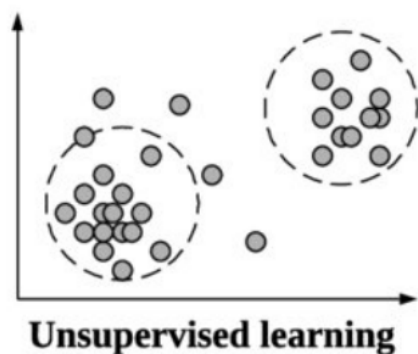


Figure 14: Unsupervised Learning

Key Differences	Supervised Learning	Unsupervised Learning
Goals	Predict outcome for new and unseen data by learning from training dataset	Get insights from large volumes of new data
Data	Labelled data with human intervention	No labelled data; Work independently to discover inherent structure of data
Applications	Spam detection, sentiment analysis, weather forecast, price prediction	Anomaly detection, recommendation engines, customer personas, medical imaging
Complexity	Simple method typically calculated through data-driven programming languages such as R and Python	Computationally complex due to large training dataset
Drawbacks	Time-consuming to train and requires expertise for input and output variables	Highly likely to generate inaccurate results without human intervention to validate output variables

Table 2: Supervised vs Unsupervised Learning: A comparison

idea of how unsupervised learning clusters data points that seem to show similarities. Under the umbrella of unsupervised learning [43], fall:

1. **Clustering:** This involves discovering sets of similar data and grouping them on the basis of a defined criteria. For example, K-means clustering algorithms assign similar data points into groups, where the K value represents the size of the grouping and granularity. This technique is helpful for market segmentation, image compression, etc.
2. **Dimension reduction:** This reduces the number of features in a model through techniques like Principal Component Analysis (PCA). Often, this technique is used in the preprocessing data stage, such as when autoencoders remove noise from visual data to improve picture quality.

2.6.4 Applications of Machine Learning

Of the many technological advances that have gained momentum as part of the industrial revolution, ML is undoubtedly one that keeps finding ways to expedite its learning capabilities. With its foothold

in various industries and application areas, ML algorithms, tools and techniques are leveraged in an optimal way to make informed decisions [43].

1. **Decision-making:** ML algorithms coupled with data-driven analytics form the foundation for organizations to make intelligent and informed decisions in their business processes. Nowadays, industries and organizations including government agencies, e-commerce, telecommunications, financial services, healthcare and others rely on the results of business intelligence combined with algorithmic decisions. These data-driven predictions and insights give business experts an estimate on how future outcomes would look like for the business and the impact they can expect as a result of them.
2. **Cybersecurity and threat intelligence:** Data that is sensitive, confidential and protected such as Personally Identifiable Information (PII), Protected Health Information (PHI) and intellectual property are susceptible to data breach campaigns and threats that can potentially disrupt and derail the functioning of organizations. As part of Industry 4.0, cybersecurity is an important and integral area as it handles the protection of networks, systems, hardware and all categories of data from theft, damage and other digital attacks. With a rise in data breaches, businesses are running a high risk of suffering irrecoverable reputational damage. ML has proven to be advantageous in the cybersecurity domain by constantly staying up to speed with capabilities to detect malware in encrypted traffic, identify insider threats and secure data in the cloud. For instance, ML-based clustering algorithms can be employed in organizations to detect cyber-anomalies, policy violations. On the other hand, ML-based classification algorithms can also be used to detect various types of cyber-attacks and intrusions by taking into account the impact of security features.
3. **Internet of things (IoT) and smart cities:** IoT is commonly defined as the collective network of connected devices that transmit data and automate tasks without the need for human interaction. The premise of IoT is connecting 'everyday' things with the internet. For instance, everyday devices like toothbrushes, vacuums, cars and machines have sensors embedded in them for the purpose of collecting data and responding intelligently to users. With the advent of computer chips and telecommunication, IoT aims to enhance activities at home, education, transportation, retail, healthcare, agriculture, etc. Since the core of ML is to learn from experience, identify trends and patterns based on observations, it has become a crucial technology for applications that make use of IoT. In the context of smart cities, ML algorithms coupled with IoT technology can help with predicting traffic, availability of parking spots, total energy expended by users, etc. An intelligent transportation system based on ML algorithms can also predict future traffic by analyzing travel trends and history of customers. Further to this, they can assist them in recommending alternate routes to combat traffic and peak hours.
4. **Healthcare:** ML algorithms are also widely deployed in the healthcare industry to solve diagnostic and prognostic problems such as disease prediction, medical knowledge extraction, patient management, etc. For instance, with the ongoing COVID-19 crisis, ML algorithms help scientists and researchers forecast when and where the virus is likely to spread along with classifying patients at high risk, their mortality rate, the level of infection, etc.
5. **E-commerce:** Sales and marketing is table-stakes in the e-commerce industry to drive overall efficiency and ensure customer retention. ML comes in handy by using algorithms to provide businesses an overview of consumer behavior and their purchase preferences. These algorithms can also help retailers stay up to speed with consumer purchase histories and create personalized

product recommendations for the next purchases. Retailers can make use of these advancements to manage their inventory, logistics, warehousing and protect customer retention.

2.7 Summary

The chapter gives an overview of the systematic literature review conducted as part of the research to understand the landscape of the domain of interest. It also presents the methodological approach followed to be able to achieve inclusive search results and consolidate the findings. The chapter also includes supporting literature background that helps the research as well as the audience understand the landscape of machine learning, its applications and related fields. As part of next steps, the researcher consolidates the findings of the literature review to support the development of a methodology that can meet the research objectives stated in Chapter 1.

3 Data Collection

This chapter discusses the methodological approach that has been followed before the design and development of the artifact for the thesis. It introduces the steps taken to collect the data required to design the artifact and then synthesizes the results to confirm the need to develop a methodology.

3.1 Need for Qualitative Method

The results obtained through the literature review serve as the preliminary source by discussing the best practices and current state-of-the-art approaches for ML in a business context. The literature review has helped with understanding the landscape of interpretability and why business stakeholders often struggle to make sense of data-heavy outcomes. However, the literature review lacks the discovery of formal applications of methodologies that can be applied to a business context to drive interpretability in ML models. Most of the existing literature highly focuses on the need for a governance framework, for instance, without going into the specifics of how this can be formally applied to a business context. To understand the applicability of the research in a business context, the literature review is supplemented by a series of interviews conducted with relevant stakeholders. The interviews are conducted to collate the experiences and viewpoints of various stakeholders to support the need for a standardized methodology. The synthesis of the results obtained as part of the data collection process results in the development of the artifact. Table 3 shows the summary of primary and secondary data collection methods used for the thesis.

3.1.1 Qualitative Research Goal

The methodological approach for understanding the research context to design and develop an artifact for the thesis is based on the guidelines of Design Science Methodology for Information Systems and Software Engineering by Wieringa [45]. Artifacts are potentially constructs, models, methods, or instantiations or “new properties of technical, social, and/or informational resources” [11]. Conceptually, a design research artifact can be any designed object in which a research contribution is embedded in the design. Wieringa asserts that a study can have a knowledge goal and may or may not have an improvement goal. If an improvement goal is missing, it is a curiosity-driven project study that aims to increase our knowledge about the topic. Using this qualitative research, the researcher would like to identify the artifact’s desired functionality, its architecture and then create the actual artifact. Resources required for moving from objectives to design and development include knowledge of theory that can be brought to bear in a solution. According to Wieringa, the guidelines to establish the research context are as follows-

1. **Knowledge goal(s)** - *What does the researcher want to know?*

For this thesis, the researcher wants to understand the landscape of stakeholders’ existing knowledge, the current methodology in practice, and the gaps that can be bridged with the support of the new methodology.

2. **Improvement goal(s)** - *Are there credible application scenarios for the project results?*

The improvement goal for the research is the proposal of a new methodology that can address the gaps and limitations observed in the methodology that’s currently being followed. There are credible application scenarios for the project results. The resulting methodology based on its usability can be adopted in forthcoming projects at the organization.

Research Method	Primary or Secondary	Rationale
Literature Review	Primary	To situate the research in an existing body of work and evaluate trends within a research topic.
Observation	Secondary	To understand how something occurs in real-world setting.
Interview/Focus Groups	Secondary	To gain more in-depth understanding of a topic from a stakeholder point of view.

Table 3: Summary of Data Collection Methods

3. **Current knowledge** - *Why is the research needed? Do you want to add anything, e.g. confirm or falsify something?*

The research is needed to tap into the potential of building ML models that are better and interpretable for businesses to take decisions and understand their insights better. By means of this research, the author also wants to confirm the need to develop a standardized methodology that businesses need to adopt to make their ML models more business-friendly and interpretable.

3.2 Qualitative Interviews

Interviews are one of the qualitative research methods that rely on asking questions in order to collect data. In [46], the authors say that *"The purpose of the qualitative research interview is to contribute to a body of knowledge that is conceptual and theoretical and is based on the meanings that life experiences hold for the interviewees."* For this research, interviews have been coupled with the process of data collection in order to provide the researcher with a well-rounded collection of information for analyses. The interviews involve two or more people, one of whom is the interviewer asking the questions and the other(s), the respondent. According to Daniel Turner in [47], there are several types of interviews differentiated by their level of structure. They are as follows:

- **Informal Conversational Interviews**

With this approach, the researcher and the respondent interact in an 'off the top of the head' style that allows for plenty of flexibility in the structure. Since there is no pre-defined structure for this interview format, questions are created on the go while keeping the central theme of the research in place. Although this format offers flexibility in carrying a conversation for both the researcher and the respondent, it suffers from a lack of structure as the answers received are mostly arbitrary.

- **Open-ended Interviews**

With this interview format, all participants are always asked identical questions but the responses are expected to be open-ended. Participants get the liberty to answer questions, express viewpoints and experiences in a way that they find comfortable, keeping them as detailed as possible. Since the nature of the interview is open-ended, researchers also have the liberty to frame follow-up questions and prompts based on previous responses of the respondents.

- **General Interviews**

This approach is more structured than the informal format in the sense that questions are pre-structured before the actual interview. The structure of the interview depends entirely on the

researcher conducting the interview. The interview environment allows the researcher to develop a rapport with the respondents and create follow-up questions based on a set of pre-constructed questions. Unlike the informal conversational approach, here, the questions are structured but can be adapted explore a more personal approach with the respondents.

Keeping the scope of the thesis, main research, and sub-research questions in mind, the open-ended standardized interview approach has been selected as the format to be followed. The reason is, the researcher wants to explore the benefits of having a certain amount of flexibility during the interview process. Moreover, since the stakeholders involved in the interviews are already known to the researcher beforehand, it is easier for the researcher to frame a general set of questions and then construct follow-up questions along the way as the conversation picks up. The conversational manner of the open-ended interview approach encourages both the researcher and respondents to engage in knowledge transfer in a detailed way.

3.2.1 Interview Design

The structure of the interview for the most part has been designed by taking inspiration from the three-step process by Creswell mentioned in [48]. It includes interview preparation, the construction of effective research questions, and interview implementation. Although the three-step process by Creswell is extensive and includes multiple sub-steps that can provide the researcher with the tools needed to conduct well-constructed professional interview with their participants, the interview structure conducted as part of the thesis require the following steps:

1. **Interview preparation-** The researcher should select the appropriate candidates for interviews who will be willing to openly and honestly share information or “their story”.
2. **Construct effective questions-** The questions should: a.) include wordings that are open-ended i.e respondents should be able to choose their terms when answering questions b.) should be as neutral as possible i.e devoid of phrases that might influence answers c.) be asked one at a time to ensure respondents feel at ease and give as much information as possible for each question
3. **Interview implementation-** a.) Ask for permission to record b.) Explain the purpose of the interview and establish the necessary context c.) Explain the format of the interview d.) Ask interviewees if they have any questions before they get started with the interview

The interview has been conducted in three parts and the interview questions can be found in the Appendix. Each part of the interview has a learning goal for which the questions have been framed accordingly.

Vision and Goals

The learning goal of the first part is to learn more about who the stakeholders are, their responsibilities, and what they are looking to get out of the research. Since the focus of the thesis revolves around understanding business goals, for each question, the stakeholders discuss responses that can help the researcher understand the essence of driving business efficiency.

Existing Knowledge

The learning goal of the second part is to assess what the stakeholders know currently, where the gaps in knowledge are and how this can be bridged as part of the thesis. For this, the state-of-the-art methodology followed by the organization is discussed along with the challenges and problems that they cross paths with.

Methodology (System under Development - SUD)	Stakeholder Description	Stakeholders in the Organization
Stakeholders directly interacting with the output of the methodology	Functional beneficiaries benefit from the methodology	End-users
Stakeholders in the wider environment of the methodology	A financial beneficiary benefits from the methodology financially, such as a shareholder or director of the company	Business Leaders
Stakeholders involved in the development of the methodology	The sponsor initiates and provides a budget for developing the methodology Important source of goals and requirements for the methodology	Business Lead, PMO, Project Lead
	Developers such as requirements engineers, designers, programmers, and testers build the system	Data Scientists

Table 4: List of Stakeholders

Unicorn Wishes & Wind Down

The learning goal of the third and final part of the interview is for the researcher to identify and dive into potential opportunities. For this, the questions were more focused on what the stakeholders expect of the outcome and how the to-be methodology can arrive at that.

3.2.2 Stakeholder Analysis

According to Design Science Methodology for Information Systems and Software Engineering by Wieringa [11], a stakeholder of a problem is defined as a person, group of persons or institution affected by treating the problem. Stakeholders are the source of goals and constraints of the project, which are in turn the source of requirements in the treatment, so it is important to identify relevant stakeholders. The System Under Development (SUD) in this context is a methodology. To collect data and understand the existing practices followed to develop the methodology, relevant stakeholders need to be identified.

It is also important to note that stakeholders may have different levels of awareness of a problem and its possible treatments. At the highest awareness level, a stakeholder is completely aware of the problem context, the need of a treatment and is actively involved in the efforts to develop a treatment to address it. At a lower awareness level, a stakeholder is not aware of the problem nor of the need of a treatment. At the lowest level of awareness, a stakeholder is aware of an improvement possibility but is not interested in actually carrying out the improvement. From Table 4, it can be concluded that Business lead, PMO, Project lead and Data scientists are the stakeholders that belong to the highest awareness level as they are the ones actively engaged in the process of developing the methodology i.e. SUD. Fig 15 gives a visual representation on how the stakeholders interact with the proposed artifact. While the data scientists are actively involved in the solution building process, the business lead and PM support and facilitate the project. The end-users are stakeholders who benefit from the outcome and business leaders in the outermost circle are the decision-makers.

Sampling Method

Creswell [48] asserts that while selecting relevant participants for the interview, the researcher should utilize one of the various types of sampling strategies such as purposeful, criterion-based, critical case

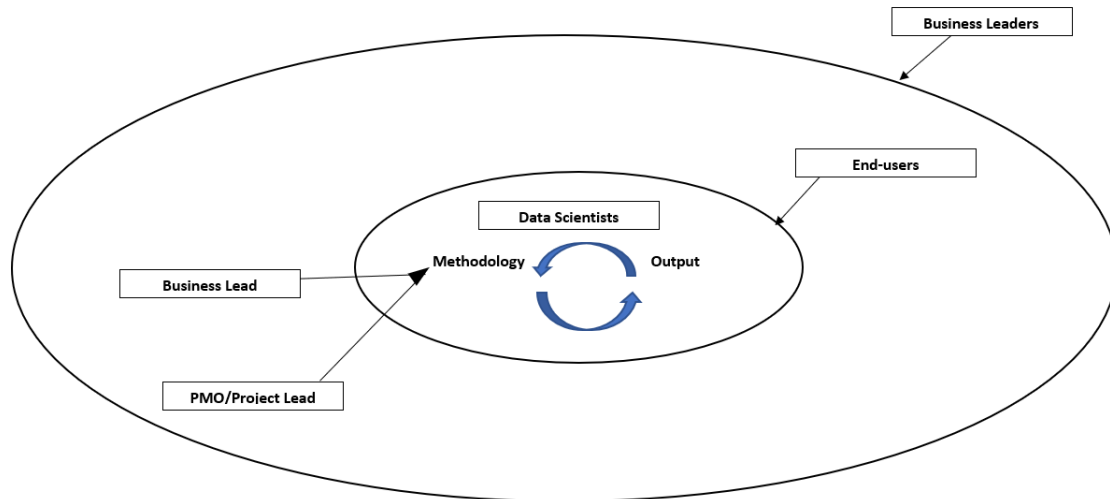


Figure 15: Stakeholders interacting with the Methodology(SUD)

sampling among others in order to obtain qualified candidates that will provide the most credible information to the study. For the purpose of this thesis, the purposeful or selective sampling technique has been selected. According to Patton [49], the logic and power of purposeful sampling lie in selecting information-rich cases for study in depth. Information-rich cases are those from which one can learn a great deal about issues of central importance to the purpose of the inquiry, thus the term purposeful sampling. Studying information-rich cases yields insights and in-depth understanding rather than empirical generalizations.” Qualitative researchers use this technique to recruit participants who can provide in-depth and detailed information about the phenomenon under investigation.

While considering the sample size of the research, several considerations have been looked into that are subjective to the goals of the research. For this research, a sample size of 4 has been chosen by taking inspiration from two of the seven factors proposed by the authors in [50] that can possibly affect the sample size. The first one being, the population considered for this research is relatively heterogeneous given the fact that the inputs of multiple stakeholders across cadres, both business and technical, need to be considered while building the methodology. This also helps in identifying stakeholder goals and benefits that can be derived as a result of the research. Secondly, keeping time and budget constraints in mind, it has been decided to not increase the sample size, rather make the most use of gathering information from the selected sample size.

3.2.3 Synthesis of Results

This section explains the results obtained as part of the researcher’s observation and initial qualitative interviews conducted to understand the landscape of stakeholders’ existing knowledge, the gaps identified as part of the existing methodology, the motivation and objective to develop a new methodology and what the proposed methodology should contain.

Qualitative Data Analysis

According to [51], validity and reliability relate to quantitative studies and is now slowly making the paradigm shift towards qualitative research as well. However, qualitative studies also rely on establishing a measure of ‘trustworthiness’. This can be established by triangulating data through

Stakeholders	Responsibilities	Observations of Researcher
Data Scientist I, II	Clean the data, pre-process and eventually build the model. Get the approval from business stakeholder if the data is good enough to be fed as input to the model.	<p>There have been gaps identified with how the stakeholders view the big picture:</p> <ol style="list-style-type: none"> 1. Data Scientists are highly focused on the challenges that come with the data. In terms of not having a single source of truth within the organization, access control and management to reports and data sources, understanding of the business such as the semiconductor ecosystem, fabrication process of wafers, etc. 2. Project Manager is focused on the strategic direction of the outcome. The project manager believes that the technical advances may come and go but what's important is engagement, trust and change management that need to be established across the organization. 3. The tech lead is highly focused on automating human touch points within the organization. The focus is more towards operationalizing ML by adopting best practices and understanding business and data tradeoffs within the organization.
Project Manager	Facilitate, direct, assist whenever required and break down challenges for the team. Create a network to support data scientists in finding the right sources within the organization for domain-related clarifications.	
Senior Applied Data Scientist (External Consultant)	Tech lead for the team. Understands both the business and technical standpoint.	

Table 5: Observations of the Researcher from a Stakeholder Standpoint

different sources and/or different kinds of data, and then consolidating the analysis to check if they provide a consistent meaning or interpretation in relation to the research question. To extract insight from the knowledge collected as part of the interviews conducted, a part of the qualitative data analysis guide by Dey [52] is used. The approach includes the process of reading and annotating transcripts. Through this step, the researcher reads the interview transcripts, annotates important points that are relevant to the research goals and context. This is closely followed by categorizing the data. Since the interview itself has been categorized in three different parts as mentioned in Section 3.2.1, this step is fairly easier to analyze. Lastly, the collation and categorization of information, confirms the need for building a methodology for the organization. The following sub-sections categorize the findings from the interviews and then summarizes the need to develop a methodology.

(a.) Stakeholder Understanding and Gaps Observed

There are four stakeholders in total who volunteered to participate in the interviews. The interview has been guided in a way that the researcher is able to extract as much information as they can from the stakeholders regarding their role and responsibilities in the organization, their areas of focus and how they navigate them in their projects. The researcher has identified gaps in the responses and categorized them on the basis of their areas of focus. Based on the observations, it is commonplace for data scientists to be highly focused on data sources, quality and how they can best optimize the

data to get accurate model outcomes. Similarly, business representatives are more focused on being able to drive efficiency with the results obtained and are not highly invested in solving data-related issues. Such gaps could lead to losing sight of the overall objectives that are tied to the business goals. The gap between business and data stakeholders increase as the idea moves from top executives all the way down to data-focused stakeholders. The back and forth between the data team and business executives might lead to misconstrued definition of objectives, assumptions and context that could potentially decrease the efficiency within the organization. The results can be seen in Table 5.

(b.) Shared Objectives

The four stakeholders believe that business success would be measured in terms of usage of the model outcome i.e. if the end-users find the outcome of the model beneficial to make decisions and how it can in turn drive overall efficiency within the organization. The true success of the model outcome is when a business stakeholder who hasn't been particularly involved in the solution building process can still interpret and understand the outcomes to make actionable decisions.

They are also of the shared belief that the intervention of ML technologies can never replace the experience of humans. They agree on the fact that it will never be fully automated, rather a source of support as human translators within the ongoing decision-making process. Two of the respondents also highlighted the fact that the combination of the experience of business stakeholders along with the data-driven insights from the model can also facilitate building trust easily among top-level business executives. Having a human facet constantly involved in the decision-making process brings about ease in understanding the results better, deriving insights, and making better business decisions.

The stakeholders foresee challenges for the business in terms of operationalizing ML and getting the results integrated into existing day to day processes within the organization. They fear that business leaders can become overwhelmed by the pace of change with the adoption of ML model-based decision making. Another challenge is getting business executives and leaders to trust the outcomes that the model generates so that as an organization they can focus on the wins.

Need for a methodology

Stack Exchange [53] defines framework as a loose but incomplete structure that leaves room for other practices and tools to be included. A methodology on the other hand is defined as a set of principles, tools and practices to guide processes to achieve a particular goal. Simply put, frameworks double as a structured problem-solving approach so that stakeholders can focus on aspects of a problem. Methodologies on the other hand are stringent documented approaches intended to accomplish an outcome. Among consultants in particular, frameworks are more popular as they are fluid and can be modified on the go. The disadvantages of adopting a framework within an organization could be the possibility of process defects, inconsistency and lack on standardization in outcomes, difficulty in enforcing performance and compliance metrics, etc.

Since methodologies are more systematic and structured, it has been decided by the researcher that the scope of this thesis will result in the development of a methodology. The methodology does not stop with simply presenting a high-level outline of the steps that a business needs to adopt but also the rationale behind the steps to achieve a desired and actionable outcome.

As a result of the literature review and qualitative interviews conducted, there is a need to develop a methodology that can help build ML models that are interpretable for the business stakeholders. Some commonly used ML algorithms such as decision trees and rule-based algorithms offer inherently and methodologically transparent outcomes. However, with the existence of a plethora of ‘black box’ ML models, such as deep neural network, tree and network ensembles, support vector machines among others, which do not provide embedded interpretability, it is difficult to translate their outcomes to business stakeholders who often lack the technical understanding of the logic underpinning these models. For the purpose of high-stake decision making in organizations and to arrive at a strategic direction, business leaders need to be in the know of how the models make a certain decision and why. The remainder of the chapter will drill-down on the development of a standardized methodology that can encapsulate the business logic, understanding, sensitivity and impact along with the technical drivers that the models possess.

3.3 Summary

The chapter explains the qualitative approach carried out by the researcher to understand the scope of the research objective, identify existing knowledge and improvement goals that need to be achieved as a result of the research. It also provides readers with a detailed view of the qualitative interviews conducted with relevant stakeholders and the findings that have been documented as a result of the observations. The results of the data collection process serve as the foundation for the design and development of the methodology to meet the research objectives.

4 Design and Development

The results obtained as part of Chapter 3 are used to design and develop the methodology. This chapter first discusses the existing methodology in use, the gaps and limitations observed, and then details the design of the newly proposed methodology. The design of the proposed methodology is then further broken down as per the various phases that are integrated within it.

4.1 Methodology Design

As mentioned in Section 3.1.1, the researcher will first map out the existing methodology that the team and organization follow at present with the information collated as part of the literature review, observations, and interviews. This section will first briefly describe the conceptual model of the existing methodology and then proceed to explain the gaps and limitations that have been identified.

4.1.1 Existing Methodology

The core concept of the existing methodology in practice has been inspired by the Cross-Industry Standard Process for Data Mining (CRISP-DM), a methodology that focuses on aligning data mining processes with business goals. It is an iterative approach with opportunities to evaluate the progress of the project and ensures business goals remain the core of the project rather than an afterthought. In addition to this, it serves as a roadmap by offering best practices for data engineers, analysts, and business stakeholders to adhere to and carry out a project [54]. In practice, many of the tasks can be performed in a different order and it will often be necessary to backtrack to previous tasks and repeat certain actions. Hence, the sequence of the phases is not adhered to in that specific manner. However, the arrows indicate the most important frequent dependencies between phases but are subject to change in a particular project based on the outcome of each phase.

[6] says the CRISP-DM model can be seen as a hierarchical process model with four levels of abstraction: phases, generic tasks, specialized tasks, and process instances where the phases and generic tasks are intended to cover all possible data mining situations for a given context. The generic tasks are designed to be complete and stable - complete to cover the whole process of data mining and all possible data mining applications while stable means that the model will be valid for unforeseen developments like new modeling techniques. The specialized task goes one level further and explains how the actions in the generic tasks should be carried out in specific scenarios. The fourth level, process instance, is more of a record of actions, decisions, and results of the whole project.

The lifecycle of a typical data mining project as specified in CRISP-DM can be seen in Fig 16 with 6 phases.

1. **Business Understanding:** The first phase focuses on understanding the core problem, objectives, and requirements and aligning them with the strategy and goals of the business. This phase aims to get a hold of the business perspective and then convert the knowledge into a data mining problem. A detailed plan can also be created during this phase.
2. **Data Understanding:** The second phase revolves around getting familiar with the data, discovering insights, and checking the quality. Usually, the business and data understanding phases work alongside each other so that the business context of the data can be obtained for data scientists to get a clearer idea.

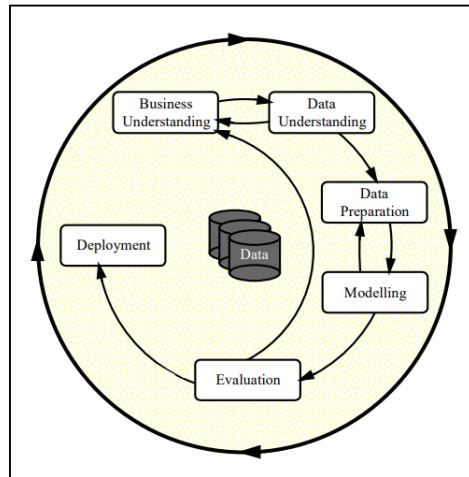


Figure 16: Conceptual Model of CRISP DM [6]

3. **Data Preparation:** In this phase, the raw data will be manipulated into a form that can be analyzed and used as input to the ML model. This phase aims to construct the final dataset that will feed into the to-be-built model. Data preparation tasks such as attribute selection, data cleaning, construction of new attributes, and transformation of data are performed in no particular order.
4. **Modeling:** This phase is to select various modeling techniques to be applied to the prepared and pre-processed data. During this phase, the parameters are calibrated to optimal values to fit with the model specifications. Often, data scientists may have to backtrack to the data preparation phase while modeling due to constraints and problems with the data.
5. **Evaluation:** Prior to deploying the model, it's important to evaluate the model by reviewing the steps executed and checking if business objectives have been met. This phase determines if there are gaps that the model fails to bridge while aligning with business goals. The aim of this phase is to arrive at a decision that the stakeholders are satisfied with to proceed to deployment.
6. **Deployment:** In the last phase, the models are pushed into production by running them on a live environment. Depending on the deliverable requirements, the deployment phase can involve generating an interactive dashboard or implementing a repeatable data mining process itself. This makes the model's outcomes available to users, developers or systems, so they can make business decisions based on data, interact with their application and so on.

4.1.2 Observations

More often than not, while embarking on a project that relies heavily on data and analytics, data engineers and analysts tend to lose sight of the overall business goal. While it's implied that data-centric approaches can unearth questions that weren't explored previously, innovative solutions to address them and opportunities that could drive efficiency, it is also easy to get lost in the vastness of it. This holds true for a typical ML solution building process since most of the phases are iterative in nature. This means, we often need to review the business goals, KPIs and available data from previous steps to adjust the outcomes of the ML model results. While CRISP-DM enables data scientists to facilitate model building in an iterative yet flexible manner, it still has its many limitations. The results of this section have been consolidated on the findings from the literature review, observations within

the team and discussions with relevant stakeholders. The gaps and limitations of the methodology are described in detailed below:

- Although CRISP-DM is positioned as an agile-focused, iterative process model for data scientists to make use of, many within the data community view CRISP-DM as a rigid waterfall process. This can be explicitly seen in the business understanding phase that mandates “the project plan contains detailed plans for each phase” which is the trademark of a typical waterfall approach that requires detailed, upfront planning. With advances in leveraging tools and techniques for big data, there can be additional effort spent during the data understanding phase, for example, as the business grapples with the additional complexities that are involved in the structure and shape of these data sources. CRISP-DM with its waterfall-like approach and slow iteration, despite having a flexible way of moving between phases, can fall short of accommodating complex projects.
- CRISP-DM lacks the human component in its workflow. This is a setback in the process model as stakeholder identification and mapping respective responsibilities is the foundation of any successful methodology. By not taking into account the human component of a project, the methodology fails to align with the objectives of building a solution that’s focused on stakeholder goals, objectives and in turn improving business efficiency. This also neglects the aspects of informed decision-making which should ideally be the core of the methodology that needs to be arrived at. What’s far more common in any process model is often the misalignment between the expectations set at the top of the organization by business leaders and the foundation of what the data scientists can realistically deliver. The methodology fails to create the first-level understanding of what the project entails and how stakeholder objectives and expectations can be mapped.
- In continuation to the previous point, since CRISP-DM lacks the focus on stakeholder goals and understanding, there is almost no focus on testing the level of interpretability for an ML model. This can give way to the likelihood that the team will build an impressive analytic solution that will not add business value. When developing an ML model, the consideration of interpretability as an additional design driver can improve its implementation for inducing stakeholder trust and satisfaction, in turn leading to better business decisions.
- CRISP-DM is a process model that’s strictly limited to data scientists, data engineers, etc. However, with the advancement in cross-functional way of working across teams and functions in the organization in addition to the need to involve stakeholders since the inception of a project and initiative, it suffers from a degree of stakeholder inclusivity for better decision-making in businesses.

4.2 Inclusive & Interpretable Cross-Industry Standard Process for ML in Business (iCRISP-ML)

ML models must be adaptable to a changing business environment and data scientists developing them should be cognizant of the business objectives that drive the model building process. Moreover, there needs to be a standardized methodology that is all-inclusive and captures objectives and requirements from business and technical standpoints and stakeholders. This warrants the requirement of not only building successful ML models but making sure they’re easily interpretable enabling better business decisions. While CRISP-DM focuses on data mining exclusively, it does not cover the

application scenario of ML models inferring real-time business decisions. Secondly, the core concept of stakeholder understanding and the need for interpretability is dismissed in CRISP-DM.

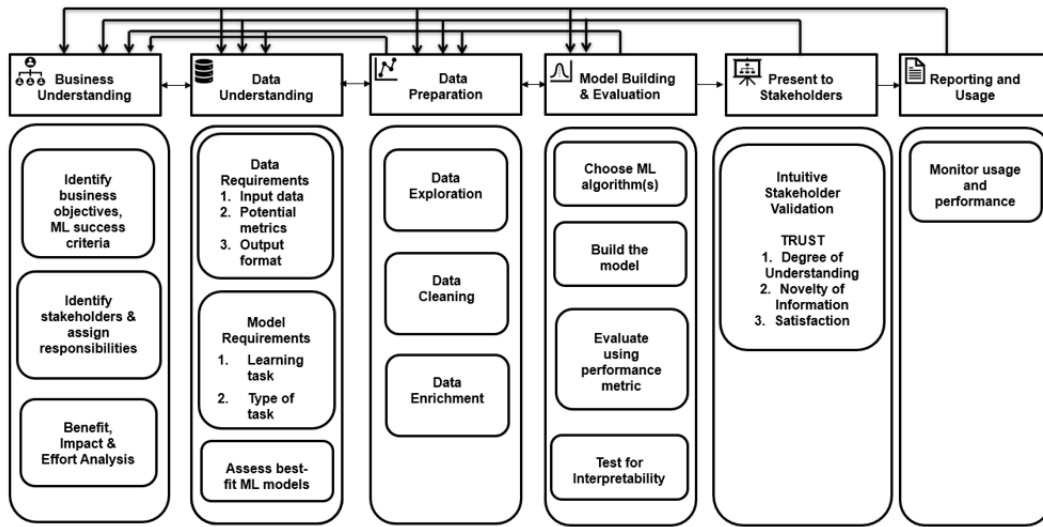


Figure 17: Proposed Methodology: iCRISP-ML

Based on the findings collated and observed as part of the literature review and qualitative interviews, the proposed methodology i.e. iCRISP-ML (Inclusive and Interpretable Cross-Industry Standing Process for ML) has been designed and developed. As seen in Fig 17, the iCRISP-ML is a six-phase methodology that resembles the CRISP-DM process model. However, the phases have been restructured and tailored to fit the problem context, keeping in mind the gaps and limitations that need to be bridged. It emphasizes the need for incremental deliverables and provides feedback loops between phases and steps to backtrack whenever required.

4.2.1 Objectives and Goals

The presented work relies on the core concept that interpretability should not be considered as an afterthought, rather, a core component that is incorporated from the beginning with the cooperation of all relevant stakeholders. This makes it easier while presenting final outcomes to stakeholders during a report-out so they can quickly grasp the core concepts of the project, objectives, scope and assumptions to eventually make informed-decisions. iCRISP-ML aims to serve as a detailed step-by-step guideline for the process of building ML models in a business context. At each stage of the solution building process, a sufficient degree of interpretability is defined on the basis of the stakeholders involved so that everyone is constantly aware of the bigger picture and what the objectives are right from the inception of the project. The methodology is consistent with the best practices of project management in the context of ML and the business in itself. One of the main objective of iCRISP-ML is to stay inclusive to all stakeholders and not just limit its workflow to data engineers and data scientists. This is ensured right from the requirements collection phase and participants establish activities for each stakeholder group at each phase to achieve a sufficient level of interpretability. Moreover, the methodology also aims to present itself as an industry and application-neutral methodology that can be adopted cross-functionally across various business domains.

The main objectives and goals of iCRISP-ML are as follows:

- Establish shared understanding across key stakeholders about the solution and its intended use i.e. the relevant stakeholders involved in the phases of the project are kept abreast of what the project is intended to achieve for the business and how it will drive impact.
- Build stakeholder trust in solution outputs i.e. the methodology focuses on including all relevant stakeholders throughout the lifecycle of the project. This makes it easier to build trust among stakeholders in solution outputs and make better business decisions.

4.3 Phases of the Methodology

This section details the six phases of iCRISP-ML. Each phase will first have an explanation on its purpose within the methodology, the objective to achieve interpretability along with stakeholder goals and then discuss each step within the phase.

4.3.1 Business Understanding

Understanding the requirements of the business is of paramount importance. While initiating any project it is important to first identify the scope of the project, the success criteria, stakeholders involved, etc so as to avoid potential scope creep and business project failure. This is carried out by gathering success criteria along with business, machine learning, and economic success criteria. It does not include a finalized project plan step as that would make it look like a rigid waterfall approach. Instead, this phase is open to being fluid and includes only high-level details that are important to make decisions and progress towards the next phases of the project lifecycle.

Overall goal

The objective of the first phase is to ensure the feasibility of the project and how it can impact the efficiency of the business. This phase is key to determine the drivers for the overall project success and to build an explainable and interpretable model. The business understanding phase covers the following activities:

1. Identification of business objectives and success criteria for ML models
2. Identification of project sponsors and key stakeholders
3. Assign roles and responsibilities to identified stakeholders
4. Estimate costs, benefits and effort associated with the success of the project

Stakeholder goals

All stakeholders involved need to be able to:

- Gain an understanding of project objectives, scope and assumptions tied to the overall business goal
- Gather requirements, document them as a project plan

Question to ask	Description	Possible Stakeholders
<i>Who is affected by the project?</i>	To identify stakeholders who are directly and/or indirectly impacted by the outcome of the project.	<ul style="list-style-type: none"> • End-users who directly interact with the outcome of the project • Financial beneficiaries: Business leaders, directors, shareholders of the organization
<i>Who is involved in the development of the project?</i>	To identify stakeholders who are involved in the development or implementation required for project success.	<ul style="list-style-type: none"> • Developers such as requirements engineers, designers, programmers, and testers if there is a system to be built. • Sponsors such as project/product manager initiate and provide a budget • Consultants to facilitate the development of the project

Table 6: Questions to identify stakeholders

(a.) Determine business objectives and success criteria for ML models

The primary activity is to first identify the rationale behind the project and why it's important for the business. It is also important to go into a high-level detail on the issues associated with business goals by means of fact-finding and coming up with a detailed explanation of the business and strategy goals. It is commonplace to first start gathering background information about the current business situation, documenting specific business objectives decided upon by key decision makers and agreeing upon pre-defined criteria to determine ML-related success from a business perspective.

Success criteria for ML

Success criteria for ML in a theoretical and qualitative approach can be quite tedious to identify in the early phases of the project. Hence, it is recommended to identify a Minimal Viable Product (MVP) [55] that can help stakeholders achieve an acceptable level of performance to meet the business objectives.

(b.) Identify stakeholders

A stakeholder in any project is identified to be an individual, department or organization that may be affected by the results of a project or have an effect on how the project needs to be carried out. Identifying stakeholders determines the relationship they have with the project and business objectives. The ideal way is to identify the stakeholders in the early phases of a project and manage them throughout the entire project life cycle. By identifying them earlier, they're given the freedom to facilitate effective participation throughout the project and create better perspectives. The questions that can help identify stakeholders easily are summarized in Table 6.

(c.) Assign responsibilities using RACI matrix

Once the key stakeholders have been identified, it is also important to assign responsibilities so that the stakeholders involved are aware of their roles in the lifecycle of the project. This is important to clarify roles and eliminate confusion between stakeholders. Organizations can also keep projects on schedule, transition between project handoffs and prioritize communication between teams internally and cross-functionally.

Design of RACI Matrix

The RACI matrix is recommended to map out tasks, milestones, key decisions involved in completing a project and assign stakeholders to each action item. The acronym RACI stands for the four roles that stakeholders might play in any project - Responsible, Accountable, Consulted and Informed. If the matrix is defined during the early stages of the project, the stakeholders are aware of the roles and responsibilities they need to take on through the lifecycle of the project. The standard design of RACI matrix for the purpose of this thesis has been slightly modified as show in Table 7. The project stages along with the relevant stakeholders are outlined as titles.

The project stages are further broken down into three categories. They are as follows:

1. **Planning and Definition** - This phase defines two stages of the methodology - Business and Data Understanding. It is important for the stakeholders involved in these stages to dive deep into understanding how the business objectives of the project are aligned with the overall strategy and goals of the organization. Moreover, the plan and definition also applies to the data they're working with for the project, the tasks that are necessary to meet the objectives, etc.
2. **Execution** - This phase defines the stages of the methodology where stakeholders need to be involved in the actual data-centric implementation tasks such as.
3. **Evaluation and Closure** - This phase concludes the lifecycle of the project where in the results generated are presented to high-level stakeholders and the deliverables are handed over to the respective end-users.

The four roles that stakeholders might play in any project include the following:

Responsible: These stakeholders are responsible to do the work or implementation required as part of the project. They must complete the task, meet objectives ad also make decisions. Every task should have at least one responsible stakeholder. Example: Developers, data scientists, etc.

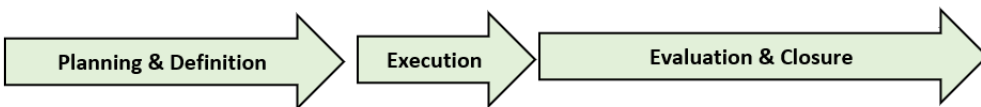
Accountable: These stakeholders are essentially considered as the 'owners' of the task or objective who delegate and review the progress of the project. They monitor the tasks, decisions and must sign off or approve when the task, objective or decision is complete. They need to also ensure the other stakeholders involved know the expectations of the project. Every task should have only one accountable stakeholder. Example: Project managers

Consulted: These stakeholders are active participants throughout the lifecycle of the project and are constantly kept updated on the progress. They don't necessarily have a task assigned to them but the outcome of the project will have an impact on their work and role in the organization. Throughout the lifecycle of the project, they provide feedback and input on the work being done. Example: Business lead, business analyst, etc.

Informed: These stakeholders need to stay up to speed with the project progress but not formally consulted or overwhelmed with the granularity of details. There are no dependencies on these stakeholders for the tasks throughout the project rather double as decision-makers once the project is completed. Usually, these stakeholders are positioned outside of the core project team. Example: Director, shareholder, business leaders, management board, etc.

(d.) Benefit, Impact and Effort Analysis

An estimation of the initiative's direct business value to an organization brings in many insights.



Project Stage & Stakeholders	Business Understanding	Data Understanding	Data Preparation	Modeling & Evaluation	Present to Stakeholders	Reporting and Usage
Business Leaders						
Business Lead						
PMO/Project Lead						
Data Science Engineer						
End-Users						

Table 7: RACI Matrix to assign stakeholder responsibilities

These insights might be reflected in terms of acquiring new customers, retaining existing customers, the ability to upsell customers, and the anticipated new revenue the initiative will bring. Typically, in any business case, it is important to analyze critical parameters such as costs, benefits, business impact and effort required. This process not only helps with determining the economic benefit of an initiative but also avoids the possibility of bias in decision-making. Designing a standardized way of doing this can be beneficial while developing a new business strategy, making resource allocation or purchase decisions, comparing investment opportunities, etc.

Initial Steps

1. Understand the overall business goal that needs to be achieved through the realization of the initiative or the project.
2. Parameters to be considered:
 - Stakeholder engagement
 - Team competence and availability - Tentative timeline (in days, weeks, months)
 - Potential risks (Business and/or Technical)
 - Outsourced skills, if required (ML experts, consultants)
3. Tabulate the potential benefits, measures to evaluate them and respective benefit owners mapped against each of them. This gives insights into how the project outcomes could potentially bring in benefits from an organization point of view.
4. For each stage of the project, estimate the FTE of each stakeholder involved in the project.

Team Evaluation

The initial steps will help with high-level clarity on what the expectations are for the project and how this can impact the business in the long run. Based on the results, the project can be positioned in one of the four quadrants. As shown in Fig 18, the four quadrants are positioned across 'Impact' and 'Effort' as X and Y axis respectively.

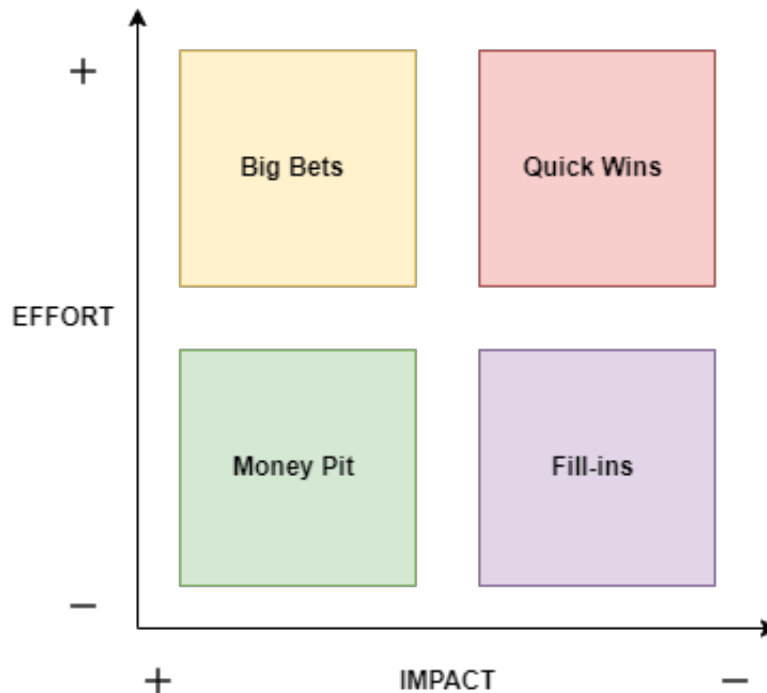


Figure 18: Impact-Effort Matrix Template

- **Big Bets (High Impact, High Effort)** - Focuses on long-term returns with more investments going into the initiatives
- **Quick Wns (High Impact, Low Effort)** - This quadrant focuses on initiatives that can bring in maximum benefits to the organization without having to invest much
- **Money Pit (Low Impact, High Effort)** - Initiaitves that require lot of investments but bring in very less benefits. Typically, these projects are prioritized less as it doesn't prove to be beneficial in any way.
- **Fills ins (Low Impact, Low Effort)** - Initiatives that fall under this category are very budget-friendly and doesn't cost much to implement. However, they also come with cons where in the returns are not as high as expected.

4.3.2 Data Understanding

The initial step in building a machine learning model for the business is to understand the need for it in the organization. Since the development process by itself can be resource intensive, there is a need to set clear objectives from the start since a deployed model will bring much more value when it's fully aligned with the objectives of the business. The data understanding phase works alongside the business understanding phase and steers the focus in a direction to ascertain, assemble, and scrutinize the data sets required to realize project goals. While the business understanding phase establishes the stakeholders involved, the intended problem the project aims to solve and the definition of project success, the data understanding phase goes one level deeper to define the type of problem the model will need to solve, the source of input data and whether it is of sufficient quantity and quality, the format in which the output will be finally presented to the end-users, etc. It's understood that data

guides any business process which includes data collection and data quality verification to achieve business goals. An overview of the steps involved in this phase can be seen in Fig 19.

Overall goal

The key to establishing trust among stakeholders is by giving a clear and detailed understanding of the requirements collection, determining the data to be employed on the model and the type of implementation that's going to follow. Hence, the objective of this phase is to document statistical properties of the data and the data generating process in itself. This doubles as a foundation for data quality assurance during the operational phase of the project.

Stakeholder goals

All stakeholders involved need to be able to:

- Gain a high-level understanding of ML solution building process and potential outcomes
- Support data scientists to gather data requirements, establish assumptions, define constraints, etc
- Gain a high-level understanding of domain aspects relevant to the goals and objectives of the project

(a.) Data Requirements:

First, the requirements related to data are collated and documented. (i) Input data - Assess if they are reliable, of suitable quality and representative of real-world data

(ii) Potential metrics and solution format - Assess if they are consistent with the project goals in terms of accuracy, format, ease of understanding for the end users, level of potential business insight, and their validity from the ML and business perspectives

(iii) Output format to present to the end user - Tables, visualizations, graphs, dashboards, etc.

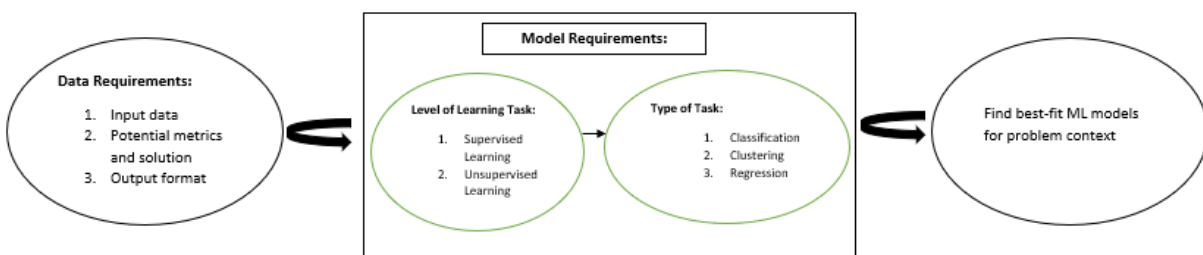


Figure 19: Steps in Stage 2: Data Understanding

(b.) Model Requirements:

Next, the requirements related to the ML model such as the type of learning task required for the problem is collated and documented. With the initial exploratory data analysis performed in the previous phase, the data scientists already have an initial understanding of the dataset, including its features and attributes.

(i) Learning task - The learning task categories considered within the scope of the thesis are Supervised Learning and Unsupervised Learning. Stakeholders need to make a choice between the two

depending on their use case and requirements. Supervised machine learning models require labeled datasets that have been prepared by a data scientist which means the training dataset will include input and labeled output data. The model then learns the relationship between input and output data so that the ML model can predict outcomes and classify new data. Unsupervised learning tasks unlike supervised learning do not require labeled data.

(ii) Type of task - The learning task is further categorized into classification, clustering and regression. Stakeholders need to make the choice of which type of task needs to be employed for their context.

(c.) Assess best-fit ML models for the context The type of machine learning algorithm chosen will be based on an understanding of the core data and the problem that needs to be solved. This step is essential to choose a high-level modelling approach, test its validity and whether it is proven and likely to work in this industry. Typically, data scientists spend some time analyzing best practices followed to understand which ML model(s) could potentially work for the use case at hand.

4.3.3 Data Preparation

Data audit, exploration and cleansing play a key role in the development of stakeholder trust in the approach and ultimately in the solution to achieve stakeholder trust. The reason being introducing error-prone data in the model building phase could introduce unnoticed and skewed output values. The ML model learns from the data so poor quality training data quality may mean the model is ineffective once it is deployed. Therefore, it is important that the data is checked and cleaned to ensure a standardized format and missing data if any should be detected along with outliers. The objective of the stage is to develop the final dataset from the initial raw data for the purpose of modelling. The steps involved in this stage are likely to be performed multiple times and not in any prescribed order. *The steps involved in this stage remain unchanged as seen in ??.*

Overall goal

The goal of this phase is to ensure that data scientists and accompanying stakeholders are able to trust, understand, and ask better questions of their data, making insights more accurate, meaningful and trustworthy.

Stakeholder goals

Business stakeholders (Business Lead, Project lead, PMO) need to:

- Provide domain and business related knowledge to help data scientists understand the business context of the data
- Consult and clarify clarification of any domain-related aspects
- Support with identifying data sources for enrichment

Data scientists need to:

- Develop the required understanding of business data and conduct data quality checks
- Assess the potential of the data for modelling purpose

Data Exploration

During the data exploration step, an in-depth analysis and visualization of the dataset is performed to uncover insights and assess whether the initially identified data is sufficient to achieve the business goals. This is typically considered as the first step that the data scientists perform to identify significant relationships among the data.

Data Cleaning

Regardless of how sophisticated and high-performing the ML algorithm is, good results cannot be obtained from unclean data. This step is essential to verify the quality of the data that is being dealt with. It is important to identify missing values, attributes, data types, duplicates, etc. In addition, this step also involves splitting the data into training and test sets. To ensure an accurate outcome, ML models generally need large arrays of high quality training data since the model learns the relationship between input and output data from the training dataset.

Data Enrichment

In the event that initially identified data needs to be enriched by external data, additional analysis can be performed by adding new data (both internal and external data sources). The new data is extracted, audited, cleansed and added to the previously used data. This step is repeated until the necessary level of clean data is achieved to progress to the next stage or, if it has been established that achieving it is impossible, this finding is further discussed with the key stakeholders and the relevant decisions are made.

4.3.4 Modeling and Evaluation

This phase of the methodology entails selecting the modeling technique that needs to be used, followed by the actual build of the ML model, the evaluation using a statistical performance metric and the test for model interpretability. The translation to the ML task depends on the business problem that we are trying to solve which means the constraints and requirements identified as part of the business and data understanding phases will set the foundation for this stage.

Overall goal

The goal of the model building and evaluation phase is not only to merely build an accurate model following the best practices of modeling in ML but also evaluate its level of interpretability so that business stakeholders can better understand and grasp the nuances of what the model is trying to convey.

Stakeholder goals

Business stakeholders (Business Lead, Project Lead, PMO) need to:

- Support data scientists by engaging in regular consultations to clarify domain-related aspects
- Provide feedback and domain insight

Data scientists need to:

- Select the ML technique to be used for modelling
- Build the model and evaluate its performance - both statistically and on the basis of its interpretability level

Choose ML algorithms

On the basis of the list of best-fit ML algorithms that have been previously determined as part of the data understanding phase, the data scientists will narrow-down on the which algorithm to experiment with depending on the problem context. It is crucial that the right technique is chosen to balance the required outcome and level of interpretability that needs to be arrived at.

Build the model

Once the data is in good shape and the algorithms have been selected, the data scientists can get started on building and describing the model, its special features, behavior and interpretation. During this step, the initial model parameters are chosen and the selected techniques are run on the input dataset. This step is iterative as it involves a lot of experimentation and discovery from selecting the most relevant features to testing multiple algorithms. Until the best model is arrived at, this step follows an iterative pattern.

Evaluate using performance metric

At this step, the data scientists will perform actions that ensure the final model is as good as it can be. The models will be assessed based on pre-determined performance metrics and the values will be tabulated or visualized to find the best performing model. Model parameters that affect model performance will also be checked to eventually test the performance of the final model. More often than not, during this phase, the performance of the trained model needs to be validated on a test set.

Test for Interpretability

This is the final step that is implemented before presenting the results to the stakeholders. Here is where the model is tested against various approaches to check for the level of interpretability it brings about in its outcomes. The primary step in getting started with this phase is identifying the model interpretability which is the level of interpretability of the model used to solve the original task of the system. A model is interpretable if it gives rise not only to mechanistic understanding (transparency) but also to a functional understanding by a human (Paez, 2019). By the following choices, data scientists and others involved can first identify the level of interpretability in the model.

- **Intrinsically interpretable model or ante-hoc model** - The models that fall under this category can be understood as a whole, i.e., a human can digest the results in a comprehensive manner without having to require additional support. Alternatively algorithmic transparency is considered, which means the model is mathematically understood. For example, in the case of a linear model, the regression weights is model-specific and requires understanding of model internals such as weights and structural information. Some of the commonly known intrinsically interpretable models are decision trees, bayesian networks, naive bayes model, linear & logistic models.
- **Post-hoc models** - These models use a helper model from which to derive an explanation. The helper model often also called as surrogate or proxy model can completely simulate the behavior of the main model or approximate certain aspects. While working with post-hoc models it is also important to understand how explanations can be provided and how much of it should eventually be presented to the target audience.

Once the data scientists have identified it is a post-hoc model that needs further support using helper models, then they can begin to employ interpretability techniques. To compartmentalize this process for ease of use, this step is further broken down into 3 categories namely: (i.) Input (ii.) Explanatory

medium (iii.) Presentation. Each of these categories detail the necessary conditions that need to be looked into while determining the right technique to employ to bring about interpretability in the model outcomes. An overview of this process can be seen in Fig 34.

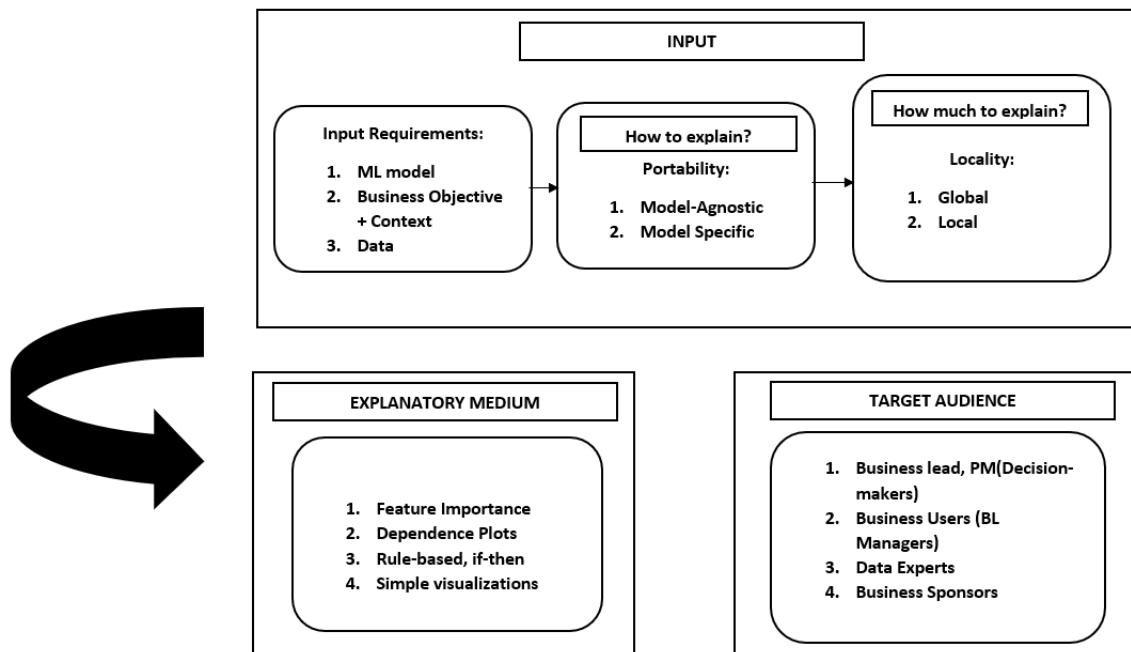


Figure 20: Test for Interpretability

Input

The first part of this three-step process details the input requirements and first-level understanding of the techniques that need to be employed on the model that is trained and ready to be interpretable.

(a.) Input Requirements The primary input requirements are the ML model itself, business context of the problem the interpretability technique aims to solve and the data sources. Since the model is already trained at this point, these requirements are just collated to re-establish the overall objective that needs to be achieved.

(b.) How to explain?

Model-Agnostic - As the name suggests, these techniques can be used on any model and are applied after the model has been trained. These agnostic methods usually work by analyzing feature input and output pairs. By definition, these methods cannot have access to model internals such as weights or structural information.

(c.) How much to explain?

- **Global** - This level of interpretability is about understanding how the model makes decisions, based on a holistic view of its features and each of the learned components such as weights, other parameters, and structures. It takes into account the features within the model that are important and the kind of interactions between them that take place. With this level of interpretability, the distribution of the target outcome can be clearly understood by the stakeholders.

- **Local** - With this level of interpretability, the behavior of an individual prediction in a dataset can be examined. A single instance can be examined closely to understand what the model predicts as an outcome for the input.

Explanatory Medium

The explanatory medium is the second part of the process towards creating interpretable models. Here, the ways in which interpretations are shown as outcomes can be decided by the data scientists. As shown in Fig 34, some of the most commonly used explanatory mediums are feature importance where the most important features that contribute to high prediction probability are shown to business stakeholders so that they can understand the rationale behind a certain outcome. Similarly, dependence plots show whether the relationship between the target and a feature is linear, monotonic or more complex in model's outcome. Rule-based and if-then based on business rules and natural language can also easily convey useful meaning regarding the model outcome to business stakeholders. Lastly, simple visualizations that have a combination of the above can also help stakeholders get an insight into what the model is trying to convey to them.

Target Audience

The third and last part is to identify relevant stakeholders who will be at the receiving end of the interpretable outcomes. The identification of stakeholders also ties back to the business understanding phase and depending on the problem context, can be modified. As mentioned in Fig 34, the rationale behind choosing the stakeholders are described below. While the list shown in Fig 34 is exhaustive enough, the methodology is flexible enough to accommodate more relevant stakeholders, if there are any, or remove them from the list, if required.

Business Users (End-users): These users are eventually the consumers of the insights and at the receiving end of an explanation on a decision, action or recommendation. In a business setting, these users should be able to infer the insights provided as part of the outcome, interpret them and make actionable decisions that could potentially drive efficiency across the business.

Business sponsors: These users are business executives that make decisions, actions, or recommendations causing an impact across business lines, product lines, functional units and finally the customers themselves. The primary concern of business executives and sponsors is the contribution of the model outcome and its alignment with the business strategy and objective.

Data Experts: These users are the data scientists and data engineers who actively contribute by designing, training, testing, deploying, monitoring decisions, actions and recommendations of the ML model. The outcome is beneficial to them as well in the sense that it could help them understand debugging and testing requirements to enhance the results.

Regulators: These are regulators, who may want to ensure that the AI system does not discriminate or harm any individual or group of individuals. Depending on both the sector where the AI system is being used and the nature of the AI application, the rigor required for the explanations may vary. The primary concern of regulators would be to ensure that consumers are provided adequate explanations that are actionable and there is no harm to consumers.

Business Lead, PM: These users are actively involved in the project from the inception and are also the decision-makers. It is important for them to be able to understand the model outcomes primarily

to provide business context to the data scientists. They can also let the data scientists know if the the results are good enough to be presented to the higher level management.

Interpretability Checker

Summarizing all of the above steps, a systematic interpretability checker has been designed and can be seen in Fig 21. Using the interpretability checker, the data scientists can assess if the model chosen is required to go through the process of an additional layer of approximation for interpretability. The interpretability checker poses as a simple flowchart representation of the steps that are necessary to be carried out while identifying if an ML model requires further interpretability techniques or can remain inherently interpretable.

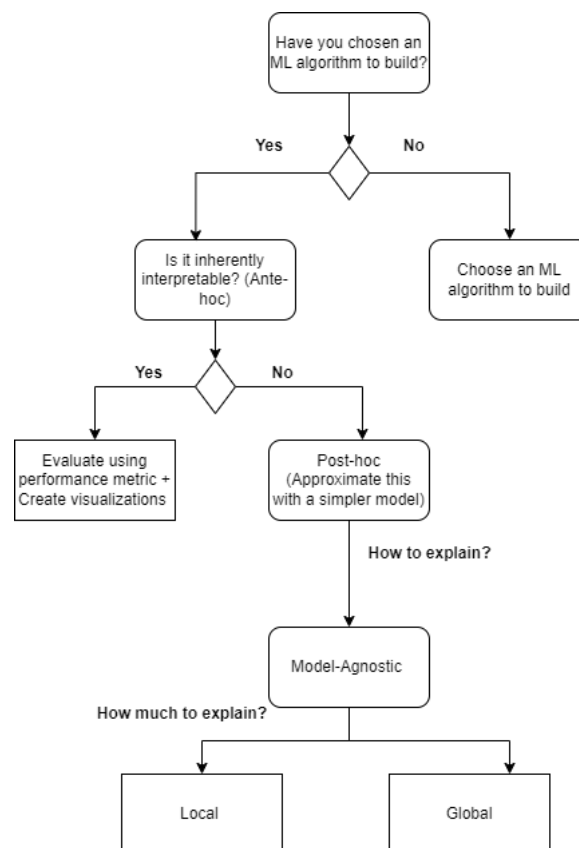


Figure 21: Interpretability Checker

4.3.5 Present to Stakeholders

This phase in the methodology is when the outcomes of the model are presented to the stakeholders to get their insights and check if the outcomes align with the business objective. This phase is crucial to the management of the organization as well as the team involved in the development as it all comes down to understanding what's more valuable for the business - maximizing performance metrics, understanding input-output relationships, the data that is valuable to the business and the insights that can help make better decisions.

Overall goal

The goal of this phase within the methodology is to ensure business executives and end-users gain

an understanding of the business insights derived from the outcome and are able to align with how it could possibly benefit the organization.

Stakeholder goals

Business stakeholders (Business Lead, Project Lead, PMO) need to:

- Support with presenting the final insights to executives and end-users
- Clarify domain-related aspects

Data scientists need to:

- Actively document feedback and prepare for roll-out of final deliverable to end-users

Present Insights

During this step, the insights are presented by the business lead, PM, and/or the data scientists to relevant stakeholders such as business executives, sponsors, end-users, etc. The presented insights can be positioned in a formal or informal setting with an interpretable, comprehensive, easy readability-induced presentation. A business readout should be planned in such a way that the information conveyed to the business stakeholders is succinct, clear and drives takeaways for the business in a strategic direction. This means overwhelming technical information that may be irrelevant to the stakeholder's knowledge can be dismissed from the final presentation.

Stakeholder Validation

The ML model in itself has been evaluated on the basis of the performance metric but the measure of how good the outcome is from a stakeholder standpoint is still a relative unknown, unless measured qualitatively or intuitively. As discussed in [56], if the results of an ML prediction are presented to the end-users, there needs to be a way in which the user's pragmatic understanding of the results are measured. The stakeholder validation is recommended to be done intuitively by the person conveying the explanation to the stakeholders. Following the suggestions by Doshi-Velez and Kim [57], the metrics used for this methodology have been selected based on the level of human involvement that is required to measure them. This kind of human-grounded evaluation is most appropriate to test more general notions of the quality of an explanation. Evaluation that is centered around human involvement is about conducting simpler human-subject experiments that maintain the essence of the target application. Through the validation, the person presenting the outcome can gauge the stakeholders' goals and objectives such as long-term goals (building trust and understanding the model), immediate objectives (debug and improve the model, ensure regulatory compliance, take actions based on the model output, justify actions influenced by a model, understand data usage, learn about a domain, contest model decisions), and specific tasks to perform with the explanations (assess the reliability of a prediction, detect mistakes, understand information used by the model, understand feature influence, understand model strengths and limitations) [58].

The feedback sessions organized after the business readout is a good setting to keep these metrics in mind while engaging with stakeholders. For the scope of this thesis, the metrics have been chosen and created as follows:

- **Novelty of Information based on stakeholder's mental model** - To measure this metric, the person conveying the explanation can compare the stakeholder's knowledge before knowing the explanation and after.

- **Degree of Understanding** - This metric can be measured by the degree to which stakeholders feel that they have understood the explanations provided to them and how easy they are to fit with the business process.
- **Satisfaction** - This metric is measured on the basis of how satisfied stakeholders are with the explanations provided to them, urging them to take next steps as decisions within the business process.

4.3.6 Reporting and Usage

This phase of the methodology represents the final deliverable being accepted and used by the end-users. It is characterized by its practical use in the designated field of application.

Overall goal

The goal of this phase is to ensure user acceptance and usability. Despite the validation of stakeholder satisfaction, degree of understanding and novelty of information in the previous phase, there might be discrepancies in understanding the usability of the deliverable. This phase ensures there is constant monitoring and engagement with the stakeholders to ensure ease of use and quality assurance. *This phase of the methodology is discussed in limited detail within the scope of the thesis.*

Stakeholder goals

Business stakeholders (Business Lead, Project Lead, PMO) need to:

- Actively engage with end-users to clarify domain and context related questions
- Support with relevant documentation such as user guides and manuals

Data scientists need to:

- Write technical report and/or manual to serve as a user guide for navigation through the deliverable

Monitor usage and performance

The risk of not maintaining the model is the degradation of the performance over time which leads to false predictions and could cause errors in subsequent systems [55]. Usage and performance of the deliverable can be monitored actively with weekly or bi-weekly feedback meetings to document feedback and concerns end-users cross paths with the system. These should be communicated comprehensively with the business lead who can further keep the data experts in the know to find solutions. Depending on the level of urgency with the user requests, the solutions can be incorporate either immediately or pushed back to a later time.

4.4 Overview of the Methodology

Fig 22 organizes and makes explicit the core concepts of how iCRISP-ML relates to the stakeholder desiderata thus enabling informed business decisions. iCRISP-ML intends to be iterative and agile to help deliver informed and interpretable solutions that are inclusive to all stakeholders within the organization. This is ensured through team collaboration from the inception of the project lifecycle to introduce a sufficient level of interpretability among stakeholders. As discussed earlier, it is important

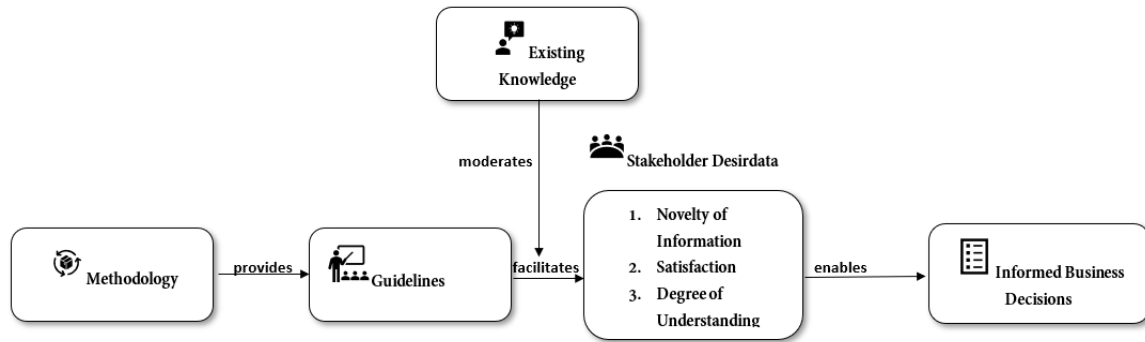


Figure 22: An Overview of iCRISP-ML and its goals

for the methodology to provide detailed guidelines that facilitate stakeholder desiderata such as novelty of information, degree of understanding and satisfaction. The methodology provides guidelines to build a satisfying model outcome encapsulating both business and ML-related objectives to facilitate stakeholder desiderata. Stakeholders' existing knowledge on the problem context is achieved by including them from the inception of the project in order to set a well-established context and not take them by surprise. The researcher believes that stakeholders believe outcomes that are consistent with their prior beliefs and assumptions. According to Nickerson in [59], this effect is called confirmation bias. This essentially means that explanations are not spared by this kind of bias and people will tend to devalue or ignore explanations that do not agree with their beliefs. As a consequence, the existing knowledge of the problem context for a stakeholder affects the extent to which their desiderata are satisfied. Therefore, as the methodology proposes, it is important for all relevant stakeholders to be engaged in understanding the context, assumptions and scope discussions of a given project to make the most out of the outcomes and drive overall efficiency within the organization.

4.5 Summary

The design and development chapter presents the process carried out by the researcher to design and develop the artifact i.e. the methodology. It first presents the existing methodology in use followed by the gaps and limitations observed as part of it. These observations are then studied further to be able to propose a methodology that can bridge the gaps and future objectives. The remainder of the chapter presents the newly proposed methodology i.e. iCRISP-ML, its phases and sub-steps.

5 Case Study Evaluation

In this chapter, the designed methodology is demonstrated using a case study at NXP Semiconductors. This case study demonstration intends to show that the methodology can work in a real-world setting and can be adopted in forthcoming projects.

The case study takes a portion of a large project that is currently undertaken at NXP and demonstrates the working of the methodology. The chapter starts with a description of the case study, the problem context, the team overview, and the work environment i.e. the tools and resources used. Then, the designed methodology is executed on a portion of the project using real-world data from NXP. The chapter concludes with the outcome of the case study, the learnings, and the limitations.

5.1 Case Study Description

Predictive Analytics is one of the fields of advanced analytics that makes use of AI and ML models for data-informed decision-making. The core of predictive analytics is to make predictions, and forecast activity, behavior, and trends about future outcomes using real and historical data coupled with data mining and machine learning techniques. The approach not only helps find useful patterns in data but also identifies risks and opportunities for the business. The research hypothesis in [18] claims that predictive analytics can support strategic decision-making by detecting projects endangered with misappropriate budget execution. This helps the board of management in making decisions that focus on minimizing negative consequences and liabilities. Likewise, predictive analytics is also widely used in forecasting sales predictions for a business by analyzing historical data, trends, behavior, etc. Organizations rely on forecasts to make informed decisions, foresee market performances, and manage resources like cash flow, project funds, plans, inventory, workforce, inventory, etc [60]. Since organizations typically rely on software tools, expert systems, and algorithms the quality of a business decision and what it entails depends on the quality of the forecast [61]. Moreover, forecasting also involves analyzing critical and sensitive business data. It is of paramount importance to bring in a component of trust so that business leaders can make the right decisions to drive overall business performance. To demonstrate the methodology in a real-world setting, an ongoing project that is currently undertaken at NXP Semiconductors as a collaborative effort between the CTO office and R&D IT teams has been selected.

The success of any organization is focused on its capacity to continually improve its products to stay up to speed with market trends and growth. Likewise, the future of NXP's success depends on its ability to improve existing products and develop new products for both existing and new market segments within the semiconductor ecosystem. NXP directs its R&D efforts and investment largely to the development of new semiconductor solutions to outpace market trends and target significant opportunities for growth and efficiency. For this engagement, they sought to generate monthly sales forecasts for new and existing products, and different material groups and business lines for the next 5 years. This also includes predicting the overall sales of each end-market in general on a macro level that includes the sales of each Business line(BL) and each Main Article Group (MAG) code. Furthermore, NXP is also interested in predicting the product lifecycle of New Product Introduction (NPI) i.e. 12NCs. NPI is the process that takes an idea from an initial working prototype to a thoroughly refined and reproducible final product. Every NPI product has sales associated with it which directly impacts the company's growth and revenue. The product life cycle prediction will enable NXP to identify the revenue generated by each 12NC that will further help in allocating R&D funding to the

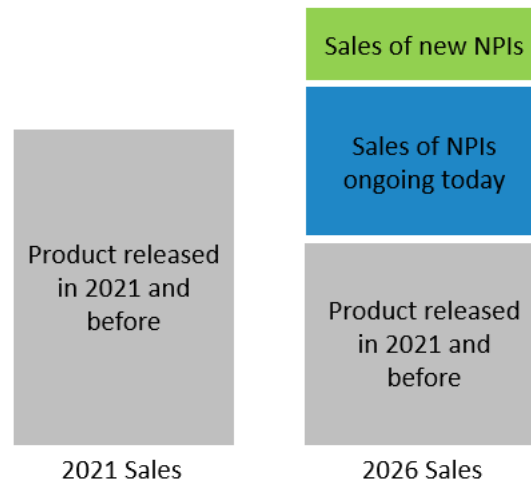


Figure 23: Overview of the plan for the project *[internal]*

products that generate the highest amount of sales to maximize the ROI for R&D efforts and investment.

As shown in Fig 23, there are four boxes in total. The grey standalone box on the left indicates the sales of ongoing products before the year 2021. The boxes stacked on the right represent the objectives of the project i.e. to predict the sales of ongoing products, products released before the year 2021, and NPIs that are in the process of being released by the year 2026.

5.1.1 Problem Context

At present, the analysis conducted to understand the company revenue, trends, and forecast future outcomes has been undertaken using conventional and heuristic-based approaches. These predictions rely on General Managers (GM) grouping together every year to predict sales forecasts using MS Excel and PowerBi. Since the decision-making process is highly dependent on human experience and domain expertise as a baseline to make actionable decisions for the business, there is a lot of manual effort and time expended to get outcomes.

The objective of the project is to utilize advanced analytical and machine learning approaches to analyze the sales patterns of several NPI products released by NXP and forecast the sales for the year 2026. Within the scope of the thesis and the case study, the core problem that needs to be addressed is the importance of introducing interpretability throughout the lifecycle of the project using a standardized methodology. It also attributes to identifying the most important features of the business that stakeholders can make interpret and make use of for higher predictive performance in the ML model. Currently, the team of data scientists working on the project has built ML models with exhaustive datasets and multiple features. The key is to help stakeholders understand which features are most influential in the model's predictions and can provide valuable insights into the problem being solved along with the data being used.

5.1.2 Team Structure and Overview

The project is undertaken by representatives from the CTO office and R&D IT teams within NXP. The team comprises of a Business Lead who is also the business and operational Subject Matter Expert (SME) for other stakeholders, a PMO/project lead, two data scientists, and the researcher who works alongside the team as an intern.

The team has content discussion meetings scheduled twice a week to discuss specific project-related tasks, milestones, and new insights to get feedback from the business lead and project manager. During these meetings, data scientists typically present the results of the work carried out over the week and discuss domain-related aspects with the business stakeholders for clarity.

The team also follows the Kanban methodology to manage the workflow based on a continuous improvement format. This eases the burden on data scientists and developers and ensures productivity and efficiency overall. As part of this format, the team also has weekly standup calls to track progress, and action items that are waiting on approval and discuss roadblocks with the project manager.

5.1.3 Product Hierarchy within NXP

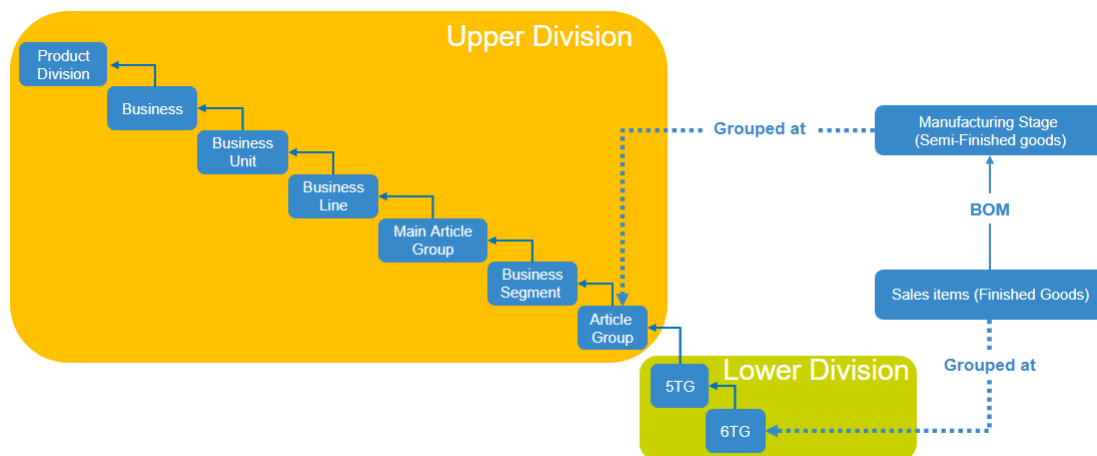


Figure 24: Classification of Products within NXP [internal]

The Product Hierarchy (PH) within NXP groups sales items i.e. finished goods and manufacturing stages i.e. semi-finished goods. The grouping is primarily done as part of the planning control process. There are two divisions of product classification - Upper and Lower classification. The upper classification is governed by the finance team as it is primarily used for financial purposes in the organization. The lower classification is governed by business lines for the purpose of planning and reporting. This can be seen in Fig 24.

Granularity of products decrease by moving up the hierarchy. The granularity can be seen in Fig 27 where Part 12 NC is the lowest level of product grouping. 12NCs are further grouped into 6TG and 5TG for the purpose of demand and forecast. AG goes all the way to Business Segments that are defined by the Business Units in the organization. These are further grouped into MAG codes that belong to Business Lines such as Advanced Analog, Automotive Process, Connectivity and Security, Edge Processing, Radio Power and RF Processing. The top-most hierarchy is the End Market which

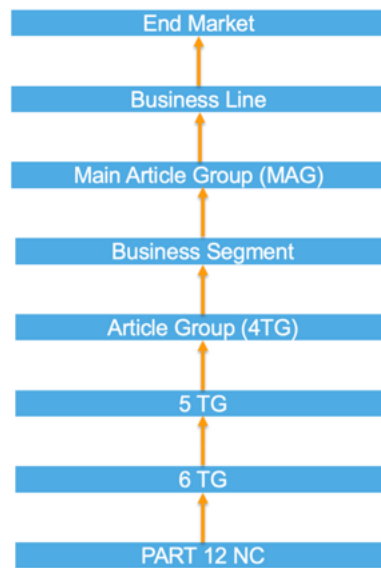


Figure 25: Product Hierarchy [internal]

are four in total - Automotive, Industrial & IoT, Mobile and Communication Infrastructure. For example, Automotive end market consists of 5 business lines, 20 MAG codes, 312 4TGs, 1010 5TGs, 2302 6TGs and 12001 Part 12 NCs.

5.1.4 Work Environment - Tools and Technologies

A dedicated Amazon Web Services (AWS) account with limited user access is provisioned for the purpose of this project. S3 is the primary data storage service and default data source for Amazon managed AI services. The business owner directly uploads the required dataset into an S3 bucket with SSE-S3 encryption. Sagemakerstudio that also supports Jupyter Lab interface within AWS has been chosen as the platform to perform all of the tasks revolving around data preparation, preprocessing, cleaning, enrichment and model building. Fig 26 shows the architecture of the work environment used for the purpose of this project. As shown in Fig 26, each user profile works on code-related operations using the Jupyter lab interface and stores the data in S3 buckets.

Amazon Sagemaker is a fully managed service that provides data science and machine learning capabilities for data scientists and others to make use of within the AWS platform. Sagemaker provides business users with the option of creating individual accounts since a generic sagemaker role would eliminate the possibility of tracking individual user operations and activities within the environment. Every individual user created in the studio domain has been assigned a custom sagemaker role, e.g. `rd-cto-cfn-sagemaker-rols-username-xxxxxxx`.

Open-source python libraries have been used by the data scientists to perform all the relevant data transformation operations for the data. The in-built data science kernel with instance type of `ml-t3-xxlarge` has been used for the purpose of operations with the necessary libraries installed as and when required.

Further, data scientists also make use of Quicksight to visualize data, create reports and dashboards for

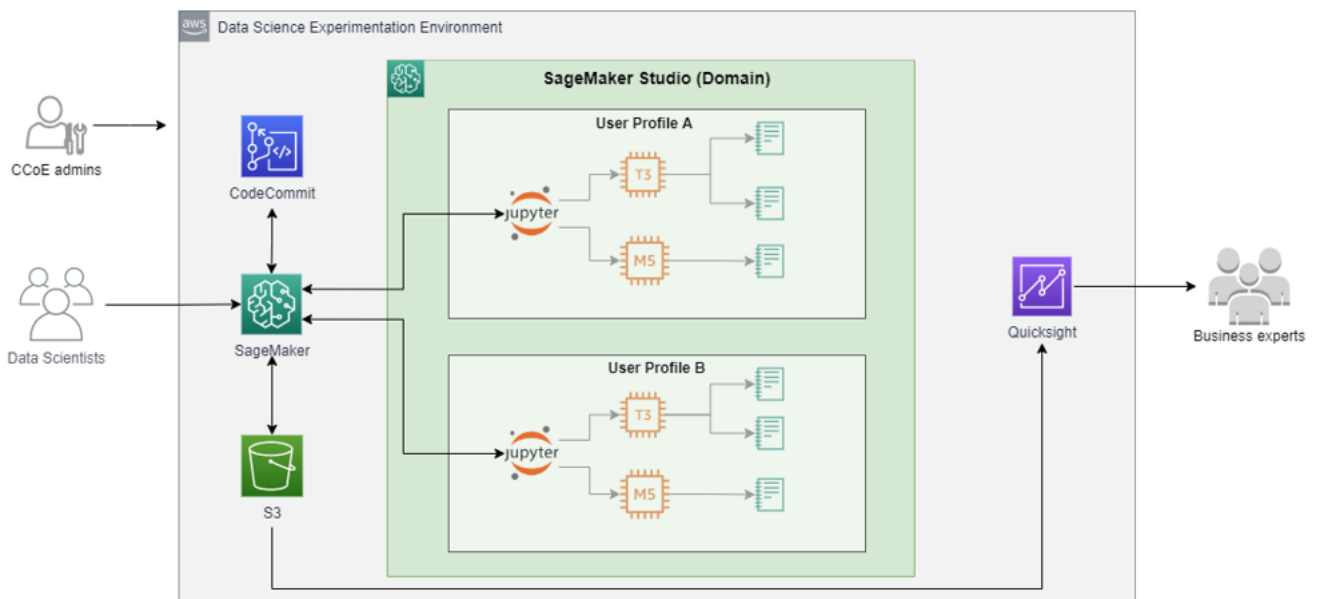


Figure 26: Work Environment [internal]

business experts to analyze insights. It provides the capability to connect the dataset directly from S3 and collaborate on the same platform to build visualizations.

5.2 Case Study Execution

NXP has provided the researcher with resources in terms of the tools and data required to execute the methodology. In this section the designed methodology is applied to and executed on a portion of the case study to demonstrate the level of interpretability at every phase.

Note: As discussed with relevant stakeholders, the scope of the case study does not cover the sixth phase of the methodology - 'Reporting and Usage' owing to time and effort constraints.

5.2.1 Business Understanding

This phase is critical to understand the business objectives along with the ML success criteria that needs to be achieved in order to meet business goals. It also means the objectives established within this phase is common to all stakeholders and to get an overall understanding of the project at large. On a high-level, the broad objectives that need to be achieved as part of this initiative is to build scalable solutions and consume business data, generate insights-driven actions for the business and optimize R&D investment by forecasting revenue expectations and estimating ROI. With support from the business lead of the team, the researcher has drafted the following as part of the business understanding phase to be able to achieve sufficient level of interpretability.

(a.) Business Objectives

While establishing business objectives for decisions-making, the direction in which the business is planning to move forward becomes the main focus. Here, the context is for the team or the business lead to show business stakeholders what the project is trying to achieve for the business. The idea is to make them more likely to support new projects and initiatives.

Leverage enterprise data: Build models to leverage enterprise data using advanced analytics and ML algorithms. This will help NXP derive actionable insights to make data-driven decisions.

Operations - Demand and Supply: Use data-driven methodologies and advanced forecasting techniques to plan ahead, manage demand and supply gaps and decide/take action in time.

Research and Development (R&D) Investment: With long-term forecasts and revenue, NXP can make better, data driven decisions on which product to develop and which product lines to fund.

Finance: With a more accurate forecast, NXP can give investors a more robust forecast of the future and NXP's value proposition.

The bigger objectives for the business that need to be met as part of the case study undertaken are mentioned above. For the scope of the thesis, the objectives are to be able to determine the most important features that the business can make use of, interpret and implement in an ML model to improve the predictive performance.

(b.) ML Success Criteria

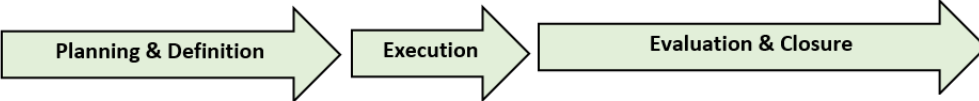
The overarching success criteria considered within the scope of this thesis and the demonstration is *Usability*. Since usability is an inherent feature of any good application or process, the success criteria for an ML-related project depends on the usability of the outcome and the impact it brings about within the organization. Eventually, a usable ML model outcome benefits the business in the bigger picture in terms of greater chance of market success and achieving business goals. Further, usability is attributed by three other supporting metrics such as:

- Impact - Ability of the ML model to deliver insights that can be used for decision-making
- Scalable - Ability of the ML model to take in high data volume across multiple domains within the organization
- Interpretable - Ability of the ML model to be understood by stakeholders in a digestible format

(c.) Identify stakeholders and assign responsibilities

The stakeholders are first identified on the basis of the team involved in the development of the project and the others that are impacted and interested in the outcome of the project. Using the RACI matrix that stands for Responsible, Accountable, Consulted and Informed, the stakeholders and end-users are mapped to a matrix with their respective roles. In doing this, not only is the first-level context established among stakeholders, but everyone is already aware of their positions within the scope of the project and what needs to be done going forward. For the project, the stakeholders are as follows:

1. Business Leaders - The stakeholders who are the high-level decision makers. They need to be updated on the context of the project and how it impacts the business.
2. Business Lead - The lead of the project and the decision maker for the team involved in the solution building process
3. PMO/Project Lead - Alongside the business lead, the PM is also involved actively in making decisions for the team to make progress throughout the project lifecycle



Project Stage & Stakeholders	Business Understanding	Data Understanding	Data Preparation	Modeling & Evaluation	Present to Stakeholders	Reporting and Usage
Business Leaders	I	I	I	I	A	I
Business Lead	A	C	C	A	R	A
PMO/Project Lead	R	A	A	C	A	A
Data Science Engineer	C	R	R	R	C	R
End-Users (BL CFOs)	I	I	I	I	I	I

Table 8: Stakeholder Roles using RACI Matrix

4. Data Scientists - The stakeholders involved in the design, development and solution building process
5. End-Users - The stakeholders who are directly impacted by the outcome of the project

The RACI Matrix with the assigned responsibilities can be seen in Table 8.

(d.) Impact, Effort and Benefit Analysis

This step in the methodology determines the benefits and their corresponding measures along with the effort that is required to meet the objectives of the project. As shown in Table 9, the benefits of the project that impact the overall efficiency of the business growth have been tabulated with the measure that can be used to evaluate the benefit. Along with the business lead and the PM of the project, the researcher has tabulated the benefits and measures to be able to present to a higher level of stakeholders such as business leaders for feedback and establish context on the scope of the project, assumptions and the potential outcome it could bring for the business.

The business lead together with the research has plotted the impact and effort estimate for the project on the basis of the following considerations:

1. Stakeholder Engagement - The stakeholders who would potentially be invested in the outcome of the project are the end-users (BL CFOs) and the business leaders who can make actionable decisions with the results presented.
2. Team Competence and Availability
 - (a.) Data Scientists (start off with two and can be expanded as and when the solution building process grows)
 - (b.) PM - To direct and streamline the project
3. Tentative timeline - The project is assumed to build iteratively as more and more use cases keep adding as layers to the initial use case and objective. There is no hard deadline in the bigger picture, but there could be short-term goals in terms of sprint and quarterly deadlines.

BENEFIT	MEASURE
Data-driven actions/decisions for the business	<ol style="list-style-type: none"> 1. Do end-users actually use the model in their business process? 2. Do decisions get better with the usage of the model outcome i.e. dashboard? 3. Usage Confidence - Usage of model by end-users to make predictions 4. Usage satisfaction
Better R&D funding decisions to newly introduced products	Accurate prediction of NPI Lifecycle Model
Collaboration with domain-level experts such as AWS to introduce a standard way or technology to build and deploy models	<ol style="list-style-type: none"> 1. Productivity Metrics: <ol style="list-style-type: none"> 1. Adoption Rate 2. Effectiveness 2. Good models with results that can be trusted
Automation of forecasts as opposed to manual predictions for better business decisions	<ol style="list-style-type: none"> 1. Reduction of human errors, false-positives and guesstimates 2. Human vs Machine prediction comparison - Show that the model forecast is better than human forecast

Table 9: Benefits and Measure

4. Potential risks - An exhaustive risk estimation has not been conducted for this project. However, the business risk could be the probability of the model outcome not being used in the business process by the end-users.
5. Outsourced consultants - Onboarding the AWS team to be a part of the ML solution building process.

The impact-effort plot has been used as part of use-case discussions with higher-level business stakeholders and is not included within the scope of the thesis as it involves future use cases that are sensitive to be shared.

5.2.2 Data Understanding

This phase of the methodology requires in-depth understanding of the data in relevance to the business context and the objective that the stakeholder needs to meet. Since the scope of the project is large by itself, the evaluation of the case study has been focused on a subset of the master dataset. This phase entails identifying the input data used for the demonstration, followed by a detailed explanation on the data and the ML learning task that has been identified.

The researcher, as part of the case study, has also consolidated a set of considerations which can be seen in Table 10 as a result of several discussions with relevant stakeholders involved in the project. This is to understand the various considerations that need to be looked into while understanding the data prior to the ML project. For the case study demonstration, the data is classified as 'M' sized with data issues such as data merge and semantic data quality issues. The ways in which these issues have been resolved can be seen in the forthcoming sections.

Input Data

The input data is a master dataset which is a consolidated dataset with data columns that contain infor-

Size quantified as S, M, L, XL	Potential data quality issues ranked in increasing order of difficulty	Knowledge of Business
1. Kb to Mb - S 2. Mb to 2GB - M 3. Gb to Tb - L 4. TB - XL	1. Data merge - multiple data sources 2. Semantic data quality issues 3. Constraints to find the right source of data 4. Missing data [Eg: Year of product release] 5. Security-related issues: Access management, permissions, confidentiality	Data and domain knowledge 1. Additional training, support, focus group sessions 2. Finding the right SMEs for domain knowledge

Table 10: Data-related considerations

mation about the product fabrication process, semiconductor market shares from external providers, actuals i.e. real sales of NXP in dollar value and the product hierarchy within the organization. However, for the case study within the scope of the thesis, relevant columns have been extracted from the master dataset and data preparation has been performed there after. The columns extracted from the master dataset belong to the market segment and NPI sales datasets.

(a.) Market Segment data

Market segment data is one of the data sources in the master dataset used within the scope of this thesis by the researcher is the market segment data. The dataset is a combination of both external data coming from a data provider and internal data within NXP. It includes semiconductor market sales (SAM) by market segment and strategic vertical, and also NXP's biggest competitor based on market share of previous year. The dataset is spread across different columns that are analyzed on the basis of the market growth of semiconductors industry, NXP's industry competitor growth and also includes the main end-markets of NXP such as:

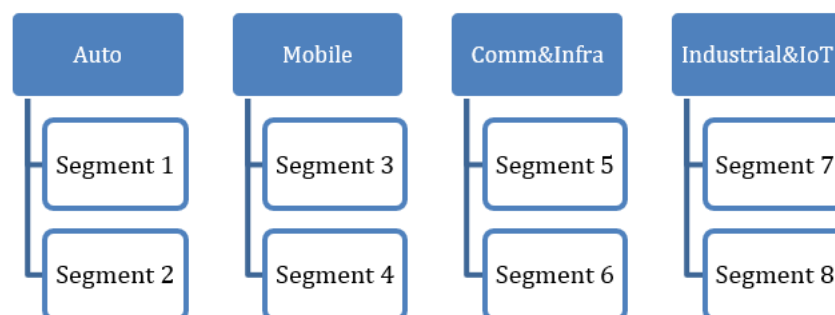


Figure 27: End-market to Market Segment

- **Automotive** - The automotive end-market offers a solution portfolio to enable the capabilities of sensing and thinking while ensuring the safe and secure operation of the vehicle through automotive radar and vision systems. It also facilitates high performance, next generation in-fo-tainment systems such as eCockpit, connected radio, automotive sound system, etc. in vehicles. The automotive end-market facilitates cars to seamlessly connect to the outside world and

supports vehicle-to-vehicle communication by means of Vehicle Network Processing (VNP), smart car access, etc.

- **Industrial and IoT** - This industrial and IoT end-market enables industrial applications at all layers of factory automation from manufacturing level up to the cloud. It also provides solutions to control lighting, HVAC, monitoring and safety systems. The processors serve a range of medical imaging, monitoring, diagnostic and treatment equipment.
- **Mobile** - The mobile end-market is recognized as the application leader that enables analog and charging solutions to add more capabilities to mobile phones, notebooks, tablets, etc. by means of NFC, eSE, eSIM and UWB solutions. The anti-counterfeiting technology, among others products, supports charging cables, power adapters and wireless charging pads for mobile phones and provide safety to their customers by making trusted accessories.
- **Communication Infrastructure** - The communication infrastructure end-market helps with connecting industrial IoT and personal devices to wireless networks. It also facilitates enterprise networking with processors and software to improve performance, flexibility and security of LAN, wireless and WAN connections.

Each of these end-markets are further categorized into appropriate subsets known as 'Market Segments'. The market segments have historical market share data in dollar value ranging from 2013 until 2021 along with the predicted market share for the next five years and the biggest competitor of NXP with their total sales in dollar recorded for the previous year.

Since the market segment data comes from an external data provider, it needs to be mapped to the corresponding NXP data. To understand the mapping solution, informal discussions were set up with the data scientist who worked on the mapping. It has been understood that the relationship between market segment and MAG code results in a many to many relationship which means a product with a specific MAG code can belong to multiple market segments. This also means that the product contributes to the sales of all those market segments. Unfortunately, the ratio of contribution cannot be inferred as a result of the many to many mapping.

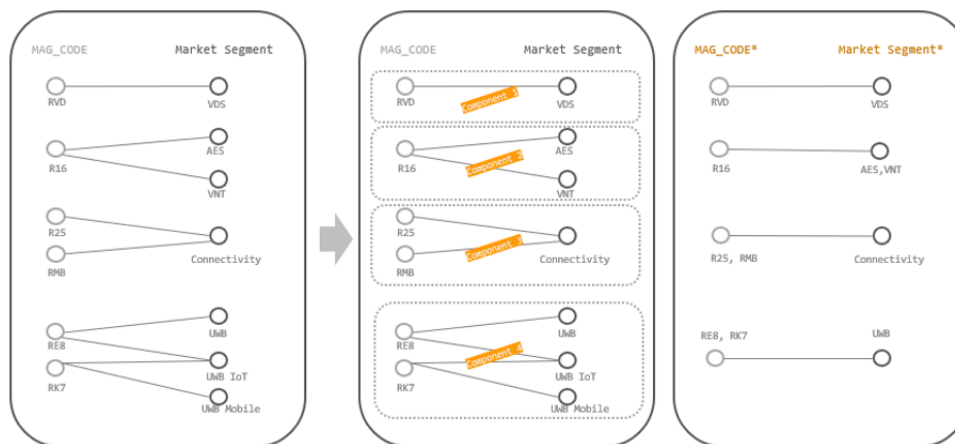


Figure 28: Newly proposed mapping between MAG code and Market segment

As shown in Fig 28, the new mapping solution is to split the old mapping into compartments such that each market segment and MAG code belong to that component and not outside. This resolves the issue of many to many relationship and ensures MAG codes are mapped to their respective market segments.

(b.) NPI Sales Data

The NPI sales dataset contains the different NPI products released and their sales in dollars. Every product in the dataset is identified by a unique 12-digit part number (PART 12NC). The dataset is a CSV file, few MB's in size and is manually uploaded to the respective AWS account. In total, there are 31577 product-related data in the dataset.

Metric for Evaluation

The data scientists along with the business lead and PM have decided to evaluate the model performance using the Weighted Average Percent Error (WAPE). For business stakeholders, the understanding is that lower the WAPE value, the better the model outcome is. The WAPE value of a forecast model is generally between 0 and infinity. The key advantage of using WAPE over other error metrics such as Mean Absolute Percent Error (MAPE) is that it weighs the individual impact of each product's sale. This means it takes into accounts each product's contribution to the total sale while calculating the overall error. Keeping the business context in mind, it is important to use a metric that conveys business insights while predicting the best possible model outcome in terms of its outcome. For the scope of the case study, the researcher needs to understand the metric that is currently in use since the final outcome should be able to improve the predictive performance of the model.

5.2.3 Data Preparation

The data preparation phase includes identifying the data types in the dataset, performing an exploratory data analysis to check the quality of the data, scaling the data, if required, and dropping irrelevant columns.

Columns present

From the master data that has been shared for the purpose of the research, columns relevant to the market segment data has been selected. The relevant columns are then aggregated on NXP sales data by performing a simple group by operation. The columns present in the dataset are as follows:

1. Market Segment - Level of grouping projects in a portfolio
2. NS_BL_Code - Business Line Code
3. NS_MAG_CODE - Main Article Group (MAG) Code
4. NS_MAG_DESCRIPTION
5. NS_PLAN_DATE - The quarter of sales
6. YEAR
7. SAM_Growth - Semiconductor Application Market (SAM) growth computed as time series data
8. NS_SALES_IN_DOLLARS

9. competitor_ratio_segment
10. nxp_ratio_segment
11. INNOVATION_AREA-Connectivity & RF
12. INNOVATION_AREA-High End Compute
13. INNOVATION_AREA-MCU
14. INNOVATION_AREA-RF specialty
15. INNOVATION_AREA-Sensor, HV, AMS & Smart Power

Dropping irrelevant columns

To understand a high-level view of the columns that need to be dropped from the dataset in order to reduce the dimensionality, the in-built feature importance using a random forest regressor has been performed. This method is available in scikit-learn package in Python for both regression and classification purposes. The random forest regressor makes use of gini importance or mean decrease impurity technique to compute the feature importance. The model is a combination of multiple decision trees where each tree is a set of internal node and leaves.

For the internal nodes, each selected feature is tasked with the decision of dividing the dataset further into two separate sets with similar responses. Using this technique, the measure of how each feature decreases the impurity of the split can be computed i.e. the feature with the highest decrease is selected for the internal node. Therefore, the more a feature decreases the impurity, the more important the feature is. In random forests, the impurity decrease from each feature can be averaged across trees to determine the final importance of the variable.

RandomForestRegressor - Feature Importance:		
	Variable	absCoefficient
10	Feature_10	0.390406
11	Feature_9	0.325148
0	Feature_6	0.105324
1	Feature_11	0.100432
2	Feature_12	0.042832
9	Feature_2	0.023041
8	Feature_13	0.008754
3	Feature_14_a	0.001501
7	Feature_14_b	0.001422
5	Feature_14_c	0.000495
4	Feature_14_d	0.000429
6	Feature_14_e	0.000215

Figure 29: First-level feature ranking using Random Forest Regressor [anonymized]

As shown in Fig 29, columns Feature_14.a to Feature_14.e are encoded values of the same column and ranked lowest in the list. On presenting the initial results to the team, it has been decided to

combine the encoded values as a single column and recompute the features to check if the rank goes up. In Fig 30, the consolidated column 'Feature_14' is still ranked lower than the other features. It has been decided to drop the column from the dataset for the remainder of the process.

Likewise, despite the fact that FEATURE_10 and FEATURE_9 are ranked high among others, on discussion with business stakeholders, it has been identified that the columns unlike other relevant features are not in time-series format i.e. they are limited to a particular year and in this context, the year 2019. This led to the decision of removing the two columns from the dataset as it is natural for them to be ranked higher than the other columns in the dataset leading to wrong conclusions.

The columns 'Feature_13', 'Feature_9' have also been removed from the dataset as it is important to identify the co-relation of the products with the sales data independent of time.

	Variable	absCoefficient
6	Feature_9	0.267108
0	Feature_6	0.243585
5	Feature_10	0.238182
1	Feature_11	0.169809
2	Feature_12	0.048733
4	Feature_2	0.017800
3	Feature_13	0.007635
7	Feature_14	0.007149

Figure 30: Second-level feature ranking using Random Forest Regressor [anonymized]

Additional features

Although the columns Feature_9 and Feature_10 have been dropped owing to the fact they are not in time-series format as opposed to the other columns in the dataset, it doesn't change the fact that they are important features. On discussion with the data scientists and business lead within the team, it has been decided to create a new feature called nxp_market_percentage which is the NXP market share. The rationale behind creating this additional feature is to support the idea that it can be created as a time-series data, with the NXP sales and SAM sales. Using the following formula, the new column has been created and added to the existing dataset.

$$\text{NXP Ratio Percentage} = 100 * (\text{NXP_Actuals} / \text{SAM_in_dollars})$$

where NXP_Actuals refers to the sales of NXP in dollars and SAM_in_dollars refers to the sales of the market share in dollars

In addition to this, other columns that belong to the NXP product hierarchy have also been added to the dataset to check for which level of grouping can be most optimal when being fed to the ML model to get a higher accuracy model score.

The final list of columns in the dataset are:

1. NS_BL_Code
2. NS_MAG_CODE
3. SAM_Growth
4. NS_SALES_IN_DOLLARS
5. nxp_market_percentage
6. SAM_IN_DOLLARS
7. NS_TYPE_GROUP_6TG_CODE
8. NS_TYPE_GROUP_5TG_CODE
9. NS_AG_CODE

Dealing with categorical variables

The dataset is highly populated with features that are categorical and they become incomprehensible when passed directly to an ML model. This is because ML models cannot process categorical variables and require both the independent and dependent variables i.e. input and output features to be numeric. The process of converting categorical data into numerical data for the machine to learn from the data and give the right model outcome is called Encoding. While there are many encoding techniques available within the ML community to make use of, it is also important to identify the type of categorical data that is being dealt with. However, regardless of the encoding method, they all aim to replace instances of a categorical variable with a fixed-length vector. In this dataset, the categorical data are of nominal data type which means they don't have any particular order, rather, they are simply categorized within a particular column. The categorical columns in the dataset are identified as follows:

1. NS_BL_Code
2. NS_AG_CODE
3. NS_TYPE_5TG CODE
4. NS_TYPE_6TG CODE
5. NS_MAG_CODE

Encoding techniques

The first encoding attempt using the one-hot encoding technique resulted in a really sparse matrix that inflated the number of dimensions within the dataset for the model to work with. On observing the data further, it has been identified that the categorical variables in the dataset are of high cardinality features i.e. there are too many unique values within the same column. Essentially this means that as the number of feature grows, the amount of data needed to accurately distinguish between these features in order to give a prediction and generalize the model also grows exponentially. This is a

dimensionality-related problem.

Target Encoding

To combat the issue of dimensionality caused by one-hot encoding, target encoding has been performed. It involves replacing a categorical feature with average target value of all data points belonging to the category. The target encoder performs an encoding on the categorical data by replacing them for a measurement of the effect they might have on the target. According to [62], using target encoding, features are replaced with a blend of posterior probability of the target given particular categorical value and the prior probability of the target over all the training data.

Normalization of data

Features in the dataset that are measured at different scales do not contribute equally to the model fitting and model learned function eventually leading to a certain degree of bias while making the final outcome. To overcome this issue, the input features in the dataset are normalized using the Min-Max scaler prior to model fitting. By doing so, all the features will be transformed into the range [0,1] i.e. the minimum and maximum value of a feature is going to be 0 and 1, respectively. Scaling is also an important step that needs to be performed before performing Principle Component Analysis (PCA) which is explained in detail in the next sub-section.

Cumulative Variance by Principle Component Analysis

As a precursor to the model building and evaluation phase, Principle Component Analysis (PCA) has been performed on the dataset. PCA is not a feature selection method but a dimensionality reduction method. However, the objective remains the same i.e. to reduce the amount of data needed to compute the model. The explained variance is a statistical measure of how much variance in a dataset can be attributed to each of the principle components generated by the PCA method. This is important as it allows the ranking of components in the order of importance and focusing on the most important feature when interpreting the outcome.

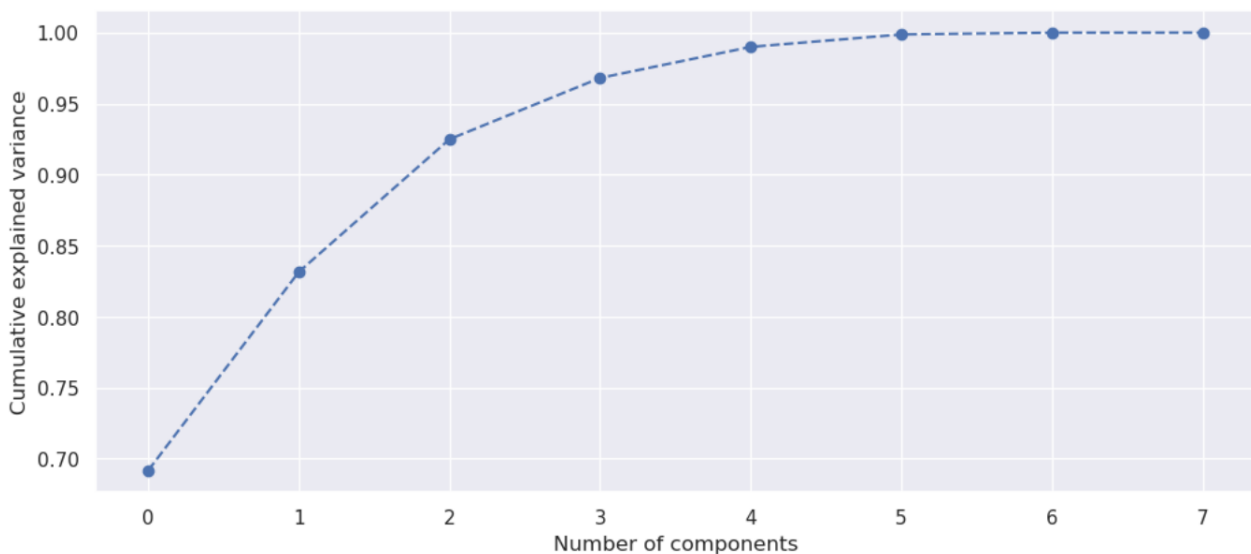


Figure 31: Cumulative Variance explained by PCA

As shown in Fig 31, the larger the variance explained by a principle component, the more important

that component is in the dataset. Explained variance can be used to choose the number of dimensions to keep in a reduced dataset. It can also be used to assess the quality of a machine learning model. In general, a model with high explained variance will have good predictive power, while a model with low explained variance may not be as accurate. Fig 31 shows that 95% of the model variance can be explained by the first four principle components.

5.2.4 Model Building and Evaluation

The model building and evaluation of the phase has been performed for the purpose of testing the interpretability for an existing ML model. The model that is in existence has been identified as a black-box model that needs to be approximated further to make the outcome more interpretable. Moreover, the aim of this phase is to ensure the model is tested for interpretability using one of the techniques mentioned in Fig 34. The following sections will explain in detail the process that has been performed to obtain the results of the interpretable model.

Existing ML Model to test Interpretability

The Neural Field (NF) approach for the product lifecycle model is the ML model that exists for which the interpretability needs to be tested. The data scientists in the team have built the model on the basis of the textual description of a product i.e. 12NC and the release-for-sale date of the product. This can be seen in Fig 32 where the yellow box titled 'Product description' represents the textual description of a product and the yellow box titled 'Time dimension' represents the release-for-sale date of the product. The output of the neural field is a curve that is generated with the help of the parameters i.e. latent vectors. The neural field then outputs a curve in the form of a latent vector (basically the parameters to generate the curve). The Curve Generator (CG) consumes this latent vector and uses it as the weights and biases of its Multilayer Perceptron (MLP) layers to enable the CG to generate lifecycle curve for this product.

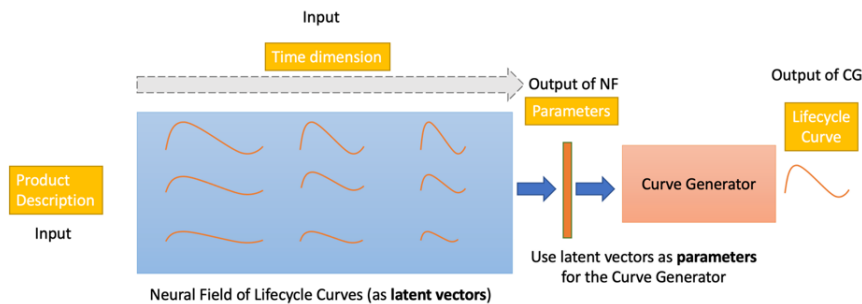


Figure 32: Neural Field for Product Lifecycle [Internal]

Using this neural field approach the lifecycle curves are predicted in a cold-start setting i.e. new products have no previous sales data for the model to draw references. The model is trained in such a way that as and when new sales data comes in, it can adjust itself to fit the new data for higher predictive performance.

Test for Interpretability

This step in the model building phase involves performing a check for interpretability. As explained

in Chapter 4, the interpretability checker is made use of to understand the type of technique that can be employed on an ML model in order to obtain interpretable outcomes. Fig 33 shows the steps that are involved in the interpretability check specific to the case study. The green tick marks indicate the following:

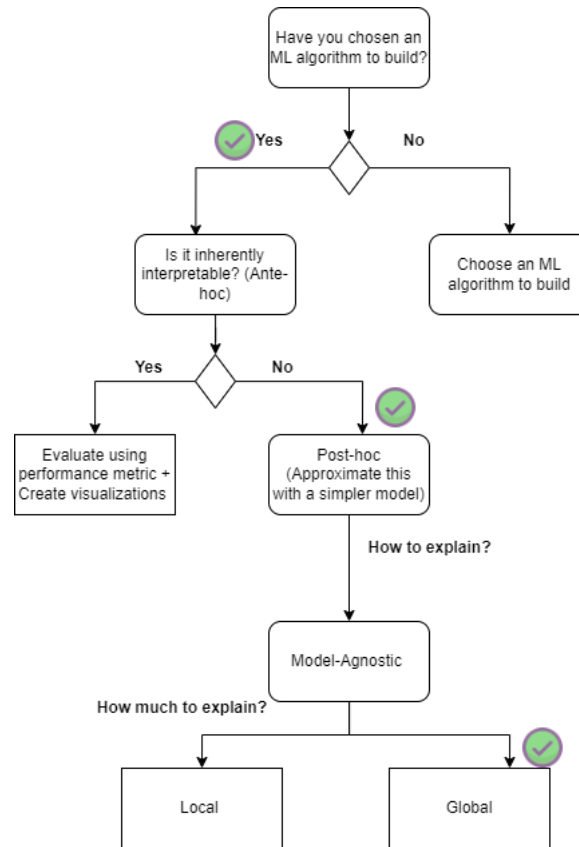


Figure 33: Interpretability Checker for the Case Study

1. The data scientists have already chosen an ML algorithm to build
2. It has been identified that the ML model is not intrinsically interpretable i.e. it is a black-box model that requires post-hoc interpretability
3. Global interpretability has been chosen as the suitable locality for the ML model

Rationale behind choosing the techniques

The final Input, Explanatory Medium and Target Audience selected for the interpretability test is summarized in Fig 34. The text highlighted in yellow indicates the process that is followed for this case study. This section further explains the rationale behind choosing specific techniques that are employed on the ML model for interpretability.

(a.) Input: The input requirements have been identified and are all important for the context i.e. the existing model for which interpretability needs to be tested, the business objectives and the context that needs to be achieved as well as the data that is required to implement the interpretability process on the ML model.

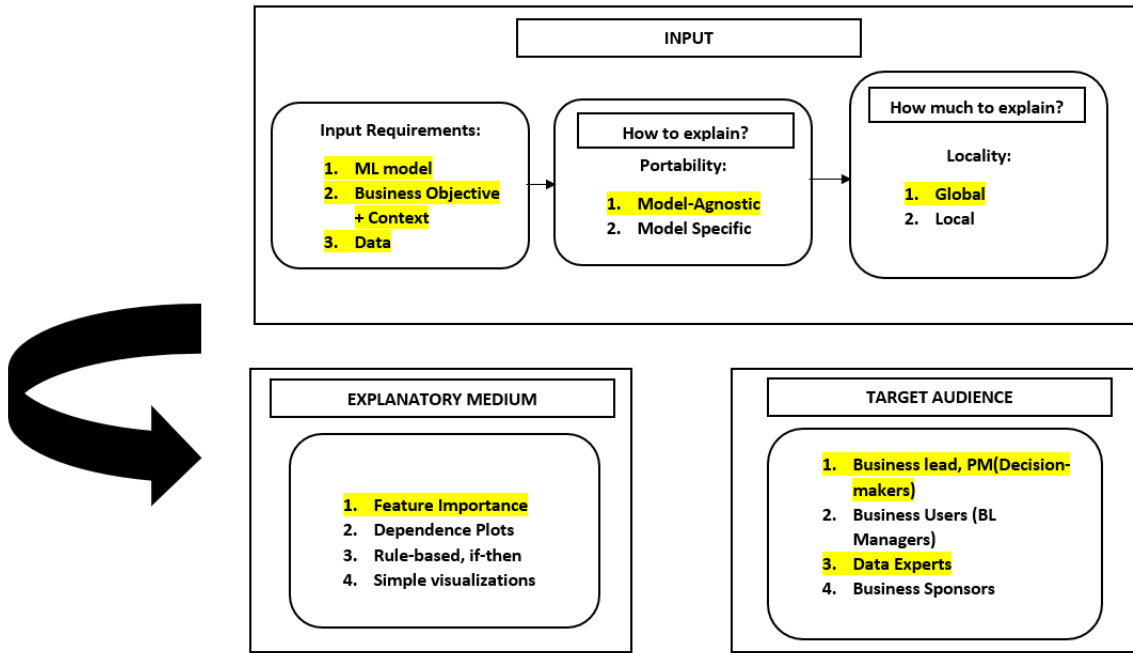


Figure 34: Test for Interpretability

In terms of locality, the global interpretability technique is chosen so that stakeholders can have access to interpretability across all predictions to attribute which features contribute the most to the model's decisions. In other words, unlike local interpretability, global interpretability allows the stakeholders to determine to what extent the features in the dataset contribute to how the model makes its predictions over all of the data and is not limited to specific instances. Global interpretability is also chosen keeping in mind that the target audience also includes non-data science experts. It gives an overall view on how the model makes predictions and which features in a business context contribute to a high predictive performance. Specifically in this case study, the model provides global insights to stakeholders outside of data science — it's more of an analytical function to provide business stakeholders more insights.

(b.) Explanatory Medium: For an ML model, a 'feature' is typically considered as a measurable input that is used in predictive models to predict future outcomes. It could be the color of an object or the age of a person. These features, when dealt with the right way, can be very beneficial leading to improvements such as boosting predictive results, decreasing computational times, reducing excessive noise and increasing decision-making transparency [63], [64]. Simply put, it is the process of converting raw observations and insights into desired features using statistical and machine learning approaches. For the purpose of this case study, keeping in mind the objective of the project i.e. to determine the important features from the market segment and product hierarchy data to improve the model performance metric, it has been decided to perform feature importance using various interpretability techniques. On comparing this technique with dependence plots, it has been identified that the latter despite being intuitive and comparatively easy to implement, only displays the average marginal effects. As a result, heterogeneous effects in the dataset may be hidden i.e. one feature might show a positive relationship with a prediction for half of the data, but a negative relationship for the other half.

(c.) **Target Audience:** The target audience such as business lead, PM and data experts have been chosen keeping in mind the internal core team who will eventually make the decisions of including the features in the ML model to improve the predictive performance. The results are presented to the identified target audience for validation on the next steps i.e. to be included in the final model outcome.

Choosing ML models for feature importance

The best-fit ML models such as Random Forest and Light Gradient Boosting have been analyzed for the purpose of the feature importance. The process and rationale behind choosing the two models are explained in the sections below.

(a.) Random Forest

Random Forests can defy the interpretability-accuracy tradeoff to a certain extent. It has been chosen for this demonstration as one of the two ML models owing to its good predictive performance, low overfitting and easy interpretability. It is straightforward to derive the importance of each variable on the decision i.e. it makes it easy for stakeholders to understand how much each variable is contributing to the decision. As mentioned in Section 5.2.3, random forest has its own built-in feature selection

Feature Importance using Random Forest Regressor

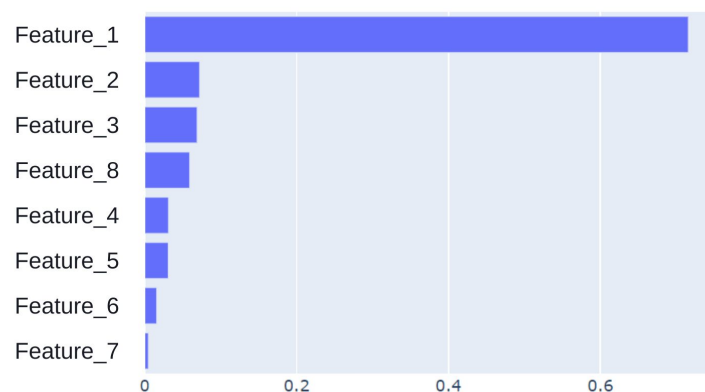


Figure 35: Feature Ranking using Random Forest Regressor[anonymized]

technique on the basis of gini impurity. The idea behind the technique employed by this algorithm is to reduce variance in the prediction of several noisy decision trees by averaging their results. Fig 35 shows the results of the features ranked in order of their importance using the built-in feature importance technique of the random forest regressor.

(b.) Light Gradient Boost (LGBM)

Light GBM (LGBM) is a gradient boosting framework that uses tree based learning algorithm. LGBM is prefixed as 'Light' because of its high speed and can handle large data set with relatively lower computation time. In comparison with random forest, the LGBM model has been widely used across industries and business use cases to get accurate outcomes.

Computing Feature Importance

The features and their importance are computed using two techniques: SHAP and Permutation Feature Importance. For random forest, SHAP has not been computed due to the fact that it suffers from a high computation time to give accurate outcomes. LGBM, on the other hand, supports both SHAP and Permutation feature importance with a relatively lower computation time.

1. Shapely Additive Explanations (SHAP)

SHAP is considered supremely better than the traditional scikit-learn methods because they can be inconsistent, which means that the features that are most important may not always be given the highest feature importance score. For instance, in the case of a tree-based model, two equally important features may be given different ranks based on the level of splitting employed on the features. SHAP estimates the importance of a model by checking the performance of the model with and without the feature for every combination of features.

Global interpretation using SHAP

Global interpretation with SHAP can be done by examining the mean absolute SHAP value for each feature across the whole dataset. By doing this, the magnitude of each feature's contribution towards the target variable i.e. sales is quantified. This can help with anticipating the behaviour of an ML model with respect to the whole distribution of values for its input variables.

They're also intuitive to understand which features impact the predictions most. The features with higher mean absolute SHAP values are considered most influential in terms of the ML model's predictive performance. With SHAP, this is achieved by aggregating the SHAP values for individual instances across the entire population.

Summary plot using SHAP

The summary plot using SHAP gives an overview of which features are most important for a model. The model has been analyzed at a global level by estimating feature importance along with how the features affect the prediction and model outcome. The plot in Fig 36 sorts features by the sum of SHAP value magnitudes over all samples, and uses SHAP values to show the distribution of the impacts each feature has on the model output.

The plot as shown in Fig 36 shows how much each predictor contributes, either positively or negatively, to the target variable. The colors, red and blue, represent the feature value, high and low, respectively.

The plot is able to present to the stakeholders the following information:

- Feature importance: Displays the variables ranked in descending order.
- Impact: The horizontal axis shows whether the effect of that feature is associated with a higher or lower prediction.
- Original value: The color indicates whether that variable is high (in red) or low (in blue) for that observation in relation with the model output.
- Correlation: A high value of 'Feature_1' indicates a high and positive impact on the model output i.e. sales. The "high" comes from the red color, and the "positive" impact is shown on the X-axis.

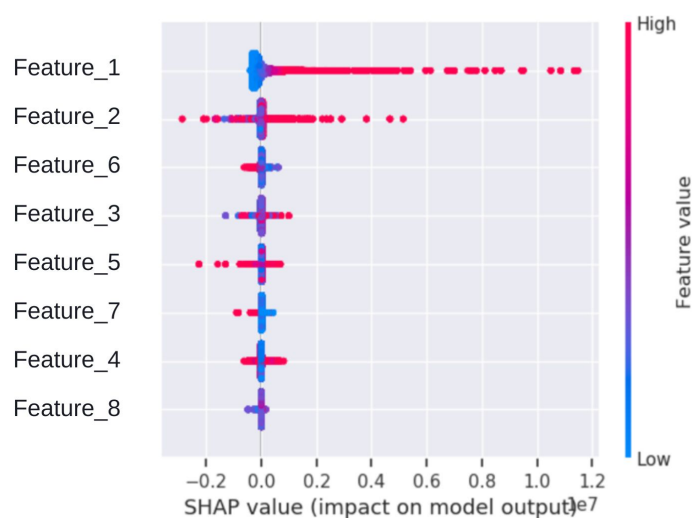


Figure 36: Summary Plot using ShAP for LGBM Model [anonymized]

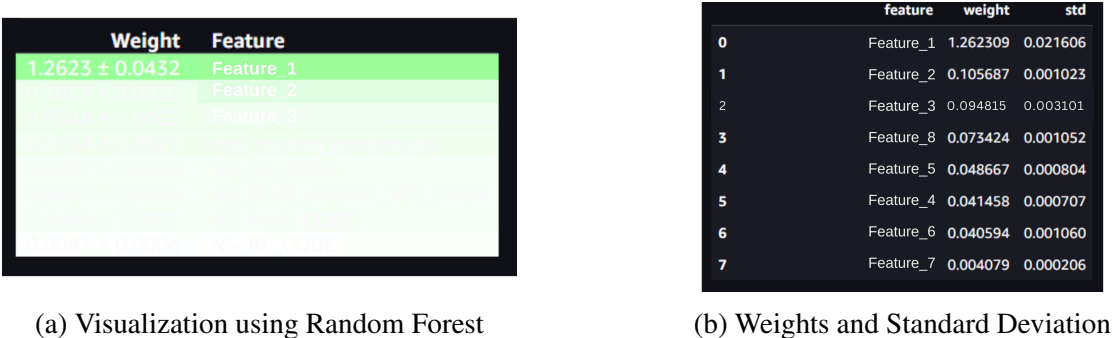


Figure 37: Permutation Feature Importance using Random Forest [anonymized]

2. Permutation Feature Importance

The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled in a dataset [65]. The core concept of this process is to break the relationship between the feature and the target. Therefore, the drop in the model score is indicative of how much the model depends on the feature.

Permutation feature importance makes use of the ELI5 library. It enables extraction and visualization of feature weights and their contribution from the model as a form of global explanation [66]. The importance of a feature is estimated by calculating the increase in the model’s prediction error after permuting the feature. A feature is considered as “important” if shuffling its values increases the model error i.e. the model relied on the feature for the prediction. Likewise, a feature in the dataset is considered “unimportant” if shuffling its values leaves the model error unchanged i.e. the model ignored the feature for the prediction.

For each column in the dataset, permutation feature importance process changes the data in the column and then tests how much that affects the model accuracy.

- The data is shuffled in random order while keeping the values of other columns constant.
- New predictions are then generated based on the shuffled values and the quality of the new

Weight	Feature
1.1496 ± 0.0332	Feature_1
0.2417 ± 0.0053	Feature_2
0.0854 ± 0.0035	Feature_3
0.0559 ± 0.0026	Feature_4
0.0487 ± 0.0024	Feature_5
0.0344 ± 0.0017	Feature_6
0.0325 ± 0.0012	Feature_7
0.0279 ± 0.0005	Feature_8

(a) Visualization using LGBM

	feature	weight	std
0	Feature_1	1.149607	0.016624
1	Feature_2	0.241663	0.002659
2	Feature_3	0.085446	0.004876
3	Feature_4	0.055979	0.000524
4	Feature_5	0.050344	0.000673
5	Feature_6	0.033951	0.001257
6	Feature_7	0.027967	0.000538
7	Feature_8	0.008276	0.000736

(b) Weights and Standard Deviation

Figure 38: Permutation Feature Importance using LGBM [anonymized]

predictions are evaluated.

- The feature importance score is computed by calculating the decrease in quality of the new predictions in comparison with the original predictions.

Permutation feature importance includes a visual representation by means of a list that tabulates the features and their corresponding weights. The gradient of green and slowly fading into white indicates the positive or negative impact on the model decisions. As shown in Fig 37 and Fig 38, the visualizations can be seen with Feature_1 as the most important feature displayed in green and then as the importance of the features on the model outcome decreases, the color turns white. The values at the top of the table are the most important features in the model, while the features at the bottom are considered least important to the predictive performance of the model. The first number below the 'Weight' column in each row indicates how much the performance of the model performance has decreased with random shuffling, using the same performance metric as the model. In the case of Random Forest and LGBM, RMSE has been used as the performance metric. The number after the '±' measures how the performance has varied from one-reshuffling to the next, i.e. degree of randomness across multiple shuffles.

The results have also been visualized as dataframes in Fig 38b and Fig 37b. The standard deviation is calculated to measure the amount of randomness in the permutation importance calculation by repeating the process with multiple shuffles. The 'weight' column that represents the importance of the particular feature represents the importance of the feature accumulated in multiple shuffles instead of in a single trial.

5.2.5 Present to Stakeholders

In this phase, the results are presented to the identified stakeholders to validate their interpretation and test the quality of the outcomes. This has been conducted as part of the bi-weekly content meeting to get feedback. Within the scope of this case study, the main stakeholders for the researcher are:

- Business Lead and/or PM - To approve and validate the features that have been ranked as important from a business standpoint
- Data Scientists - To confirm the methodology employed for feature importance and apply the results in the existing ML model to check improvement of WAPE score

The validation has been conducted with an intention to convey that the outcome of an ML model for a business does not depend solely on the accuracy score or its predictive performance. It is also

attributed by meeting stakeholder requirements and satisfaction in adopting the outcome of an ML model as part of the business process to make informed decisions. Trust has been identified as the overarching metric to measure the stakeholder's requirements. The three metrics that attribute to developing stakeholder trust in outcomes are defined as follows:

- Degree of Understanding
- Novelty of information based on stakeholder's mental model
- Satisfaction

As shown in Table 11, the questions that have been asked to measure the qualitative outcome of the feature importance performed as part of the interpretability step are documented. The table also shows the consolidated responses from the stakeholders at the receiving end of the outcome. Based on the results of the features presented, the stakeholders i.e. data scientists, business lead and PM trust the outcome to an extent and are convinced that they would contribute to the improvement of the model's predictive performance. As a result of the feature importance that has been performed, the following features have been selected to be fed into the model as input:

- Feature_1
- Feature_6
- Feature_5
- Feature_7

Metrics	Questions	Helper Questions	Respondents
Trust	Is the data used in generating important features suitable for your decision-making process?	Do you trust the market segment data?	Trust level for market segment data is slightly low as the data provider is external. However, the belief that it will improve the model score and also estimate NXP sales in comparison with market growth is very much there.
Mental model, Degree of Understanding	Did you infer something new from the feature outcome as compared to what you knew before?	Has your mental model & degree of understanding improved?	The feature ranked at the top is indeed a surprise and a value-add to the business context.
Satisfaction	Do you think the feature outcome will impact the model outcome?	Will the WAPE score be impacted?	Yes. It will be included in the existing product lifecycle model to check for improvement.

Table 11: Questions for stakeholders based on metrics

Test the outcome

Based on the insights gathered from the stakeholders, the features have been included in the existing ML model by the data scientist to test the validity and quality of the features generated. Fig 16 shows the results in a visualized graph format. The graph compares the product lifecycle with and without the inclusion of features on the basis of its context ratio. The 'Y' axis represents the value of the WAPE i.e. performance metric and value on the 'X' axis represents the context ratio i.e. NXP actuals/sales in dollars. From the graph, it is inferred that the model with the inclusion of features performs better when there is less context ratio i.e. the context between 0 and 50% indicate that the model can improve the WAPE score and potentially decrease it to improve the predictive performance.

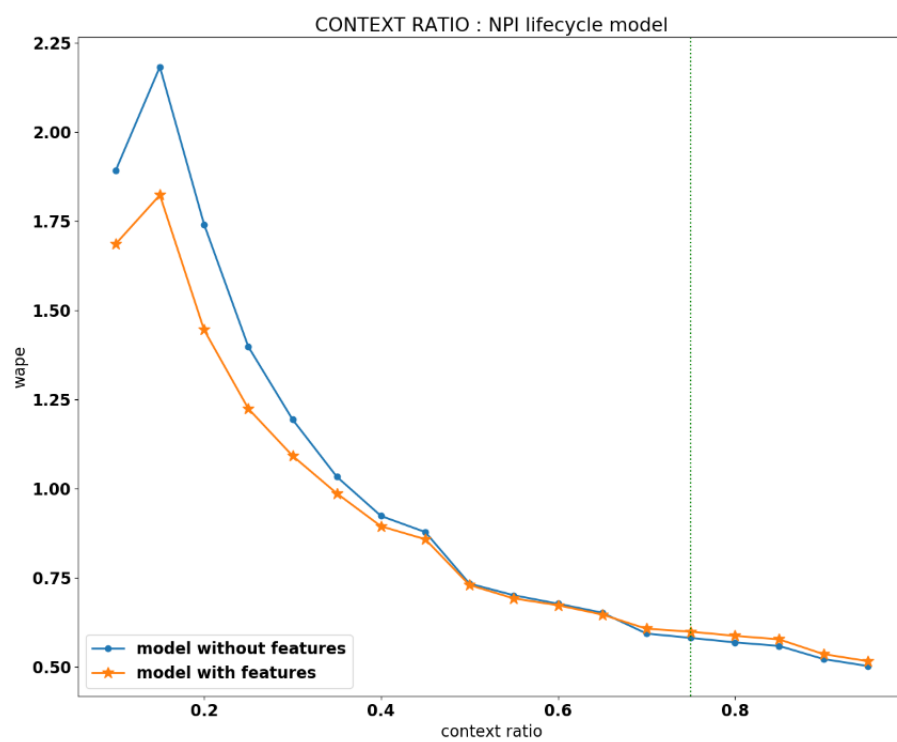


Figure 39: ML Model Results - Comparison of Model with and without Features

5.3 Outcome of Case Study

Interpretability is not merely an afterthought and cannot be limited to those that are building the solution alone i.e. data scientists, engineers and developers. Rather, it needs to be enforced explicitly since the beginning, getting the right stakeholders on-board so that when the outcomes are finally presented, they're able to ease-in to make informed decisions. The case study demonstration by the research has been successful in driving this method across to the board of stakeholders who have been involved in design decisions and solution building process. The results obtained as part of the case study have been demonstrated in this chapter. Though the case study has not been demonstrated as an end-to-end process i.e. from the first phase until the sixth phase, it attempts to meet the objective defined at the start of the project lifecycle.

- The case study demonstration is a portion of a larger project that is currently underway at NXP. To demonstrate the effectiveness of the methodology in a real-world context, the researcher has

attempted to leverage ML and interpretability techniques to show the important features in a dataset to meet business objectives and in turn, improve the predictive performance of the ML model.

- The choice of techniques for interpretability has been concluded using the interpretability checker that's embedded within the methodology.
- The validation of the case study can be seen as part of the fifth phase i.e. present to stakeholders where the results are presented to the relevant stakeholders to make decisions further. The outcome is then tested on the basis of the three metrics to check if there is a level of trust achieved among stakeholders.

5.3.1 Limitations

The working prototype of iCRISP-ML in a business use can be seen as part of the case study demonstration. However, limitations have also been identified as part of the case study demonstration.

- While iCRISP-ML is more focused on the workflow and approach to execute an ML project, it does not pay much attention to the ways in which a team should collaborate. Since the demonstration of the case study was not bound by time or sprints, it's difficult to check the applicability of the methodology in a larger use case that has strict timelines and deadlines to meet.
- The methodology has a sub-step within the fifth phase to measure the outcomes using metrics from a stakeholder standpoint. Although these metrics are intuitive and human-grounded, they cannot be measured quantitatively.
- iCRISP-ML also lacks a clear definition of what interpretability means in the context of the business. This could involve factors such as the complexity of the model, the transparency of the results, and the ease with which the results can be understood and used by stakeholders.
- The methodology also misses a sub-step in the fifth phase 'Present to Stakeholders' to address the feedback given by business stakeholders. This sub-step can involve revisiting the ML model or using different interpretation techniques to make the results more transparent and easier to understand.

5.4 Summary

The chapter presents the execution of the designed and developed methodology, iCRISP-ML, within an existing use case within the organization. The chapter gives an overview of what the case study entails, the problem context within the scope of the research along with a general understanding of the data and production environment. The case study is explained and executed with the help of the phases of the methodology along with inferences from each phase and the final outcome. The chapter concludes with the final outcomes of the case study execution and the limitations that have been observed as part of it.

6 Expert Interview and Validation

This chapter presents the validation of the proposed methodology using qualitative interviews conducted with the experts in the organization. The goal of this phase, by [45], is to validate the usability of the methodology. First, the final methodology and results are presented to the experts as part of a presentation by the researcher. This is followed by individual interviews set up with each of the experts to evaluate the usability of the methodology.

6.1 Overview of Validation Process

Wieringa [45] explains that the simplest way to validate an artifact is by expert opinion. Through this, the design of an artifact i.e. the methodology is presented to the relevant panel of experts who imagine how such an artifact will interact with problem contexts and then predict what effects they think this would have in a specific context. The panel of experts is used as instruments to “observe,” by imagining the artifact in a real-world context. However, it is important to note that validation by expert opinion is only possible if the experts have context on the purpose of the artifact and can imagine realistic problem contexts to make reliable predictions about the usability of the artifact. For this reason, the experts have been carefully chosen based on how well explained they are with the design and development of the methodology throughout the researcher’s work.

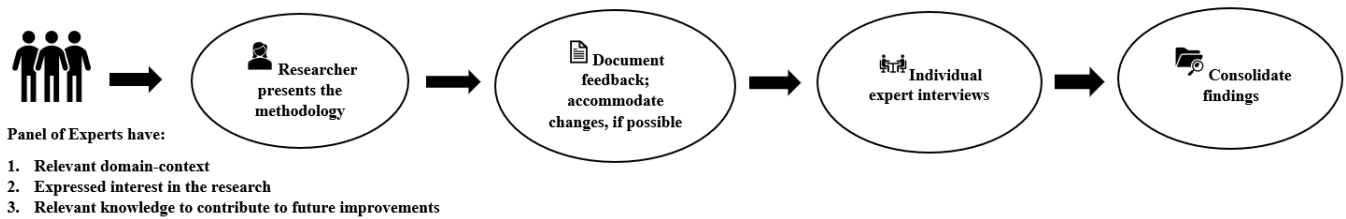


Figure 40: Overview of Validation Process

The validation process of the methodology is two-fold. Fig 40 summarizes the validation process carried out as part of this research with the panels. First, the methodology has been presented to the stakeholders who gather to understand the fully designed and developed methodology with its sub-phases, objectives, scope, and assumptions. This is followed by individual expert interviews with each panelist to dive deep into their feeling towards the methodology and assess the possibility of future adoption within the organization.

6.1.1 Participant Selection

The participants for the expert interview have been selected based on their involvement with the scope of the thesis, role in the PoC at NXP that also aligns with the research results, and areas of interest. Most of the participants have been actively involved in the scope discussions since the inception of the research within the organization. The interviews have been conducted in such a way that the participant profiles are first understood and documented to get an overview during the process of evaluating the responses. Table 12 gives an overview of the participants’ current role in the organization, a brief description, and the length of the interviews.

Participant	Role in the Organization	Role Description	Length of Interview
A	Data Scientist	Work on different use cases to identify the potential of data and the value it can bring to the organization.	00:46:30
B	Data Scientist	Work on different use cases to identify the potential of data and the value it can bring to the organization.	00:37:37
C	Portfolio Manager	Facilitate, direct and assist, break down challenges for the team.	00:24:02
D	Data Analyst	Exploratory data analysis, data visualizations and build machine learning models.	00:26:37
E	Senior Principal Data Scientist	Enable data scientists within NXP to do data science better, improve data literacy and increase productionalization of the models.	00:50:11
F	Project and Program Manager - Analytics	Data-driven analytics on data that matters for R&D to set a strategic direction for the organization.	00:33:38
G	Principal Data Scientist	Enable different profiles within NXP to start using data science and and support with productionalizing their work.	00:30:20

Table 12: Participant Profile

To analyze the results easily, participants have been grouped together on the basis of their role and specialization in their organization. This makes it easier for the researcher to understand their feedback and pain-points towards adopting and eventually using the methodology.

- **Group 1 - Data Experts:** Participants A, B and D are involved in the solution-building process, data engineering, visualizations and generating insights.
- **Group 2 - Management:** Participants C and F lean towards understanding where the organization needs to invest engineering efforts and focus on the bigger picture i.e. to make informed decisions.
- **Group 3 - Operational Experts:** Participants E and G are highly focused towards operationalizing models across the organization and facilitating ML related operations efficiently

6.1.2 Methodology Presentation

The evaluation phase of the methodology starts by first conducting a presentation for the panel in a moderated setting. It is regarded as a precursor to the expert interview and is qualitative in nature. The researcher has presented the results in order to obtain nuanced and unfiltered feedback from the experts on the outcome of the findings. The presentation re-emphasizes the research objective of the thesis, the existing methodology, gaps and limitations identified, the newly proposed methodology and an exhaustive understanding of each phase. The presentation also gives an overview of the next steps that would follow for the experts in terms of individual interviews for the purpose of validation. The presentation has been conducted online via Microsoft Teams by keeping in mind the availability and schedule of the participants. The presentation has also been recorded and can be accessed on

request if need be.

At the end of the presentation, a round-table discussion has been conducted to allow experts to share their thoughts and feedback on the methodology. Most of the experts have been actively involved in the design decisions since the inception of the research so most of the feedback is directed toward improving the visuals of the methodology. The discussion also had questions regarding the iterative process of the methodology and how it can be made explicit that it follows an agile style rather than a waterfall approach.

6.1.3 Interview Format

The individual expert interviews have been designed in a way that the researcher can get the most optimal input from the experts regarding the design, development, and usability of the methodology. As discussed in Chapter 3, the interview style followed is the open-ended interview. Through this format, participants are interviewed with the same set of questions but the flow of the interview can be open-ended i.e. there is flexibility for individuals to be themselves and answer liberally.

The format of the interview is as follows:

- **Get acquainted with Participants** - The researcher asks the participants questions on their role, professional background, years of work experience, importance of having ML tools and technologies from a business standpoint.
- **Design and Development** - The researcher would like to know the experts' opinion and their thoughts on the design and development process towards building the methodology.
- **System Usability Scale (SUS)** - The SUS scale is a questionnaire that determines the experts' perceived usability towards the methodology.
- **Synthesis of Results** - The final findings as a culmination of the literature review, design, development and implementation of the methodology is presented on a Miro board for the experts' to review and add insights, if there are any.

The following sections will go into detail for each of the steps mentioned above with the findings observed at every stage.

6.2 Get acquainted with Participants

The first part of the interview focuses on getting to know the participant, their role in the organization and years of work experience. The results of these observations are tabulated in Table 12. Apart from this, the researcher would also like to understand their perspective on why ML as a concept is important for business stakeholders to understand. The insights gathered from each participant is discussed in this section. The analysis is also made easier by following the grouping of participants discussed in Section 6.1.1.

Group 1 - Data Experts:

According to participant A, business stakeholders more often than not tend to get the false impression that any use case can be solved with support from ML but that is not necessarily the case always. Sometimes simple statistics can be employed on a problem context to solve the issue rather than

building a sophisticated ML algorithm. This also makes business stakeholders underestimate the effort expended by data scientists on use cases to get a satisfactory result.

For participant B, the question of 'how' is more important when it comes to business stakeholders. The 'how' is highly co-related with trust as a factor. There needs to be a certain amount of trust between the business people and the models that are built. If there is no trust, the models remain as PoCs and will never make it to the production environment.

Participant D thinks that ML is a concept that a lot of business people want to start working with hands-on but they suffer from a global understanding on the basic workflow and how it can be used efficiently. It's also easier for a data related stakeholder to be able to have conversations with a business person if they have a high-level understanding on what an ML model does and how it can lead to good predictive performance.

In conclusion, all three participants are of a similar mindset that the key to establishing trust among business stakeholders for informed decision making is if they have a certain degree of understanding on ML as a concept.

Group 2 - Management:

According to participant C, when it comes to business stakeholders, it is eventually all about the ROI for the organization. ML needs to be understood by business stakeholders to be able to make decisions on the basis of their outcome in order to invest in the right area within the organization. The participant also thinks it's about commitment from the business stakeholder to be able to learn as and when technology advances to drive success within the organization.

Participant F thinks that trust is the main factor to be established and requires a certain degree of understanding. A well-established engineering company with smart stakeholders be it engineers or managers, working across all domains need to be able to make sense of the technology advances.

In conclusion, Group 2 participants are highly focused on the business value that can be extracted from ML to be able to drive efficiency and growth within the organization and in the industry.

Group 3 - Operational Experts:

Participant E believes that there exists a lot of misconception on ML as a concept. It can be perceived in two ways - one, where business stakeholders think that ML can be used for anything and can give actionable results. On the other hand, there are stakeholders who are deemed as non-believers and think that ML can never surpass human knowledge. The participant thinks that an understanding of the foundation of ML is key to setting the right expectations for business stakeholders.

According to participant G, business people have a stake in the design and development of an ML model even though they are not actively involved in the solution building process. For this reason, they need to know what they're working with so that they can eventually trust the outcome to make better decisions.

In conclusion, Group 3 participants are similar to other participants where they believe understanding ML on a high-level leads to better decisions because of trust in model outcomes.

6.3 Design and Development

The design and development section of the interview is focused on understanding what the different experts' think of the newly designed methodology in comparison with CRISP-DM and identifying if there are any critical interventions or steps that are missing from the researcher's design decisions. For this section, there are three sub-steps to address these questions. They are as follows:

- Adequacy of the methodology design
- Comparison with existing methodology
- Critical interventions

6.3.1 Adequacy of the methodology design

Through this part of the interview, the researcher wants to understand from the experts if they believe the design of the methodology is adequate enough to be incorporated as part of their workflow.

According to participant A, the methodology has been well-designed and encapsulates important components of a typical project workflow that involves ML. In addition to this, there are new phases that have been added that complement the need to establish interpretability among stakeholders specifically the fifth phase i.e. present to stakeholders. The participant also re-iterates that the methodology seems complete but the actual usability can be tested once it is implemented in a use case. The participant also explains this constraint with the help of an example - in the case of choosing the right ML algorithms. According to one's expertise in forecast-related use cases, one would easily be able to identify the available algorithms in the market with minimal research on how they can be utilized. However, in the case of an image processing use case that one has never worked on, more time would be spent on assessing the best model for that particular use case. Without knowing when to make a decision, one would make assumptions and end up choosing something, which may not be able to give adequate results to be able to present to the stakeholders. These kinds of experiments need to be tested with the design of the methodology.

Participant B also agrees that the methodology has been designed well keeping in mind the core concept i.e. interpretability and can be applied to all use cases going forward. Likewise, for participant C, the methodology design is well-structured and can be embedded in use cases.

Participant D agrees that the design of the methodology looks visually appealing and gives a good overview of the necessary steps that need to be done. They also think that the well-integrated steps makes conversations on expectations between business stakeholders and data scientists, engineers, etc. easier to facilitate.

For participant E, the methodology is an extension of CRISP-DM and is adequate in terms of the additional steps that have been integrated to address interpretability within the organization. They also mention that in terms of design, it is an improvement, adequate enough to be able to incorporate but not dramatically different. According to participant F, there is always room to learn while putting things into practice. The design of the methodology implies that it can work on use cases, so they think it's adequate enough.

Participant G thinks that the visual design of the methodology comes across as complex but what really helps is the fact that there is a clear connection with a common and well-know framework such as CRISP-DM. The design of the methodology also includes additional components for the purpose of interpretability which makes it complete and addresses the necessary steps.

6.3.2 Comparison with existing methodology

This part of the interview focuses on extracting insights from the experts on how the newly proposed methodology is more advantageous than what is currently in use i.e. CRISP-DM. Since most of them are familiar with the current methodology and the various phases involved, the insights are valuable to validate the design of the proposed methodology.

Both participants A and B agree that the design of the methodology is an extension that is well laid on CRISP-DM. They think it is appropriate and can be an improvement to the workflow, if incorporated, to introduce an agile mindset and constant engagement with relevant stakeholders.

For participant C, the value-add through this methodology is the fact that there is better business alignment and interpretability of components that are usually complex in nature. They also think that CRISP-DM lacks these essentials i.e. interpretability, human involvement, stakeholder validation, etc. CRISP-DM maybe project management driven but it suffers from business alignment driven, so the newly proposed methodology fills that gap.

Participant D believes the changes in the proposed methodology have simplified the process in general while keeping in mind the core concept i.e. interpretability. The newly structured phases and sub-steps with their respective objectives make it easier to communicate with a business stakeholder. They also think that CRISP-DM is very data engineering focused with phases like evaluation, deployment - these are essentials that a data scientist needs to do. However, the newly proposed methodology goes beyond this and emphasizes on the fact that the responsibilities extend beyond a data scientist and data engineer i.e. more stakeholders are involved.

According to participant E, the methodology highlights the interpretability concept, enforces presentation to stakeholders, gathers feedback and monitors performance after it's deployed. This is implicit in the traditional CRISP-DM but in the new methodology, it's more explicit and structured.

Participant F and G believe the methodology is business-focused to ensure there is interpretability at every phase and a commonly agreed upon success criteria from a business and ML standpoint. This is because it's difficult to estimate or commit to accuracy or a quantitative metric, but it's important to estimate in terms of business perspectives i.e. what the model should ideally be for the business to make better decisions.

6.3.3 Critical interventions

The last part of the design and development section aims to understand if there are any critical steps that have not been addressed by means of the methodology. Table 13 consolidates the various suggestions given by the experts to be included in the methodology.

Participant A mentions that monitoring the model to check if it deviates from the earlier set baseline

is an important step that could be included. This can be introduced by means of A/B testing to check if the new model based on feedback is performing better than the existing model after presentation to stakeholders. The aim of this step is to be able to replace the existing model with the better model, if there is better performance.

For participant B, the critical step that could be a value-add is a step to prepare the environment to work on. The rationale being data scientists and data engineers often spend a lot of time deciding the environment such as the cloud environment and tools. To avoid delays during the solution building process, having such a step would be able to save time and effort.

Participant C thinks the methodology doesn't cover operational support which includes incident management, debugging among other things that are more operational focused. Questions such as 'Are the servers still running?', 'Do we have new change requests?', 'Is there a migration plan that the team needs to take into account?' could be answered with support from an operational phase within the methodology.

In participant D's opinion, adding an explicit deployment phase to make it end-to-end would be considered a value-add.

Participant E believes in general none of the existing methodologies in the industry focus on having peer reviews for data scientists, data engineers. The team by itself may not have all the expertise on certain topics so reviews can also be extended to getting opinions from external experts at the early, mid and end phases of the project to get approval and valuable insights.

Participant G thinks that The output of any data science effort is a running service for end-users to interact with. Therefore, a missing component could be a phase or sub-step to productionalize the ML model for end-users to get a feeling of how the model would actually look like.

Suggestions from Experts	At which stage of the methodology can it be included?
A/B Testing - To compare new model performance with baseline model	6th phase - Monitoring and Usage
Prepare production environment	1st and/or 2nd phase - Business and Data Understanding
Peer and Expert Reviews	3rd, 4th, 5th phases - Data Preparation, Model Building & Evaluation, Present to Stakeholders
Operational Support - Incident management, debugging, etc.	After 6th phase - As part of ML-Ops
Productionalize ML model	After 6th phase - As part of ML-Ops

Table 13: Overview suggestions from experts for improving the methodology

Intention	Points
Strongly disagree	1
Disagree	2
Neutral	3
Agree	4
Strongly agree	5

Table 14: Intention and Points associated with SUS

6.4 System Usability Scale (SUS)

The System Usability Scale (SUS) is a reliable method of measuring the perceived user experience when the subject interacts with a digital product, tool, website, methodology or framework. The technique employed by the SUS is to identify extreme expressions of the participants' attitude [67]. It consists of ten statements with five response options for respondents; from "strongly agree" to "strongly disagree" ranked on a scale from 1 to 5. Participants will rank each statement from 1 to 5 based on how much they agree with the statement. 5 means they agree completely, and 1 means they disagree vehemently. It contains alternating statements to avoid response biases among participants.

Questionnaire

The questionnaire consists of ten statements that focus on measuring the perceived user experience. The ten statements used for the purpose of this validation process are as follows:

1. I think that I would like to use this methodology frequently.
2. I find the methodology unnecessarily complex.
3. I believe the methodology would be easy to use.
4. I think that I would need the support of a user manual and/or technical guidance to be able to use this methodology.
5. I think the various phases and sub-steps in this methodology are well integrated.
6. I think there is too much inconsistency in this methodology.
7. I would imagine that most people would learn to use this methodology very quickly.
8. I find the methodology very cumbersome to use.
9. I would feel confident using the methodology in forthcoming projects.
10. I need to learn a lot of things before I could get going with this methodology.

Calculation of SUS score

The participants will have ranked each of the 10 statements on a scale from 1 to 5, based on their level of agreement. Table 14 summarizes the points mapped against the intention of the respondent's feeling towards a particular statement.

Step 1 - Calculate odd numbered statements

The first step is to calculate the score for all odd-numbered statements. This is done by first looking at the number of points scored by the participant and then deducting 1 from it i.e. $(X-1)$.

Example: If a participant has ranked an odd-numbered statement as 'Agree' i.e. $X=4$ according to Table 14, then one point is subtracted from this. This makes the score for that statement 3 points.

Step 2 - Calculate even numbered statements

The second step is to calculate the score for all even-numbered statements. This is done by looking at the number of points scored by the participant and then subtracting that from 5 i.e. $(5-X)$.

Example: If a participant has ranked an even-numbered statement as 'Disagree' i.e. $X=2$ according to Table 14, then subtract two from five. This makes the score for that statement 3 points.

Step 3 - Calculate the SUS Score - $(\text{Score step 1} + \text{score step 2}) * 2.5$

The final step to calculate the SUS score is done by adding up all the points from Steps 1 and 2. The result is then multiplied by 2.5.

Example: The total number of points obtained from step 1 is 15 ($3 + 4 + 2 + 4 + 2 = 15$). And from step 2 is 10 ($2 + 4 + 3 + 1 + 0 = 10$). Add 15 and 10 together to get a total of 25 points. This is then multiplied by 2.5 $(15+10)*2.5 = 62.5$. The SUS score is 62.5. The researcher has automated the calculation of SUS score using Google Sheets and Java Script [68].

Interpretation of SUS score

The SUS score explains the usability performance of the methodology in terms of effectiveness, efficiency, and overall ease of use. Although the responses from participants are calculated as a score that's between 0 and 100, it is not a percentage or percentile score. The average SUS score is 68 [67]. This means that a score of 68 is ranked at the 50th percentile. According to the Proprietary scale developed by Jeff Sauro [69], a percentile and letter grade are used as indicators to report an SUS score. Table 15 summarizes the grade and percentile associated with an SUS score.

6.4.1 Results of SUS

The results of SUS have been tabulated in Table 16. The table shows the ten statements that have been used to ask the participants about their opinions and the points each participant has assigned to the statements. The last row in the table shows the calculated SUS score for each participant. Comparing the results with the percentile and grade in Table 15, we can see that four participants have an SUS score that is associated with 96th percentile and an A+ grade. Two participants have SUS scores that are in the 90th percentile with an A grade. One participant has an SUS score that is ranked in the 50th percentile with a C grade. Figure 41 shows the separation of odd and even numbered statements with the average point given by all the stakeholders for each statement. Since the questionnaire consists of ten statements with alternating items, the odd numbered statements are formulated towards evaluating

Grade	SUS	Percentile
A+	84.1 - 100	96 - 100
A	80.8 - 84.0	90 - 95
A-	78.9 - 80.7	85 - 89
B+	77.2 - 78.8	80 - 84
B	74.1 - 77.1	70 - 79
B-	72.6 - 74.0	65 - 69
C+	71.1 - 72.5	60 - 64
C	65.0 - 71.0	41 - 59
C-	62.7 - 64.9	35 - 40
D	51.7 - 62.6	15 - 34
F	25.1 - 51.6	2 - 14

Table 15: Percentile and Grade associated with the SUS score

the positive aspects of the methodology i.e. higher the point associated with the statement, the better the overall SUS score would be. Similarly, even numbered statements target the participants feeling towards the methodology in a negative light i.e. lower the point, the better it would be for the overall SUS score. It can be seen that the odd numbered statements have a color hue of green that indicate relatively high responses which means the participant feel confident to a good extent. The even numbered statements are also performing well except for the second statement that has a dark red associated with it indicating that a certain participant feels the methodology would require additional support in the form of a manual and/or technical guidance.

The next section describes individual statements in detail along with the participants' feeling towards it. It also discusses feedback given by participants that stand out and need more attention.

ODD NUMBERED STATEMENTS -> HIGHER THE BETTER

Questions	Average
I think that I would like to use this methodology frequently	4.29
I believe the methodology would be easy to use	3.86
I think the various phases and sub-steps in this methodology are well integrated	4.43
I would imagine that most people would learn to use this methodology very quickly	4.14
I would feel confident using the methodology in forthcoming projects	4.57

EVEN NUMBERED STATEMENTS -> LOWER THE BETTER

Questions	Average
I find the methodology unnecessarily complex	1.14
I think that I would need the support of a user manual and/or technical guidance to be able to use this methodology	3.00
I think there is too much inconsistency in this methodology	1.29
I find the methodology very cumbersome to use	1.14
I need to learn a lot of things before I could get going with this methodology	1.43

Figure 41: Odd and Even Numbered Statements - Average Scores

1. I think that I would like to use this methodology frequently

While participants B, C and F strongly agree towards using this methodology frequently in their use cases, participants A, D, E and G are either neutral or agree to a certain extent. The participants

Statements/Participant	A	B	C	D	E	F	G
I think that I would like to use this methodology frequently	3	5	5	4	4	5	4
I find the methodology unnecessarily complex	1	1	1	1	1	1	2
I believe the methodology would be easy to use	5	3	4	4	4	3	4
I think that I would need the support of a user manual and/or technical guidance to be able to use this methodology	2	2	4	2	5	3	3
I think the various phases and sub-steps in this methodology are well integrated	3	5	5	5	3	5	5
I think there is too much inconsistency in this methodology	1	1	1	1	3	1	1
I would imagine that most people would learn to use this methodology very quickly	4	4	4	5	4	4	4
I find the methodology very cumbersome to use	2	1	1	1	1	1	1
I would feel confident using the methodology in forthcoming projects	5	5	5	5	4	5	3
I need to learn a lot of things before I could get going with this methodology	1	1	2	1	2	2	1
SUS SCORE	82.5	90	85	92.5	67.5	85	80

Table 16: Participants and SUS Results

feel the methodology would be beneficial to them as it's an extension of CRISP-DM but they're unable to estimate the extent to which it would be useful without using it in a use case at least once. Participant A also added that ad-hoc requests from stakeholders related to various aspects such as visualizations, business-related clarifications, data engineering, model performance cannot be limited to a single iteration. Likewise, participant D assures that personally, they wouldn't know how the interaction of the methodology would be with a business stakeholder who needs a grasp of some of the data engineering-related concepts. All the participants believe that the methodology addresses interpretability as the core concept and it would be a value-add in their use cases.

2. I find the methodology unnecessarily complex

This statement has similar responses across all participants ranked at strongly disagree i.e. '1' with the exception of one participant ranking it at disagree i.e. '2'. All of the participants agree that the methodology by itself is not complex as it has been visualized clearly, well-explained and includes necessary steps to bring about a degree of interpretability. Participants A and B also add that it depends on the stakeholder using the methodology i.e. for non-technical stakeholders, the methodology might warrant additional support. As for data related stakeholders, the phases in the middle are quite important and well-known except for the additional steps introduced in this methodology. Participant D also adds that ML as a concept in itself is overtly complex, so a methodology would be helpful in giving all the stakeholders an overview on what needs to be done and at which phase.

3. I believe the methodology would be easy to use

The general tonality while answering this statement is the same across all participants. All of them think that the methodology is straightforward and easy to use but since it involves multiple stakeholders across all domains, there may be some apprehensions in getting them all board to follow a structured way of working. Participant C also thinks this can be mitigated with the help of easy-to-use templates to support stakeholders in getting familiar with the methodology in practice.

4. I think that I would need the support of a user manual and/or technical guidance to be able to use this methodology

This statement has mixed feelings associated with the points across all participants. Participants A, B and D who are also grouped together as data experts believe that the need for a user manual and/or technical guidance is subjective based on the stakeholder expertise in working with agile methodologies. They also think that as data scientists, they may not need additional support in working with the methodology, but having an explanation in some form i.e. text description and/or visuals would help. Participant C believes to be successful in using the methodology, it would be helpful to have both a manual and guidance on technical skills as a pre-requisite. Participant E, who falls under the category of operationalizing ML models, strongly agrees with a user manual to better understand the objectives and goals necessary at each phase to ensure interpretability.

5. I think the various phases and sub-steps in this methodology are well integrated

For this statement, participants B, C, D, F and G strongly agree that the various phases and sub-steps are well-integrated. They think that the phases make sense chronologically and are positioned well within the methodology. There is also flexibility for the phases to be iterative in nature because of the feedback loops giving leeway to be able to move from one phase to another without having to follow a particular order. However, participant A explains that though there is a high-level understanding of the workflow, there is still a gap in understanding what each phase essentially needs to do. This can be mitigated with the help of a manual in place. Participant E also agrees that having a global and unified view of what needs to be done at each phase would be a value-add.

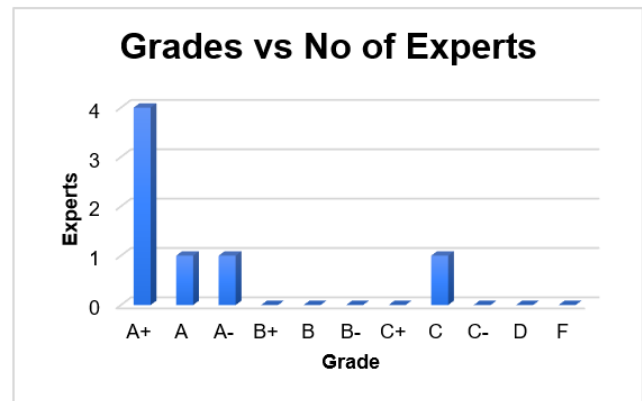
6. I think there is too much inconsistency in this methodology

The responses to this statement are ranked at the same point except for one participant who feels neutral towards it. All the participants think the methodology is clear in terms of the overall flow and the sub-steps that belong to each phase. They also see themselves doing each of those phases in forthcoming use cases. Participant E feels there needs to be more consistency in terms of interpretability as a component for everyone to easily understand and adhere to.

7. I would imagine that most people would learn to use this methodology very quickly

For this statement, most of the participants strongly agree to the fact that most people would learn to use the methodology very quickly. They think that stakeholders with an agile mindset and those that have already worked with CRISP-DM will not struggle too much to adapt themselves to this methodology. Participant D also adds that stakeholders with a background in the basics of data science and

Grade	Experts
A+	4
A	1
A-	1
B+	0
B	0
B-	0
C+	0
C	1
C-	0
D	0
F	0



(a) Overall grade associated with the no. of participants (b) Visualization of the no. of experts vs overall grade

Figure 42: No. of participants vs the overall grade

ML-related concepts shouldn't find the methodology and phases too surprising. Participants A and F are of the same mindset that conceptually the methodology makes sense but putting it into practice and getting everybody to understand could be a challenge to a certain extent.

8. I find the methodology very cumbersome to use

This statement has uniform responses across all participants. All of them agree that the methodology requires effort to understand and be able to apply in practice but it is not cumbersome to use.

9. I would feel confident using the methodology in forthcoming projects

While most of the participants strongly agree to use the methodology in forthcoming projects, some of them are a little apprehensive as 'confidence' is a broad term. They agree that it would be beneficial in many ways, but usage confidence is something that cannot be measured well in advance.

10. I need to learn a lot of things before I could get going with the methodology

The common feeling towards this statement is that if it's strictly about learning, they wouldn't have to as the researcher has kept them well-informed on the methodology, the phases, sub-steps, and what it entails. However, one of the participants believes having a well-structured user manual in place would make the process of adopting the methodology within the team easier.

6.4.2 Summary of SUS Findings

Fig 42 shows a visualization of the number of experts that are associated with a certain grade. This shows that of the seven experts who participated in the validation process, four of them (B, C, D and F) with SUS scores of (90, 85, 92.5 and 85) respectively are fully satisfied with the usability of the methodology and are confident in adopting the best practices going forward. Participant A with SUS score of 82.5 is a close second who agrees to an extent the methodology is fully capable of being

adopted in forthcoming projects but they wouldn't know for sure without having implemented it at least once. Similarly, participant G with an SUS score of 80 believes the confidence level of using the methodology is subjective even though the methodology covers almost all instances for an ML project to achieve the necessary level of interpretability. Finally, participant E with an SUS score of 67.5 strongly thinks that a user manual and/or technical guidance is mandatory to be able to make sense of the methodology requirements and its capabilities.

The overall SUS score by taking an average of all individual scores is 83.21 i.e. it is associated with grade A with a percentile between 90 and 95. Collectively, the results seem to be in the positive direction with respect to the effectiveness, efficiency and satisfaction of the methodology.

6.5 Synthesis of Results

The final portion of the interview is when the researcher presents a Miro board that has three boxes. The Miro board has been constructed as a result of the observations documented throughout the duration of the thesis complemented by the findings from the literature review, the design, and development of the methodology, and interactions with different stakeholders. Each box has sticky notes with findings that have been observed throughout the duration of the thesis. Fig 43 shows the Miro board representation with the findings. The three boxes represent the following:

1. **Continue-** This box represents retrospective. The sticky notes in this box indicate the steps that need to continue to achieve better results. Identifying the right team and experts in a project is something that is working well as part of the current use cases. Similarly, decision-making always has a component of human factor involvement and does not rely solely on model outcomes.
2. **Stop-** This box represents items and activities that need to stop or slowly fade out in time. The sticky notes in this box will complement the sticky notes in the going forward box. The lack of a standardized methodology could be potentially disadvantageous to driving results as there won't be a single source of truth or structure for the stakeholders to fall back on. Similarly, relying only on heuristic-based decision making i.e. the experience of stakeholders, and dismissing data-driven insights could bring in bias in results. Lastly, interpretability being considered as an afterthought could prove to be disadvantageous where a model with high predictive performance is built without adding value to the context of business objectives.
3. **Going Forward-** This box represents the next steps. Once the items that need to continue and stop are understood by the stakeholders, the sticky notes in this box will be easier to map and relate to. First, involving all relevant stakeholders from project inception to introduce interpretability is important i.e. identify the right stakeholders. Similarly, it is important to identify ML-related objective in addition to what the business expects as an end result. This is to ensure the people involved in the solution-building process understand the objectives of a model outcome from a business standpoint. Third, decision-making should always be seen as a combination of human experience and data-driven insights and not a standalone process. Last, the outcomes presented to stakeholders need to be validated to ensure they understand and are satisfied with the outcomes.

Participant A believes that the proposed methodology can contribute to achieving the sticky notes in the yellow box, provided there is a fit between the team and the process. It also highly depends on the

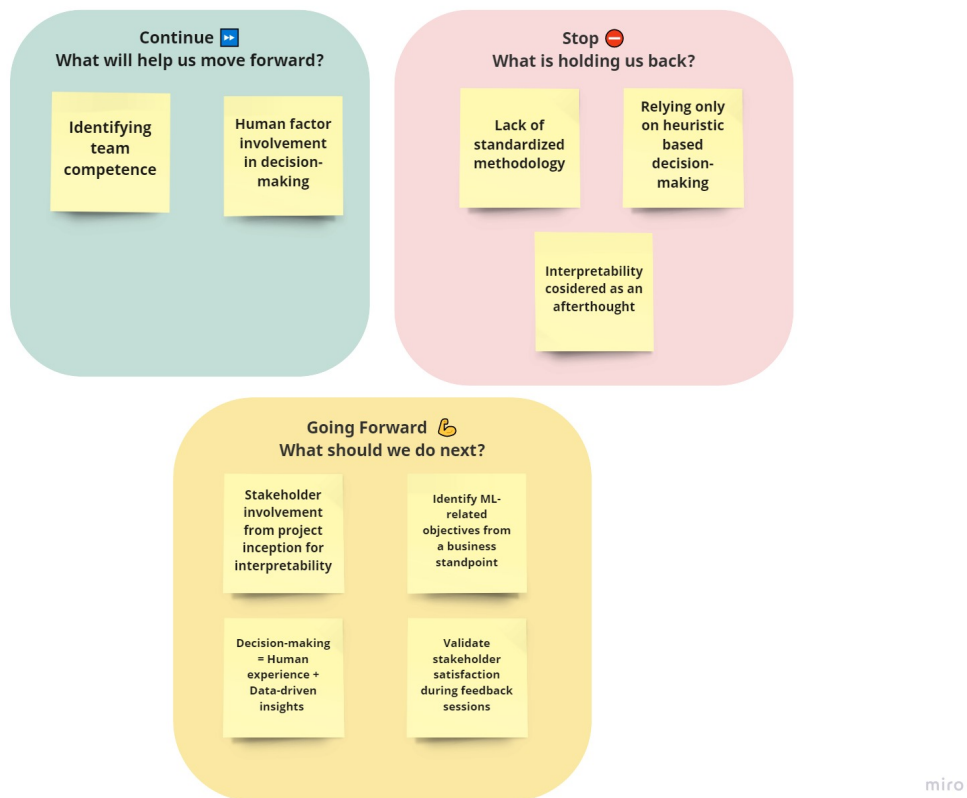


Figure 43: Miro Board - Representation of Findings

use case and the model that needs to be built.

Participant B is also of the belief that the stakeholder involvement since the inception of the project can add value along with the phase where the results need to be presented and validated by stakeholders. There will also be a structured way of working that will lead to achieving the items within the yellow box.

According to participant D, when considering the size of a team, as the team grows, there will be difficulties in managing and staying up to speed with tasks. Utilizing the methodology would be beneficial internally to prioritize and direct tasks. For an organization-wide perspective, it depends on how the company has structured its data. In the case that the data is siloed, getting everybody to follow such a structured methodology could become tedious but there could be progress. Participant F also agrees that there are standards in place to choose from for a workflow but the formal application of them in practice and getting everybody involved is where the real challenge lies. If that can be achieved with the help of this methodology, the items in the yellow box can be achieved within the team and eventually in the organization. Similarly, participant G adds that the yellow box would be helpful in bringing the business on board with the core concept of data science and machine learning in a way that they can trust the outcome and make informed business decisions.

6.6 Summary

This chapter gives a detailed explanation of the process followed to validate the findings and methodology through expert interactions and interviews. From the results, it can be observed that the researcher has made an effort to validate the overall methodology qualitatively by means of expert interviews but also add a quantitative evaluation method in the form of SUS. The findings have been documented in detailed descriptions and visualizations for the audience to understand the results.

7 Conclusion and Recommendations

This chapter first presents the key takeaways of the thesis and the contributions to both science and practitioners along with practical recommendations to the organization. It gives an overview of the insights that have been gathered as part of the research and answers the research questions that have been formulated in Chapter 1. The chapter also includes a discussion on the threats to the validity of the research. Lastly, a practical overview of the limitations and future work is discussed.

7.1 Key Takeaways and Contributions

The thesis presents a methodology i.e. iCRISP-ML for relevant stakeholders to understand the impact of ML models and their outcomes in a business context. However, ML models by themselves are not a panacea. Understanding how they work to an extent, their limitations, and the impact of outcomes on the business can help stakeholders make informed decisions about when and how to use them, and how to properly interpret and act on the results. For instance, if a business stakeholder understands that the quality of an ML model is only as good as the training data i.e. the data it is trained on, they may be more cautious about relying on the model's predictions or recommendations if they suspect that the data used to train the model may be biased or otherwise flawed.

iCRISP-ML has been developed as an inclusive and business-friendly methodology for developing interpretable ML models. The methodology has a structured approach that is accessible and understandable to business stakeholders to ensure that the models being developed are fit for their intended purpose and that they are being used appropriately to support the business objectives. Involving all relevant stakeholders in the development of an ML model, from the business understanding phase through to deployment, can be an effective way to ensure that the model is interpretable and meets the needs of the business. iCRISP-ML can also help facilitate communication and collaboration between data scientists and business stakeholders, which can be critical for the success of an ML project. It provides a common language and framework for discussing the problem, the data, and the model, and for identifying and addressing any issues or concerns that may arise throughout the lifecycle of the project.

7.1.1 Contributions to Science

The thesis contributes to the knowledge gaps of the problem context in different ways. The contribution as part of the thesis to the academic community is two-fold. First, the thesis presents and demonstrates a methodology specific to the problem context by taking into consideration existing frameworks and methodologies. Secondly, the thesis introduces a methodology that not only addresses the technical standpoint but aligns with the best practices of project management.

The research on the importance of interpretability in ML models is widespread given the novelty of the concept within the ML community of practitioners. The systematic literature review performed as part of the thesis gives an overview of the existing landscape of research and literature conducted on this topic. While there are studies that focus on frameworks as potential means to overcome this challenge, none of the literature studies examine the application of interpretability in ML from a business standpoint. The existing literature to an extent emphasizes strongly the need to explore the spectrum of interpretability techniques while building ML models, but there are no formal applications in a practical and scientific context for an organization to adopt.

The systematic literature review provides insights into the lack of a standardized methodology that can potentially be used to encourage stakeholder involvement, both business and technical, to build interpretable ML models right from the inception of a use case. To bridge the research gap, the present research builds on top of an existing methodology and knowledge such as CRISP-DM which is widely used within the academic community. This makes it easier for relevant practitioners to adopt the methodology without having to learn something new extensively.

7.1.2 Contributions to Practice

The contribution of the research findings to practice and industry practitioners has been consolidated based on the researcher's observations throughout the research and the results of the validation with experts.

Industry giants such as IBM and Microsoft have iterated on CRISP-DM to produce their variants with more iterative loops between phases such as data preparation and modeling. However, these refined methodologies aim to deliver their premium service engagements or are considered product-centric to meet specific needs. Therefore, these vendor-centric, proprietary methodologies suffer from general-purpose adoption by other organizations that do not typically have diverse technology needs. The thesis introduces a methodology that includes the best practices of project management while including the necessary level of interpretability at each phase. Practitioners can make use of the methodology for a given use case to align business objectives and leverage capabilities of advanced analytics, ML, etc. to improve efficiency and provide actionable insights. The general architecture of the methodology that is presented can be used as a starting point for practitioners to build interpretable models to make informed decisions. On a high level, the methodology presents a set of phases and sub-steps that are integrated and can be augmented to the existing business process and workflow of the organization.

The methodology also proposes an interpretability checker to guide the stakeholders involved in the solution-building process to choose the right techniques for the right ML model while testing for interpretability. Practitioners can directly make use of the checker to be able to assess the techniques that can be used, how much to explain, and the target audience to whom the results need to be presented. Stakeholder validation with metrics has been included as part of this methodology i.e. to facilitate the evaluation of how stakeholders feel towards the outcome and not just quantitatively determine the ML performance through a performance metric. The methodology can also be made compatible with commonly used agile processes such as Scrum and Kanban to structure the workflow within the team.

The validation of the usability of the methodology leans towards positive responses from experts in the domain. Business stakeholders, data science experts such as analysts, and applied scientists have assigned satisfactory scores using the SUS to use the methodology going forward. These points indicate to the audience that the methodology is not only easy to use but can be satisfactory, effective, and efficient in a real-world use case.

7.2 Research Questions

The main research question and sub-research questions discussed in Chapter 1 will be answered in detail as part of this section.

RQ: What is an appropriate methodology to build interpretable ML models for business stakeholders to trust and make data-informed decisions?

The definition of 'appropriate' relates to the usability of the methodology in a real-world context i.e. the level of effectiveness, efficiency, and satisfaction it can bring about for the stakeholders who are directly interacting with the methodology in practice.

The thesis successfully presents a conceptual and visual representation of a methodology that explains the process of building better and interpretable ML models from a business perspective. Within the methodology, recommendations have been proposed to identify success criteria that pertain not just to the business but also to the ML model that needs to be able to meet those objectives. In addition to this, the methodology also encapsulates the importance of interpretability at every phase by recommending stakeholders be aware of their responsibilities and action items that need to be completed. The core of the methodology lies in the concept that an ML model with high predictive performance is not enough to establish interpretability, but it also depends on the reception of the model outcomes by the stakeholders. The stakeholder validation with the three proposed metrics such as quality of the mental model, stakeholder satisfaction and degree of understanding help address this concern with an intuitive validation at the end of receiving a model outcome.

Moreover, iCRISP-ML can also be adopted within agile practices such as Scrum, and Kanban and can be tailor-made to suit the workflow of the team and the use case at hand.

SQ1: What are the analytical methods in practice for business decision-making?

There are four commonly used analytical methods in practice for business decision-making. They are - descriptive, diagnostic, predictive, and perspective analytics.

Descriptive analytics is the most common type of analytical method found in business that uses historical data and instances from the past to derive insights for the business. The insights are mostly represented in dashboards and in reports which are convenient and easy ways to consume the data and make informed decisions.

Diagnostic analytics focuses on understanding the rationale behind why a certain instance has occurred. This type of analytical method focuses on drilling down into the data to understand patterns, identify correlations and explain anomalies. Through diagnostic analytics, causal relationships can also be determined using advanced statistical techniques to search for patterns and correlations.

When organizations are inundated with data from different sources that reside in various silos, predictive analytics helps finding ways to gain insights by combining statistical and machine learning techniques. It shifts the focus from understanding historical events and instances of the past to creating insights about future outcomes in a real-world context. Through predictive analytics, organizations can be in the know of what will most likely happen in the future, the likelihood of potential outcomes and streamline the best course of action to steer the organization towards that direction.

Prescriptive analytics yields recommendations and best course of action as next steps. The optimal course of action is recommended on the outcomes from predictive and descriptive analytics that are

conducted prior to prescriptive analytics. With efficient use of prescriptive analytics, organizations can make decisions based on informed facts rather than assumptions. It can also help organizations simulate the probability of various outcomes to understand the level of risk, uncertainty and the likelihood of worst-case scenarios.

SQ2: Why do business stakeholders find it difficult to understand the outcomes that ML models generate?

Business stakeholders, unlike those that are involved actively in the solution building process, often lack the technical knowledge underpinning an ML model. Moreover, since AI/ML technology suffer from the black-box problem i.e. lack of visibility into the internal working of the model, it makes it all the more difficult for data scientists themselves to interpret and there after explain their results. This makes business people reluctant to trust the outcomes that an ML model makes. Despite ML models generating predictions that have a profound predictive performance and impact, it's not easy for decision-makers to trust them easily.

For business stakeholders, it's often hard to grasp the concept of entrusting a model with the task of decision-making, let alone make them better than a human expertise can, especially when there is insufficient understanding of how an outcome is made.

SQ3: What is the difference between explainability and interpretability according to the literature?

The terms 'explainability' and 'interpretability' are often used interchangeably within the ML community and practitioners. While they are very closely related, the two of them do have differences that are key to identifying what it is that one wants to achieve with a model outcome.

Interpretability refers to the extent to which a cause and effect can be observed within an ML system or a model outcome i.e. the extent to which one can predict what is going to most likely happen. It's being able to discern the outcome of an ML model without necessarily knowing the why behind the decision.

Explainability, on the other hand, is the extent to which the internal workings of an ML model can be broken down in simple and digestible format to humans i.e. to explain quite literally how the ML model makes a particular decision.

SQ4: Why is it important to introduce a degree of interpretability in ML model outcomes?

The reasons to introduce a degree of interpretability in ML model outcomes are twofold -

1. First, to increase trust among business stakeholders. This can be achieved by giving them more visibility into the ML model outcomes and the best course of action that can be recommended as a result of them. By following the best practices of interpretability techniques such as showing the most important features that has high business impact supported by the existing knowledge they possess on the feature, makes them trust the outcome easily.
2. Having interpretability techniques in place, can also improve overall troubleshooting to check how well a model is work. With increased visibility in the way a model makes a certain decision, data experts such as scientists, engineers and developers can also try and change input parameters or debug error-prone variables to give better predictive performance and increase

the efficiency of the model.

SQ5: How can a methodology be designed to build ML models that are interpretable?

This sub-research question seeks a way to design and develop a methodology to ensure ML models are interpretable when they're presented to stakeholders and/or decision-makers. The design decisions and process of developing such a methodology have been backed by extensive literature and existing knowledge. By taking inspiration and identifying knowledge gaps from a methodology that is popular among academic and industry practitioners i.e. CRISP-DM, the researcher designs a methodology that is well-laid on top of the layers of CRISP-DM. The methodology is designed in a way that stakeholders who potentially interact with it can easily understand the various phases and sub-steps that are integrated. The methodology keeps in mind the core concept i.e. to establish a sufficient level of interpretability for stakeholders at each phase so that when the actual outcome is presented during a report-out, a certain degree is already established to make informed decisions.

SQ6: How can the developed methodology be evaluated in the organization?

iCRISP-ML has a well-laid foundation in terms of insights gained as part of the systematic literature review and design decisions being scientifically backed by existing knowledge and methodology. The evaluation conducted as part of this thesis is twofold. First, the methodology is executed in a real-world case study by the researcher to check if the phases and sub-steps are well-integrated and can be applied in future use cases, going forward. The result of the case study execution by the researcher shows only a portion of the large use case being demonstrated within the scope of the methodology. To validate this further, the design, development, and perceived usability of the methodology are evaluated using qualitative interviews with relevant experts within the organization. The responses obtained from the experts are measured using a scale and discussed extensively. They are overall positive in terms of usability, satisfaction, effectiveness, and efficiency of the methodology. There are some outliers in the responses, but still positive nonetheless.

Since this is a relatively broad scope, within the time frame given to the researcher, the methodology has been executed and validated to support the adoption in forthcoming use cases. The purpose of the twofold validation process conducted for the thesis is to demonstrate that the methodology is suitable for practical usage with sufficient theoretical backing as well.

7.3 Recommendations

The research has been conducted in collaboration with NXP Semiconductors and the following recommendations have been formulated to encourage stakeholders and others to adopt the best practices and contributions of the research. First, the researcher recommends developing a plan of action to drive change management to successfully adopt the methodology within the workflow. There is sufficient technical expertise and knowledge to leverage ML in the as-is scenario, as well as the resources, data, and production environment. However, the scope goes beyond that; it includes engagement and commitment toward accepting a new methodology and adapting to it. Given that the experts in the panel agree strongly that the methodology would be of high value if adapted, the researcher recommends NXP detail a change management plan that includes a quick onboarding and structuring within the team to make use of the methodology.

Secondly, the team aspires to work towards an agile and iterative approach. The researcher recommends identifying a structured agile approach to recognize the volatility of developments in the project, and respond to changes without going off the rails. In this regard, the researcher recommends the team to choose an agreed upon process such as Kanban or Scrum that can work alongside iCRISP-ML. While Kanban is a continuous flow of tasks or user stories with visualization of the workflow on a board, Scrum has regular, fixed-length sprints. Kanban can be used when there is no hard deadline mapped against a user story and the team is open to ad-hoc requests. Scrum is more structured and disciplined towards accepting tasks and focuses on smaller increments of work to deliver at the end of each sprint.

Thirdly, the researcher also recommends NXP ensure all stakeholders are actively involved at various checkpoints within the phases of iCRISP-ML to conform to the core requirements of the methodology. Even if it's practically challenging to get everybody on board to follow something disciplined and structured, a high-level understanding of the methodology requirements can help with achieving the desired outcome.

7.4 Threats to Validity

The researcher confirms that the thesis has been implemented and documented carefully utmost agreed upon. While the scope of the thesis revolves around highly sensitive data, the researcher has made sure confidentiality-related constraints have been discussed carefully before documenting them in the report. However, there are possible threats to the validity of the research and they've been discussed explicitly in this section.

First, there is a threat to the validity of iCRISP-ML from an organization standpoint. Although the researcher herself has demonstrated the methodology in practice with the help of a case study, the expert interviews conducted only measure the perceived usability of the methodology in a future context. Unless the methodology has been tried and tested by one of the experts in the panel, there cannot be an assurance that the study has proposed a viable methodology for practice. However, one way to mitigate this threat is with the assumption that, of the seven experts in the panel, at least five of them have observed the researcher directly interacting with the methodology and producing viable results for the team. Moreover, the design choices that led to the development of the methodology have been validated by seven experts and the team is already underway with the usage of the methodology in the organization.

Secondly, the researcher has chosen the experts for the purpose of validation on the basis of their existing knowledge and involvement to an extent throughout the duration of the research. This can also be considered as a case of bias in choosing the sample for the validation. While this may be the case, the researcher also assumes that these experts are the actual users of the methodology going forward. The possibility of bias has been mitigated by:

- Recording interviews and documenting transcripts (Available on request)
- Following a structured set of interview questions for all participants
- Using a reliable usability scale to derive a quantitative metric that can be assigned to the validation process

Finally, the researcher has limited the scope of the research to the first five phases of the methodology and not considered the last phase. While this is also attributed due to constraints such as time and effort, the threat here could be the lack of an end-to-end overview of how the methodology looks like in a real-world context. According to the feedback given by the experts, the scope of operationalizing the ML model is broader and goes beyond building an ML model and deploying. In the instance that the team responsible for deploying the model dissolves, then comes the question - 'What happens to the ML model?'. This also brings in a factor of change management and ML-Ops to understand what needs to happen after the model is up and running in the production environment. However, since the scope of the thesis and research work has been limited to five phases and clearly been established since the beginning, this threat to validity is mitigated to a certain extent.

7.5 Limitations and Future Work

One of the main limitations observed as part of this study is that it doesn't demonstrate an end-to-end overview of what the methodology entails owing to constraints such as time, resources and effort. In this regard, audience and practitioners can get a high-level understanding of how the first five phases would look like in a real-world use case but not the deployment of the outcome. As part of future research, the researcher would also like to include components of how ML models can be operationalized, as suggested by experts during the validation process. This explores the area of ML-Ops and can be a great value-add to the design of the methodology and in practice.

Secondly, the methodology has been demonstrated by the researcher herself in a portion of a larger case study within the organization. Following which, expert opinions have also been taken into consideration on the design, development and perceived usability of the methodology. However, since they haven't actually used the methodology in a use case from scratch, it may seem like an unknown. However, future research should focus on the shortcomings that have been identified during the validation process with the experts. The methodology presented as part of this research can serve as a starting point to build extensively as part of future research.

Third, the scope of interpretability in ML by itself is extensive and widespread across various topics, domains and techniques. The research has attempted to make the scope of the study as concise as possible and limit to what's necessary for the business context, but there is a possibility that there could be the existence of other possible techniques to perform interpretability related tasks such as feature importance, feature engineering and visualizations. As part of future research, an exhaustive manual with potential techniques can be created and documented to use as a guideline while deciding appropriate techniques to use on ML models.

Bibliography

- [1] J. McNellis, “You’re likely investing a lot in marketing analytics, but are you getting the right insights?” <https://blogs.gartner.com/jason-mcnellis/2019/11/05/youre-likely-investing-lot-marketing-analytics-getting-right-insights/>.
- [2] B. van Giffen, D. Herhausen, and T. Fahse, “Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods.” <https://doi.org/10.1016/j.jbusres.2022.01.076>.
- [3] Rosenfeld, A. . Richardson, and Ariella, “Explainability in human-agent systems. autonomous agents and multi-agent systems.” <https://doi.org/10.1007/s10458-019-09408-y..>
- [4] A. B. Arrieta, N. D´iaz-Rodr´iguez, J. D. Sera, A. Bennetotb, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai.”
- [5] M. E. Morocho-Cayamcela and W. L. H. Lee, “Machine learning for 5g/b5g mobile and wireless communications: Potential, limitations, and future directions.” <https://doi.org/10.1109/ACCESS.2019.2942390>.
- [6] R. Wirth and J. Hipp, “Crisp-dm: Towards a standard process model for data mining.”
- [7] Q. V. Liao and K. R. Varshney, “Human-centered explainable ai (xai): From algorithms to user experiences.”
- [8] <https://www.mckinsey.com/capabilities/operations/our-insights/operationalizing-machine-learning-in-processes>.
- [9] W. Fan, J. Liu, S. Zhu, and P. M. Pardalos, “Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (aimdss).” <https://doi.org/10.1007/s10479-018-2818-y>.
- [10] N. Bevan, J. Carter, and S. Harker, “Human-computer interaction: Design and evaluation.” https://doi.org/10.1007/978-3-319-20901-2_13.
- [11] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research.”
- [12] B. Kitchenham and S. Charters, “Guidelines for performing systematic literature reviews in software engineering.” <https://doi.org/10.1016/j.infsof.2008.09.009/>, 2007.
- [13] “Zotero.” <https://www.zotero.org/>.
- [14] D. Gunning, E. Vorm, J. Y. Wang, and M. Turek, “Darpa’s explainable ai (xai) program: A retrospective.” <https://doi.org/10.1002/ail2.61>.
- [15] V. Attri, I. Batra, and A. Malik, “A relative study on analytical models.” <https://doi.org/10.1109/ICIEM51511.2021.9445372>.
- [16] R. Mello and R. A. Martins, “Can big data analytics enhance performance measurement systems?.” <https://doi.org/10.1109/EMR.2019.2900645>.

- [17] A. I. Khan and A. Al-Badib, "Emerging data sources in decision making and ai." <https://10.1016/j.procs.2020.10.042>.
- [18] M. Wach and I. Chomiak-Orsa, "The application of predictive analysis in decision-making processes on the example of mining company's investment projects." <https://doi.org/10.1016/j.procs.2021.09.284>.
- [19] B. Unhelkar and T. Gonsalves, "Enhancing artificial intelligence decision making frameworks to support leadership during business disruptions." <https://doi.org/10.1109/MITP.2020.3031312>.
- [20] D. D. Shin and Y. J. Park, "Role of fairness, accountability, and transparency in algorithmic affordance," *Comput. Hum. Behav.*, vol. 98, pp. 277–284, 2019.
- [21] J. Rožanec, E. Trajkova, I. Novalija, P. Zajec, K. Kenda, B. Fortuna, and D. Mladeni, "Enriching artificial intelligence explanations with knowledge fragments." <https://doi.org/10.3390/fi14050134>.
- [22] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)." <https://doi.org/1109/ACCESS.2018.2870052>.
- [23] M. V. Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior." <https://doi.org/1109/ACCESS.2018.2870052>.
- [24] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for explainable ai." <https://uxplanet.org/easily-calculate-sus-score-a464d753e5aa>.
- [25] Z. C. Lipton, "The myths of interpretability." <https://doi.org/10.1145/3233231>.
- [26] B. K. Finale Doshi-Velez, "Towards a rigorous science of interpretable machine learning." <http://arxiv.org/abs/1702.08608>.
- [27] H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan, "Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs." <https://doi.org/10.1145/3411764.3445088>.
- [28] S. Chatterjee, A. Das, C. Mandal, B. Mukhopadhyay, M. Vipinraj, A. Shukla, R. N. Rao, C. Sarasaen, O. Speck, and A. Nürnberger, "Framework for interpretability and explainability of image-based deep learning models." <https://doi.org/10.48550/arXiv.2110.08429>.
- [29] S. Chatterjee, A. Das, C. Mandal, B. Mukhopadhyay, M. Vipinraj, A. Shukla, R. N. Rao, C. Sarasaen, O. Speck, and A. Nürnberger, "Torchesegeta: Framework for interpretability and explainability of image-based deep learning models." <https://doi.org/10.3390/app12041834>.
- [30] "Automated individual decision-making, including profiling." <https://gdpr-info.eu/art-22-gdpr/#:~:text=22%20GDPR%20Automated%20individual%20decision,significantly%20affects%20him%20or%20her>.
- [31] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"." <https://doi.org/10.48550/arXiv.1606.08813>.

- [32] D. Zhdanov, S. Bhattacharjee, and M. A. Bragin, "Incorporating fat and privacy aware ai modeling approaches into business decision making frameworks." <https://doi.org/10.1016/j.dss.2021.113715>.
- [33] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model." <https://doi.org/10.1155/2021/6634811>.
- [34] O. Loyola-Gonzalez, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view." <https://doi.org/10.1109/ACCESS.2019.2949286>.
- [35] R. R.Hoffman, S. T. Mueller, G. Klein, and J. Litman, "A novel model usability evaluation framework (muse) for explainable artificial intelligence." <https://doi.org/10.48550/arXiv.1812.04608/>.
- [36] R. Alhomsy and A. S. Vivacqua, "The explainable business process (xbp) - an exploratory research." <https://doi.org/10.22456/2175-2745.107964>.
- [37] Sokol and Flach, "One explanation does not fit all: The promise of interactive explanations for machine learning transparency." <https://doi.org/10.1007/s13218-020-00637-y>.
- [38] G. El-khawaga, M. Abu-Elkheir, and M. Reichert, "Xai in the context of predictive process monitoring: An empirical analysis framework." <https://doi.org/10.3390/a15060199>.
- [39] C. Molnar, "Interpretable machine learning: A guide for making black box models explainable."
- [40] A. M. Turing and J. Copeland, "The essential turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life plus the secrets of enigma," 2004.
- [41] M. Bohanec, M. K. Borstnar, and M. Robnik-Sikonja, "Integration of machine learning insights into organizational learning - a case of b2b sales forecasting." https://doi.org/10.1007/978-3-319-38974-5_7.
- [42] H. Assem and D. O'Sullivan, "Towards bridging the gap between machine learning researchers and practitioners." <https://doi.org/10.1109/SmartCity.2015.151>.
- [43] Sarker and I.H, "Machine learning: Algorithms, real-world applications and research." <https://doi.org/10.1007/s42979-021-00592-x>.
- [44] B. Qian and J. Su, "Orchestrating development lifecycle of machine learning based iot applications: A survey." <https://doi.org/10.1145/3398020>.
- [45] R. J. Wieringa, "Design science methodology for information systems and software engineering." <https://doi.org/10.1007/978-3-662-43839-8>.
- [46] B. DiCicco-Bloom and B. F. Crabtree, "The qualitative research interview." <https://doi.org/10.1111/j.1365-2929.2006.02418.x>.
- [47] D. W. T. III, "Qualitative interview design: A practical guide for novice investigators." <https://doi.org/10.46743/2160-3715/2010.1178>.
- [48] J. W. Creswell, "Qualitative inquiry and research design choosing among five approaches."

- [49] P. M, “Qualitative evaluation and research methods.”
- [50] J. Ritchie, J. Lewis, C. M. Nicholls, and R. Ormston, “Qualitative research practice -a guide for social science students and researchers.”
- [51] N. Golafshani, “Understanding reliability and validity in qualitative research.” <https://doi.org/10.46743/2160-3715/2003.1870>.
- [52] I. Dey, “Qualitative data analysis: A user-friendly guide for social scientists.”
- [53] <https://thinkinsights.net/consulting/framework-methodology/#:~:text=According%20to%20Stackexchange%3A,to%20achieve%20a%20particular%20goal>.
- [54] M. Bohanec, M. Robnik-Sikonja, and M. K. Borstnar, “Organizational learning supported by machine learning models coupled with general explanation methods: A case of b2b sales forecasting.” <https://doi.org/10.1515/orga-2017-0020>.
- [55] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, S. P. Ludwig Winkler, and K.-R. Müller, “Towards crisp-ml(q): A machine learning process model with quality assurance methodology.” <https://doi.org/10.3390/make3020020>.
- [56] R. R.Hoffman, S. T. Mueller, G. Klein, and J. Litman, “Metrics for explainable ai: Challenges and prospects.” <https://doi.org/10.48550/arXiv.1812.04608/>.
- [57] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning.” <https://doi.org/10.48550/arXiv.1702.08608>.
- [58] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences.”
- [59] R. Nickerson, “Confirmation bias: A ubiquitous phenomenon in many guises.” <https://doi.org/10.1037/1089-2680.2.2.175>.
- [60] Ramani and A. Venkata, “Comparative analysis of different forecasting techniques for ford mustang sales data.”
- [61] R. Lackes, M. Siepermann, and G. Vetter, “What drives decision makers to follow or ignore forecasting tools - a game based analysis.” <https://doi.org/10.1016/j.jbusres.2019.02.036>.
- [62] “Target encoding.” <https://contrib.scikit-learn.org/categoryencoders/targetencoder.html>.
- [63] A. Chatzimpampas, R. M. Martins, K. Kucher, and A. Kerren, “Featureenvi: Visual analytics for feature engineering using stepwise selection and semi-automatic extraction approaches.” <https://doi.org/10.48550/arXiv.2103.14539>.
- [64] J. Krause, A. Perer, and E. Bertini, “Infuse: Interactive feature selection for predictive modeling of high dimensional data.”
- [65] <https://scikit-learn.org>.

-
- [66] M. Gashi, M. Vukovi, N. Jekic, S. Thalmann, A. Holzinger, C. Jean-Quartier, and F. Jeanquartier, "State-of-the-art explainability methods with focus on visual analytics showcased by glioma classification." <https://doi.org/10.3390/biomedinformatics2010009>.
- [67] J. Brooke, "Sus: A quick and dirty usability scale."
- [68] J. Guerci, "Easily calculate sus score." <https://uxplanet.org/easily-calculate-sus-score-a464d753e5aa>.
- [69] J. Sauro, "System usability scale - interpretation." <https://about.gitlab.com>.

Appendices

A List of Acronyms

ML	Machine Learning
AI	Artificial Intelligence
RQ	Research Question
IS	Information Systems
DSRM	Design Science Research Methodology
SLR	Systematic Literature Review
XAI	Explainable Artificial Intelligence
DARPA	Defense Advanced Research Project Agency
GDPR	General Data Protection Regulation
PCA	Principal Component Analysis
IoT	Internet of Things
CRISP-DM	Cross Industry Standard Process for Data Mining
KPI	Key Performance Indicator
RACI	Responsible Accountable Informed Consulted
FTE	Full Time Equivalent
CTO	Chief Technology Officer
BL	Business Line
MAG	Main Article Group
NPI	New Product Introduction
SME	Subject Matter Expert
WAPE	Weighted Average Percent Error
MAPE	Mean Absolute Percent Error
SAM	Semiconductor Application Market
NF	Neural Field
CF	Curve Generator
MLP	Multilayer Perceptron
LGBM	Light Gradient Boost Method
SHAP	Shapley Additive Explanations
SUS	System Usability Scale
PMO	Project Management Office

B First-level Stakeholder Interview Questions

Introduction

1. Permission to record the meeting
2. Establish context for research objective and master thesis
 - a.) Msc Business Information Technology - Programme specifications
 - b.) Scope: Research the need for a methodology
3. Discuss findings from literature review

(a.) Vision and Goals

Learning goals: Learn more about what stakeholders are looking to get out of the research.

1. What is your role in this project?
2. Why do you think this project is important for NXP?
3. What would success on this project look like to you?
4. What challenges do you foresee this project possibly running into?

(b.) Existing Knowledge

Learning goals: Find out what stakeholders know and where the gaps in knowledge are.

1. What problem(s) are we trying to solve for the business?
2. In your opinion, what defines success for customers?
3. What would success on this project look like to you?
4. What are the biggest challenges we face as an org while trying to make business decisions? Do you think they are data-informed?

(c.) Unicorn Wishes and Wind Down

Learning goals: Dive into potential opportunities.

1. How would you ideally trust a prediction that a model makes?
2. How will AI/ML/Advanced analytics be integrated with NXP's overall strategy?

C Expert Interview: Questions

Introduction

1. Permission to record the meeting
2. Results of the interview will be used as a source of data for the thesis document ensuring anonymity of participants
3. Establish context for research objective and master thesis
 - a.) Msc Business Information Technology - Programme specifications
 - b.) Scope: Validation of research from an expert standpoint

C.1 Agenda

- Introduction and Context – 5 mins
- Overview – Revised Methodology – 5 mins
- Design and Development – 10 mins
- System Usability Scale (SUS) – 10 mins
- Synthesis of Results - 10 mins
- Outro - 5mins

C.2 Getting acquainted

1. Could you please introduce yourself?
2. Can you tell me something about your professional background at NXP?
 - a. Can you elaborate on your current role in this PoC?
3. How many years of work experience do you have?
4. How familiar are you with Machine Learning related concepts?
5. Why do you think it is important for business stakeholders to understand the foundation of Machine Learning?

C.3 Design and Development

1. Is the design of the methodology adequate for the team to incorporate as part of their workflow?
2. Do you agree with the newly structured 6-phase methodology as compared to the existing methodology i.e. CRISP-DM?
3. Are there any critical interventions or steps missing?
 - a. What would you change/add/eliminate from the methodology?

C.4 Synthesis of Results

1. Do you agree with the synthesis of results?
2. Do you believe adopting the proposed methodology would help achieve the ‘yellow box’ in the organization?
3. Do you think this methodology can be adopted as part of NXP’s business process in forthcoming projects?

C.5 Outro

1. Would you like to add something to the discussion, in case we did not address it yet?
2. Do you have any questions for me?

D System Usability Scale - Discussion

1. I think that I would like to use this methodology frequently.

Participant A: We haven't experimented with this new methodology. Initially, when we used CRISP-DM, we found it useful but at a later point we moved to Kanban because of ad-hoc requests from stakeholders related to visualizations, business understanding, data engineering or filtering, performance of the model - they were mostly ad-hoc and not part of a cycle/loop that we could follow. We had to pull in requests and work on them based on some priority. Unless we experiment with this and a well-structured team and use case, we wouldn't know.

Participant B: Yes, I would like to use it for all use cases.

Participant C: I spoke about this at length. This is definitely something we can use and I intend to use it going forward.

Participant D: I think it is useful and I don't give it a 5 because I haven't used it yet so I'm wondering how it will work for the user and stakeholder side. I can imagine how it would affect my work as a data analyst or scientist. Extra points are not given because I wonder how it's going to be for the user and stakeholder side.

Participant E: It's an extension of CRISP-DM so people are familiar with it; I would like to use it because interpretability is important and this has that as an element.

Participant F: Strongly agree. No additional comments here.

Participant G: I would like to use it and I think it's beneficial.

2. I find the methodology unnecessarily complex

Participant A: I don't think it's complex, it's simple to understand and something every data scientist would understand as long as we are the ones who are going to use it. For non-technical people, there may be questions that are raised. For data scientists, the middle blocks are important but for business people the first and last blocks are important.

Participant B: Ultimately, it depends on the stakeholder who is using the methodology.

Participant C: Well, unnecessarily complex, I do not think. It is complex and definitely not simple but I do think all the steps are needed. I do not think it is unnecessarily complex.

Participant D: No, I don't think so. It's not overtly complex. Doing ML is overtly complex so it's nice to have an overview like this to fall back on and understand what needs to be done in general.

Participant E: I disagree that it's unnecessarily complex; It is a simple extension of an existing method.

Participant F: I don't find the methodology complex at all.

Participant G: I can clearly relate it with CRISP-DM and I fully understand the purpose. This methodology has additional steps focused towards a more corporate setting.

3. I believe the methodology would be easy to use.

Participant A: Strongly agree here.

Participant B: Yes, when the methodology includes different stakeholders, it would be difficult to force them into following all the phases. It's not complex, but it's not easy to get everybody to follow the methodology.

Participant C: I think so. If you understand the gist of it, yes. I think if you are well-explained and you have some templates to use - then yes. If you can have some tooling, design around it, then it should be easy to use.

Participant D: Same reasoning as the first statement. It would be easy from my side to work on it. But I don't how it would be to interact this methodology with the stakeholder.

Participant E: Indeed. It's a simple extension of an existing framework that can be easy to use.

Participant F: It sounds straightforward but when we put it in practice it may not be that easy.

Participant G: Yes, I agree. No additional comments.

4. I think that I would need the support of a user manual and/or technical guidance to be able to use this methodology.

Participant A: This is subjective based on the expertise in working with agile methodologies.

Participant B: I think with the description and the visuals, it's enough to follow.

Participant C: To be successful in using it, it would be good to have a manual and technical guidance. For me, it's both. This would be a 4. I strongly agree you need to have it in place because it is a pre-requisite.

Participant D: I don't think I would particularly require a user manual. But I could imagine it being useful having some explanation on how you can use it. I've had some things in the past that seem logical to me but doesn't make sense to someone else - for example, phrasing can be interpreted differently.

Participant E: I would need a user manual. Because at each step, from an interpretability perspective, I would need to remember the objectives and goals for each stakeholder.

Participant F: I feel neutral about this. I wouldn't know yet.

Participant G: I don't think I would need user manual. Maybe an example or a blog post that explains an example.

5. I think the various phases and sub-steps in this methodology are well integrated.

Participant A: You have the keywords in there, but there is no detailed description unless there is a manual in place to explain what are the steps involved in each phase: for example, what does choosing an ML algorithm mean? What are the key steps involved? When do we know it's completed? Approximate timeline to finish? Who are the parties involved? I agree there is a high-level understanding and in that sense, it is well-integrated.

Participant B: Yes, they are.

Participant C: Yes, this is strongly agree. I think it all flows nicely.

Participant D: The various phases and sub-steps are well integrated, they make sense chronologically but also if you iterate and you jump from one to another - it makes sense.

Participant E: I would like to see a unified view of what needs to be done at each phase.

Participant F: The phases and the sub-steps make sense together and are well-integrated.

Participant G: the feedback loops are clear to me and also by nature the steps are placed where they belong.

6. I think there is too much inconsistency in this methodology.

Participant A: No additional comments. I strongly disagree here.

Participant B: Not at all.

Participant C: No. Strongly disagree.

Participant D: No. Quite clear and also clear why and from where all these steps are coming from. I recognize those steps in me doing data projects. Not just ML, but many of those steps are useful in a data analysis topic not just ML itself.

Participant E: Definition and what needs to be done with interpretability may not be clear to everyone. That can be improvised.

Participant F: Not at all.

Participant G: I don't think it is inconsistent.

7. I would imagine that most people would learn to use this methodology very quickly.

Participant A: I wouldn't completely agree, unless we actually use it. It's also subjective and depends on the structure of the team and the organization of the project.

Participant B: People who have already worked with CRISP-DM will not find it difficult to work with this. Same goes with people who have worked with agile. But those that have worked with waterfall - this may be difficult. Depends on the stakeholders.

Participant C: Yes, if you have an agile mindset, it won't be that difficult.

Participant D: I strongly agree. I do think most people would learn to use it very quickly. Especially if they have a background in data science, I don't think any of those steps there should be a surprise for someone with that background. And for a stakeholder, maybe that would be different but let's put a preconception that they have some understanding of data science.

Participant E: Easy to use and adopt. How effectively they can use it is a different question but they can adopt it for sure.

Participant F: Conceptually, it all makes sense. It's also easy to understand but to put it into practice can be a challenge.

Participant G: I agree that it would be easy for people to adopt.

8. I find the methodology very cumbersome to use.

Participant A: I don't think so. No additional comments, I disagree.

Participant B: Not at all.

Participant C: It takes a bit of effort, but it's not cumbersome.

Participant D: I don't find it cumbersome to use.

Participant E: I wouldn't say so. More and more interpretability is being built into the data science tools. I don't think it's very cumbersome to use. I disagree with it.

Participant F: Not at all. I disagree.

Participant G: I disagree here.

9. I would feel confident using the methodology in forthcoming projects.

Participant A: I am confident in using the methodology, I don't find any reason why we shouldn't use this within the team.

Participant B: Yes, absolutely.

Participant C: Strongly agree.

Participant D: I do feel confident using it in forthcoming projects because I'm going to try and implement it in the project that I'm working on currently.

Participant E: I don't think it would cause any issues while working on projects, so I'm confident about using it.

Participant F: Yes, strongly agree.

Participant G: It feels like everything is in place but I am unsure of how confident I would be using it. I'm neutral on this.

10. I need to learn a lot of things before I could get going with this methodology.

Participant A: There are not many things to learn and these are things that need to be followed anyway.

Participant B: I've already learned everything with your help and the presentation. It's not a lot, but it is important to understand the different types of interpretability techniques.

Participant C: Learning, I don't think. If it's strictly about learning, no, because you explained it very well. But to support, I would need technical and user guidance.

Participant D: No, I don't think I would need to learn a lot of things before I get started.

Participant E: I don't think I would need to learn a lot. I disagree.

Participant F: I don't think I would need to learn a lot. I disagree.

Participant G: No. Personally, I don't have to learn new things.

Code for SUS score calculation

```
function sus(...answers) {  
  let soma = 0;  
  for (let indice = 0; indice < answers.length; indice++)  
    if (indice % 2 === 0)  
      soma += answers[indice] - 1;  
    else  
      soma += 5 - answers[indice];  
  
  }  
  return (soma * 2.5);  
}
```