

# Engineering Efficiency: Leveraging Data for Optimized Man Hour Estimation

Master Thesis

Author: Fabian Jansink  
University of Twente  
Master Business Administration: Digital Business & Analytics

Date: 14-07-2023

Version: Final



**UNIVERSITY  
OF TWENTE.**

# COLOPHON

Master Thesis

DATE

14-07-2023

VERSION

Final

AUTHOR

Fabian Jansink  
Master Business Administration: Digital Business & Analytics  
Faculty of Behavioural, Management and Social Sciences  
University of Twente

UNIVERSITY SUPERVISORS

dr. A.B.J.M. Wijnhoven (Fons)  
Associate Professor  
Faculty of Behavioural, Management and Social Sciences  
University of Twente

dr. M. Renkema (Maarten)  
Assistant Professor  
Faculty of Behavioural, Management and Social Sciences  
University of Twente

## ABSTRACT

Accurate cost estimation is crucial in an engineering company's project management as it involves forecasting project costs, enabling effective resource allocation, project feasibility assessment, and informed decision-making. Traditional cost estimation methods rely on expert-driven, lengthy, and subjective estimates, which underutilize data. While data-based cost estimation methods have shown great potential in academia, their practical application is lacking. This study explores the gap between academic potential and practical implementation to bridge the divide aiming to enhance the practical application of data-based approaches in real-world cost estimation scenarios using a design science methodology that combines theory and practical relevance to generate actionable knowledge and practical solutions.

Based on the literature review, a regression model is selected for effective cost estimation utilizing available data. Based on literature and collaboration with tender management, requirements prioritize explainability, model variables, and specific performance criteria. While a significant amount of projects were available in the dataset, due to limited and unstandardized data, only 71 projects were included in the model. Several regression methods were tested to identify the best-performing model, in which the best-performing regression model achieved an adjusted R-squared of 0.807 using four significant variables. However, the model's high mean absolute percentage error of approximately 40% indicates its instability and unsuitability for practical implementation at present.

This study shows that the main challenges encountered consisted of fragmented data management practices, resulting in poor data quality, establishing model trust, and cultural adoption of a bottom-up decision-making approach towards top-down. Nevertheless, the utilization of data for cost estimation has been widely acknowledged by both employees and research papers as holding significant potential value for the future. Addressing challenges that impede the practical use of cost estimation models is crucial to gain a better understanding of the gap between academic potential and practical application, potentially enhancing their effectiveness and real-world utility.

# TABLE OF CONTENTS

|   |      |
|---|------|
| Abstract .....                                    | iii  |
| List of Figures .....                             | vi   |
| List of Tables .....                              | vii  |
| List of Equations .....                           | vii  |
| List of abbreviations .....                       | viii |
| 1. Introduction .....                             | 1    |
| 1.1 Research background.....                      | 1    |
| 1.2 Problem statement.....                        | 1    |
| 1.3 Research Goal.....                            | 2    |
| 1.4 Research context.....                         | 3    |
| 1.5 Research relevance.....                       | 4    |
| 1.6 Research questions .....                      | 5    |
| 2. Literature review.....                         | 6    |
| 2.1 Data-driven decision-making (DDDM) .....      | 6    |
| 2.2 Cost estimation methods.....                  | 8    |
| 2.2.1 Parametric estimations.....                 | 9    |
| 2.2.2 Analogous estimations .....                 | 10   |
| 2.2.3 Engineering build-up estimations.....       | 10   |
| 2.2.4 Probabilistic estimations.....              | 10   |
| 2.2.5 Overview of methods .....                   | 11   |
| 2.3 Machine learning cost estimation methods..... | 13   |
| 2.3.1 Machine learning methods .....              | 14   |
| 2.3.2 Discussion of machine learning methods..... | 17   |
| 2.4 Regression methodology .....                  | 18   |
| 2.4.1 Regression formula & assumptions .....      | 18   |
| 2.4.2 Regression models .....                     | 20   |
| 2.5 Variable selection: .....                     | 21   |
| 3. Methodology .....                              | 24   |
| 3.1 Research Strategy .....                       | 24   |
| 3.2 Ethics .....                                  | 32   |
| 4. Results.....                                   | 33   |
| 4.1 Design.....                                   | 33   |
| 4.1.1 Interviews tender department.....           | 33   |
| 4.1.2 Requirements .....                          | 34   |
| 4.1.3 Impact work processes.....                  | 35   |
| 4.1.4 Data gathering .....                        | 36   |
| 4.2 Development .....                             | 39   |
| 4.3 Demonstration .....                           | 44   |

|       |   |    |
|-------|---|----|
| 4.4   | Validation .....                                  | 47 |
| 4.4.1 | Interview & Discussion .....                      | 47 |
| 4.4.2 | Challenges & Opportunities.....                   | 49 |
| 5.    | Discussion & Conclusion .....                     | 52 |
| 5.1   | Discussion results.....                           | 52 |
| 5.1.1 | Theoretical implications.....                     | 53 |
| 5.1.2 | Practical implications.....                       | 54 |
| 5.2   | Limitations .....                                 | 55 |
| 5.3   | Future research .....                             | 56 |
| 5.4   | Conclusion.....                                   | 57 |
|       | Bibliography.....                                 | 58 |
|       | Appendices.....                                   | 62 |
|       | Appendix A: BPMN models .....                     | 62 |
|       | Appendix B: Data schema .....                     | 66 |
|       | Appendix C: AACE classification .....             | 67 |
|       | Appendix D: Interviews Requirements .....         | 68 |
|       | Appendix E: Regression output .....               | 69 |
|       | Appendix F: Python code .....                     | 75 |
|       | Appendix G: Interview impact work processes ..... | 77 |
|       | Appendix H: Interview validation phase .....      | 78 |
|       | Appendix I: Discussion validation phase .....     | 79 |

# LIST OF FIGURES

- Figure 1: BPMN proposal process.....3
- Figure 2: Cost estimation methods division .....8
- Figure 3: AI, ML, and Data science relationship .....13
- Figure 4: Trade-off between model interpretability and performance (Barredo Arrieta et al., 2020)..17
- Figure 5: Design science research (Hevner et al., 2004). .....25
- Figure 6: Research strategy phases .....26
- Figure 7: Proposed method.....30
- Figure 8: Adjusted BPMN proposal process .....36
- Figure 9: Man-hour distribution per bin size of the 71 projects .....39
- Figure 10: Distribution man-hours population (71) vs sample (████) .....39
- Figure 11: Distributions of variables related to man-hours .....40
- Figure 12: OLS regression model 1 .....43
- Figure 13: OLS regression model 4 (Total resources variable) .....44
- Figure 14: Project phase average man-hours distribution (series 1) with count projects (series 2).....46
- Figure 15: Tree diagram challenges overview.....50
- Figure 16: Tree diagram opportunities overview .....51
  
- Figure A 1: Tender Process.....62
- Figure A 2: Proposal writing process.....63
- Figure A 3: █████ proposal process.....64
- Figure A 4: █████ proposal process.....65
  
- Figure C 1: AACE framework (Christensen & Dysert, 2005) .....67
  
- Figure E 1: OLS regression all variables.....69
- Figure E 2: Correlation all variables .....70
- Figure E 3: OLS regression significant variables.....71
- Figure E 4: Correlation significant variables .....72
- Figure E 5: OLS regression Total resources variable .....73
- Figure E 6: Correlation Total resources variable .....74
  
- Figure F 1: Python standardization .....75
- Figure F 2: Python variable selection .....76

# LIST OF TABLES

- Table 1: Organizational cost estimation methodologies .....8
- Table 2: Overview of advantages and disadvantages of cost estimation methods .....12
- Table 3: Machine learning cost estimation methods .....13
- Table 4: Overview of advantages and disadvantages of machine learning cost estimation methods ..16
- Table 5: Regression vs ANN for cost estimation.....18
- Table 6: Regression assumptions.....19
- Table 7: Cost-affecting variables in an engineering company context .....21
- Table 8: Overview conducted Interviews/ discussion .....25
- Table 9: Tender department Interview summary .....33
- Table 10: Meta-requirements, meta-design, and kernel theory.....35
- Table 11: Variables overview from the dataset.....37
- Table 12: Input variables metrics .....38
- Table 13: Linear regression of variables related to man-hours: R squared & F statistic .....40
- Table 14: OLS regression significant variables.....42
- Table 15: Lasso regression significant variables .....42
- Table 16: Ridge regression significant variables.....42
- Table 17: OLS regression with total resources variable .....44
- Table 18: Lasso regression with total resources variable .....44
- Table 19: Ridge regression with total resources variable .....44
- Table 20: Test cases of best-performing regression model .....46
- Table 21: Comparison with earlier work.....47
  
- Table B 1: Data schema .....66

# LIST OF EQUATIONS

- Equation 1: General regression formula.....19
- Equation 2: OLS Regression model formula.....45

# LIST OF ABBREVIATIONS

|       |  |
|-------|--|
| AI    | Artificial intelligence                              |
| ANN   | Artificial Neural Network                            |
| Capex | Capital expenditures                                 |
| CBR   | Case-Based Reasoning                                 |
| CER's | Cost Estimating Relationships                        |
| DDDM  | Data-Driven Decision making                          |
| DT    | Decision Tree  |
| E     | Engineering  |
| EPC   | Engineering, Procurement and Construction            |
| EPCm  | Engineering, Procurement and Construction management |
| IA    | Intelligence Amplification                           |
| ML    | Machine Learning                                     |
| MAE   | Mean Absolute Error                                  |
| MAPE  | Mean Absolute Percentage Error                       |
| MSE   | Mean Square Error                                    |
| RF    | Random forest algorithm                              |
| RFI   | Request for information                              |
| RFQ   | Request for quotation                                |
| TM    | Tender Management                                    |
| WBS   | Work Breakdown Structure                             |



# 1. INTRODUCTION

## 1.1 RESEARCH BACKGROUND

Cost estimation is a critical aspect of project management in many industries, including engineering. It involves forecasting the cost required to perform the work within the scope of the project (Leonard et al., 2005). Companies engaged in construction projects rely on accurate cost estimation for effective resource allocation and project feasibility, influencing decisions at every stage of the planning, bidding, design, and construction management processes (Flyvbjerg et al., 2003).

When it comes to construction projects, it is important to differentiate between the contractor and engineering companies. Contractors are responsible for the actual construction while engineering companies provide services (e.g. mechanical design, project management, structural calculations). Engineering consulting firms in construction primarily incur expenses for the engineering services they offer for different projects. For engineering companies, poor cost estimation can lead to delays, cost overruns, and other project management issues that can impact the overall success of a project. As a result, accurate cost estimation is essential for organizations to manage risk and achieve their project objectives (Troost & Oberlender, 2003).

Data plays a crucial role in improving the accuracy of cost estimation in engineering companies. By leveraging historical project data and using analytical techniques, organizations can develop more accurate and reliable cost estimation models (Doloi, 2011). Data-driven approaches also help identify patterns and trends in project cost drivers, enabling organizations to make more informed decisions about resource allocation and project feasibility (He et al., 2021). According to a survey conducted by PwC (PricewaterhouseCoopers), data-driven organizations are more likely to improve decision-making processes than those who do not (PwC, 2019).

A report by IDC, a global market intelligence firm, states that enterprise data is projected to increase at a 42% annual growth rate, which brings various opportunities to utilize this data (Agarwal et al., 2016; Reinsel et al., 2018). Numerous methods currently exist, like artificial intelligence (AI) and machine learning (ML) methods to exploit data in decision-making processes. Companies are however failing to become data-driven, regardless of this being an objective (Bean & Davenport, 2019).

## 1.2 PROBLEM STATEMENT

Tender departments need to perform proposals for several different industries under increasing time pressure while the expected number of proposals is increasing (Matel et al., 2022). For tender departments, this means that per different fields of engineering (e.g. Civil, Electrical, Piping, etc.) The hours, and thus costs<sup>1</sup> have to be estimated per project. Many papers have addressed the importance of the cost estimation methodology since insufficiencies remain to persist in more traditional cost estimation methods (Doloi, 2011). These insufficiencies stem from the fact that traditional methods fail to utilize data from previous projects (Matel et al., 2022).

Additionally, during the tendering phase of a project, estimators can face a scarcity of information necessary for accurate cost estimation. In the absence of information, they rely on their expertise, experience, and intuition to make informed judgment calls to estimate the costs (Cheng et al., 2010;

---

<sup>1</sup> Note: In the context of engineering services, the number of man-hours required for a project is inherently linked to the costs of the project. This means that man-hour estimates and cost estimates can be used interchangeably when referring to cost estimation for engineering services within this research.

Matel et al., 2022). This leads to the use of more traditional methods which results in the counseling of engineers to estimate the man-hours for their department. However, this also leads to a subjective (i.e., defined to an extent by one's personal opinion) man-hours estimate (Cheng et al., 2010; Matel et al., 2022).

Due to the necessity of engineers to familiarize themselves with the project's new context, proposals are inclined to be slow-moving. Furthermore, engineers adopt distinct working methods, resulting in notable variations in estimates among them. Moreover, in a highly competitive environment where several other competitors are bidding for the same project, there is a risk of the client rejecting the offer, potentially opting for a competitor's bid. This creates a dynamic landscape in which tender departments must navigate. If the estimated costs are deemed excessive, the project may be lost to a competitor. Conversely, if the cost are too low, a project might result in a financial loss (Flyvbjerg et al., 2003). The lack of consistent methods in cost estimation poses a challenge for estimators, in providing accurate and effective methods. The absence of a systematic approach to minimize estimation errors has led researchers to explore mathematical models, machine learning techniques, and other methods to address the issue of inaccurate or erroneous predictions in cost estimation (Tayefeh Hashemi et al., 2020).

AI and ML techniques offer the ability to extract insights from data, which can be leveraged to create predictive models, as evident in numerous studies conducted in engineering service cost estimation research and the construction industry (Bilal et al., 2016; He et al., 2021). However, despite their envisioned theoretical potential, the practical application of these techniques remains constrained in these industries (Abioye et al., 2021; Shoar et al., 2022). Several challenges hinder this practical application, these challenges consist of cultural issues, high initial costs, security, ethics, and data fragmentation (Abioye et al., 2021; Bilal et al., 2016).

Moreover, the construction industry is one of the least digitized industries and struggles to fully adopt the benefits of AI and ML, including engineering companies' services (Regona et al., 2022). Even with progress in the field of AI and ML, obstacles persist in implementing these techniques, such as the challenge of the black box element in several ML techniques (Abioye et al., 2021). It has been stated that it is crucial to develop a model capable of justifying its outcomes and providing explanations for predicted costs, which in turn is essential for the successful implementation of AI models (Tayefeh Hashemi et al., 2020; Elmousalami, 2021).

### 1.3 RESEARCH GOAL

The goal of this research is to bridge the existing gap between the envisioned theoretical potential and real-world application by investigating the feasibility, addressing the challenges, and exploring the opportunities associated with the development and use of a data-driven model for cost estimation. The research aims to overcome the limitations of traditional cost estimation methods by leveraging previous project data. Specifically, the focus will be on creating a practical, efficient, and accurate data-driven model that can provide timely estimates of man-hours. The ultimate objective is to enhance the decision-making process during the proposal stage by offering a reliable and transparent tool that supports informed cost estimations and increases the competitiveness of the Tender department in bidding for projects.

## 1.4 RESEARCH CONTEXT

Company X is an engineering company providing comprehensive engineering solutions. They offer services related to project planning, design, and construction management. Company X conducts tenders as a strategic approach to secure projects. Tenders allow them to competitively bid for contracts by submitting proposals outlining their expertise, capabilities, and cost-effective solutions. The tendering process typically involves identifying project requirements, preparing bid documents, evaluating competitors, estimating costs, and presenting compelling proposals to potential clients. This enables Company X to demonstrate its qualifications and win projects based on their value propositions.

Company X has invested a significant amount of hours in drafting these proposals. A part of these hours is required by engineering to calculate the expected required project hours in so-called man-hour estimates. The current cost calculation methods used by Company X are slow, mainly intuitive, and capacity demanding of engineers, which leads to a high financial impact.

Currently, Company X approaches tenders using various methods; however, all of these methods underutilize the data of previous projects. They rely on engineers for providing input on estimating man-hours and associated costs for projects. This estimation process follows a bottom-up approach, utilizing a work breakdown structure to identify the tasks and activities involved in the project. Based on this breakdown, engineers calculate the required man-hours. While this approach allows for detailed estimation, it heavily relies on human judgment and expertise, which can be subjective and time-consuming. For a simplified overview of the proposal phase see Figure 1, in Appendix A the total overview of the tender process is given. This figure shows that the current methods utilized by Company X primarily rely on engineers providing their expertise in estimates for different stages of the proposal phase.

By incorporating data-driven techniques and leveraging historical project data, industry benchmarks, and advanced analytics, Company X can enhance their tendering process, improving accuracy, efficiency, and competitiveness. However, as mentioned, several papers have investigated the use of AI in construction and engineering companies, but real implementation and application still lack, which is also the case within Company X. An ANN cost estimation model has been developed, however, the application of this model is missing due to several constraints, mainly because the model has been developed almost 5 years ago on a limited amount of data. On top of that, the model gives only one output value, with no further information; the model thus lacks any form of explainability.



*Figure 1: BPMN proposal process*

## 1.5 RESEARCH RELEVANCE

This research holds immediate practical relevance for Company X, particularly in their tender department, as it investigates the potential benefits of developing and implementing a data-driven model for cost estimation in a decision-making process of an engineering company. The primary objective of this study is to address practical challenges and opportunities associated with the adoption and development of such models in an engineering company, to enhance accuracy, efficiency, and speed in the estimation process.

Lost tenders directly translate to sunk costs, representing the expenses incurred by employees during the bidding process. Consequently, these costs contribute to the overhead expenses of the firm. By creating a model that can rapidly and accurately estimate the man-hours required for different departments, the company can offer more precise and cost-effective proposals to potential clients, thereby enhancing competitiveness and profitability (Matel et al., 2022).

While previous studies in engineering companies have focused on the development of various types of models, such as case-based reasoning (CBR), regression, and artificial neural networks (ANN) (Cheng et al., 2010; Chou et al., 2009; Matel et al., 2022), this research contributes to the existing body of knowledge by examining the challenges encountered during the development and potential implementation of these models in practice. Additionally, this research explores and provides an overview of machine learning-based methods for cost estimation in engineering companies, outlining their advantages, disadvantages, and their suitability for practical implementation.

The practical implementation of cost estimation models is a crucial aspect that determines their effectiveness and real-world utility. Despite the established academic potential of these models, their limited practical implementation suggests that challenges or barriers are preventing their widespread adoption in industry settings (He et al., 2021), emphasizing the existing gap between theoretical understanding and practical application, and the need to bridge this divide. This gap exists not only in the construction industry at large but also within the engineering sector (He et al., 2021).

By acknowledging this gap, the research aims to address these challenges and explore opportunities to overcome them, thereby facilitating the implementation of data-driven models for decision-making in engineering companies. This theoretical understanding can then serve as a foundation for future research and development in the field, leading to improved practices, enhanced accuracy, efficiency, and speed in cost estimation, and potentially enabling engineering companies to offer more precise and cost-effective proposals to clients. Additionally, this contributes to the broader disclosure of data-driven decision-making in an engineering company context.

## 1.6 RESEARCH QUESTIONS

The following research question is established based on the problem statement and the research goal.

“What are the perceived challenges and opportunities experienced by decision-makers in the Tender department of an engineering company regarding the development and use of a data-driven model for cost estimation?”

Based on this research question, several sub-questions are formed to answer the main research question. The research questions are answered through five different research phases, this is elaborated upon in chapter three, the methodology.

1. What are the potential benefits and limitations associated with the use of data-driven decision-making for cost estimation?
2. What are the most effective cost estimation methods in the engineering industry and how can they be used to improve the cost estimation practices in a decision-making process?
3. How can a data-driven model be developed to accurately estimate the costs of proposals in the Tender department, taking into account available input data and relevant variables?
4. What are the requirements and criteria for implementing the proposed model in the Tender department and how can the current work processes be adapted to integrate the model effectively?
5. What are the performance and limitations of the developed model and how can it be improved to better support decision-making in cost estimation for proposals?

To answer the research question design science research is applied. Design science is a research methodology that focuses on the creation and evaluation of innovative artifacts as a means to address identified problems or opportunities in a specific domain (Hevner et al., 2004). It is applied when there is a need to develop new knowledge through the design and creation of novel artifacts that provide practical solutions to real-world problems. Design Science combines theoretical foundations with practical relevance, aiming to contribute to both research and practice by generating actionable knowledge through the development and evaluation of artifacts (Hevner et al., 2004). Design science is applied in this research due to the existing gap between academic potential, and practical application. Design science is applied in response to identified business needs within the research environment, utilizing a knowledge base that forms the foundation for applicable knowledge. Through the conduct of design science, this research not only contributes to the research environment but also enriches the existing knowledge base. This theoretical understanding serves as a solid groundwork for future research and development, possibly leading to enhanced practices. This is further elaborated upon in Chapter 3.

The next chapter consists of the literature review in which the first three research questions are answered. Followed by that, the methodology is described. Chapter four consists of the results in which research questions four and five are answered. Finally, in chapter five the discussion and conclusion of this research are given.

## 2. LITERATURE REVIEW

In this chapter, a literature study is performed for research questions one, two, and three. The first part describes DDDM and its relevance in the construction industry. The second part consists of commonly used cost estimation methods in engineering companies and the construction industry. Because of the limited number of studies in engineering companies, cost estimation methods from the construction industry are also reviewed since there is a substantial amount of overlap between the two industries. The third part consists of ML-based methods used for cost estimation. Limitations and problems with these methods are discussed and a best practice will be researched regarding these methods. Additionally, the chosen model is elaborated on, and significant cost factors are outlined.

### 2.1 DATA-DRIVEN DECISION-MAKING (DDDM)

In recent years, data-driven models have experienced a surge in popularity across diverse sectors and domains, including the construction industry (Bilal et al., 2016; He et al., 2021). These models leverage data to uncover valuable insights, make accurate predictions, and drive informed decision-making. By harnessing the vast amount of available data and employing sophisticated algorithms, data-driven models have revolutionized the way organizations approach problem-solving and strategic planning, and decision-making (Provost & Fawcett, 2013).

The digital age's arrival has exponentially amplified data generation across domains, presenting organizations with vast opportunities for data-driven decision-making (DDDM) (Provost & Fawcett, 2013). According to Provost & Fawcett (2013), DDDM in companies is associated with higher productivity and market value. DDDM can be defined as, *“the practice of basing decisions on the analysis of data rather than purely on intuition”* (Provost & Fawcett, 2013). DDDM involves systematically collecting, analyzing, and interpreting data to gain valuable insights that can drive strategic, operational, and tactical decisions. By leveraging the vast amounts of data now at their disposal, organizations can uncover patterns, trends, and relationships that were previously hidden, thereby enabling more informed and evidence-based decision-making (Provost & Fawcett, 2013).

Ever since the inception of data-driven decision-making, the analytics of decision-support systems have evolved significantly, incorporating a fusion of operational research, machine learning, and information systems (Provost & Fawcett, 2013). This integration has paved the way for enhanced capabilities in extracting insights from data and leveraging advanced techniques such as machine learning (ML) and artificial intelligence (AI) within the realm of DDDM. ML algorithms, for instance, can be employed to optimize or even automate decision-making processes by learning patterns from historical data and making predictions or recommendations based on new information. AI-powered systems can further augment DDDM by enabling intelligent data processing, natural language processing, and cognitive capabilities. This synergy between analytics, ML, and AI in DDDM not only enhances the accuracy and efficiency of decision-making but also opens up new avenues for organizations to leverage the potential of emerging technologies in gaining a competitive edge (Provost & Fawcett, 2013).

Another related concept is intelligence amplification (IA), which refers to the augmentation or enhancement of human intelligence using technology or tools. In the context of data-driven decision-making, intelligence amplification plays a crucial role in empowering individuals and organizations to make more informed and effective decisions by leveraging data and analytical insights (Wijnhoven, 2022). IA aims to amplify human cognitive abilities by integrating technologies such as machine learning, natural language processing, or data visualization into the decision-making process

(Wijnhoven, 2022). The objective is to leverage technology to augment human thinking, reasoning, and problem-solving skills, rather than relying solely on autonomous AI systems (Wijnhoven, 2022).

In the construction industry, the adoption of data-driven decision-making (DDDM) techniques has become increasingly prevalent presenting many opportunities through leveraging various intelligent data-driven approaches such as natural language processing (NLP), machine learning (ML), and data mining (Bilal et al., 2016). One specific area within the construction industry where DDDM plays a significant role is cost estimation (Bilal et al., 2016; Regona et al., 2022). By employing ML techniques, construction companies can utilize historical project data, identify patterns, and make accurate cost predictions, enabling informed decision-making based on data-driven insights (He et al., 2021).

It is worth noting that cost estimation in the construction industry is predominantly expert-driven rather than data-driven (Doloi, 2011). Traditionally, cost estimators heavily rely on their expertise, domain knowledge, and intuition to estimate project costs (Doloi, 2011). While this approach may have been effective in the past, the increasing availability of data and advancements in analytical techniques present an opportunity to enhance cost estimation practices (He et al., 2021).

However, transitioning from expert-driven to data-driven cost estimation requires overcoming several challenges. These challenges include data availability, quality, security, and integration from disparate sources (He et al., 2021; Regona et al., 2022). The data management practices are fragmented in the construction industry which makes extracting data a difficult task (Regona et al., 2022). Furthermore, it is crucial to acknowledge that successful implementation and adoption of DDDM or IA in the construction industry, particularly in the context of cost estimation, extend beyond the development of accurate predictive models. Social factors, organizational learning, trust, human in the loop, and effective change management play a vital role in realizing the full potential of DDDM and IA (Grønsund & Aanestad, 2020; Wijnhoven, 2022).

Organizational learning and intelligence amplification (IA) are closely interconnected in the context of DDDM. Organizational learning refers to the process by which an organization acquires, creates, shares, and utilizes knowledge to improve its performance and adapt to changing environment (Wijnhoven, 2022). It involves individuals within the organization collectively gaining new insights, understanding, and skills through a continuous cycle of socialization, externalization, combination, and internalization. Socialization is the process of sharing tacit knowledge among individuals within the organization, creating a collective understanding and group-level tacit knowledge. Externalization involves transforming tacit knowledge into explicit and codified knowledge. Combination refers to the integration of explicit knowledge from different sources, leading to the creation of new knowledge. Internalization occurs when individuals incorporate explicit knowledge back into their personal understanding, evaluating it in the context of their own beliefs, values, and decision-making processes (Wijnhoven, 2022). The concepts of socialization, externalization, combination, and internalization in organizational learning are closely related to the concepts of triple-loop, single-loop, and double-loop learning.

Triple-loop learning refers to the integration of human learning processes with ML processes. It involves the iterative process of individuals and organizations learning from the outcomes of AI systems and integrating those insights into their existing knowledge and decision-making processes (Wijnhoven, 2022). This integration occurs during the internalization phase, where individuals incorporate explicit knowledge back into their personal understanding, evaluating it in the context of their own beliefs, values, and decision-making processes. Triple-loop learning involves leveraging AI systems to enhance organizational learning and decision-making capabilities (Wijnhoven, 2022).

## 2.2 COST ESTIMATION METHODS

There is a significant amount of research regarding cost estimation in the construction industry. Cost estimation methods specifically for engineering services are however more limited. Differences are present between engineering companies and construction cost estimation practices. Costs for engineering companies are less material based and have a higher level of abstraction compared to construction costs because of the inherent difference between offered products and services (Matel et al., 2022). Nevertheless, many of the methods used in the construction industry can also be used in the estimation of engineering services. There are several different cost estimation methods found in the literature (ICEAA, 2009; NASA, 2015). A division is made between the different estimation methods, these are, parametric, engineering build-up, comparative and probabilistic. Another classification can be made between a deterministic model, which creates one single cost estimate, and a probabilistic model, which creates an output range (see Figure 2). First, the different estimation methods are discussed that are commonly applied in cost estimation practice. In the next subchapter, ML-based cost estimation methods are discussed.

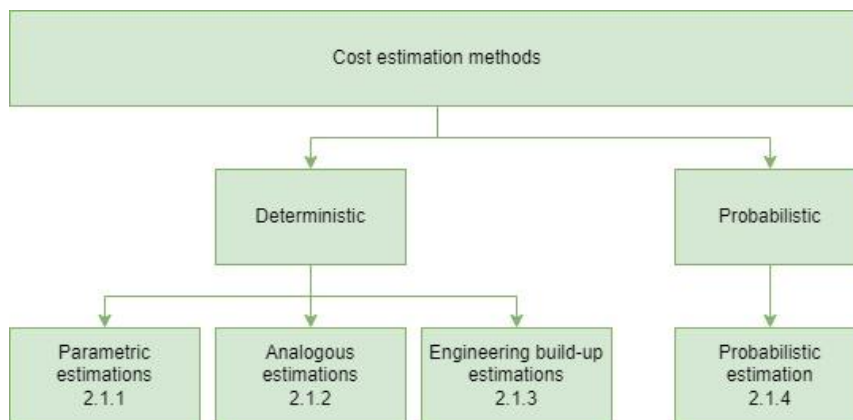


Figure 2: Cost estimation methods division

While many different methodologies exist, some organizations established standard cost estimation methodologies. The literature ranges from best practices, total costs frameworks, cost and value datasets, cost estimations guides, and cost estimations assessments guides. The literature regarding these organizations is summarized below in Table 1. These resources serve as a starting point for understanding cost estimation methodologies.

Table 1: Organizational cost estimation methodologies

| Organization   | Application  | Title                                     | Source        |
|--|--|---|---------------|
| NASA (National Aeronautics and Space Administration)         | Guide for cost estimation                                | Cost Estimating Handbook (CEH)            | (NASA, 2015)  |
| DACE (Dutch Association of Cost Engineers)                   | Cost engineering database                                | Cost and value                            | (DACE, n.d.)  |
| ICEAA (International Cost Estimating & Analysis Association) | Complete guide for cost estimation practice              | Cost estimating body of knowledge (CEBoK) | (ICEAA, 2009) |
| GAO (U.S. Government Accountability Office)                  | Best practices for developing and managing program costs | Cost estimating and assessment guide      | (GAO, 2020)   |



|  |  |                                     |   |
|--|--|-------------------------------------|---|
| AACE (International Association for the Advancement of Cost Engineering) | Guidelines for applying principles of project cost estimates   | Cost estimate classification system | (Christensen & Dysert, 2005)                            |
| CIOB (Chartered Institute of Building)                                   | Guide to essential principles for estimation for building work | Code of Estimating Practice         | (Chartered Institute of Building (Great Britain), 2009) |
| RICS (Royal Institute of Chartered Surveyors)                            | Set of rules for construction cost estimation projects         | New Rules of Measurement            | (RICS, 2020)  |

### 2.2.1 Parametric estimations

The parametric estimations technique, also called the feature-based method, uses statistical relationships between historical data and other variables to calculate an estimation based on parameters (Society of Parametric Analysts, 2008). Parametric estimations look at the relationships between a project's cost characteristics from factual data. Characteristics can include physical attributes, performance specifications, or functions, these relationships are also called Cost Estimating Relationships (CERs). Based on these relationships the costs are estimated (Kwak & Watson, 2005). Techniques like multiple linear regression, factor analysis, and principal component analysis are commonly utilized in parametric methods to pinpoint significant parameters influencing construction costs (Swei et al., 2017).

Parametric approaches are transparent, explicitly account for uncertainty, and can be integrated with analytical tools, which makes them very popular for cost estimations (Swei et al., 2017). Swei et al. (2017) researched a parametric approach to enhance estimated expected costs, in which the parametric model has been shown to decrease cost overruns for large-scale construction projects. Lowe et al (2006) also showed that using multiple regression techniques has higher accuracy than more traditional cost estimation methods (Lowe et al., 2006).

Parametric estimations can be a quick way to accurately estimate costs, even in the preliminary stages of a project, and can be easily replicated (NASA, 2015). It bases this estimation on data and is not susceptible to subjectivity, this is a top-down approach. Furthermore, a preliminary estimate can be created quickly which reduces the cost of preparing a proposal. The dataset on which this model is based is however the most important aspect of the modeling. If the dataset is not large or reliable enough it can create significant estimation errors (NASA, 2015). This is also the case if the relationships of CERs are not valid. Moreover, CERs should be continually revised because new information about projects and cost alters the relationships and thus the entire model. This task is a time-consuming activity, which also requires a significant amount of statistical knowledge about the creation of these models. CERs can be linear as well as non-linear. Creating non-linear CERs requires an algorithm to be created, and the relationships between the variables might be hard to establish.

### **2.2.2 Analogous estimations**

Analogous estimations rely on the comparison of similar projects to derive cost estimates. If a new project is similar to another project a quick comparison can be made on crucial project features. An analogy uses actual costs from similar projects and adjusts these actual costs based on differences between the existing and new projects, thus based on historical data (NASA, 2015). These types of estimations are typically used in the preliminary stage of a project's life cycle (GAO, 2020). Analogous estimations require expert opinions to assess the adjustment level to modify the analogous data to fit the new project. Different types of analogous cost estimation methods exist, of which several are based on machine learning (Cheng et al., 2010).

One of the key advantages of analogous estimates is that they can be employed before the required details are fully known. If there is a strong argumentation for the similarity between projects, an estimate can be justified more easily. Next to that, an analogous estimation can be developed quickly which reduces the cost of proposals. ML methods are the most promising since data is utilized in powerful tools that can quickly and accurately estimate costs. There are however numerous disadvantages, namely, the analogous estimations primarily rely on only one or a few point estimates (GAO, 2020). The analogous estimations induce subjectivity because of the expert opinion that is required to estimate the adjustment level in non-ML methods. Furthermore, in non-ML methods, the examination regarding a strong analogy might be time-consuming and require technical knowledge about program data and projects. Subsequently, the absence of similar historical project data can impede accuracy.

### **2.2.3 Engineering build-up estimations**

Engineering build-up estimations also referred to as detailed, are estimations developed by estimating the cost per activity based on the project's structure (activity-based costing). The structure of projects is commonly presented in a Work Breakdown Structure (WBS) (GAO, 2020). For the estimation, engineers are typically consulted to give insight into the amount of work that needs to be carried out on which a cost estimate is made. The costs are often estimated on the lowest level of detail which is also referred to as the work package. The cost estimator's responsibility is to review the estimated costs of the engineer for validity, logicity, completeness, and overall view (NASA, 2015). Based on the total an additional amount of costs is added which consists of the overhead costs. Engineering build-up estimations are commonly used in more mature projects (NASA, 2015).

The build-up estimate can also be reused to give insight into individual project budgets. Moreover, the amount of detail in the estimate also makes it easier to negotiate with clients because cost details can be defended. The intuitive aspect does however induce a level of subjectivity in the estimation. Because of this, the estimation cannot offer any statistical confidence. Furthermore, the detailed cost estimate has the disadvantage of being labor-intensive because it requires a substantial effort from engineers to create a build-up estimate. Susceptibility to human errors is also introduced because calculations are done manually. New estimates also need to be build-up for each alternative scenario (GAO, 2020).

A survey conducted in the UK showed that major causes of inaccuracy in cost estimations come from the lack of knowledge of engineers, insufficient time, poor documentation, and broad variability in subcontractors' prices (Akintoye & Fitzgerald, 2010). This shows more pitfalls for engineering build-up estimations.

### **2.2.4 Probabilistic estimations**

In the aforementioned methods, the outcome is deterministic, producing one single-point estimate. The probabilistic estimation includes giving a range of possible outcomes. Probabilistic cost

estimations attempt to quantify the risks and uncertainty within cost estimation. Techniques are employed to consider a range of estimates, to account for different potential outcomes, rather than solely relying on a point estimate. A probabilistic estimate is coupled with some commonly used distributions, normal, lognormal, beta, triangular, and Weibull (Chou et al., 2009; Zhu et al., 2016).

A popular used probabilistic distribution is the Monte Carlo distribution (Chou et al., 2009). A Monte Carlo distribution simulates a large volume of randomized numbers within a defined distribution to simulate possible outcomes. In its essence, a Monte Carlo simulation provides the ability to map and handle the uncertainty associated with cost estimation practice (Zhu et al., 2016). A confidence level can be selected to decide the amount of uncertainty that users of the model are willing to handle. Based on this choice a probability distribution is made in which the range of cost estimation is visible.

The primary benefit of probabilistic estimates is that they offer insight into risks, uncertainties, and the precision of the estimate. The range also helps communicate the impact of changes by way of quantification effects (NASA, 2015). Using probabilistic estimating is based on the idea that it is more reasonable to take into account a range of potential outcomes, rather than a single-point estimate. This is because a probabilistic range acknowledges that results can vary (Elkjaer, 2000).

Creating probabilistic models, however, does prove some challenges. Establishing cost distributions for each cost component can be a difficult task to accomplish. To maintain accuracy, a probabilistic model must be updated as new data becomes available, a task that also is time-consuming (Chou et al., 2009).

### **2.2.5 Overview of methods**

In this section, a final overview is given of the four different methods to summarize all the requirements, advantages, and disadvantages, see Table 2. The most generally used methods in cost estimation for engineering services, as well as construction costs, are a combination of the analogous method and engineering build-up method based on the reviewed papers (Doloi, 2011). The methods however, usually fail to capitalize data, which can result in a lengthy and subjective estimation. Through the application of data from previous projects, database models (e.g. ML methods) present the potential to overcome this gap (Matel et al., 2022). Although ML methods exhibit some similarities with the parametric and analogical methods, they are described separately to allow for a comprehensive comparison of all the different techniques.

Table 2: Overview of advantages and disadvantages of cost estimation methods

| Methods              | Requirements  | Advantages  | Disadvantages   |
|----------------------|---|---|---|
| Parametric           | <ul style="list-style-type: none"> <li>- Historical dataset for statistical analysis</li> <li>- Statistical knowledge</li> </ul>  | <ul style="list-style-type: none"> <li>- A quick way for initial estimations</li> <li>- Easy to replicate</li> <li>- Based on data instead of intuition</li> <li>- Reduced costs for preparing estimation</li> </ul>  | <ul style="list-style-type: none"> <li>- Needs a sufficient amount of historical data</li> <li>- The dataset and model need to be maintained</li> <li>- Cost-estimating relationships can be hard to determine</li> <li>- The traceability of CERs is challenging</li> </ul>  |
| Analogous            | <ul style="list-style-type: none"> <li>- Expert knowledge required about previous projects</li> <li>- Comparison factors</li> </ul>   | <ul style="list-style-type: none"> <li>- Can be done with a limited amount of information about the project</li> <li>- Through reasoning more easily defensible</li> <li>- Can give quick first insight</li> <li>- Easy to understand</li> <li>- Accurate if comparative data is available</li> </ul> | <ul style="list-style-type: none"> <li>- Accuracy is limited if no suitable data is available</li> <li>- Intuitive adjustment factors with non-machine learning methods</li> <li>- Requires knowledge about previous projects and data</li> <li>- Needs to be normalized</li> <li>- Difficulty identifying similar project(s)</li> </ul>  |
| Engineering build-up | <ul style="list-style-type: none"> <li>- Expert knowledge</li> <li>- Work breakdown structure</li> <li>- Sufficient amount of time</li> <li>- Sufficient amount of information about project</li> <li>- Validity check by cost estimator</li> </ul> | <ul style="list-style-type: none"> <li>- Very detailed estimate</li> <li>- More easily defensible during negotiations</li> <li>- Insight into key cost components</li> <li>- Reusability for future projects</li> <li>- All cost components are taken into account</li> </ul>                         | <ul style="list-style-type: none"> <li>- Intuitive, thus subjective</li> <li>- Prone to human error</li> <li>- Time-consuming for engineers</li> <li>- High costs to determine cost estimate</li> <li>- Every project needs a new estimate</li> <li>- Expert knowledge may not always be readily available</li> <li>- The scope of projects needs to be defined sufficiently</li> </ul> |
| Probabilistic        | <ul style="list-style-type: none"> <li>- Probabilistic distribution model based on historical data</li> <li>- Statistical knowledge</li> <li>- Software (Monte Carlo Simulation)</li> </ul>   | <ul style="list-style-type: none"> <li>- Ranges of outcomes</li> <li>- Gives insight into risks, uncertainties, and the precision</li> <li>- Helps communicate impact of changes in parameters</li> <li>- Improves accuracy and reliability in estimates</li> </ul>                                   | <ul style="list-style-type: none"> <li>- Each cost component needs a distribution (should first be identified)</li> <li>- Difficulty in correlations between cost components</li> <li>- Can be computationally inefficient (Monte Carlo simulation)</li> </ul>  |

## 2.3 MACHINE LEARNING COST ESTIMATION METHODS

The task of cost estimating can be transformed by ML methods towards a data-driven approach (Tayefeh Hashemi et al., 2020). In the previous sub-chapter, cost estimation methods are described along with corresponding advantages, disadvantages, and requirements according to the industry standards subdivision. In this part, ML-based methods are reviewed to gain insight into future trends and possibilities regarding cost estimating to utilize data for cost estimation. A best practice is proposed based on the literature review. To give an overview of the relationships between data science, AI, and ML, see Figure 3. While there is overlap, the primary difference between data science and ML is that data science examines data and tries to extract meaning from it, whereas the objective of machine learning is to comprehend and construct methods that utilize data to build models that can make predictions (Tayefeh Hashemi et al., 2020).

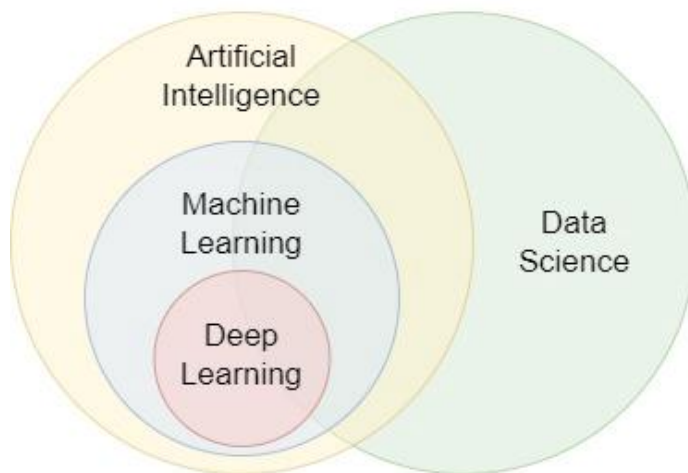


Figure 3: AI, ML, and Data science relationship

In Table 3 below the different ML estimation methods are shown with the corresponding sources to which they are applied. Several methods are reviewed based on literature reviews on the below-stated methods.

Table 3: Machine learning cost estimation methods

| Machine learning methods         | Sources  |
|----------------------------------|--|
| Artificial Neural networks (ANN) | (Hyari et al., 2016; Matel et al., 2022)   |
| Regression models                | (G. H. Kim et al., 2004; Smith & Mason, 1997; Trost & Oberlender, 2003) (B. Kim & Hong, 2011; Swei et al., 2017) |
| Case-based reasoning (CBR)       | (Ji et al., 2011; B. Kim & Hong, 2011; G. H. Kim et al., 2004; Zima, 2015)                                       |
| Random forest algorithm (RF)     | (Shoar et al., 2022)   |
| Decision Tree (DT)               | (Tayefeh Hashemi et al., 2020)   |

### 2.3.1 Machine learning methods

Several studies exist on construction cost estimation utilizing ML-based methods. These methods are mainly focused on the entire cost estimation of a project, this includes materials, construction services, etc. Only a few studies exist on the estimation of engineering services cost. The section below explores several ML approaches to select the most relevant one for the current situation.

**Artificial neural networks (ANN):** Artificial neural networks (ANN) are powerful tools that have been widely applied to estimate the cost of construction projects as well as engineering services costs (Cheng et al., 2010). ANN operate based on the imitation of the functioning of the brain by applying neurons in layers that are equipped with mathematical functions. Based on data input ANN can self-learn through which it can recognize patterns and correlations. A significant drawback of ANN is the black box element, there is no way to gain insight on how the model has estimated a given number. Explainability is thus lacking which makes the model not applicable in every situation. Furthermore, the optimization of ANN is a time-consuming task since it requires trial and error to find the optimum in the number of layers and neurons for the best performance (Tayefeh Hashemi et al., 2020).

Matel et al. (2022) conducted research on an ANN for the cost estimation of engineering services (Matel et al., 2022). While the model showed great potential, the data limitation was the most constraining factor for real applications. The performance of an ANN is dependent on the quality of its input data, reflecting the fundamental "Garbage In, Garbage Out" principle. In other words, the output of an ANN is only as good as the quality of the data it receives. The "Garbage In, Garbage Out" principle is however also applicable to various other ML models.

ANN are one of the most popular applied ML methods for cost estimation (Tayefeh Hashemi et al., 2020). According to Hashemi et al. (2020), hybrid models outperform standard ANN by solving the inherent limitations of ANNs. An example of these hybrid models is the fuzzy evolutionary neural network system. Nevertheless, creating such a system requires advanced technical knowledge and expertise, making it a challenging and time-consuming task to undertake. Despite the demonstrated theoretical advantage of ANN, practical limitations remain a challenge (Matel et al., 2022).

**Regression modeling:** Another popular method that has been widely adopted is regression models. While there is an ongoing discussion about whether regression is classified under ML or statistics (or both), the model is described in this part since several methods that utilize regression are part of ML (e.g. MARS, GAM, and ridge regression).

Many studies have shown the theoretical as well as the practical contribution of parametric estimation methods (G. H. Kim et al., 2004). Regression analysis are also one of the most popular applied methods for parametric cost estimation (Tayefeh Hashemi et al., 2020). As explained in the previous section of the parametric estimation, it is based on the relationship between CERs and tries to predict the dependent variable ( $y$ ) based on the independent variable(s) ( $x$ ). A disadvantage of regression modeling is however the establishment of the relationships between the variables. The more variables there are, the more complex the model becomes, making it less interpretable and harder to establish.

According to (G. H. Kim et al., 2004), the performance of regression models is only slightly inferior to ANN when CERs are known. Regression models have however the advantage that it is more interpretable than ANN which makes them more practical for usage. Furthermore, regression models have the ability to provide certainty about the predicted outcome (Swei et al., 2017).

**Case-based reasoning:** Case-based reasoning (CBR) is a machine learning method that can be seen as a form of an expert system. It is based on rule-based reasoning, which is derived from experience or memory. The objective of CBR is to leverage past problem-solving experiences to tackle current cases,

achieved through associating or comparing (Zima, 2015). CBR works according to the following steps: Retrieve, Reuse, Revise, and Retain. These steps contain observing key attributes, identifying these attributes in similar problems, predict the direction of new problems based on similar experience with adjustments (G. H. Kim et al., 2004). An advantage of CBR models is that they can explain how the cost estimation is made. CBR models are more user-friendly for updates compared to ANN, as including new cases in a CBR model only involves adding the information, whereas ANN updates necessitate a full retraining process (Zima, 2015).

Nevertheless, a CBR still requires a domain expert to assess the adjustment level and project similarity selection (G. H. Kim et al., 2004). Other challenges consist of acquiring and organizing a comprehensive case base, which can be time-consuming and resource-intensive. The quality and relevance of cases are crucial for effective CBR, requiring careful selection and representation, requiring an expert to assess which cases are benchmark. Additionally, defining an appropriate similarity measure to match new cases with past cases is a non-trivial task, as it should capture relevant features and relationships (Ji et al., 2011; Zima, 2015). Another challenge lies in the adaptation process, where adapting past cases to new situations requires domain expertise and can be subjective (Zima, 2015).

**Decision trees:** Decision trees (DT) are a method predominantly used for classification. A DT typically starts with one node branching into different possible outcomes each of which has additional nodes that branch off into other possibilities. A DT divides data into hierarchical rules on each tree node which is split based on an algorithm (Elmousalami, 2021). Because it is primarily used for classification, its strength does not lie in the prediction of a continuous output, thus making it less applicable as a cost estimation method.

**Random forest algorithm:** A random forest algorithm (RF) is another popular supervised machine learning method that utilized multiple classification and decision trees for prediction. The construction of decision trees in the random forest algorithm involves randomly selecting subsets of features and training data. This process is repeated multiple times to create a collection of decision trees. It is a non-parametric approach that can handle unbalanced data both numerical and categorical variables. Furthermore, an RF provides the user to determine the importance or contribution of variables and thus limits the black box problems (Shoar et al., 2022). However, interpreting the way an outcome is composed remains challenging. On top of that, a random forest is also not designed to predict a probability for a continuous output.

Table 4: Overview of advantages and disadvantages of machine learning cost estimation methods

| ML method                       | Requirements   | Advantages  | Disadvantages  |
|---------------------------------|--|---|--|
| Artificial Neural Network (ANN) | <ul style="list-style-type: none"> <li>- Large dataset</li> <li>- Generalizable dataset</li> <li>- Applicable software support</li> <li>- Numeric dataset</li> </ul>   | <ul style="list-style-type: none"> <li>- Can learn from itself</li> <li>- High accuracy</li> <li>- Very flexible models</li> <li>- Powerful tool</li> </ul>                               | <ul style="list-style-type: none"> <li>- Black box</li> <li>- Difficult to establish model</li> <li>- Updating requires entire retraining which is a difficult and time-consuming task</li> <li>- Not easy to optimize model (requires trial and error)</li> <li>- Deterministic model (difficult to give probability/ certainty about the outcome)</li> </ul> |
| Case-based reasoning (CBR)      | <ul style="list-style-type: none"> <li>- Requires experts' knowledge during establishment of CBR</li> <li>- Time-consuming to establish</li> <li>- Features need to be known</li> <li>- Numeric and symbolic data</li> </ul> | <ul style="list-style-type: none"> <li>- Explainable</li> <li>- Reasonable high accuracy</li> <li>- Easy to update with new cases</li> <li>- More applicable for long-term use</li> </ul> | <ul style="list-style-type: none"> <li>- If no similar case exists it can have high deviations</li> <li>- Relationships can be hard to establish (if-then rules)</li> <li>- Time-consuming process</li> <li>- Dependency on experts</li> </ul>   |
| Decision tree (DT)              | <ul style="list-style-type: none"> <li>- Sufficient amount of data</li> <li>- Numeric and symbolic data</li> </ul>   | <ul style="list-style-type: none"> <li>- Interpretable</li> <li>- Can give a quick estimation</li> <li>- Able to handle large datasets</li> </ul>   | <ul style="list-style-type: none"> <li>- Primarily used for classification problems</li> <li>- With increasing complexity trees can be hard to interpret</li> <li>- Less effective in predicting continuous outcome</li> </ul>   |
| Random Forest (RF)              | <ul style="list-style-type: none"> <li>- Combination of several decision trees</li> <li>- Numeric and symbolic data</li> </ul>   | <ul style="list-style-type: none"> <li>- Non-parametric</li> <li>- Combination of several decision trees</li> </ul>   | <ul style="list-style-type: none"> <li>- Not easy to interpret</li> <li>- Relative slow model</li> </ul>   |
| Regression                      | <ul style="list-style-type: none"> <li>- Statistical software</li> <li>- Linear data</li> <li>- Requires significant data prerequisites before applicable use</li> <li>- Numeric data</li> </ul>                             | <ul style="list-style-type: none"> <li>- Interpretable</li> <li>- Easy to make</li> <li>- Ability to give range output (confidence interval)</li> </ul>                                   | <ul style="list-style-type: none"> <li>- CERs need to be established</li> <li>- As complexity increases the models become harder to make (primarily with non-linear relationships)</li> <li>- Susceptible to outliers</li> </ul>   |



### 2.3.2 Discussion of machine learning methods

This research aims to develop a data-based cost estimation method that utilizes the emergent data captured rather than relying on intuition. Three methods were considered based on Table 4: regression modeling, ANN, and expert systems (CBR). Probabilistic methods were favored over deterministic ones because they provide a range outcome that is easier to communicate to model users (Chou et al., 2009). This research aims to develop a model that relies less on expert knowledge and instead focuses on data-driven approaches. Unlike expert systems like CBR, which heavily depend on expert knowledge for rule definition and solution adaptation, this research aims to reduce reliance on explicit expert knowledge. By leveraging patterns and insights obtained directly from the data, the goal is to create a more autonomous and scalable model capable of making predictions and decisions based on the inherent information within the data, thereby reducing the need for expert knowledge.

The two models that remained are ANN and regression, which showed great practical potential for predictive analysis and integration into decision-making processes (G. H. Kim et al., 2004). In Table 5 criteria for the model are given in which regression and ANN are compared.

The model should help tender departments with preliminary decision-making, so creating a black box will limit its explainability and justification of results. ANN can be incredibly flexible and accurate with adequate and dependable data (Chou et al., 2009), but the development of an explainable ANN may not be practical due to the significant technical knowledge required, as explainable AI is still a novel area of research (Barredo Arrieta et al., 2020). Although neural networks tend to achieve high accuracy, inherent limitations, primarily the black box problem, impede their practical application, see Figure 4.

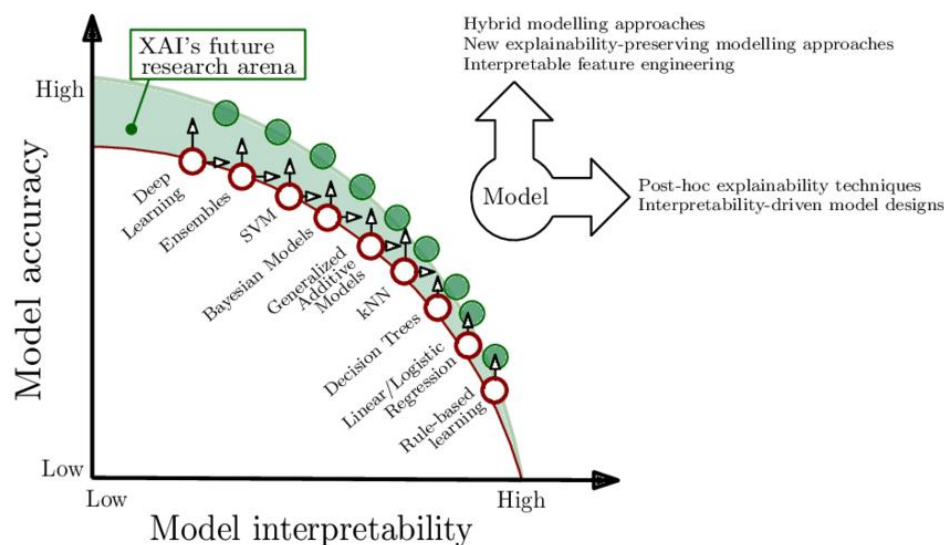


Figure 4: Trade-off between model interpretability and performance (Barredo Arrieta et al., 2020)

The tradeoff between accuracy and interpretability<sup>2</sup> is an important consideration, and although regression models perform slightly worse than neural network models, the difference is minor according to (G. H. Kim et al., 2004). In professional scenarios, the direct impact of artificial intelligence taking over human decision-making is often impractical because it requires decision accountability,

<sup>2</sup> Interpretability, explainability and transparency are all related concepts, while there is overlap, they differ in their specific focus and implications (Doran et al., 2018; Müller & Guido, 2017). In this research explainability is used as the encompassing concept for insight into the model because it provides a systematic framework and taxonomy for understanding and categorizing different explainability techniques (Barredo Arrieta et al., 2020). The operationalization of explainability is contextually dependent based on the created artifact and the specific requirements and goals of the project (Wieringa, 2014), and thus elaborated upon in chapter 4.

involves dealing with ambiguous problems, and entails managing decisional uncertainty (Wijnhoven, 2022). Including scientific justification in cost estimation methods, by adequately describing the technical processes behind the achieved results, not only increases confidence but also enhances transparency and maintainability (Elfaki et al., 2014).

Both techniques' models compare favorably to traditional cost estimation methods. Regression is more advantageous than an ANN when Cost Estimating Relationships (CERs) are known according to G.H. Kim et al (2004)(G. H. Kim et al., 2004)(G. H. Kim et al., 2004) since the determination of uncertainty in predictions is much easier to achieve in regression models compared to neural networks. The tradeoff between the neural network and multiple regression lies in the model's complexity as well as the model's prediction accuracy (Cheng et al., 2010; Zima, 2015). The model is used in a decision-making process, in which a deterministic black box model is not ideal. Furthermore, ANN requires a significant amount of data to come to their strength, and developing the model on only a limited amount of data can result in poor performance (Tayefeh Hashemi et al., 2020).

Due to the limited number of variables available in the dataset, which are also known during the proposal phase, and the scarcity of information in the earlier stages of the estimation process, the estimation process can be challenging, highlighting the significance of demonstrating the level of uncertainty. Based on these arguments, regression modeling is deemed the most suitable for predicting the cost of projects for an engineering company, for this research context.

Table 5: Regression vs ANN for cost estimation

| Criteria                                     | Regression Model  | Artificial Neural Network   |
|--|---|---|
| Accuracy                                     | Lower   | Higher  |
| Complexity                                   | Lower   | Higher  |
| Flexibility                                  | High  | High  |
| Explainability                               | High  | Low   |
| Training time                                | Short   | Long  |
| Performance on limited data                  | Fair  | Poor  |
| Ability to determine uncertainty             | Good  | Poor  |
| Data type (nominal, ordinal, ratio, ordinal) | Suitable for continuous data, (dummy variables for nominal and categorical data), more difficult with (many) categorical data | Suitable for all types of data (dummy variables for nominal and categorical data) |

## 2.4 REGRESSION METHODOLOGY

### 2.4.1 Regression formula & assumptions

An important remark to make is that this research is focused on multiple regression models, which incorporate two or more independent variables for the prediction of the dependent variable. De Veaux et al. Stats Data and Models book is used as a basis theory of regression modeling (De Veaux et al., 2021). Furthermore, papers are consulted on different types of applicable methods used for cost

estimation. While there are different formulas for different regression modeling types (logistic, quadratic function, etc.), the general formula for multiple regression is shown below, Equation 1 (De Veaux et al., 2021).

Equation 1: General regression formula

Formula:

$$y(i) = b_1x_1(i) + b_2x_2(i) + \dots + b_nx_n(i) + c + e(i)$$

where:

- **y(i)** represents the value of the dependent variable for the ith observation in the dataset.
- **x1(i), x2(i), ..., xn(i)** represent the values of the independent variables for the ith observation in the dataset.
- **b1, b2, ..., bn** represent the coefficients or slopes associated with each independent variable.
- **c** represents the constant or intercept term.
- **e(i)** represents the error or residual associated with the ith observation.

An important aspect to take into account is the assumptions, which need to be fulfilled before using conducting the regression analysis. When these assumptions are not satisfied, the model might produce inaccurate and unreliable results (De Veaux et al., 2021). These assumptions are presented in Table 6. When some assumptions are not met, there do however exist ways to fulfill these, this is also presented in this table. However, it is important to note that there are several alternative methods available to address these assumptions in case of violation.

Table 6: Regression assumptions

| Assumption        | Definition  | Method to Fulfill  |
|-------------------|---|--|
| Linearity         | The relationship between the independent and dependent variables needs to be linear   | Non-linear transformations, such as polynomial transformations or using spline functions, can be employed to achieve linearity.  |
| Normality         | The errors (residuals) in the regression model are assumed to be normally distributed with a mean of zero   | A log transformation or square root can be utilized to satisfy this assumption   |
| Multicollinearity | The statistical phenomenon where two or more independent variables in a regression model are highly correlated with each other, making it difficult to determine the individual effects of each variable on the dependent variable. | Methods to address multicollinearity include removing one of the correlated variables, performing dimensionality reduction techniques (e.g., PCA), or using regularization techniques like ridge regression. |
| Homoscedasticity  | The variance of the errors (residuals) should be constant across all levels of the independent variables, indicating consistent variability in the residuals across the range of independent variables.                             | Using heteroscedasticity-consistent standard errors or weighted least squares regression can help address heteroscedasticity and achieve homoscedasticity.   |

### 2.4.2 Regression models

In the overview below different types of regression models are shown. While there consists a myriad of different regression types, a few popular examples that can be utilized are shown below. Different types of regression models are evaluated, consequently feature selection methods are reviewed to give an overview of possible regression methodologies (De Veaux et al., 2021; Müller & Guido, 2017).

- OLS (Ordinary Least Squares) is a method used in linear regression to estimate the coefficients that best fit the observed data by minimizing the sum of squared differences between the predicted values and the actual values. It provides a closed-form solution for finding the coefficients that create the best-fit line or hyperplane to describe the relationship between the independent variables and the dependent variable.
- Ridge regression is a linear regression technique used to address multicollinearity, which occurs when predictor variables are highly correlated. It accomplishes this by adding a penalty term to the ordinary least squares regression objective function. This penalty reduces the magnitude of regression coefficients and shrinks them toward zero. By striking a balance between model complexity and overfitting, Ridge regression improves the model's ability to generalize to new data.
- Lasso regression, also known as L1 regularization, is a linear regression technique that serves as both a feature selection and regularization method. Similar to Ridge regression, it adds a penalty term to the ordinary least squares regression objective function. However, the penalty in Lasso regression is based on the absolute values of regression coefficients. This leads to sparsity in the model by driving some coefficients to exactly zero. As a result, Lasso regression is particularly useful for high-dimensional datasets and prioritizing the most important predictors.
- Polynomial regression: This is a type of regression analysis in which the relationship between the variables is modeled as an  $n$ th-degree polynomial and is better suited for non-linear data. Polynomials are more flexible and can fit more complex data, this flexibility can also lead to overfitting, where the model fits too closely to the noise in the data, rather than the underlying patterns.
- S-curve regression: in the S-curve regression the relationship between the variables is shaped by an S-curve function. This is used for non-linear relationships.
- Generalized Additive Models (GAM): are regression models that extend linear regression by incorporating non-linear techniques. They allow for flexible modeling of various data types and distributions. Although based on the generalized linear model (GLM), GAMs relax some of its assumptions. While overfitting is a potential concern, regularization methods like smoothing parameters can help address it. Interpretability of GAMs can be more challenging compared to linear regression due to the inclusion of non-linear effects and interactions. Adequate data is necessary for accurate estimation of GAM functions
- Partial least squares regression (PLS): Tries to explain the maximum amount of variance in the independent variables while also capturing the maximum amount of covariance between the independent and dependent variables. PLS is useful for handling multicollinearity in multiple linear regression.
- Principal component regression: is a multivariate statistical technique that involves reducing the dimensionality of a set of correlated predictors through principal component analysis (PCA) and using these principal components as input variables in a linear regression model. PCR can be useful for handling multicollinearity and improving the performance of linear regression models when there are many correlated predictors.

In practice, a regression model can have combinations of linear and non-linear components. Choosing the right model is highly dependent on the input data. Feature selection is another important step for

which different techniques can be applied. These techniques vary in their complexity and assumptions, three different techniques are discussed below. Forward selection is a method that starts with no variables in the model and tests each variable one at a time, adding the variable with the best fit at each step until no additional variables significantly improve the regression model. Backward elimination consists of starting with all variables available in the dataset (or selection beforehand) and removing variables that contribute the least to the model until no additional variables can be removed without significantly decreasing the performance of the model. Stepwise selection combines both forward and backward selection adding and removing variables based on statistical criteria such as the F-test. Stepwise regression tries to find the best subset of variables without overfitting the model or limiting the performance (De Veaux et al., 2021).

## 2.5 VARIABLE SELECTION:

As mentioned, a myriad of research has been conducted towards cost estimation of construction projects and only a limited amount of studies has researched cost estimation for engineering companies, which has led to a limited amount of research on variable selection for such firms. The variables (CERs) used in the cost estimation fundamentally differ from the variables used in the estimation of the cost of engineering services. To establish the most significant variables a few studies regarding the estimation of engineering services cost are reviewed. These studies consist of the following. Hyari et al. (2016) conducted research regarding the conceptual cost estimation model for engineering services in public construction projects which used a total of five variables (Hyari et al., 2016). Shoar et al. (2022) researched the application of the RF model to predict cost overruns in high-rise residential building projects and identified 12 variables affecting cost overruns (Shoar et al., 2022). In a separate study, Matel et al. (2022) employed an ANN to estimate the cost of engineering services. Through their analysis, they identified a total of 16 variables.

A total of 16 cost factors are established based on previously conducted studies on which the following, Table 7 is created, which are relevant factors for estimating the costs of engineering services.

*Table 7: Cost-affecting variables in an engineering company context*

| Variable number | Cost factor                                      | Type of variable | Source   |
|-----------------|--|------------------|--|
| 1               | Scale of work                                    | Ratio            | (Hyari et al., 2016; Matel et al., 2022)                     |
| 2               | Project phase                                    | Ordinal          | (Hyari et al., 2016; Matel et al., 2022)                     |
| 3               | Project duration                                 | Ratio            | (Matel et al., 2022; Shoar et al., 2022)                     |
| 4               | Type of work                                     | Nominal          | (Hyari et al., 2016; Matel et al., 2022; Shoar et al., 2022) |
| 5               | Project scope                                    | Nominal          | (Matel et al., 2022)   |
| 6               | Level of experience on the client's side (scale) | Ordinal          | (Matel et al., 2022; Shoar et al., 2022)                     |
| 7               | Quality of information (scale)                   | Ordinal          | (Matel et al., 2022; Shoar et al., 2022)                     |
| 8               | Number of project team members                   | Ratio            | (Matel et al., 2022; Shoar et al., 2022)                     |
| 9               | Collaborating disciplines (number)               | Ratio            | (Matel et al., 2022)   |
| 10              | Type of client and requirements (scale)          | Ordinal          | (Matel et al., 2022; Shoar et al., 2022)                     |

|    |   |         |  |
|----|---|---------|--|
| 11 | Main market type                                | Nominal | (Hyari et al., 2016; Matel et al., 2022) |
| 12 | Client's attitude toward design changes (scale) | Ordinal | (Matel et al., 2022; Shoar et al., 2022) |
| 13 | Project manager experience (scale)              | Ordinal | (Matel et al., 2022; Shoar et al., 2022) |
| 14 | Pre-contract design (scale)                     | Ordinal | (Matel et al., 2022; Shoar et al., 2022) |
| 15 | Contract type                                   | Nominal | (Matel et al., 2022)                     |
| 16 | Intensity                                       | Ratio   | (Matel et al., 2022)                     |

- The scale of work: This is defined as the total investment costs (TIC), which is expressed as the CAPEX. The higher the CAPEX value, the more work usually needs to be done by the engineering firm.
- Project phase: There are four different types of phases for engineering firms; these consist of feasibility, conceptual development, basic engineering, and detailed engineering. The amount of work and level of detail per phase differs substantially which thus affects the amount of time spent on the project.
- Project duration: The duration of the project (lead time) has an impact on the amount of work that needs to be performed in a certain period. A limited amount of time might require more people or disciplines to work alongside which requires a lot of information sharing, thus also more coordination of project managers. A consequence of mistakes can be rework that amplifies the costs.
- Type of work: Engineering firms can carry out three distinct types of roles, which are Engineering (E), Engineering Procurement and Construction (EPC), and Engineering Procurement and Construction Management (EPCm). There are significant differences between the financial risks for the different roles, and thus important to include them.
- Project scope: The project's nature is concentrated on whether it is a new build (greenfield) or an extension, or maintenance of existing construction (brownfield). Typically, a new build requires more effort as many aspects must be defined.
- Level of experience on the client's side: This affects the project significantly since more experienced clients mean a more fluent progression of the project. Furthermore, it also affects the amount and quality of information the client delivers.
- Quality of information: When there is insufficient or unreliable information available, the potential for increased risk is heightened, which can lead to unanticipated expenses throughout the project.
- Number of project team members: The number of project team members required is expected to affect the costs when more members are required to work on it since this it is expected that this will influence the amount of hours worked on the project. The number of project team members consists of the total number of employees that have worked on a project (thus not FTE).
- Collaborating disciplines: This concerns the number of different disciplines (e.g. electrical, mechanical, piping, etc.) that work on the project. The more disciplines work on the project, the more coordination is required between the disciplines which affects the project's management. Additionally, the number of disciplines involved also influences the project's magnitude, as more disciplines often necessitate work in various fields, resulting in a larger project.
- Type of client and requirements: The level or amount of requirements that the clients request significantly impact the total amount of work that needs to be performed. Clients might require specific guidelines or models that need to adhere to strict rules which can impact workload, in turn affecting the work hours.

- Main market type: The market type is important since the work that is performed in the different markets (e.g., Pharma, Food, Oil & Gas, etc.) differ substantially. Different standards are applied for the different markets that can affect the requirements, type of work, drawings, etc., and this can potentially influence the cost.
- Client's attitude towards design changes: The client's attitude towards design changes could affect cost based on the cooperativeness when changes occur, from either side. When clients request additional work, this can affect the cost further.
- Project manager experience: The project manager's experience is of importance, which has been shown to be significant by a previous study conducted internally within Company X. Cost is affected by the level of experience due to the amount of time, coordination, and collaboration required, the level of experience is measured as a tile of project management from A to D.
- Pre-contract design: The amount of work needed to achieve the project deliverables depends on the pre-design completion level. The quality and extent of the pre-design phase may vary, which can affect subsequent project phases. If the pre-design is incomplete or of low quality, more effort may be required in the next phase to ensure project success
- Contract type: Generally, there are three different types of contracts fixed price, reimbursable ceiling, and reimbursable no ceiling. A fixed price has more risks and thus a higher contingency, which enlarges the total price. For reimbursable contracts, the level of risk is lower which lowers the contingency. The contract type might potentially affect the cost.
- Intensity: The amount of work that must be done in a short time is typically higher in high-intensity projects, which are defined by the number of hours the team works on the project per week. As a result, errors made during the project can have a substantial cost impact.

### 3. METHODOLOGY

In this chapter, the methodology of this research is described. First, the research strategy is described which entails the proposed method (Figure 7) for developing the model. The different phases of design science are described independently.

#### 3.1 RESEARCH STRATEGY

The research strategy outlines the different steps that are taken to conduct and answer the research question. This research employs a design science research approach. Design Science is a research methodology that focuses on creating innovative artifacts to address practical problems and improve the understanding of these artifacts (Hevner et al., 2004; Wieringa, 2014). According to Hevner et al. (2004), design science combines the knowledge from both the design and science disciplines to develop effective and usable solutions. The goal of design science is to create new knowledge through the design, development, and evaluation of artifacts that can be applied in real-world contexts, see Figure 5. In this research context, there exists a problem concerning the practical implementation of cost estimation models in the research environment, which significantly affects their effectiveness and real-world applicability. Despite their acknowledged academic potential, the limited practical use of these models highlights the presence of challenges or barriers that hinder their widespread adoption in industry settings (He et al., 2021). This emphasizes the existing gap between the envisioned theoretical potential and practical application, underscoring the need to bridge this divide and develop usable solutions through the application of design science.

By adopting a design science approach, this research integrates theoretical foundations and practical considerations specific to engineering companies, creating a synergistic blend of theoretically grounded and practically applicable knowledge. Building upon the knowledge base established in the previous chapter through a comprehensive literature review, design science is instrumental in addressing specific business needs and leveraging existing knowledge to drive practical applications. Through the conduct of design science research, this study not only contributes valuable insights to the research environment but also expands the knowledge base, providing a robust theoretical framework for future research and development initiatives that propel advancements in real-world practices.

The design science follows the following phases in this research, these are; Problem identification & motivation, Define objectives for possible solutions, Design & Development, Demonstration, and Validation (Hevner et al., 2004; Wieringa, 2014). In this research, the final phase differs from the typical evaluation phase described in the research phases of design science by Hevner et al. (2004). Instead of evaluation, the focus is on validation due to the non-implementation of the developed model. Validation research involves subjecting an artifact prototype to various scenarios presented by a contextual model to observe its response. On the other hand, evaluation research examines how implemented artifacts interact with their real-world context. Validation is performed before implementation, whereas evaluation occurs after implementation (Wieringa, 2014).



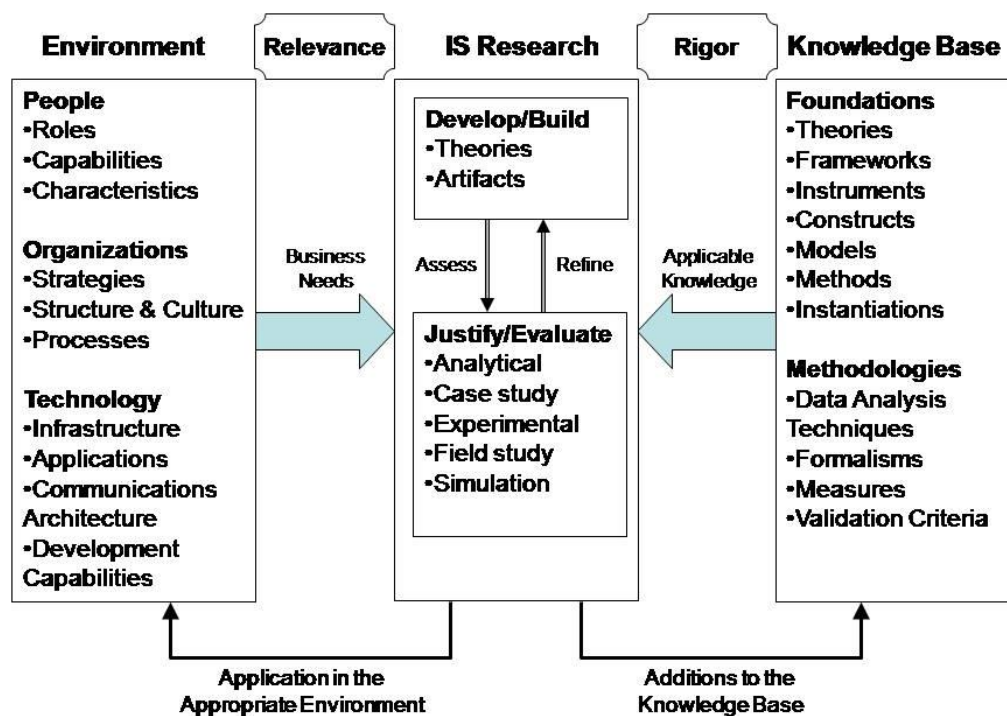


Figure 5: Design science research (Hevner et al., 2004).

This research uses qualitative and quantitative data. The qualitative data emerges from interviews and discussions held with stakeholders and domain experts. The structure of the interviews was based on pre-defined questions, but it remained flexible, enabling a more open conversation (semi-structured). This approach facilitated a deeper understanding of the perspectives and experiences of the interviewee(s), informing future decisions and strategies. It encouraged the sharing of insights and knowledge, leading to the identification of potential challenges and opportunities that may have otherwise been missed (Loubser, 1968). For an overview of all the interviews held for which purpose, see Table 8. Quotes are not used in this research due to confidentiality reasons, as discussed with the company supervisors of this research project.

Table 8: Overview conducted Interviews/ discussion

| Interview/<br>discussion<br>number | Goal interview           | Function<br>interviewee        | Duration                                       | Appendix   |
|------------------------------------|--------------------------|--------------------------------|--|------------|
| Interview 1                        | Meta-Requirements        | Tender department              | 26 min   | Appendix D |
| Interview 2                        | Meta-Requirements        | Tender department              | 23 min   | Appendix D |
| Interview 3                        | Meta-Requirements        | Tender department              | 16 min   | Appendix D |
| Interview 4                        | Meta-Requirements        | Tender department              | 24 min   | Appendix D |
| Interview 5                        | Impact Work<br>Processes | Tender department              | 23 min   | Appendix G |
| Interview 6                        | Validation               |                                | 40 min   | Appendix H |
| Interview 7                        | Validation               |                                | 22 min   | Appendix H |
| Discussion 1                       | Validation               | Various functions<br>(6 total) | 15 min<br>presentation<br>40 min<br>discussion | Appendix I |

The interviews are transcribed, coded, and summarized (Huberman & Miles, 2014). Huberman and Miles (2014) state that transcribing and coding interviews in qualitative data analysis are crucial. Transcribing ensures accuracy and provides a foundation for analysis, while coding helps identify themes, patterns, and relationships, leading to a comprehensive understanding of the phenomenon under study (Huberman & Miles, 2014). A thematic analysis is conducted by examining the transcriptions and identifying recurring themes, topics, or concepts that arise from the interviews. These themes are utilized as initial categories for the coding schema. Subsequently, an iterative process is employed to revise and refine the coding schema.

The Gioia method provides a framework for organizing the data into first-order, second-order, and aggregate dimensions. First-order themes capture specific concepts or ideas that emerge directly from the data. These themes are identified by highlighting interesting parts of the transcribed interviews. Second-order themes, on the other hand, involve grouping related first-order themes to create broader categories or dimensions. These dimensions provide a higher-level understanding of the data by capturing commonalities and connections among the first-order themes. Finally, aggregate dimensions further consolidate the second-order themes to form overarching concepts or constructs that represent the essence of the data (Gioia et al., 2013).

In the reviewed papers within this research, the research area under investigation is primarily focused on quantitative analysis, this research utilizes a holistic approach to gain a more comprehensive understanding of the topic. Consequently, the coding schema does not rely on priori codes. Instead, an inductive coding approach is adopted, where codes are derived from patterns and themes that emerge from the interview data. This allows for the discovery of novel insights and perspectives. The iterative process of inductive coding facilitates a comprehensive exploration of the research topic, ensuring a rich and nuanced analysis. Appendix B provides more insight into the data schema.

The quantitative data gathered and analyzed in this research is done in Excel and Python. The data cleaning and outlier detection are done in Excel since the received data from IT also uses Excel as an output. For the development of the model Python is used, see Figure 7 for more information. In Figure 6 the corresponding chapters are shown per research phase with the corresponding sub-research questions.

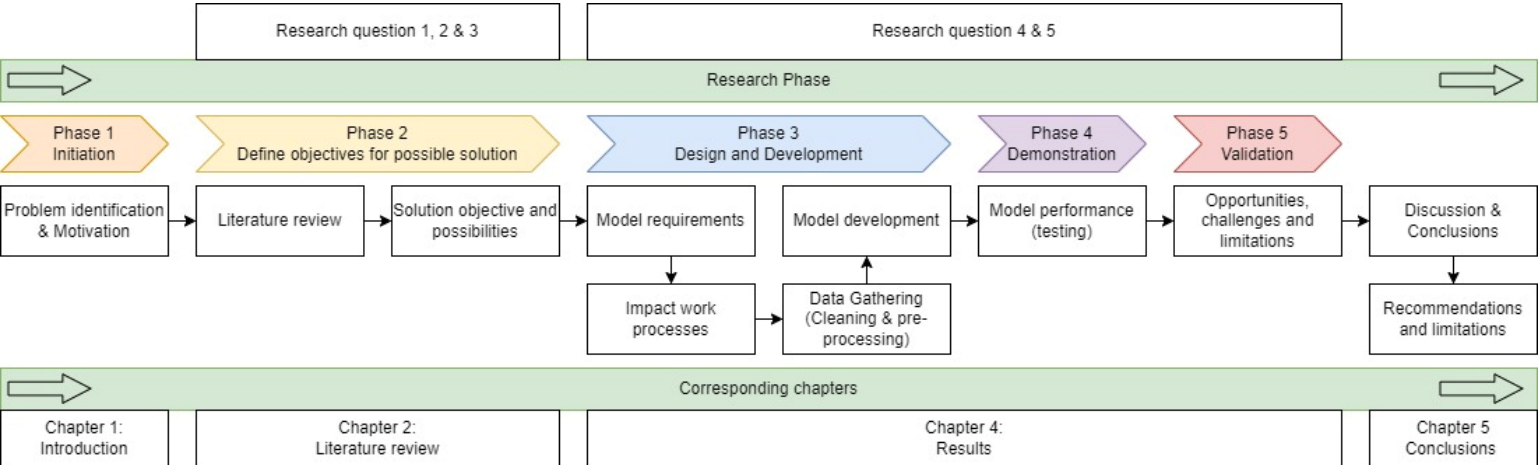


Figure 6: Research strategy phases

**Phase 1: Initiation**

The first phase consists of starting the research and defining the problem. It is important to start by knowing what the details and the exact problems are. This phase is done by performing an initial literature review and consulting domain experts. Different perspectives are taken into account to

clarify the most compelling problem. The problems with the current method used by Company X are addressed, as well as trends in current literature and why it is important to conduct this research. This part is described in chapter one, the introduction.

### **Phase 2: Define objectives for possible solution**

In phase two, possible solutions are defined and described. This is presented in chapter two, the literature review. In the literature review, data-driven models and their relationship to data-driven decision-making are described, furthermore, research regarding cost estimation techniques and data-based techniques are reviewed. First DDDM is described along with opportunities and challenges with DDDM. The following key words were used “Data-driven”, “Decision making”, “Data”, “Conceptual cost estimation”, “Cost estimation”, “Engineering company”, “Construction”, “Data preparation”, “Data quality”, “Data management”, “Challenges”, “Opportunities”.

Following this, a best practice based on cost estimation organizations is researched, in which key words like; “Cost estimation”, “Cost engineering”, “Parametric estimation” were employed, which in turn resulted in Table 1, a total of seven sources. Furthermore, research regarding current commonly used cost estimation is reviewed, this is done by searching for key words like; “Cost estimation” AND OR “Construction” AND OR “Engineering company” AND OR “Costs of services” AND OR “Conceptual cost” AND OR “Data-based”, AND OR “Machine Learning”, “AI”, “ANN”, “Neural network”, “Regression”, “Case-based reasoning”, “Decision Tree”, “Random Forest”. For the construction industry key word, this search resulted in a myriad of papers, of which the conceptual cost estimate papers were primarily used due to their overlap of characteristics with engineering companies' cost estimation. In total 29 papers are used for the literature review. The search for papers was conducted primarily using Google Scholar and FindUT databases.

The literature review encompasses an evaluation of existing cost estimation methods, highlighting their strengths and limitations. It then delves into machine learning-based approaches, identifying opportunities, and challenges and exploring potential remedies to address these issues. Furthermore, variables identified in the used papers are described in chapter 2.4.

### **Phase 3: Design and Development**

The design and development phase is split up, see Figure 7. The formulation of this figure incorporates insights from the literature, with the parametric estimation model development method of ISPA serving as a key reference point (Society of Parametric Analysts, 2008).

#### **- Design**

The first step in the design phase is conducting interviews to establish the variables and model requirements. The variable selection consists of understanding the underlying mechanism and causal relationships that are likely to be important for predicting the outcome variable, Table 7. However, because of the limited number of studies, the variables need to be validated in practice, since there might be other relevant variables. To validate these variables or add other variables, interviews are held with domain experts. The validation of the variables and defining the requirements are both addressed in the interview of Appendix D.

The interviews are conducted with four employees of the tender department. The tender department is chosen based on its expertise in preparing tenders and proposals for engineering services. The interview questions are derived from the prior literature review, specifically focusing on the significant characteristics identified in the criteria column of Table 5. These characteristics have been deemed

important concerning ML-based models and their practical application. The following questions are asked:

1. For the preparation of a man-hour estimate, what approach and processes are applied?
2. What are the variables that are typically estimated in a proposal?
  - a. Which variables should/ can be used?
  - b. What level of detail should be provided? (one output, or an estimate per department?)
  - c. How many variables need to be input beforehand, and are they generally knowable in advance?
3. Should the estimate be defensible (i.e. provide some form of explanation)?
4. What is the correct balance of parameters that can be estimated in a limited amount of time? (e.g. an hour after reading the RFQ)
5. If a model exists to estimate the man-hours what is the preference, for a deterministic or probabilistic model?
  - a. Deterministic is one number as output.
  - b. Probabilistic is a range of output, so there is a degree of uncertainty.
6. Should the model provide any explanation of how the estimate was generated?
  - a. What would be other requirements regarding a model that estimates the costs of projects?

Defining requirements is a crucial step in design science research due to its relevance in guiding the development and validation of design artifacts. The design artifact refers to the developed model. By clearly articulating the requirements, researchers can align their design efforts with the identified problems or opportunities, ensuring that the resulting artifacts address the specific needs and goals. This study adopts a design science research approach that incorporates the use of meta-requirements, meta-design, and kernel theories, in accordance with the principles established by (Hevner et al., 2004; Walls et al., 1992)

The formulation of requirements for the intended artifact involves identifying meta-requirements, which represent the class of goals to be addressed by the application/use of the design artifact. These meta-requirements serve as a guide for developing the design artifact, known as the meta-design, which is hypothesized to fulfill the identified meta-requirements. They are called meta-requirements rather than just requirements because they address a generalized class of goals rather than particular, situated goals (Venable, 2006). Moreover, it is called meta-design rather than just design because the design product is not a particular instantiation, but a general approach to be used in particular occurrences of the class of goals in the meta-requirements (Venable, 2006). Furthermore, kernel theories are drawn from natural or social sciences and “govern design requirements” that support the meta-design (Venable, 2006; Walls et al., 1992). Both the theory from chapter two and the interviews with the tender department serves as the foundation for the requirements.

Moreover, the impact this model has on the work processes is evaluated, this is done through adjusting current BPMN models and evaluating these with a domain expert. An interview/ discussion is set up with the tender management department in which the following parts are discussed:

- How does the current develop model impact current work processes regarding the proposal phase?
- What are the likely perspectives of different stakeholders, such as engineers, project managers, and management, towards the model?
- How would such a model be implemented within the Tender department?
  - a) What would challenges be regarding use or implementation?

- What is the role of engineers, project managers, and tender managers now?

See Appendix F for the summary and coding schema of the interview.

**Data gathering:** For the gathering of the data, different databases internally within Company X are consulted. Within Company X different IT systems are employed and there is not one accessible database in which all the required data is available. This means different types of data (e.g. project data, proposal data, financial data, etc.) are gathered from different sources. It is crucial to gather a comprehensive amount of data, as data-based models heavily rely on the quantity and quality of data available. The effectiveness and accuracy of these models are greatly impacted by the data they are trained on, making thorough data collection an essential part of the process (De Veaux et al., 2021).

The majority of data originates from ██████████, a system that can be accessed through the IT department. The IT department provided a printout containing project and proposal data of closed and archived projects. However, some previously defined variables were not, or only limited present in the data. To further expand the dataset different experts are consulted to provide additional data. Through this process as much data as possible is gathered of the previously identified variables from the requirements part. Based on this, a single dataset is created which has all the (available) required inputs from which redundant or confidential data is removed. No external data sources are consulted, thus only internal data created and gathered by Company X is used for the development of the model.

**Data cleaning & pre-processing:** Data cleaning deals with detecting and removing errors to improve the quality of the data. Failure to ensure this can result in a model developed on erroneous data, leading to inaccurate predictions or conclusions (De Veaux et al., 2021). Errors can consist of missing values, outliers, and deviating patterns. If there is no feasible way to recover missing values, they are eliminated from the dataset. Outliers are data points that are significantly different from other observations in the dataset. Outliers can have a significant impact on the result of the regression model and thus need to be evaluated (in coordination with the proposal manager) whether they should remain or be removed from the dataset. This is done by visual inspections of the descriptive statistics (e.g., min, max, and range) or QQ plots (normal probability plots). The next step is to pre-process the data for usage in the chosen software for analysis (Python). These steps contain setting the right number of rows, saving the file in the right format, removing any other oddities, and changing categorical variables to numerical ones.

Furthermore, data is standardized which is a critical pre-processing step that involves transforming the variables to have a common mean and standard deviation, making them more directly comparable. Standardization is essential when the variables in a dataset have different units of measurement or are measured on different scales. This can make it challenging to compare the variables and can cause variables with larger values or wider ranges to dominate the analysis. The formula used for standardization is  $X_{std} = (X - \text{mean}(X)) / \text{stddev}(X)$ . Standardizing categorical and nominal variables involves a different approach, which entails converting them into binary indicator variables where each category is represented by a value of 0 or 1. See Appendix F for the snippet of Python code.

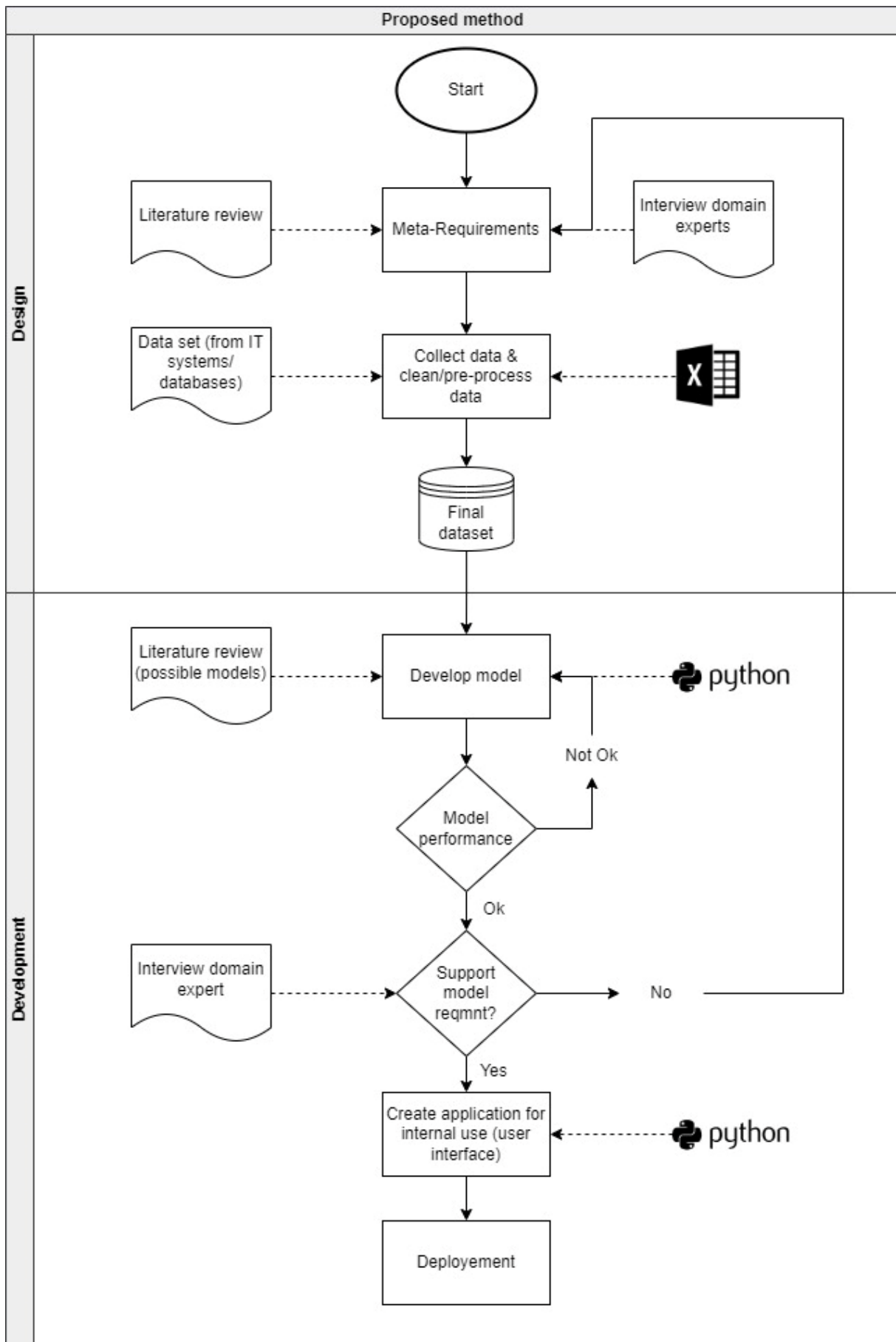


Figure 7: Proposed method

## - Development

First, the feature selection procedure is handled, which forms the basis on which variables the model is developed. The next step is the selection of the appropriate model, this is dependent on the input data and can thus only be assessed at this stage.

**Variable selection:** Variable selection is the process of identifying the most relevant variables for the model, also known as CERs or features. Its goal is to determine which variables have a statistically significant impact on the outcome. Statistical testing, like T-tests, is used to analyze the available data and identify variables that strongly influence the outcome. This entails assessing the correlation and significance values of the features and selecting those that are statistically significant.

To select variables, a stepwise regression method is used, retaining only significant variables. A Python script is written to test all possible variable combinations automatically and select the most significant model with significant variables. This code maximizes the R-squared value while maintaining significance. Additionally, another script checks for assumptions to determine if they are fulfilled or not, this code is run on the different regression models. This automation replaces trial-and-error. Refer to Appendix F for a code snippet.

**Model selection:** In the literature review, different types of models are mentioned which have different strengths. Due to the absence of a predetermined “best” approach in regression modeling, an inductive approach is employed to compare different models and determine their respective performances. Prior to this, a linear regression model is developed to determine whether the linear model provides an adequate fit. Residual, scatter plots and assumption tests are checked whether assumptions are violated, if this is the case, other regression methods can be applied to these specific variables. Based on the best-fitting model and significance test, the appropriate model is selected.

Three regression models are used; Multiple regression (OLS regression), Ridge regression, and Lasso regression. These regression models are chosen based upon that each of these techniques offers distinct advantages in dealing with the violation of certain assumptions, thereby improving the robustness and reliability of the findings (Müller & Guido, 2017).

- OLS regression assumes no multicollinearity, making it susceptible to biased parameter estimates and unstable predictions when multicollinearity is present. Using both Lasso and Ridge regression allows for more effective identification and control of multicollinearity than using OLS regression alone.
- Heteroscedasticity, which manifests as unequal variances of errors across predictor variables, violates another assumption of OLS regression. This violation can result in inefficient coefficient estimates and inaccurate inference. Lasso and Ridge regression are robust techniques to handle heteroscedasticity.

### Phase 4: Demonstration

In this phase, an assessment of the outcome and impact of the developed model during the research process are analyzed. The developed model is demonstrated to relevant stakeholders. The model is tested on cases to assess the performance of the model. The dataset is divided into training and testing sets. The training set is used to train the model, while the testing set is employed to evaluate its performance. This data splitting enables the assessment of the model's generalization to new data and helps detect and resolve overfitting issues. Typically, an 80% training and 20% testing split are employed, as done in this study.

The model's predictive performance is analyzed through two tests (MAPE, (adjusted) R squared). MAPE measures prediction accuracy as a percentage of actual values. R squared indicates the proportion of variance in the outcome explained by independent variables, with higher values indicating a better explanation. While there is no standardized way to assess the performance of cost estimation methods the MAPE and R squared are commonly used as performance measures (He et al., 2021). Following this, the meta-requirements set up in the design phases are checked to see if they fit the developed model.

### **Phase 5: Validation**

In the context of this thesis, the validation phase plays a crucial role in the research cycle as it allows for a comprehensive analysis of the data and validation of the developed model. To gather valuable insights and perspectives, the researcher conducts two interviews with key stakeholders, specifically the supervisors of the project to address challenges and opportunities. During these interviews, the developed model is presented, and its performance is thoroughly discussed, see Appendix H. The following questions are asked to delve deeper into the challenges and opportunities associated with the model:

1. Does the current model fit the requirements stated in the previous phase? If not, where does the model fall short?
2. How can the performance of the model be improved?
3. What are the challenges when developing such a model? How can these challenges be mitigated or overcome?
4. What are other opportunities that might have been overlooked during this research?
5. What are some lessons learned during this research?

Furthermore, an additional discussion session is held in which the preliminary results are presented and discussed with six employees in which challenges and opportunities are addressed. These employees may offer unique insights, alternative ideas, or critical observations that were not considered during the project with the project supervisors. The employees' participation allows for a broader and more inclusive discussion, enabling the exploration of different angles and potential blind spots. First, a presentation of the preliminary findings is given, after which the discussion is held to reflect on the decisions made and the outcome of the model. This discussion is recorded, transcribed, and coded, see Appendix I.

## **3.2 ETHICS**

The data that is gathered through IT systems or interviews are confidential within Company X. The researcher has signed an NDA (non-disclosure agreement) with Company X to ensure data protection. The confidential portions of the thesis are protected to prevent any breaches of confidentiality in the published version. Consequently, quotes are not used in this research due to confidentiality reasons, as discussed with the company supervisors of this research project. Furthermore, data is stored on the system of Company X and not on a personal computer ensuring safe storage. Prior to analysis on certain software or computers, data is anonymized to ensure complete confidentiality. Furthermore, the ethics committee of the University of Twente has evaluated and approved all of the researcher's measures for handling the provided data safely and other ethical aspects of the study.



## 4. RESULTS

This chapter presents the research results. The results consist of the design & development phase, demonstration, and validation phase. The design & development is described in Figure 7 of the methodology chapter. The design part outlines the model requirements and data-gathering process, ending with the final dataset. The development stage involves selecting the model and selecting features, and developing the model. Consequently, the model is demonstrated and validated.

### 4.1 DESIGN

#### 4.1.1 Interviews tender department

Table 9 presents the outcome of the interviews with four employees of the tender department, this table is similar to the schema of Appendix D of the interviews. Added to Table 9 is the column by whom the outcome was supported, showing a consensus of the outcome in the tender department.

Table 9: Tender department Interview summary

| Category                                  | Supported by  | Outcome  |
|---|---|--|
| 1: Processes applied                      | Differs per employee, but the process is relatively the same              | Mainly standard procedure processes are used. Some small deviations.   |
| 2: Important variables                    | All mentioned/ acknowledged these variables as being important            | Capex, collaborating disciplines, lead time, number of employees, project phase, market, type of work  |
| 3: Defensible estimation (explainability) | All   | Differs per proposal, but generally needs to be defensible. For smaller proposals of less importance. Also has to do with signing from management                                  |
| 4: Variables on a time limit              | All   | Same variables as in question 2.   |
| 5: Deterministic or Probabilistic model   | All   | The model should provide probabilities about its outcome (one output is not desired)   |
| 6: Explainable model / other requirements | All.<br>Some preferred cost as output and some preferred hours as output. | The model does need to provide some explanation, otherwise, we will probably not accept and use the model. (in line with question 5). Hours estimation preferable (per discipline) |

Table 9 provides insights from interviews conducted, revealing a consensus on the processes applied according to BPMN models from Appendix A. Additionally, the variables mentioned by the tender managers show significant overlap, with Capex, disciplines, lead-time, number of employees, project phase, and market being recognized as important for proposal creation and man-hour estimation. While not all tender managers mentioned every variable, when presented with the variables mentioned by their peers, they also acknowledged their impact on man-hour estimation. Moreover, these variables could be assessed the fastest when operating under time constraints. Based upon prior internal research, however, the following variable was also deemed important, this was; Project Manager Experience.

The significance of man-hour estimation's defensibility was found to vary based on the size of the proposal during the discussion. Larger proposals require a greater emphasis on explaining how man-hours were determined to ensure defensibility. In particular, a form of probability output is favored as it provides a greater opportunity to defend the outcome.

#### 4.1.2 Requirements

The meta-requirements are an extension based on the interviews of the previous phase 4.1.1, which is shown in Table 9, and on the literature reviewed in chapter 2. This is summarized in Table 10.

##### Meta-Requirement 1 Data:

- Meta-Design: The model should be based on the variables presented in Table 9, which were derived from the interviews and represent the most important variables in the preliminary stage of proposal making. This is based on the use of the model in the preliminary stage, allowing the tender department to quickly assess and generate man-hour estimations without the need for consultation of an engineer. It is supported by the impact of these variables on costs for engineering companies as identified in the literature review
- Kernel Theory: The variables that were identified in chapter 2.5 are used as a basis for this meta-requirement since these variables are predictor variables for estimating the cost of proposals. These variables are based on the papers of (Hyari et al., 2016; Matel et al., 2022; Shoar et al., 2022).

##### Meta-Requirement 2 Quality:

- Meta-Design: The model should achieve the highest possible performance (measured by terms of R-squared and MAPE), which is influenced by the quantity and quality of data used during the training phase. Therefore, it is important to gather a comprehensive dataset based on the identified variables to ensure optimal model performance. The performance measurements (MAPE and R-squared) are based on similar measurements used in prior studies (Cheng et al., 2010; Hyari et al., 2016; Matel et al., 2022; Badra et al., 2020; Sonmez & Ontepeli, 2009; Lowe et al., 2006)
- Kernel Theory: Practical application requires a minimum level of performance, as established by the classification model of AACE (Christensen & Dysert, 2005). This classification model, widely adopted in the construction industry, provides a theoretical foundation for ensuring the model's performance meets the desired standards. For a better understanding of this classification, see Appendix C.

##### Meta-Requirement 3 Process:

- Meta-Design: The outcome of the model needs to be explainable so that the model can be defensible. Users and stakeholders need to be able (to an extent) to assess the model's assumptions and limitations. Most proposals need to be explainable to stakeholders, if the information presented in the proposal cannot be explained, stakeholders might be less inclined to approve/ accept the proposal. The explainability requirement in this context is operationalized through the use of a confidence interval (e.g. 80%, or 90%) to provide insight for the users to get an estimate of the certainty of the prediction. Furthermore, the magnitude of the variables (ratio-wise) on the estimation of the total hours needs to be interpretable.
- Kernel Theory: Several studies highlighted the significance of explainability in practical model implementation (Elmousalami, 2021; Tayefeh Hashemi et al., 2020). By providing a scientific justification and adequately describing the technical processes underlying the obtained results, transparency, maintainability, and user confidence are improved in cost estimation methods. Incorporating these factors in the design enhances the practical application of the model (Elfaki et al., 2014).

Table 10: Meta-requirements, meta-design, and kernel theory

| Meta-Requirement | Meta-Design                             | Kernel Theory / Supporting source  |
|------------------|---|--|
| MR1: Data        | Input variables                         | Variables identified in chapter 2.5 (Matel et al., 2022; Shoar et al., 2022), also see Table 7 & Table 9 |
| MR2: Quality     | Performance-oriented model              | AACE framework (Christensen & Dysert, 2005), see Appendix C  |
| MR3: Process     | Explainable model, Confidence intervals | Model explainability (Elmousalami, 2021; Tayefeh Hashemi et al., 2020; Wijnhoven, 2022)                  |

Furthermore, an extension on the data (variables input) requirement, the decision is made to use hours as the dependent variable, based upon the interviews. As also noted in the introduction, Cost estimation in engineering companies is closely tied to man-hours estimation because the total cost of a project is directly proportional to the number of hours required to complete it. The total amount of hours is multiplied by the (average) hourly rate, which results in the total costs.

The choice for man-hour estimation is driven by the influence of multiple factors on costs, including fluctuations in the hourly rate, which itself is subject to inflation corrections depending on the year of project execution. Additionally, the strategic decision to outsource certain tasks to countries with lower hourly wage rates can affect the hourly rate. By focusing on hours as a dependent variable, the model avoids potential uncertainties associated with hourly rates and thus costs. It eliminates fluctuations caused by varying hourly rates and strategic outsourcing decisions, resulting in more reliable project expense estimates. This approach aligns with the specific needs and characteristics of the engineering sector, enhancing the accuracy and practicality of cost estimations in this context.

**4.1.3 Impact work processes**

When developing such a model, the impact this has on work processes is investigated. Conducting the interview before developing the model helps gather valuable insights and perspectives, assess the impact on work processes, address potential challenges, and ensure alignment with the current roles and responsibilities within the organization.

Results interview:

The utilization of a data-driven model for man-hour estimation holds significant implications, particularly for engineers. Presently, engineers engage in a bottom-up approach, providing detailed estimates for project work. [REDACTED]

[REDACTED]. The bottom-up approach allows to consider for various factors, such as specific tasks, technical requirements, and potential challenges, which are not captured by the model's broad overview. [REDACTED]

This cultural aspect necessitates a shift from the bottom-up approach to a top-down methodology. [REDACTED]

[REDACTED]. Additionally, the adoption of such a model lessens tender managers' dependency on engineers, allowing for faster estimation of hours.

The implementation of the data-driven model follows a top-down approach, with the tender manager and project manager assuming responsibility for estimating the man-hours. Before incorporating this

model, an external validation process should be conducted, comparing the model's estimates to current estimation methods. This validation process provides insights into the accuracy of the model's hour estimation. The interview suggested that involving engineers during the initial phase of implementing a data-driven model might enhance their acceptance of the model. Demonstrating the model's strengths, explaining its limitations, and emphasizing the engineers' continued involvement in refining and validating the estimations can help alleviate discrepancies between the current work process and data-based estimation.

The change in the BPMN model is visualized in Figure 8, the prior version is Figure 1. Regarding the proposal process, the adoption of this approach brings about specific alterations. [REDACTED]

The proposal process undergoes several notable changes as follows, see the new estimation method for man-hours in the blue square. The estimation method [REDACTED] has been replaced with an approach based on man-hour estimation using a data model. This new estimation method does not require consulting engineers. The option to prepare man-hour estimation by activity will remain available, as per the feedback from the interview, it was mentioned that certain projects may necessitate this approach. When it comes to the alignment between disciplines, engineers now play a consultative role, providing input after the completion of man-hour estimation.



*Figure 8: Adjusted BPMN proposal process*

If a data-based estimation method would be used independently, the engineers' involvement in the proposal process is significantly reduced if we exclusively rely on a data-based estimation technique. For the proposal process, the engineers are not involved if the databased method would be used independently, and the Tender Team carries all the responsibilities regarding the man-hour estimation. Previously, the engineer(s) and tender team both carried the responsibility, for which the engineer provided input for the scope baseline per discipline. Furthermore, the man-hour estimation would also be carried out by the engineer, which in turn would be communicated with the tender team.

**4.1.4 Data gathering**

Data is gathered based on the mentioned variables in Table 9 of the interviews and Table 7 in the literature review. IT provided a printout of all completed and archived projects, as well as all won proposals. In this print out the following variables are found:

Table 11: Variables overview from the dataset

| Variable                      | Available in dataset | Quantity of data  | Quality of data |
|-------------------------------|----------------------|-------------------|-----------------|
| 1. Hours spend                | Projects             | Complete          | Correct         |
| 2. Number of disciplines      | Projects             | Complete          | Correct         |
| 3. Number of employees        | Projects             | Complete          | Correct         |
| 4. Lead time                  | Projects             | Complete          | Correct         |
| 5. Total investment (CAPEX)   | Proposals            | Mainly incomplete | Unreliable      |
| 6. Project type               | Projects             | Incomplete        | -               |
| 7. Contract type              | Proposals            | Complete          | Incorrect       |
| 8. Project phase              | Project & proposals  | Mainly incomplete | Correct         |
| 9. Project manager experience | Projects             | Incomplete        | -               |
| 10. Market                    | Project & proposals  | Complete          | Correct         |

This dataset contained a total of [REDACTED] projects, dating back to [REDACTED] till [REDACTED]. Nevertheless, a substantial amount of this data was deemed irrelevant since projects [REDACTED] fall outside the scope of the tender department, [REDACTED] projects remained after this. Furthermore, some previously defined variables were not, or only limited present in the data which resulted in a huge reduction of the data size.

In the IT system database, a total of 10 variables were identified. However, for the remaining variables, efforts were made to collect data by consulting different employees within Company X, but no further information was obtainable. Out of the 10 variables, only five variables were found to be complete, and correct. Incorrect variables mean that the data does not represent the true definition of the variable as stated in the literature review, chapter 2,4.

The incomplete and incorrect variables were attempted to be collected from other systems or databases. Only for the project phase it was possible to obtain some data, this however decreased the dataset significantly. On top of that, the [REDACTED] variable was also very limited available, and not very reliable. It was necessary to evaluate the [REDACTED] variable using a set of guidelines (set up by the [REDACTED]) to determine its accuracy, which further reduced the dataset. After discussing the dataset with the proposal manager, it was decided to apply scaling to the variable [REDACTED]. This decision was based on two factors: a) limited data reliability and b) the variable's inherent variability, as [REDACTED] are not always precise and can vary significantly.


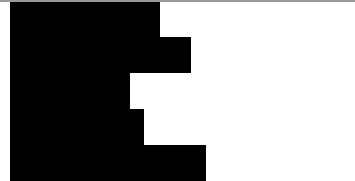
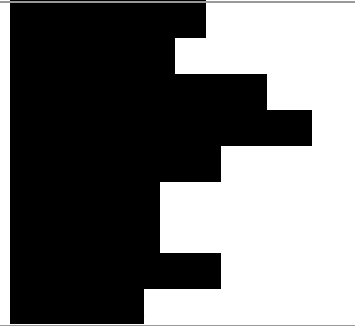
Regarding the "employees count" variable, a specific criterion was employed. It was determined that any individual who contributed to a project, regardless of whether their involvement was as short as one hour or as extensive as [REDACTED] hours, should be considered an employee. [REDACTED]  
[REDACTED].

Upon consultation with the tender management department, an adjustment was recommended. It was suggested that a minimum threshold of [REDACTED] hours worked per employee on a project be implemented to ensure more meaningful inclusion of individuals as project team members.

A total of 71 projects were included in the final dataset, out of which all variables were available except for project manager experience, contract type, and project type. This resulted in the following input

data, given in Table 12. Note, that the independent variable, the hours spent is not in this table. For the total investment variable, an interval scale is applied since there was a lot of uncertainty about the quality of the data.

Table 12: Input variables metrics

| No. | Variable              | Type of variable | Definition Scale   |
|-----|-----------------------|------------------|--|
| 1   | Number of disciplines | Ratio            | Positive real number   |
| 2   | Number of employees   | Ratio            | Positive real number   |
| 3   | Lead time             | Ratio            | Positive real number   |
| 4   | Total investment      | Interval         |    |
| 5   | Project phase         | Ordinal          |   |
| 6   | Market                | Nominal          |  |

In Figure 9 the distribution of the dependent variable, man-hours is presented per bin size of the 71 projects. The bin size ranges from 500 to >10000 hours. The bin size refers to the width of the intervals or bins used for the man-hours variable. This shows that a majority of the projects in the dataset are in the range of 1000-5000 hours, a total of 34 projects, thus more than half of the dataset. Only five projects total have more than 5000 hours.

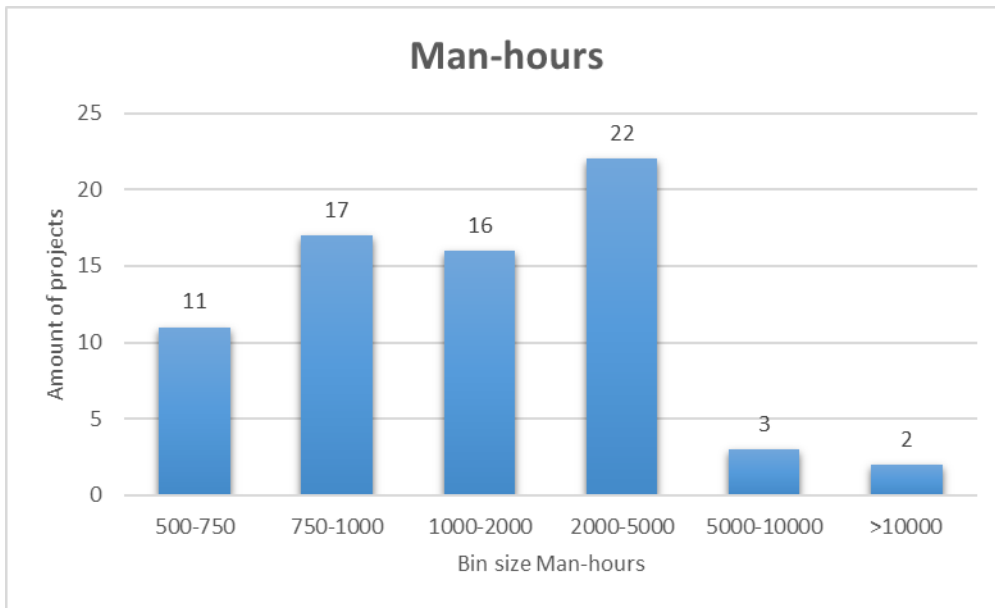


Figure 9: Man-hour distribution per bin size of the 71 projects

Having a sample that closely matches the population regarding the predictor variable (man-hours) is important for obtaining a representative sample, improving statistical power and precision, enabling robust subgroup analysis, and enhancing the stability and generalizability of the model's findings (Müller & Guido, 2017). In Figure 10 below, the sample size (final dataset size: 71) and the population (total available projects in the dataset: █████) are relatively comparable, with the largest difference observed in the bin size of 2000-5000, which has a 12% variation.

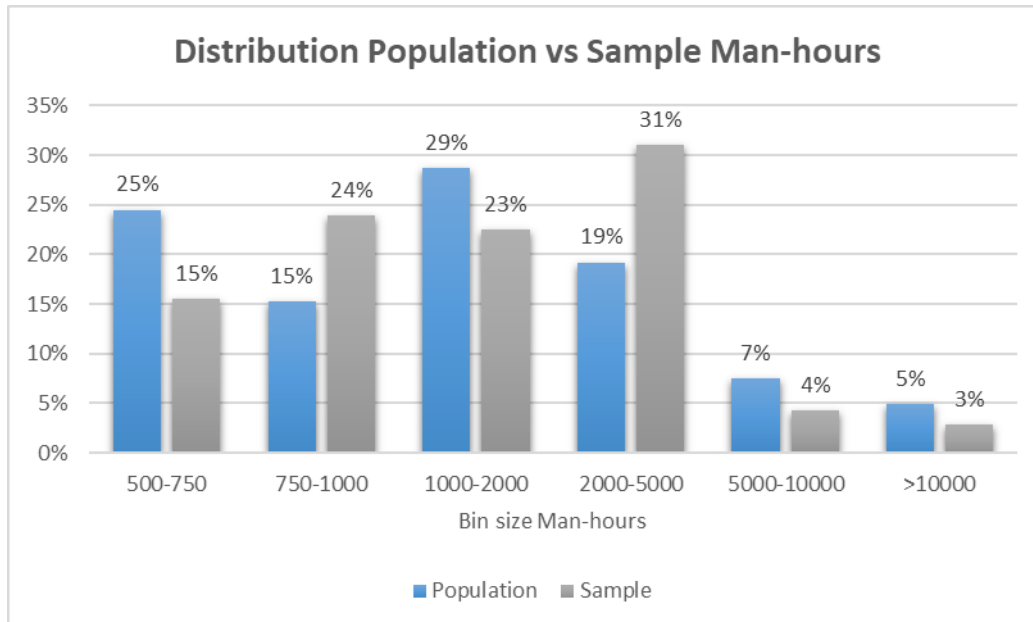


Figure 10: Distribution man-hours population (71) vs sample (█████)

## 4.2 DEVELOPMENT

In Python, the scripts are written for the development of the models. First, the linear relationship of the dependent variable (man-hours) is shown per independent variable, for an overview of the R squared (correlation) for the simple linear relationship of all the variables see Table 13. The scatterplots

are presented in Figure 11 and the average sum hours per phase or market is displayed for the market and project phase variables. The F statistic is used to assess the significance of a regression model as a whole, with a commonly used cutoff point of 0.05 to determine statistical significance.

Table 13: Linear regression of variables related to man-hours: R squared & F statistic

| Variables         | R squared | F statistic |
|-------------------|-----------|-------------|
| Count Employees   | 0.82      | < 0.001     |
| Count Disciplines | 0.44      | < 0.001     |
| Total Investment  | 0.01      | 0.40        |
| Lead time (weeks) | 0.20      | < 0.001     |
| Market            | 0.17      | 0.16        |
| Project phase     | 0.46      | < 0.001     |

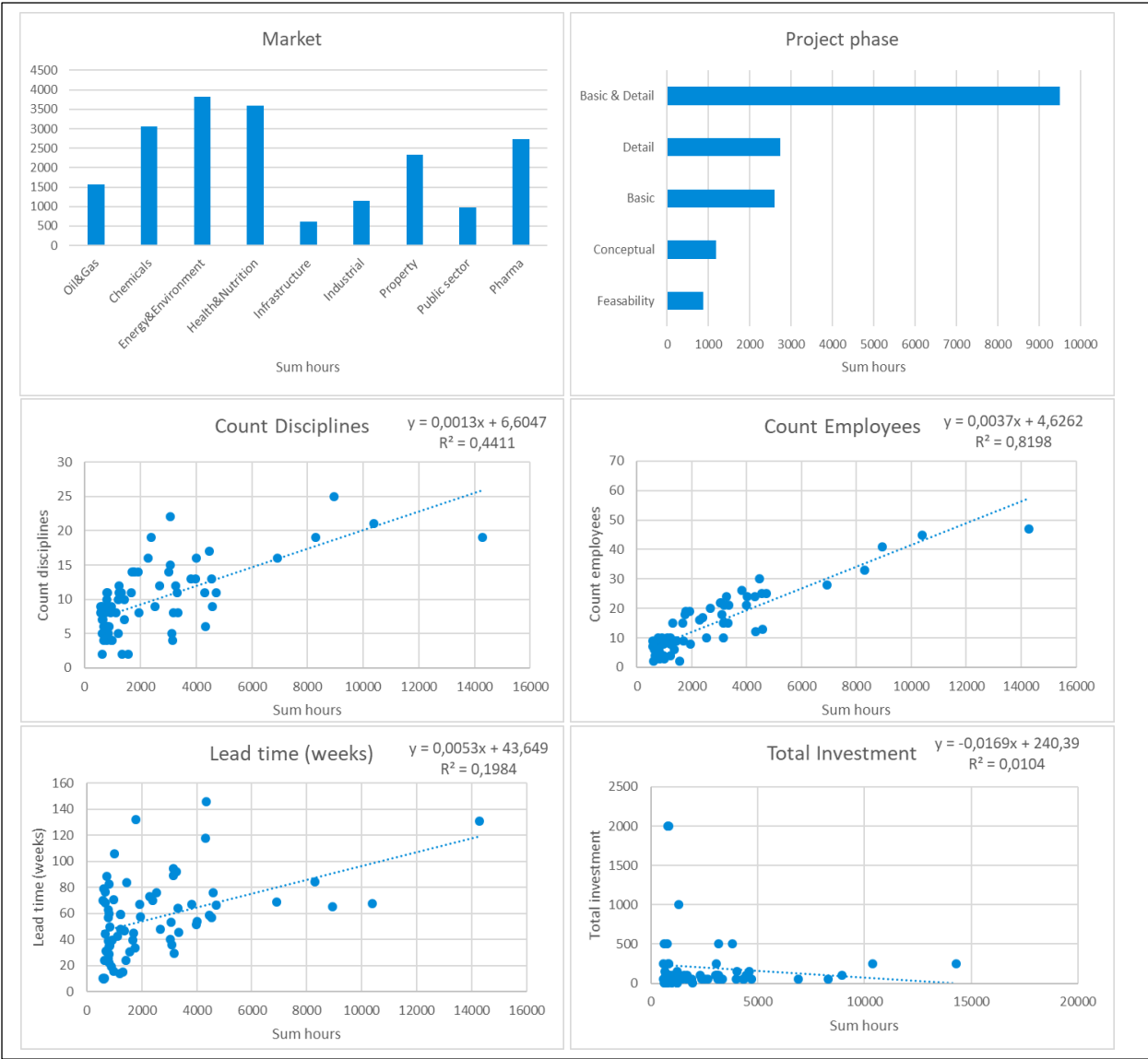


Figure 11: Distributions of variables related to man-hours

As shown in Figure 11, not every variable shows a linear relationship, or only partly, however, they do explain a proportion, see Table 13. Based upon this analysis, the Total investment and Market variables are not significant (F statistic). The inclusion of additional variables in a multiple regression model can alter the significance of a variable that was not found to be significant in a simple linear regression. It



is important to carefully interpret the results of the multiple regression model, considering the joint effects of all variables included in the analysis, thus total investment and market are used in the multiple regression model to check possible mediating effects.

The first step for the development of the model is applying a regression model containing all the variables from Table 13. In Appendix E 1 the output of this regression model is shown. The R squared of this model is 0.83, and adjusted R squared is 0.76. While R-squared simply quantifies the proportion of variance explained by the model, adjusted R-squared penalizes the addition of unnecessary predictors, providing a more accurate reflection of the model's goodness of fit. By considering both the model's explanatory power and the complexity introduced by predictors, adjusted R-squared offers a more reliable assessment of the model's generalizability and helps prevent inflated performance estimates when incorporating more variables into the model.

This output shows that the disciplines and employees are only significant, in which disciplines are negatively correlated and employees positively correlated. If the assumptions are checked, it is visible that the following assumptions are violated:

- Multicollinearity: The number of disciplines and employees are highly correlated and have a VIF value  $>5.0$  and a correlation of 0.82.
- Homoscedasticity: The Jarque-Bera test has a p-value of  $<0.05$ .
- Normality: The Breusch-Pagan test has a p-value  $<0.05$ .

As outlined in the methodology, a Python script was employed to identify the optimal combination of variables by maximizing the R squared value, while simultaneously ensuring the inclusion of only significant variables (Appendix F). It is worth mentioning that the market and total investment variables were not considered in this automated selection process, given their lack of significance in both the multiple regression model and simple linear regression analysis, see Appendix E.1.

As mentioned, three different regression methods are used for the analysis: Multiple OLS regression, Lasso regression, and Ridge regression. Ridge and Lasso regression techniques are employed alongside OLS regression due to their distinct advantages. Ridge regression is particularly effective in handling multicollinearity, which is the presence of high correlation among predictor variables. In the presence of multicollinearity, ordinary least squares (OLS) regression can produce unreliable and unstable estimates of the coefficients. Ridge regression adds a penalty term that reduces the impact of collinear variables, making the model more robust. Furthermore, Ridge and Lasso regression allows for controlling the bias-variance trade-off in the model. By introducing a penalty term, these techniques reduce the variance of the model at the cost of slightly increasing its bias. This trade-off can be adjusted by tuning a hyper parameter called the regularization parameter ( $\lambda$  or  $\alpha$ ). Higher values of  $\lambda$  or  $\alpha$  increase the regularization strength, leading to more shrinkage of coefficients and increased bias but reduced variance (Müller & Guido, 2017).

Additionally, Ridge and Lasso regression can improve the stability of the model. OLS regression can be sensitive to outliers or small changes in the dataset, leading to large variations in the estimated coefficients. Ridge and Lasso regression introduces regularization, which helps stabilize the model by reducing the impact of outliers and small fluctuations in the data, through which Ridge and Lasso regression could outperform OLS regression in terms of prediction accuracy. Regularization techniques like Ridge and Lasso can prevent overfitting and generalize better to unseen data. By controlling the bias-variance trade-off, these techniques can often lead to more robust and accurate predictions (Müller & Guido, 2017).

Applying Ridge regression, Lasso regression, and OLS regression together offers flexibility, enables comparison, addresses different data characteristics, and enhances robustness, leading to a more comprehensive analysis and validation of results.

The next step consists of developing three regression models, OLS regression (model 1), Lasso regression (model 2), and Ridge regression (model 3). In all these models, the following variables are used which were shown to have statistical significance based on prior analysis, for the Project phase a one-hot encoding is used:

1. Lead time (weeks)
2. Count disciplines
3. Count employees
4. Project phases (P1 to P5)

Table 14: OLS regression significant variables

| Model 1               | Performance |               |       | Assumptions       |           |                  |
|-----------------------|-------------|---------------|-------|-------------------|-----------|------------------|
|                       | R-squared   | Adj R squared | MAPE  | Multicollinearity | Normality | Homoscedasticity |
| <b>OLS regression</b> | 0.809       | 0.791         | 43.8% | VIF of 12 and 14  | <0.05     | <0.05            |

Table 15: Lasso regression significant variables

| Model 2                 | Performance |               |       | Assumptions                         |           |                  |
|-------------------------|-------------|---------------|-------|-------------------------------------|-----------|------------------|
|                         | R-squared   | Adj R squared | MAPE  | Multicollinearity                   | Normality | Homoscedasticity |
| <b>Lasso regression</b> | 0.829       | 0.760         | 38.9% | Less applicable in Lasso regression | >0.05     | <0.05            |

Table 16: Ridge regression significant variables

| Model 3                 | Performance |               |       | Assumptions                         |           |                  |
|-------------------------|-------------|---------------|-------|-------------------------------------|-----------|------------------|
|                         | R-squared   | Adj R squared | MAPE  | Multicollinearity                   | Normality | Homoscedasticity |
| <b>Ridge regression</b> | 0.824       | 0.746         | 39.4% | Less applicable in Ridge regression | <0.05     | >0.05            |

Based on this analysis, all regression models have relatively similar performances, as shown in Table 14, Table 15, and Table 16. For the output of this model see Appendix E.2. The following variables are significant:

1. Lead time (weeks)
2. Count Employees
3. P1 (project phase 1)
4. P5 (project phase 5)

In Figure 12 the OLS Regression result of the model is shown. In this figure, the coef (coefficient) column shows the estimated regression slopes, indicating the change in the dependent variable (Sum Hours) when the independent variable changes by one unit. The p-values (p) column indicates the probability of observing such coefficients under the assumption of the null hypothesis. A low p-value suggests a statistically significant impact on the independent variable (an alpha of 0.05 is used in this research). The t-values (t) column represents the magnitude of the coefficient relative to its variability, with higher absolute t-values indicating a more significant impact on the dependent variable. The output of the OLS regression is shown below in Figure 12 based on model 1 of Table 14.

| OLS Regression Results |               |                     |          |       |          |          |
|------------------------|---------------|---------------------|----------|-------|----------|----------|
| Dep. Variable:         | Sum Hours     | R-squared:          | 0.809    |       |          |          |
| Model:                 | OLS           | Adj. R-squared:     | 0.791    |       |          |          |
| Method:                | Least Squares | F-statistic:        | 45.14    |       |          |          |
|                        |               | Prob (F-statistic): | 3.76e-21 |       |          |          |
| Variables              | coef          | std err             | t        | P> t  | [0.025   | 0.975]   |
| const                  | -895.4199     | 395.062             | -2.26    | 0.027 | -1684.64 | -106.192 |
| Lead time (weeks)      | 12.1823       | 5.237               | 2.326    | 0.023 | 1.720    | 22.645   |
| Count Employees        | 120.7125      | 18.599              | 6.490    | 0.000 | 83.556   | 157.869  |
| Count disciplines      | -45.0970      | 53.332              | -0.84    | 0.401 | -151.639 | 61.445   |
| P1                     | 5.261e-12     | 1.4e-12             | 3.764    | 0.000 | 2.47e-12 | 8.05e-12 |
| P3                     | -201.2845     | 371.932             | -0.54    | 0.590 | -944.303 | 541.734  |
| P4                     | 327.4776      | 371.764             | 0.881    | 0.382 | -415.207 | 1070.162 |
| P5                     | 3555.1327     | 830.036             | 4.283    | 0.000 | 1896.945 | 5213.321 |

Figure 12: OLS regression model 1

Note that P2 is excluded, this is done to avoid multicollinearity for the project phase variable. Including all categories as separate variables in the regression model would create a linear dependency among them.

However, assumptions are violated, primarily in the OLS regression (model 1). Multicollinearity is one of the main assumptions that is violated, count disciplines and count employees cannot be used simultaneously in the model, while being both important variables.

Because multicollinearity is present (primarily in the OLS regression (model 1)), an interaction variable is created, which can be particularly useful in OLS regression because it allows for the direct interpretation of the coefficients associated with the interaction terms. The interaction variable "Total Resources" is formed by the multiplication of the count of employees and the count of disciplines. While as mentioned Ridge and Lasso regression are better able to deal with multicollinearity, Ridge and Lasso regression are also redeveloped with the new interaction variable to compare the performances of the three different regression methods.

After creating this new interaction variable, standardization is applied again to all the variables. This is again done for the three regression methods, OLS regression (model 4), Lasso regression (model 5), and Ridge regression (model 6). For an overview of the output of the OLS regression model see appendix E.3. For a general overview of the outcome, see Table 17, Table 18, and Table 19:

Table 17: OLS regression with total resources variable

| Model 4               | Performance |               |       | Assumptions       |           |                  |
|-----------------------|-------------|---------------|-------|-------------------|-----------|------------------|
|                       | R-squared   | Adj R squared | MAPE  | Multicollinearity | Normality | Homoscedasticity |
| <b>OLS regression</b> | 0.820       | 0.809         | 44.7% | Not violated      | >0.05     | <0.05            |

Table 18: Lasso regression with total resources variable

| Model 5                 | Performance |               |       | Assumptions                         |           |                  |
|-------------------------|-------------|---------------|-------|-------------------------------------|-----------|------------------|
|                         | R-squared   | Adj R squared | MAPE  | Multicollinearity                   | Normality | Homoscedasticity |
| <b>Lasso regression</b> | 0.762       | 0.626         | 44.5% | Less applicable in Lasso regression | >0.05     | >0.05            |

Table 19: Ridge regression with total resources variable

| Model 6                 | Performance |               |       | Assumptions       |           |                  |
|-------------------------|-------------|---------------|-------|-------------------|-----------|------------------|
|                         | R-squared   | Adj R squared | MAPE  | Multicollinearity | Normality | Homoscedasticity |
| <b>Ridge regression</b> | 0.734       | 0.626         | 44.9% | Not violated      | >0.05     | <0.05            |

The best-performing model in terms of the Adjusted R-squared is shown below in Figure 13. This is an extension of model 4 presented in Table 17.

| OLS Regression Results |               |                     |          |       |          |          |
|------------------------|---------------|---------------------|----------|-------|----------|----------|
| Dep. Variable:         | Sum Hours     | R-squared:          | 0.820    |       |          |          |
| Model:                 | OLS           | Adj. R-squared:     | 0.807    |       |          |          |
| Method:                | Least Squares | F-statistic:        | 59.40    |       |          |          |
|                        |               | Prob (F-statistic): | 6.38e-23 |       |          |          |
| Variables              | coef          | std err             | t        | P> t  | [0.025   | 0.975]   |
| const                  | -157.6796     | 311.182             | -0.507   | 0.614 | -779.152 | 463.793  |
| Lead time (weeks)      | 14.2069       | 4.950               | 2.870    | 0.006 | 4.322    | 24.092   |
| Total Resources        | 5.0000        | 0.476               | 10.496   | 0.000 | 4.049    | 5.951    |
| P1                     | 2.958e-13     | 1.62e-13            | 1.821    | 0.073 | -2.8e-14 | 6.2e-13  |
| P3                     | -99.6178      | 349.668             | -0.285   | 0.777 | -797.953 | 598.717  |
| P4                     | 515.0354      | 353.364             | 1.458    | 0.150 | -190.681 | 1220.752 |
| P5                     | 3345.1645     | 803.688             | 4.162    | 0.000 | 1740.08  | 4950.240 |

Figure 13: OLS regression model 4 (Total resources variable)

### 4.3 DEMONSTRATION

Among the six regression analyses conducted, the OLS regression model incorporating the interaction variable “Total resources” achieved the highest (adjusted) R-squared of 0.807. This implies that approximately 80% of the variance in the Sum Hours can be accounted for by the included variables, namely Total resources, lead-time, and project phases. See Equation 2 for the formula of the OLS regression model. The presence of unexplained variance in the model suggests the potential influence

of other variables or factors that were not accounted for or the limitations of the available data in fully explaining the variance.

The best performing Ridge regression model uses count employees and disciplines both in the model. The same goes for Lasso regression, which has the highest performance when using employees and disciplines together in the model. Ridge and lasso regression are better able to deal with multicollinearity than OLS regression, nevertheless, the OLS regression model employing the variable “Total resources” has the highest performance in terms of R squared. All models are relatively comparable performances in terms of R-squared (approximately 80%) and MAPE (approximately 40%).

Equation 2: OLS Regression model formula

Formula:

$$y(i) = b1*x1(i) + b2*x2(i) + b3*x3(i) + b4*x4(i) + b5*x5(i) + b6*x6(i) + c$$

where:

- **y(i)** = Sum hours
  - **b1(i)** = 14.20           = Lead time                           x1 = Lead time
  - **b2(i)** = 5.0               = Total Resources                   x2 = Total Resources
  - **b3(i)** = 2.9e-13       = P1                                   x3 = P1
  - **b4(i)** = -99.62       = P3                                   x4 = P3
  - **b5(i)** = 515.0         = P4                                   x5 = P4
  - **b6(i)** = 3345.16     = P5                                   x6 = P5
- **c** = -157,68 (represents the intercept term)

Filled in formula:

$$y(i) = 14.20*x1(i) + 5.0*x2(i) + 2.9e-13*x3(i) + -99.62*x4(i) + 515.0*x5(i) + 3345.16*x6(i) + -157.7$$

A variable that according to the proposal manager and tender managers was deemed important was the project phase. Based on discussions with employees, the project phase should have significant differences in the total amount of work required. While there may be some distinctions between the phases, the primary disparity exists between the [redacted] and [redacted] phases, which constitute the major portion of Company X’s workload. With the current dataset, it could not be shown that there was a significant difference between all the project phases. See Figure 14 for the overview of the project phases according to the count of times the phases were present in the data (series 2).

Additionally, the market was also deemed important for predicting the man-hours mentioned during the interviews. With the current dataset, this could not be shown to have a significant impact on predicting the man-hours.

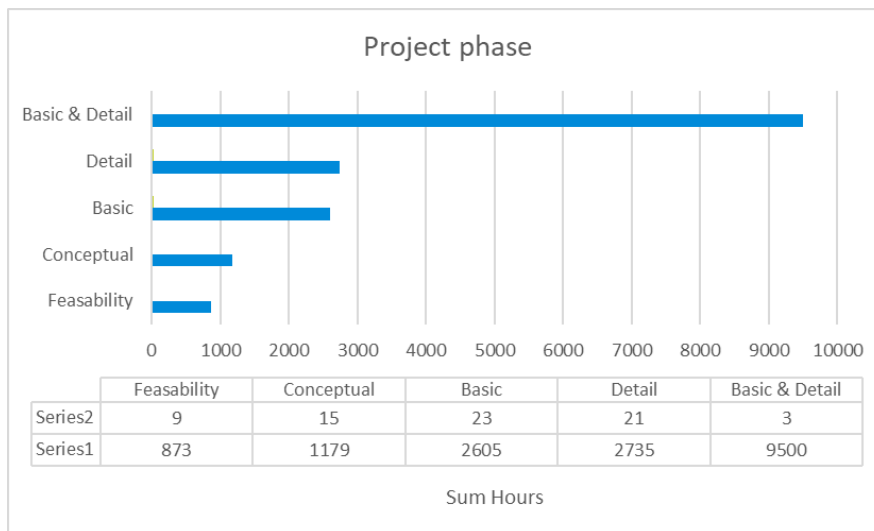


Figure 14: Project phase average man-hours distribution (series 1) with count projects (series 2)

Another limitation of the models is the presence of a consistently high Mean Absolute Percentage Error (MAPE) of approximately 40%, indicating a 40% error relative to the actual values. Moreover, the violation of the homoscedasticity assumption further underscores the need for caution when making predictions with the OLS regression model. As mentioned an 80%-20% train and test split is applied to the dataset, and to assess the model's performance, the test projects (20% of the dataset) are shown below with the model's output, as depicted in Table 20. This table shows the actual value, predicted value, and the resulting percentage error as well as the 80% confidence interval.

Table 20: Test cases of best-performing regression model

| Test case | Actual value (sum hours) | Predicted value (sum hours) | Percentage Error | 80% CI lower bound | 80% CI upper bound |
|-----------|--------------------------|-----------------------------|------------------|--------------------|--------------------|
| 1         | 3257                     | 5677                        | 74%              | 4947               | 6408               |
| 2         | 2534                     | 1803                        | -29%             | 1414               | 2192               |
| 3         | 735                      | 1675                        | 128%             | 1233               | 2116               |
| 4         | 4310                     | 3804                        | -12%             | 3348               | 4259               |
| 5         | 10381                    | 6871                        | -34%             | 4728               | 9013               |
| 6         | 8944                     | 6297                        | -30%             | 5141               | 7453               |
| 7         | 636                      | 79                          | -88%             | 0                  | 670                |
| 8         | 14292                    | 8077                        | -43%             | 5794               | 10360              |
| 9         | 788                      | 295                         | -63%             | 0                  | 801                |
| 10        | 3030                     | 3275                        | 8%               | 2618               | 3932               |
| 11        | 4465                     | 4221                        | -5%              | 3544               | 4898               |
| 12        | 1704                     | 1492                        | -12%             | 621                | 2362               |
| 13        | 572                      | 1699                        | 197%             | 1146               | 2252               |
| 14        | 670                      | 1756                        | 162%             | 1147               | 2365               |
| 15        | 1354                     | 2010                        | 48%              | 1205               | 2816               |

Other papers utilizing data to make cost estimations also employed MAPE as a performance measure. For an overview of these models see Table 21. The MAPE as well as the dataset size, variables used, and model used is depicted. This table shows that the current model underperforms in comparison with other papers.

Table 21: Comparison with earlier work

| Model used                  | Data points | Number of variables used | MAPE  | Source                    |
|-----------------------------|-------------|--------------------------|-------|---------------------------|
| Fuzzy hybrid neural network | 28          | 10                       | 10.4% | (Cheng et al., 2010)      |
| ANN                         | 224         | 5                        | 28.2% | (Hyari et al., 2016)      |
| ANN                         | 131         | 16                       | 13.7% | (Matel et al., 2022)      |
| S-curve regression          | 113         | 7                        | 25.2% | (Badra et al., 2020)      |
| Regression model            | 13          | 5                        | 35.2% | (Sonmez & Ontepeli, 2009) |
| Regression                  | 286         | 5                        | 19.3% | (Lowe et al., 2006)       |

Based on the comparison with earlier studies in Table 21, the developed model in the present work achieved a mean absolute percentage error (MAPE) of 40%. This MAPE value is higher compared to the other studies mentioned, which reported lower MAPE values ranging from 10.4% to 35.2%. This indicates that the developed model in the current study may have relatively higher prediction errors compared to the models used in the previous works. This is further elaborated on in the discussion of this research, chapter 5.

A model that accurately estimates the costs could not be established with the current regression models and dataset. Notably, the current predictions are not stable enough to be used in practice. This in turn means that the model does not yet fit the performance requirement stated in 4.1.2. There is thus no application made for internal use for the deployment of the model.

## 4.4 VALIDATION

With this current research, several challenges but also opportunities arose which are explored in this part through an interview with two stakeholders (project supervisors); see Appendix H, and through a discussion with six employees, presented in Appendix I. First, the outcome of both the interview as well as the discussion session is presented. Followed by this the challenges and opportunities are defined.

### 4.4.1 Interview & Discussion

- Interview with two stakeholders:

The interviews with two stakeholders of the project revealed that the current model does not fit the performance requirement yet when presented with the output of the model, this was especially the case for the larger and very small projects in terms of man-hours. Furthermore, when presented with the dataset, it was questioned whether the dataset on which the model was trained represented the population of all the projects. To enhance performance, the interviewee suggested gathering more data about the variables and implementing a standardized data collection method, potentially involving an obligatory data sheet after finishing a project. However, just stating that this data needed to be gathered was attempted before and this did not work. [REDACTED]

[REDACTED]. Furthermore, the underutilization of data from previous projects, in general, is emphasized. A significant portion of the

data needed for this model resides in the knowledge of employees however, since this is not captured much of this data is lost.

The challenges such as data quality and confidentiality were acknowledged, along with the time-consuming nature of manual data extraction. As also supported in the interview, a lot of time was spent during this research on gathering and cleaning the data, which was not foreseen as a major challenge at the initiation of this research. Much time was spent on intensive collaboration with IT before any data could be gathered. Other challenges persisted of data privacy and security, and unstandardized data due to which much data could not be used.

A possible overlooked opportunity is that the current model prioritized the establishment of variables within the control of tender managers. This model allows for a quick and convenient estimation of the total number of hours required. However, there is potential for improvement by developing specific relationships between different disciplines (such as electrical, piping, mechanical, etc.) and the corresponding man-hours. According to the interview, this new approach could lead to a more direct relationship with the total amount of hours, as an example, the total number of line lists for the process department could impact the total amount of hours by an x amount. While these variables are not easily estimable by tender managers, the utilization of such a model could enhance the efficiency of engineers conducting the estimate, in turn reducing the subjectivity of the engineers conducting the estimate.

To achieve this, it is crucial to gather data on the key characteristics of each discipline. Collaboration with engineers specializing in each discipline is essential to establish these relationships. Once the key characteristics are determined, relevant data can be collected.

- Discussion with six employees:

During the discussion session, the participants addressed various difficulties and challenges in developing a data-driven model for cost estimation in engineering companies. The participants discussed the performance of the current model, in which they identified limitations due to the data set, indicating that the current data might not accurately represent the project information due to a lack of standardization in the data gathering. Some variables were said to not truly represent the reality of how the project was conducted. For example, in the current dataset, no statistical difference can be established between certain project phases, which according to the experts should be present.

The model was also noted to use abstract variables, potentially missing out on capturing important nuances, through which possible risks are introduced. According to the discussion, in an industry in which many projects are unique and complex in their own way, some small aspects can have a great impact on the total amount of work we need to do, whether it is the quality of the information presented by the client, or some other key characteristics of the client or project. Capturing these nuances is important and can in some cases only be done by an expert's opinion, which is hard to capture in a model due to the many variables the model should otherwise contain.

However, they recognized that the model proves to be interesting and valuable, particularly for providing initial estimates of project hours, which could give a broad overview of the size of the project,

[REDACTED]

During the discussion session, stakeholders reiterated the potential of a data-driven model that incorporates key characteristics of main disciplines to estimate man-hours more effectively. In the



current model, the hours of all disciplines combined are estimated. Employees expressed their expectation that relationships exist between the number of hours and key characteristics from disciplines. However, as also mentioned again, the lack of supporting data poses a challenge. Nonetheless, implementing such a model would enable a more nuanced understanding by using more directly related variables related to a discipline. To develop this model, an expert in the field would be essential to establish these relationships initially. Subsequently, data can be gathered to construct the model in question.

Furthermore, to gain trust in the model a form of external validation process was proposed. Trust in the model would be important not only for management, but also for the users to know for what phase, or market the model might perform well. To gain some form of trust, the model should be used in parallel with the current estimation method to gain some insight into the performance.

[REDACTED] Currently, it is not known what the performance of the current method is, and neither is it easily possible to analyze this due to this data not being readily available. This in turn also makes it difficult to compare the developed model to the current used method. [REDACTED]

A consensus in both methods is proposed by the employees based on the discussion while emphasizing the need to gather data in a standardized way before such models and projects can be rendered feasible in practice. It was acknowledged that data and knowledge of previous projects is underutilized and should be improved. The participants emphasized the need to address these issues to develop a more effective and accurate data-driven model for cost estimation in the engineering companies' decision-making process.

#### 4.4.2 Challenges & Opportunities

**Challenges:** The challenges encountered during this research are described below. For an overview see Figure 15. The data challenge consists of data privacy and security, data gathering, data quantity, and data quality. Due to the industry's standards and contracts with clients, a lot of data is considered confidential which made some valuable data not available. Even anonymizing the data fully was not seen as sufficient, since in some cases Company X has signed an NDA with clients, rendering the utilization of such data unfeasible.

Additionally, during this research, a significant amount of time was spent on gathering the data, a complete overview of the dataset was not readily available to be explored, and it required intensive collaboration with IT to gather all relevant data. Furthermore, at the start of the research, an assumption with stakeholders was made that sufficient data was readily available, which was not the case, as noted in the interview as an important learning aspect for future projects. This further underscores the importance of effective data management concerning the internal feasibility of data-driven projects.

While several variables were set up based on the literature review, only a limited number of these variables could be established and filled in by Tender managers based on provided information in an RFQ. Based on these variables data were gathered, however, a lot of these variables were not present, or only limitedly available in the dataset. Moreover, the dataset contained variables of limited quality, particularly for Capex and the project phase. The absence of standardized guidelines for filling in project and proposal information resulted in unclearly defined data, leading to the removal of a significant portion of the dataset.

While data poses a significant barrier and challenge during the development of a cost estimation model, there are also social aspects that also need to be addressed. During the interview, the impact the model has on work processes was explored, in which some implementation constraints have been addressed. First, trust in the model should be established in which an external validation process was proposed. This consists of using the model in parallel with current work methods by the tender department and recording the output for comparison to gain insight into the performance and limitations of the model. The current approach towards man-hour estimation is mainly bottom-up in which engineers provide detailed estimates, with a utilization of the current model for DDDM this approach alters towards a top-down approach. The work processes are significantly impacted when independently using a data-driven model for man-hour estimations. This is a cultural shift since work processes regarding the proposals are affected, which has to be taken into consideration how to address this. Moreover, due to the changing industry, creating a model that can be updated and retrained was emphasized. Some industries are expected to become more important in the near future, of which not many projects are carried out yet.

Additionally, the current model has abstract input variables, which according to the discussion are not able to capture the nuance of the proposals. Nuances refer to the subtle and intricate details that can significantly affect project costs. Abstract input variables, which are generalizations of the project characteristics, may overlook critical factors that can influence costs. In this industry with many uncertain aspects in the project execution, whether it is the client's need or insufficient information provided, there is a need to mitigate these risks before bidding on a proposal which could prove to have a negative impact in the later stages of the project. In light of the constraints of both the expert-driven approach and the data-driven approach independently, a balanced model that combines the strengths of both approaches is needed.

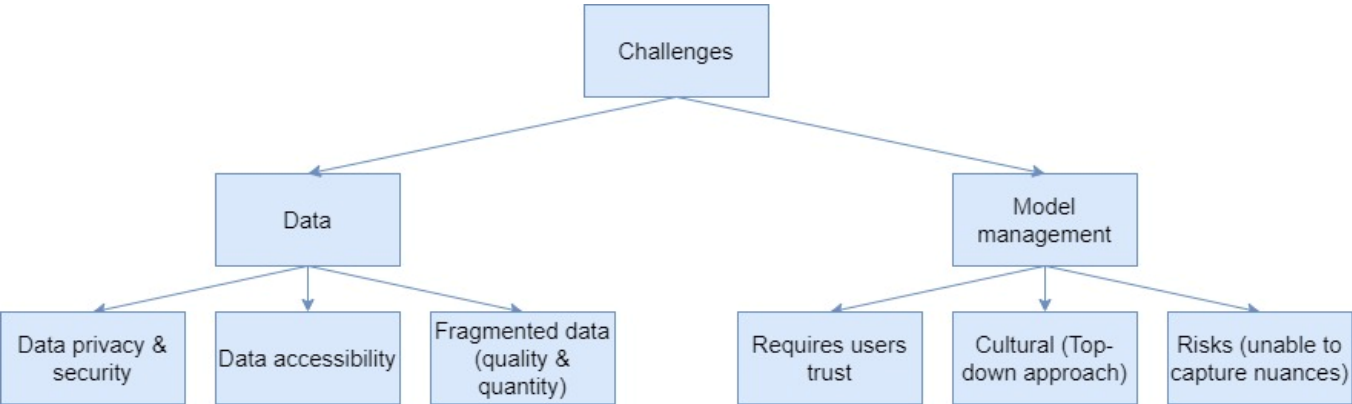


Figure 15: Tree diagram challenges overview

**Opportunities:** While several challenges were present, this is also the case for opportunities, for an overview see Figure 16. Although the currently developed model may not meet the required performance standards, there is potential for further enhancement by gathering additional data and variables. By doing so, the current model can be retrained on new data, or a new model can be created using a different ML method, expert system, or regression model, depending on the changes in the dataset and requirements. The data model can be (re)designed to align with the requirements. According to employees, this can help during the man-hour estimation to provide a quick estimation of the required resources.

During both the interview and discussion the mentioning of a system that can be updated and revised was also mentioned. Through this dynamic improvement, the model can be updated over time through

which it can learn from experience and adapt solutions based on feedback. Furthermore, a purely data-driven model is deemed unable to establish the nuances related to the man-hour estimation, since the current model utilizes abstract variables. For this reason, it was suggested to be more beneficial to adopt a model that incorporates expert input while also relying on data—a synergy between data and expertise. This approach combines the insights and knowledge of domain experts with the power of data analysis, creating a more comprehensive and robust model.

A lot of project information is known by employees, however, due to the lack of capturing this information, much of this valuable information is lost after a project has been finished. As also noted in the interview, much knowledge about projects is in employees' heads, which can be gathered through good data management practices. [REDACTED]

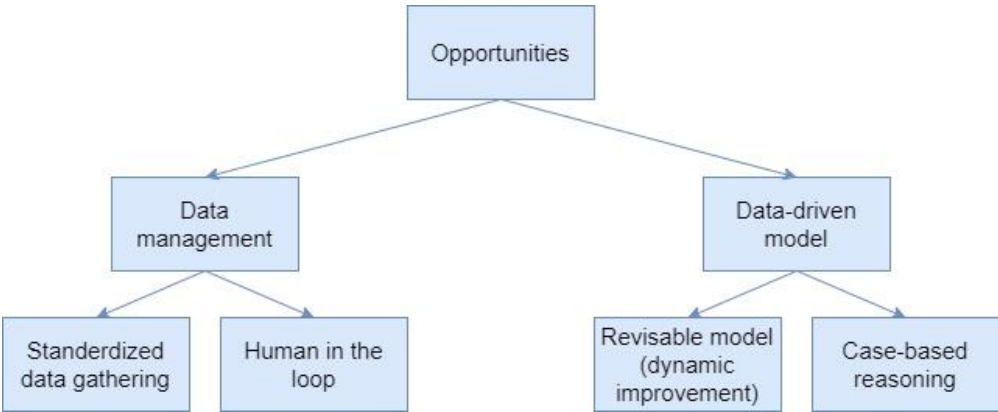


Figure 16: Tree diagram opportunities overview

## 5. DISCUSSION & CONCLUSION

This chapter discusses and reflects on the findings of the research. Following this, the implications, limitations, and future research directions are discussed. Finally, the conclusion is given.

### 5.1 DISCUSSION RESULTS

This study addresses the challenges faced by the Tender department in utilizing a data-driven cost estimation model for proposals in an engineering company. Traditional methods used for cost estimation result in an expert-driven approach which is a lengthy and detailed estimate subjective to the expertise and opinion of engineers which underutilizes data. As mentioned by (He et al., 2021) data-based cost estimation has established academic potential, however, the practical application lacks (Bilal et al., 2016; Elfaki et al., 2014). This research explores the challenges and opportunities for the development of a data-driven model for cost estimation in the decision-making process of an engineering company.

Many different models are present in the literature to develop a cost estimation method, however, due to the uncertainty involved in the preliminary stages of a proposal, a model capable of justifying the estimation was required. This is crucial because many models lack explainability and have limited capability to justify their predictions, making some unexplainable black-box AI models, such as artificial neural networks (ANNs), less practical for real-world use (Chou et al., 2009). Consequently, a regression model was chosen for development based on the available data, as regression models offer the advantage of explainability. A total of 16 variables were identified impacting the cost of engineering services. However, only six variables were present in the dataset on which the model was developed. Gathering data of good quality and sufficient quantity is one of the main challenges encountered during this research.

Furthermore, other studies focused on estimating the cost in total (Hyari et al., 2016; Matel et al., 2022), this study used man-hours as the dependent variable. Man-hours are used as the dependent variable for cost estimation due to their close relationship with project costs and their ability to mitigate uncertainties associated with fluctuations in hourly rates and strategic outsourcing decisions, thereby enhancing the reliability and practicality of expense estimates.

Given the availability of numerous regression methods, an inductive approach is employed to determine the best-performing model. Three distinct regression models are developed based on their fit to the data. The best-performing model, multiple OLS regression achieved an adjusted R-squared value of 0.807. However, the performance of the developed model, as measured by the Mean Absolute Percentage Error (MAPE), demonstrated lower results compared to similar studies. The achieved MAPE was approximately 40%. In comparison, other models such as Artificial Neural Networks (ANN), Case-Based Reasoning (CBR), and various regression models exhibited better performance in terms of MAPE.

This study used only available data from internal databases of Company X and did not extract data from external sources (outside research context), or through means of survey. This could be a factor that explains why in the current developed model a high MAPE was achieved, as the performance of the model is inherent to data quantity and quality. In contrast, other studies utilized external data or means of surveys for data gathering (Hyari et al., 2016; Matel et al., 2022).

Another possible explanation for the diminishing results of this study could be the file drawer effect, also called the publication bias. This term suggests that results not supporting the hypotheses, or have diminishing results, often go no further than the researchers' file drawers, leading to a bias in published research (Franco et al., 2014). When studies with diminishing results are not published, there is a

tendency for the literature to be skewed toward high-performing models. As a result, the overall perception of model performance may be inflated, creating a potentially misleading understanding of the effectiveness or generalizability of the models in practice (Franco et al., 2014).

During the validation of the model with several employees, the value and the potential practical use of the model were established, however, due to the instability of the current model the implementation of the model is constrained. Additionally, the current model uses four variables, which were said to not be able to capture all the nuances in the proposal phase, through which possible risks could be introduced. To capture these nuances, more variables need to be incorporated into the model according to the discussion. The data-driven estimation model could not be able to be used independently in the proposal phase, and would still require the consulting of experts (engineers) based on the currently established performance and used variables. According to the discussion due to the complex and uncertain nature in which cost estimations are made for proposals in engineering companies, there is a need for a balanced approach that incorporates data-driven models with human expertise.

### **5.1.1 Theoretical implications**

Prior research utilized a myriad of different models, of which several involve some black box ML methods (Cheng et al., 2010; Hyari et al., 2016; Matel et al., 2022), this research emphasized the importance of providing some form of explainability into the model. The preliminary stages of projects in engineering companies often involve complex and uncertain factors that can influence cost estimates. Acknowledging this uncertainty is essential in developing robust models that can account for and mitigate potential risks and fluctuations. Because of this, there is a need for explainable models that provide explanations and insights into the underlying factors and variables influencing the estimates.

Furthermore, this research suggests developing a model which can be regularly updated to maintain accuracy and relevance in cost estimation. The evolving nature of engineering projects and industry dynamics necessitates continuous updates to the model. Incorporating new data, industry trends, and feedback from experts ensures that the model remains up-to-date and reliable in supporting decision-making, as supported by (He et al., 2021). Traditional machine learning models like ANN can be time-consuming to retrain and redevelop due to their resource and data requirements (Matel et al., 2022).

Trust is an important factor that needs to be established, as well as the cultural adoption the use of such models might initiate. Decision-makers and stakeholders need to have confidence in the model's outputs and its ability to provide accurate and reliable cost estimates. This shows that only developing the model does not make it directly applicable in practice, there are social aspects that need to be considered after and during the development (Grønsund & Aanestad, 2020). This research suggests the utilization of an external validation process that can support this process involving experts and model users to assess the performance and the underlying assumptions, strengths, and limitations of the model.

However, if data-driven cost estimation models are rendered to be feasible in practice, the availability of sufficient and good-quality data remains a prerequisite. This study only utilized internal and available data sources to establish challenges when developing such models. The difficulty in acquiring this data, and the fragmented data management practices could prove to be an important factor in the discrepancy between academic potential and practical use. The challenge of data management practices has been widely recognized in the construction industry as a whole. This study not only reaffirms this challenge but also specifically highlights its significance within the context of an

engineering company. Before initiating such projects, it is suggested to critically assess the data constraints to develop possible strategies to overcome these challenges.

Nevertheless, in similar contexts as this research, CBR could prove to be a well-suited alternative method for cost estimation (He et al., 2021; Jin et al., 2012; Zima, 2015), if sufficient and relevant data is available. They capture and simulate human experts' knowledge and judgment, addressing the complexity and uncertainty in engineering companies' preliminary stages (He et al., 2021; G. H. Kim et al., 2004). These systems offer explainability by providing insights into the factors influencing cost estimation. They can be regularly updated with new data and industry trends, ensuring accuracy over time. Expert systems also establish trust through external validation, assessing performance and assumptions, and enhancing confidence in the model's reliability for decision-making.

### **5.1.2 Practical implications**

The objective of this study was to develop a data-driven model that can be utilized during a decision-making process within a Tender department of an engineering company, and research challenges and opportunities in this context. Researching the reasons behind the lack of practical use in the case of prior data-driven models developed within Company X (ANN), as well as identifying the steps that can be taken to effectively utilize such models, can have immediate practical implications.

Despite being hindered by data constraints, the model's performance is still considered relevant in providing estimations of the project size in terms of man-hours, according to the discussion session. However, caution should be exercised when utilizing the model due to the inherent instability of its estimates. The model could be retrained on more data in which the performance can be optimized, however, in the context of decision-making under uncertainty, incorporating an expert system (CBR) proves to be a valuable alternative for conducting DDDM in the current research context as supported by (He et al., 2021; Jin et al., 2012; Zima, 2015). As mentioned, CBR, which combines the power of data-driven models with the expertise of human professionals, offers a balanced and mitigated approach. This combination addresses uncertainties, enhances the reliability of decisions, and ensures a holistic approach that considers both quantitative analysis and expert judgment (He et al., 2021).

Case-based reasoning (CBR) holds practical applicability due to its alignment with triple-loop learning and organizational learning. By leveraging past cases and real-world experiences, CBR facilitates experiential learning, enabling decision-makers to adapt their approach to new situations. This alignment allows organizations to continuously improve decision-making, incorporating insights from human expertise and machine learning. CBR's focus on context-specific knowledge ensures relevant solutions, promoting knowledge transfer for continuous improvement. In the context of AI adoption, CBR fosters collaborative learning, where humans learn from AI analyses and integrate AI-generated recommendations with their expertise, leading to continuous improvement and adaptive decision-making (Wijnhoven, 2022).

However, it is equally important to address the challenges involved in developing CBR, which include the acquisition and organization of a comprehensive case base, the selection and representation of relevant cases, the definition of an appropriate similarity measure, and the subjective nature of the adaptation process (Matel et al., 2022; Zima, 2015).

By developing a model that enables rapid and accurate estimation of the man-hours required for various departments, the practical implication of this research is to assist engineering companies in offering more precise and cost-effective proposals to potential clients. In the man-hour estimation method, engineers spend a significant amount of time developing estimates for proposals. Utilizing a data-driven model can significantly reduce time spent on man-hour estimation according to the

interview. This has significant implications for the firm's competitiveness and overhead costs as it directly addresses the issue of lost tenders and the associated sunk costs incurred during the bidding process.

The exact quantification of the required amount of data before rendering DDDM projects feasible may be challenging. In general, it is important for a predictive model to accurately estimate a wide range of scenarios, which requires training data that encompasses the diversity and variability of projects it aims to estimate. Including a comprehensive representation of projects across different types, sizes, and complexities allows the model to learn applicable patterns and relationships. However, limited or incomplete training data may introduce bias and hinder the model's performance and reliability, leading to inaccurate predictions.

The task of acquiring sufficient and high-quality data has been posed as a challenge due to the fragmented data management practices in the construction industry, which in turn presents intriguing possibilities. As per Tuomi's (1999) perspective, the conventional hierarchy of data, information, and knowledge is reversed. Rather than considering data as the raw material for generating information, it is now recognized that data emerges as a product of enhancing the value of information by transforming it into a format suitable for automated processing. Specifically, data is derived from information by organizing it within a predefined data structure that defines its meaning (Tuomi, 1999).

The utilization of data has been widely acknowledged by both employees and research papers as holding significant value for the future (Bilal et al., 2016). Employees possess significant project information, as evidenced by interviews and discussions; however, without proper data capture, valuable information is lost post-project completion, emphasizing the need for systematic data gathering using standardized approaches and clear definitions for data types and designated databases. However, it is equally important to involve employees in this process, keeping them informed about the purpose behind data gathering and its intended use. By engaging employees and providing transparency, they are more likely to be motivated and inclined to actively participate in data collection efforts (Grønsund & Aanestad, 2020).

Furthermore, gathering better and more data does not only prove to hold practical value for the currently developed model but also for future data analytics endeavors regarding data-driven decision-making. The increasing adoption of AI techniques such as NLP, Computer Vision, and ML has further emphasized the importance of data as a resource. Many of these AI models heavily rely on data to generate valuable insights (Bilal et al., 2016). Effectively leveraging data as a valuable resource is crucial for organizations seeking a competitive edge, enabling them to unlock opportunities, make informed decisions, and stay ahead in a rapidly evolving business landscape (Abioye et al., 2021; Bilal et al., 2016).

## 5.2 LIMITATIONS

This research focused on developing a model based on the captured data internally available within Company X. The model was developed based on only four variables due to the limited availability of identified variables in the dataset influencing the cost of engineering services. Employing a different way of data gathering, a survey for example could have provided a different outcome of the model, however, for this research the decision was made to utilize only existing data in databases and to develop a model based on this data.

Despite the extensive number of projects in the database, only a small fraction could be utilized due to the limited availability of variables in the received dataset. To overcome this limitation, a collaborative effort was undertaken to collect supplementary data and variables from other databases.

Despite the efforts, data from a total of 71 projects were ultimately gathered for analysis. Additionally, due to the absence of post-project man-hour estimation analysis, it was not possible to assess the current performance of man-hour estimation or compare it with the developed model.

One limitation of this study pertains to the sample size used during the discussion session to validate the model. Specifically, only eight employees participated in the validation process. While every effort was made to select individuals with relevant expertise and experience in the field, the small sample size raises concerns regarding the generalizability and representativeness of the findings.

Furthermore, another limitation is regarding the specific context in which this research is conducted. The findings and conclusions drawn from this research may be influenced by the unique characteristics, circumstances, and practices of the engineering company in which the study took place. This can include factors such as organizational culture, structure, size, and other industry and or organizational-specific dynamics. This specificity limits the generalizability of the results to other organizations or settings within the engineering industry.

Moreover, the model is specifically designed for implementation within the tender department. Therefore, the utilization of variables in the model enables tender managers to estimate them based on information provided in an RFQ. If a similar model were to be developed in a different department, such as an engineering department, the variables available for estimation could vary significantly, potentially leading to a change in variables and thus utilized data, consequently affecting the model.

Finally, one challenge that has not been extensively researched in this study is the impact of the general strategy of management. It is important to consider the influence of the company's overall strategy, particularly in terms of project acquisition and bidding, on the application of data-driven models. The extent to which risk management is prioritized directly affects the practical implementation of these models. It is essential to align data-driven models with the company's strategic priorities, such as cost optimization, innovation, or customer-centricity, to ensure their relevance and effectiveness in achieving desired outcomes.

Moreover, the company's approach to risk management can play a significant role in determining the extent to which data-driven models are embraced. Companies with a strong focus on risk management may adopt a more cautious approach when implementing these models. They may strike a balance between data insights and human judgment based on experience, mitigating potential risks associated with relying solely on data-driven models. Particularly in complex or high-stakes decisions, companies may choose to leverage data-driven models as supplementary tools alongside expert judgment. Conversely, companies with a higher risk appetite may demonstrate a greater willingness to fully rely on these models, accepting the potential risks that come with it.

### 5.3 FUTURE RESEARCH

While this research addressed challenges and opportunities for the development and practical use of data-driven models for decision-making for cost estimation in an engineering company, the challenges are not addressed extensively on how to overcome these. Future research could entail the data management aspects, to resolve the fragmented data practices in the construction industry in general on how this can support DDDM. Not only in the context of cost estimation does data prove to be a valuable asset, but in many other areas data can prove to be a valuable asset. This research could entail how data should be gathered, through which means, and how data should be stored. This in turn could also explore to what extent a supporting data warehouse can be utilized to support data management.

Another future research area involves exploring the development of an expert system that combines expert judgment with data, specifically focusing on the implementation of CBR within the current



research context. This research could investigate how such a model can be effectively developed and utilized to support DDDM in the given research context. By leveraging expert knowledge and integrating it with data-driven approaches like CBR, researchers can enhance decision-making processes and improve the accuracy and reliability of decision outcomes.

Additionally, future research could also research the exploration of dynamically creating models that can autonomously update themselves over time and how this can be incorporated into existing models. This area of study holds significant importance, as evidenced by its recognition in the existing literature and the present study (He et al., 2021). The ability to dynamically update models is crucial to address the limitations associated with static models and improve their performance over time. By enabling models to adapt to real-time data and incorporate new information promptly, researchers can ensure their predictions and insights remain accurate and relevant. Furthermore, dynamic models offer the potential for continuous learning and improvement, allowing them to evolve and refine themselves as they encounter new data.

Finally, future research in the current context could also delve into the impact of the company's overall strategy, specifically in terms of decision-making processes aided by data-driven models. Although this research did not explore this aspect, further investigation is needed to comprehensively understand how the strategy influences the practical implementation and utilization of data-driven models. Specifically, exploring the implications of a risk-averse company versus a risk-appetite one in adopting and incorporating these models can provide valuable insights.

## 5.4 CONCLUSION

The main objective of this study was to develop a model with the capability to estimate man-hours for proposals inherent to costs in the engineering sector, while also addressing the associated challenges and opportunities. One of the main challenges identified in this study is the management of data, including issues related to data availability, missing values, and low data quality, due to unstandardized data. The data challenges in turn made it difficult to develop predictive models, resulting in an unstable model that is not fit for practical application.

Furthermore, creating explainable models is crucial in the uncertain environment of cost estimation. It is important to justify the predictions made at the initial stage of cost estimation. Additionally, the implementation of these models' impact on work processes poses another challenge, for employees this can be a cultural adaption, resulting in a more top-down decision-making approach than bottom-up. Moreover, trust in the model should be established prior to implementation to understand the assumptions and limitations of the developed model.

Despite the challenges that exist, there are numerous opportunities for the utilization of data-driven models in cost estimation within engineering companies. Although the developed model may be deemed unstable, it still assists decision-makers by offering an initial estimation of the project's magnitude. By leveraging data, decision-makers can harness valuable insights that contribute to more accurate and informed cost estimations, ultimately enhancing project planning and resource allocation in engineering companies.

Despite the acknowledged academic potential of these models, the limited implementation in practical settings hinders their effectiveness and real-world utility. Through researching and identifying these challenges, valuable insights into the barriers preventing their widespread adoption in industry settings are gained. Ultimately, by addressing these challenges and developing strategies to overcome them, the gap between academic potential and practice can be bridged, supporting the effective utilization of cost estimation models to improve DDDM in real-world engineering projects.

## BIBLIOGRAPHY

- Abioye, S. O., Oyedele, L. O., Akanbi, L., Ajayi, A., Davila Delgado, J. M., Bilal, M., Akinade, O. O., & Ahmed, A. (2021). Artificial intelligence in the construction industry: A review of present status, opportunities and future challenges. In *Journal of Building Engineering* (Vol. 44, p. 103299). Elsevier. <https://doi.org/10.1016/j.jobe.2021.103299>
- Agarwal, R., Chandrasekaran, S., & Sridhar, M. (2016). *Imagining construction's digital future | McKinsey*. <https://www.mckinsey.com/capabilities/operations/our-insights/imagining-constructions-digital-future>
- Akintoye, A., & Fitzgerald, E. (2010). A survey of current cost estimating practices in the UK. <http://dx.doi.org/10.1080/014461900370799>, 18(2), 161–172. <https://doi.org/10.1080/014461900370799>
- Badra, I., Badawy, M., & Attabi, M. (2020). Conceptual Cost Estimate of Buildings Using Regression Analysis In Egypt. *IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE) e-ISSN*, 17(5), 29–35. <https://doi.org/10.9790/1684-1705012935>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bean, R., & Davenport, T. (2019, February 5). *Companies Are Failing in Their Efforts to Become Data-Driven*. <https://hbr.org/2019/02/companies-are-failing-in-their-efforts-to-become-data-driven>
- Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A., & Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. In *Advanced Engineering Informatics* (Vol. 30, Issue 3, pp. 500–521). Elsevier. <https://doi.org/10.1016/j.aei.2016.07.001>
- Chartered Institute of Building (Great Britain). (2009). *Code of estimating practice*. 157. <https://www.wiley.com/en-us/Code+of+Estimating+Practice%2C+7th+Edition-p-9781405129718>
- Cheng, M.-Y., Tsai, H.-C., & Sudjono, E. (2010). *Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry*. <https://doi.org/10.1016/j.eswa.2009.11.080>
- Chou, J. S., Yang, I. T., & Chong, W. K. (2009). Probabilistic simulation for developing likelihood distribution of engineering project cost. *Automation in Construction*, 18(5), 570–577. <https://doi.org/10.1016/j.autcon.2008.12.001>
- Christensen, peter, & Dysert, L. (2005). *COST ESTIMATE CLASSIFICATION SYSTEM – AS APPLIED IN ENGINEERING, PROCUREMENT, AND CONSTRUCTION FOR THE PROCESS INDUSTRIES*. [https://www.costengineering.eu/Downloads/articles/AACE\\_CLASSIFICATION\\_SYSTEM.pdf](https://www.costengineering.eu/Downloads/articles/AACE_CLASSIFICATION_SYSTEM.pdf)
- DACE. (n.d.). *Cost Engineering - DACE*. Retrieved February 14, 2023, from <https://www.dace.nl/nl/cost-engineering/cost-engineering>
- De Veaux, Velleman, R. de, Bock, P., & David. (2021). *Stats: Data and Models* (5th ed.). Pearson.
- Doloi, H. K. (2011). Understanding stakeholders' perspective of cost estimation in project management. *International Journal of Project Management*, 29(5), 622–636. <https://doi.org/10.1016/j.ijproman.2010.06.001>
- Doran, D., Schulz, S., & Besold, T. R. (2018). What does explainable AI really mean? A new conceptualization of perspectives. *CEUR Workshop Proceedings*, 2071.
- Elfaki, A. O., Alatawi, S., & Abushandi, E. (2014). Using intelligent techniques in construction project cost estimation: 10-Year survey. *Advances in Civil Engineering*, 2014. <https://doi.org/10.1155/2014/107926>
- Elkjaer, M. (2000). Stochastic budget simulation. *International Journal of Project Management*, 18(2), 139–147. [https://doi.org/10.1016/S0263-7863\(98\)00078-7](https://doi.org/10.1016/S0263-7863(98)00078-7)
- Elmoussalami, H. H. (2021). Closure to “Artificial Intelligence and Parametric Construction Cost

- Estimate Modeling: State-of-the-Art Review” by Haytham H. Elmousalami. *Journal of Construction Engineering and Management*, 147(6). [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0002049](https://doi.org/10.1061/(ASCE)CO.1943-7862.0002049)
- Flyvbjerg, B., Holm, M. K. S., & Buhl, S. L. (2003). How common and how large are cost overruns in transport infrastructure projects? *Transport Reviews*, 23(1), 71–88. <https://doi.org/10.1080/01441640309904>
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502–1505. <https://doi.org/10.1126/science.1255484>
- GAO. (2020, March). *Cost Estimating and Assessment Guide*. <https://www.gao.gov/assets/gao-20-195g.pdf>
- Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational Research Methods*, 16(1), 15–31. <https://doi.org/10.1177/1094428112452151>
- Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *Journal of Strategic Information Systems*, 29(2), 101614. <https://doi.org/10.1016/j.jsis.2020.101614>
- He, X., Liu, R., & Anumba, C. J. (2021). Data-Driven Insights on the Knowledge Gaps of Conceptual Cost Estimation Modeling. *Journal of Construction Engineering and Management*, 147(2). [https://doi.org/10.1061/\(asce\)co.1943-7862.0001963](https://doi.org/10.1061/(asce)co.1943-7862.0001963)
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly: Management Information Systems*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
- Huberman, a. M., & Miles, M. B. (2014). Qualitative Data Analysis. *Qualitative Data Analysis A Methods Sourcebook*, 47(Suppl 4), 3–16. <http://www.uk.sagepub.com/books/Book239534?siteId=sage-uk>
- Hyari, K. H., Al-Daraiseh, A., & El-Mashaleh, M. (2016). Conceptual Cost Estimation Model for Engineering Services in Public Construction Projects. *Journal of Management in Engineering*, 32(1). [https://doi.org/10.1061/\(ASCE\)ME.1943-5479.0000381](https://doi.org/10.1061/(ASCE)ME.1943-5479.0000381)
- ICEAA. (2009). *CEBoK® – International Cost Estimating and Analysis Association* (2nd ed.). <https://www.iceaaonline.com/cebok/>
- Ji, S.-H., Park, M., & Lee, H.-S. (2011). Case Adaptation Method of Case-Based Reasoning for Construction Cost Estimation in Korea. *Journal of Construction Engineering and Management*, 138(1), 43–52. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0000409](https://doi.org/10.1061/(ASCE)CO.1943-7862.0000409)
- Jin, R., Cho, K., Hyun, C., & Son, M. (2012). MRA-based revised CBR model for cost prediction in the early stage of construction projects. *Expert Systems with Applications*, 39(5), 5214–5222. <https://doi.org/10.1016/j.eswa.2011.11.018>
- Kim, B., & Hong, T. (2011). Revised Case-Based Reasoning Model Development Based on Multiple Regression Analysis for Railroad Bridge Construction. *Journal of Construction Engineering and Management*, 138(1), 154–162. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0000393](https://doi.org/10.1061/(ASCE)CO.1943-7862.0000393)
- Kim, G. H., An, S. H., & Kang, K. I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment*, 39(10), 1235–1242. <https://doi.org/10.1016/J.BUILDENV.2004.02.013>
- Kwak, Y. H., & Watson, R. J. (2005). Conceptual estimating tool for technology-driven projects: exploring parametric estimating technique. *Technovation*, 25(12), 1430–1436. <https://doi.org/10.1016/J.TECHNOVATION.2004.10.007>
- Leonard, H., John E., S., Dennis Griffin, & Thomas, C. (2005). *Construction cost estimating : process and practices*. Pearson/Prentice Hall.
- Loubser, J. J. (1968). The Discovery of Grounded Theory: Strategies for Qualitative Research. Barney G. Glaser , Anselm L. Strauss. *American Journal of Sociology*, 73(6), 773–774. <https://doi.org/10.1086/224572>
- Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting Construction Cost Using Multiple Regression Techniques. *Journal of Construction Engineering and Management*, 132(7), 750–758.

- [https://doi.org/10.1061/\(ASCE\)0733-9364\(2006\)132:7\(750\)](https://doi.org/10.1061/(ASCE)0733-9364(2006)132:7(750))
- Matel, E., Vahdatikhaki, F., Hosseinyalamdary, S., Evers, T., & Voordijk, H. (2022). An artificial neural network approach for cost estimation of engineering services. *International Journal of Construction Management*, 22(7), 1274–1287.  
<https://doi.org/10.1080/15623599.2019.1692400>
- Müller, A. C., & Guido, S. (2017). Introduction to Machine Learning with Python: a guide for data scientist. In *O'Reilly Media, Inc.*  
[https://books.google.com/books/about/Introduction\\_to\\_Machine\\_Learning\\_with\\_Py.html?hl=nl&id=vbQIDQAAQBAJ](https://books.google.com/books/about/Introduction_to_Machine_Learning_with_Py.html?hl=nl&id=vbQIDQAAQBAJ)
- NASA. (2015, February). *NASA Cost Estimating Handbook*.  
[https://www.nasa.gov/sites/default/files/files/01\\_CEH\\_Main\\_Body\\_02\\_27\\_15.pdf](https://www.nasa.gov/sites/default/files/files/01_CEH_Main_Body_02_27_15.pdf)
- Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), 51–59. <https://doi.org/10.1089/big.2013.1508>
- PwC. (2019). *Data and analytics services: PwC*.  
<https://www.pwc.com/us/en/services/consulting/cloud-digital/data-analytics.html>
- Regona, M., Yigitcanlar, T., Xia, B., & Li, R. Y. M. (2022). Opportunities and Adoption Challenges of AI in the Construction Industry: A PRISMA Review. *Journal of Open Innovation: Technology, Market, and Complexity*, 8(1), 45. <https://doi.org/10.3390/JOITMC8010045>
- Reinsel, D., Gantz, J., & Rydning, J. (2018). *The Digitization of the World From Edge to Core*.
- RICS. (2020). *RICS cost prediction professional statement, global 1st edition*.  
<https://www.rics.org/profession-standards/rics-standards-and-guidance/sector-standards/construction-standards/rics-cost-prediction-professional-statement-global-1st-edition>
- Shoar, S., Chileshe, N., & Edwards, J. D. (2022). Machine learning-aided engineering services' cost overruns prediction in high-rise residential building projects: Application of random forest regression. *Journal of Building Engineering*, 50, 104102.  
<https://doi.org/10.1016/J.JOBE.2022.104102>
- Smith, A., & Mason, A. K. (1997). Cost estimation predictive modeling: Regression versus neural network. *Engineering Economist*, 42(2), 137–161. <https://doi.org/10.1080/00137919708903174>
- Society of Parametric Analysts, I. (2008). *Parametric Estimating Handbook © Fourth Edition-April 2008*. [www.ispa-cost.org](http://www.ispa-cost.org)
- Sonmez, R., & Ontepeli, B. (2009). Predesign cost estimation of urban railway projects with parametric modeling. *Journal of Civil Engineering and Management*, 15(4), 405–409.  
<https://doi.org/10.3846/1392-3730.2009.15.405-409>
- Swei, O., Gregory, J., & Kirchain, R. (2017). Construction cost estimation: A parametric approach for better estimates of expected cost and variation. *Transportation Research Part B: Methodological*, 101, 295–305. <https://doi.org/10.1016/J.TRB.2017.04.013>
- Tayefeh Hashemi, S., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: a systematic review on machine learning techniques. In *SN Applied Sciences* (Vol. 2, Issue 10, pp. 1–27). Springer Nature. <https://doi.org/10.1007/s42452-020-03497-1>
- Trost, S. M., & Oberlender, G. D. (2003). Predicting Accuracy of Early Cost Estimates Using Factor Analysis and Multivariate Regression. *Journal of Construction Engineering and Management*, 129(2), 198–204. [https://doi.org/10.1061/\(ASCE\)0733-9364\(2003\)129:2\(198\)](https://doi.org/10.1061/(ASCE)0733-9364(2003)129:2(198))
- Venable, J. R. (2006). The Role of Theory and Theorising in Design Science Research. *Proceedings of the 1st International Conference on Design Science in Information Systems and Technology (DESRIST 2006)*, 1–18.
- Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *Information Systems Research*, 3(1), 36–59. <https://doi.org/10.1287/isre.3.1.36>
- Wieringa, R. J. (2014). Design science methodology: For information systems and software engineering. In *Design Science Methodology: For Information Systems and Software Engineering*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-43839-8>

- Wijnhoven, F. (2022). Organizational Learning for Intelligence Amplification Adoption: Lessons from a Clinical Decision Support System Adoption Project. *Information Systems Frontiers*, 24(3), 731–744. <https://doi.org/10.1007/s10796-021-10206-9>
- Zhu, B., Yu, L. A., & Geng, Z. Q. (2016). Cost estimation method based on parallel Monte Carlo simulation and market investigation for engineering construction project. *Cluster Computing*, 19(3), 1293–1308. <https://doi.org/10.1007/S10586-016-0585-6/FIGURES/6>
- Zima, K. (2015). The Case-based Reasoning Model of Cost Estimation at the Preliminary Stage of a Construction Project. *Procedia Engineering*, 122, 57–64. <https://doi.org/10.1016/J.PROENG.2015.10.007>

# APPENDICES

## APPENDIX A: BPMN MODELS

*To ensure confidentiality, the appendix containing sensitive information has been omitted from this version of the document.*

*Figure A 1: Tender Process*

*To ensure confidentiality, the appendix containing sensitive information has been omitted from this version of the document.*

*Figure A 2: Proposal writing process*

*To ensure confidentiality, the appendix containing sensitive information has been omitted from this version of the document.*

Figure A 3: [REDACTED] proposal process



*To ensure confidentiality, the appendix containing sensitive information has been omitted from this version of the document.*

Figure A 4: [REDACTED] proposal process

## APPENDIX B: DATA SCHEMA

In this appendix the data schema of the first-order themes, second-order themes, and the aggregate dimensions are presented in Table B 1. This data schema is based upon the conducted interviews (Appendix D, Appendix G, Appendix H, Appendix I)

Table B 1: Data schema

| First-Order Themes  | Second-Order Themes           | Aggregate dimensions                      |
|---|-------------------------------|---|
| Unreliable project data   | Data Challenges               | Data management                           |
| Limited data availability   |                               |   |
| Inaccurate data   |                               |   |
| Lack of a general accessible database for data retrieval                                    |                               |   |
| Underutilization of data  | Utilization of Data           |   |
| Need for capturing precise and standardized data  |                               |   |
| Scope variability can change  | Estimation Complexity         | Estimation context                        |
| Numerous variables and complexity   |                               |   |
| Departmental differences in Man-hour estimation   |                               |   |
| Challenges in capturing nuances (small but impactful details)                               |                               |   |
| Difficulty in practical application   | Practical Application         | Implementation management (practical use) |
| Developed model usability   |                               |   |
| Importance of confidence intervals  | Model Performance and Trust   |   |
| Assessing and justifying model results (external validation)                                |                               |   |
| Comparing model performance with actual MHE   |                               |   |
| Gaining trust through parallel estimation methods   |                               |   |
| Learning from post-project calculations (nacalculaties)                                     | Learning and Improvement      | Continues improvement                     |
| Building upon previous research (internal) findings   |                               |   |
| Revisable model to dynamically improve  |                               |   |
| Considering the PLAN B initiative   | Optimization (new) model      |   |
| Predictor variables Several different predictor variables (e.g. Capex, project phase, etc.) | Requirements                  | Model development                         |
| Explainability of the outcome   |                               |   |
| Defensible model due to necessary to provide explanations management                        | Processes man-hour estimation |   |
| Standard procedures for man-hour estimation, can differ dependent on size proposal          |                               |   |

## APPENDIX C: AACE CLASSIFICATION

For the model requirement, a classification is made on how well the model should perform and in what phase the desired accuracy is. This is based upon the work of AACE, in the paper of (Christensen & Dysert, 2005). This classification is widely adopted that provides the expected accuracy for the different stages of a project. For an overview see Figure C 1. It is important to note that this classification is mainly from the contractor's point of view, for the Tender management proposals, the second class is applicable.

| ESTIMATE CLASS | Primary Characteristic   | Secondary Characteristic                 |  |   |  |
|----------------|--|--|--|---|--|
|                | LEVEL OF PROJECT DEFINITION<br>Expressed as % of complete definition | END USAGE<br>Typical purpose of estimate | METHODOLOGY<br>Typical estimating method                   | EXPECTED ACCURACY RANGE<br>Typical variation in low and high ranges [a] | PREPARATION EFFORT<br>Typical degree of effort relative to least cost index of 1 [b] |
| Class 5        | 0% to 2%   | Concept Screening                        | Capacity Factored, Parametric Models, Judgment, or Analogy | L: -20% to -50%<br>H: +30% to +100%                                     | 1  |
| Class 4        | 1% to 15%  | Study or Feasibility                     | Equipment Factored or Parametric Models                    | L: -15% to -30%<br>H: +20% to +50%                                      | 2 to 4   |
| Class 3        | 10% to 40%   | Budget, Authorization, or Control        | Semi-Detailed Unit Costs with Assembly Level Line Items    | L: -10% to -20%<br>H: +10% to +30%                                      | 3 to 10  |
| Class 2        | 30% to 70%   | Control or Bid/Tender                    | Detailed Unit Cost with Forced Detailed Take-Off           | L: -5% to -15%<br>H: +5% to +20%  | 4 to 20  |
| Class 1        | 50% to 100%  | Check Estimate or Bid/Tender             | Detailed Unit Cost with Detailed Take-Off                  | L: -3% to -10%<br>H: +3% to +15%  | 5 to 100   |

- Notes: [a] The state of process technology and availability of applicable reference cost data affect the range markedly. The +/- value represents typical percentage variation of actual costs from the cost estimate after application of contingency (typically at a 50% level of confidence) for given scope.
- [b] If the range index value of "1" represents 0.005% of project costs, then an index value of 100 represents 0.5%. Estimate preparation effort is highly dependent upon the size of the project and the quality of estimating data and tools.

Figure C 1: AACE framework (Christensen & Dysert, 2005)

## APPENDIX D: INTERVIEWS REQUIREMENTS

*To ensure confidentiality, the appendix containing sensitive information has been omitted from this version of the document.*

## APPENDIX E: REGRESSION OUTPUT

### E1

| OLS Regression Results |               |                     |          |       |           |          |
|------------------------|---------------|---------------------|----------|-------|-----------|----------|
| =====                  |               |                     |          |       |           |          |
| Dep. Variable:         | Sum Hours     | R-squared:          | 0.827    |       |           |          |
| Model:                 | OLS           | Adj. R-squared:     | 0.756    |       |           |          |
| Method:                | Least Squares | F-statistic:        | 11.64    |       |           |          |
|                        |               | Prob (F-statistic): | 3.07e-10 |       |           |          |
| =====                  |               |                     |          |       |           |          |
|                        | coef          | std err             | t        | P> t  | [0.025    | 0.975]   |
| -----                  |               |                     |          |       |           |          |
| const                  | 2257.5991     | 471.971             | 4.783    | 0.000 | 1302.948  | 3212.251 |
| Total Investment       | 32.4744       | 145.047             | 0.224    | 0.824 | -260.911  | 325.860  |
| Lead time (weeks)      | 284.7860      | 161.198             | 1.767    | 0.085 | -41.267   | 610.839  |
| Count Employees 2      | 2044.6110     | 266.537             | 7.671    | 0.000 | 1505.489  | 2583.733 |
| Count disciplines      | -498.6106     | 213.452             | -2.336   | 0.025 | -930.359  | -66.862  |
| P2                     | 184.0718      | 430.290             | 0.428    | 0.671 | -686.271  | 1054.415 |
| P3                     | 76.9864       | 445.259             | 0.173    | 0.864 | -823.635  | 977.608  |
| P4                     | 464.9541      | 488.277             | 0.952    | 0.347 | -522.679  | 1452.587 |
| P5                     | -1219.9443    | 977.933             | -1.247   | 0.220 | -3198.000 | 758.111  |
| M100                   | -386.7256     | 524.813             | -0.737   | 0.466 | -1448.261 | 674.809  |
| M200                   | 438.7726      | 386.354             | 1.136    | 0.263 | -342.702  | 1220.247 |
| M300                   | -45.4611      | 681.695             | -0.067   | 0.947 | -1424.318 | 1333.396 |
| M400                   | -187.6225     | 401.752             | -0.467   | 0.643 | -1000.242 | 624.997  |
| M500                   | -897.0605     | 719.637             | -1.247   | 0.220 | -2352.665 | 558.544  |
| M600                   | -290.3215     | 401.584             | -0.723   | 0.474 | -1102.602 | 521.959  |
| M700                   | 111.0045      | 907.147             | 0.122    | 0.903 | -1723.873 | 1945.882 |
| M800                   | -526.1671     | 929.415             | -0.566   | 0.575 | -2406.087 | 1353.753 |
| =====                  |               |                     |          |       |           |          |
| Omnibus:               | 8.881         | Durbin-Watson:      | 2.190    |       |           |          |
| Prob(Omnibus):         | 0.012         | Jarque-Bera (JB):   | 20.153   |       |           |          |
| Skew:                  | 0.077         | Prob(JB):           | 4.20e-05 |       |           |          |
| Kurtosis:              | 5.935         | Cond. No.           | 14.1     |       |           |          |
| =====                  |               |                     |          |       |           |          |

Figure E 1: OLS regression all variables

#### Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified

#### Assumptions:

- **Multicollinearity:**

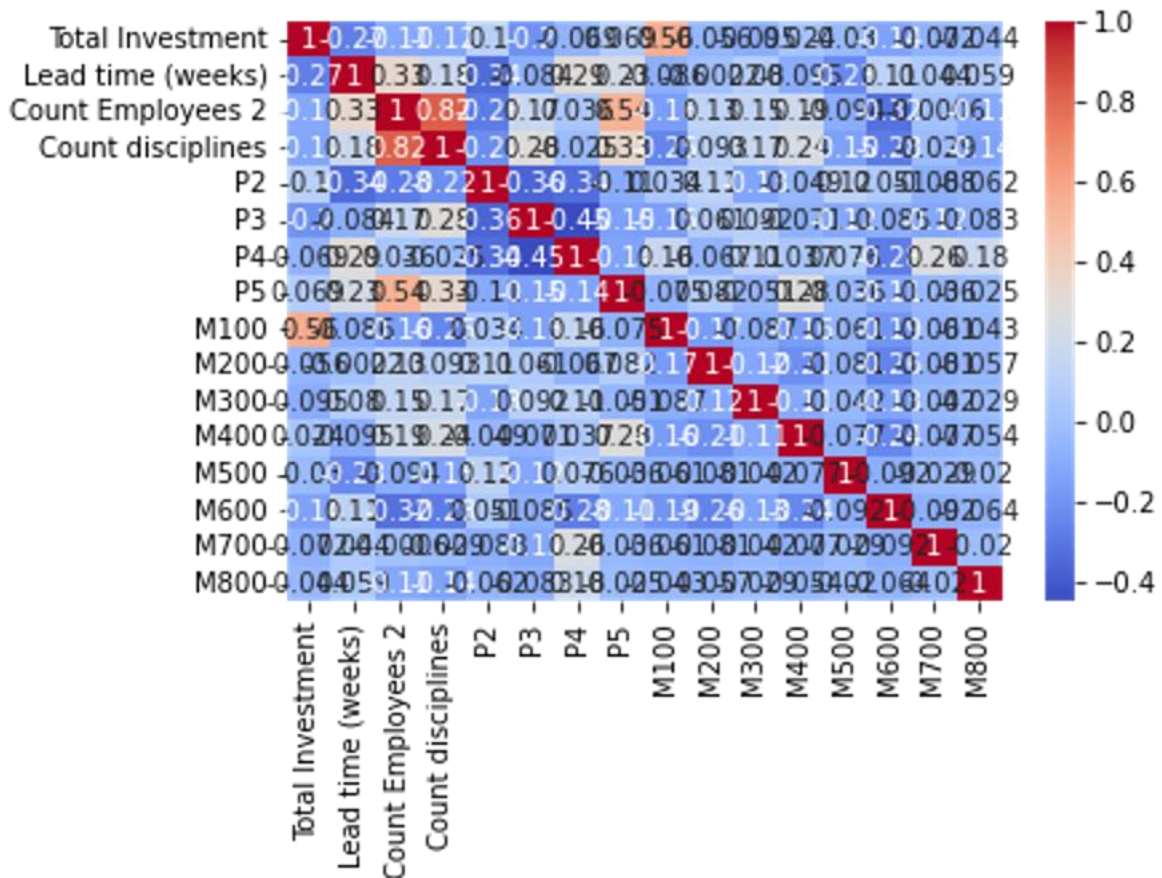


Figure E 2: Correlation all variables

| Independent Variable | VIF Factor |
|----------------------|------------|
| 0 Total Investment   | 2.102426   |
| 1 Lead time (weeks)  | 6.681139   |
| 2 Count Employees 2  | 14.028003  |
| 3 Count disciplines  | 18.019614  |
| 4 P2                 | 2.280948   |
| 5 P3                 | 3.889652   |
| 6 P4                 | 4.772515   |
| 7 P5                 | 2.407788   |
| 8 M100               | 2.470458   |
| 9 M200               | 1.852407   |
| 10 M300              | 1.296469   |
| 11 M400              | 1.911717   |
| 12 M500              | 1.202966   |
| 13 M600              | 2.044212   |
| 14 M700              | 1.224787   |
| 15 M800              | 1.147923   |

Values with multicollinearity of a value higher than 3 are problematic.

- **Normality:**

Jarque-Bera test statistic: 20.153496795291524

Jarque-Bera p-value: 4.204591184121734e-05

Significant, thus violation of normality

- **Homoscedasticity:**

Breusch-Pagan test statistic: 22.220914355709795

Breusch-Pagan p-value: 0.13620068788734918

Not significant, thus no violation

**E2: Significant variables**

In ridge and lasso regression, the interpretation of coefficient scores and p-values differs from ordinary least squares (OLS) regression due to the introduction of penalty terms. Instead of relying on p-values, these techniques focus on regularization and consider metrics such as the magnitude and relative importance of coefficient estimates (Müller & Guido, 2017). Because of this, only the OLS regression results are presented in Figure 12 based on model 1 of Table 14.

OLS Regression Results

|                   |               |                     |          |       |          |          |
|-------------------|---------------|---------------------|----------|-------|----------|----------|
| =====             |               |                     |          |       |          |          |
| Dep. Variable:    | Sum Hours     | R-squared:          | 0.809    |       |          |          |
| Model:            | OLS           | Adj. R-squared:     | 0.791    |       |          |          |
| Method:           | Least Squares | F-statistic:        | 45.14    |       |          |          |
|                   |               | Prob (F-statistic): | 3.76e-21 |       |          |          |
| =====             |               |                     |          |       |          |          |
|                   | coef          | std err             | t        | P> t  | [0.025   | 0.975]   |
| -----             |               |                     |          |       |          |          |
| const             | -895.4199     | 395.062             | -2.267   | 0.027 | -1684.64 | -106.192 |
| Lead time (weeks) | 12.1823       | 5.237               | 2.326    | 0.023 | 1.720    | 22.645   |
| Count Employees   | 120.7125      | 18.599              | 6.490    | 0.000 | 83.556   | 157.869  |
| Count disciplines | -45.0970      | 53.332              | -0.846   | 0.401 | -151.639 | 61.445   |
| P1                | 5.261e-12     | 1.4e-12             | 3.764    | 0.000 | 2.47e-12 | 8.05e-12 |
| P3                | -201.2845     | 371.932             | -0.541   | 0.590 | -944.303 | 541.734  |
| P4                | 327.4776      | 371.764             | 0.881    | 0.382 | -415.207 | 1070.162 |
| P5                | 3555.1327     | 830.036             | 4.283    | 0.000 | 1896.945 | 5213.321 |
| =====             |               |                     |          |       |          |          |
| Omnibus:          | 4.515         | Durbin-Watson:      | 2.084    |       |          |          |
| Prob(Omnibus):    | 0.105         | Jarque-Bera (JB):   | 4.399    |       |          |          |
| Skew:             | -0.295        | Prob(JB):           | 0.111    |       |          |          |
| Kurtosis:         | 4.067         | Cond. No.           | 1.50e+19 |       |          |          |
| =====             |               |                     |          |       |          |          |

Figure E 3: OLS regression significant variables

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Mean Absolute Percentage Error (MAPE): 0.4805800015221948

Root Mean Squared Error (RMSE): 1077.986033651194

**Assumptions:**

- **Multicollinearity:**

|   | Independent Variable | VIF Factor |
|---|----------------------|------------|
| 0 | Lead time (weeks)    | 4.338954   |
| 1 | Count Employees      | 14.851131  |
| 2 | Count disciplines    | 16.924715  |
| 3 | P2                   | 1.669477   |
| 4 | P3                   | 3.252624   |
| 5 | P4                   | 2.912195   |
| 6 | P5                   | 1.793778   |

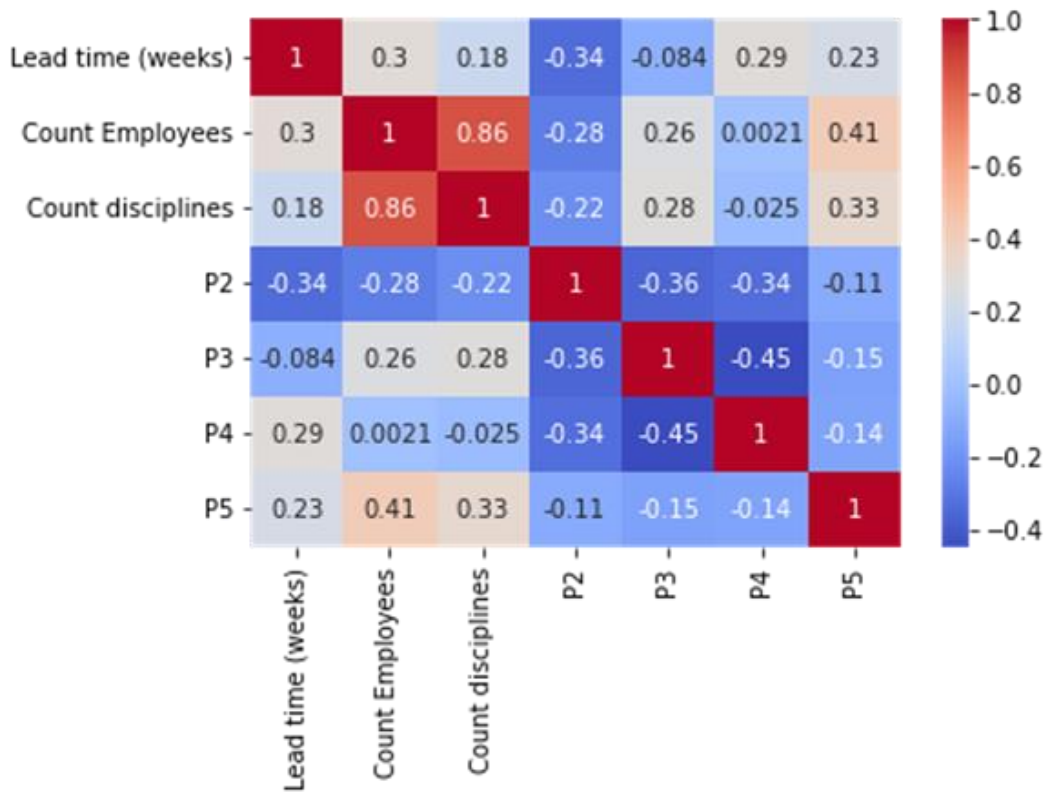


Figure E 4: Correlation significant variables

**Homoscedasticity:**

Jarque-Bera test statistic: 4.984

Jarque-Bera p-value: 0.082

Not significant, thus no violation

**Normality:**

Breusch-Pagan test statistic: 37.885

Breusch-Pagan p-value: 1.18e-06

Significant, thus violation



**E3: Total resources variable = Count employees \* Count disciplines**

| OLS Regression Results |               |                     |          |       |          |          |
|------------------------|---------------|---------------------|----------|-------|----------|----------|
| Dep. Variable:         | Sum Hours     | R-squared:          | 0.820    |       |          |          |
| Model:                 | OLS           | Adj. R-squared:     | 0.807    |       |          |          |
| Method:                | Least Squares | F-statistic:        | 59.40    |       |          |          |
|                        |               | Prob (F-statistic): | 6.38e-23 |       |          |          |
|                        | coef          | std err             | t        | P> t  | [0.025   | 0.975]   |
| const                  | -157.6796     | 311.182             | -0.507   | 0.614 | -779.152 | 463.793  |
| Lead time (weeks)      | 14.2069       | 4.950               | 2.870    | 0.006 | 4.322    | 24.092   |
| Total Resources        | 5.0000        | 0.476               | 10.496   | 0.000 | 4.049    | 5.951    |
| P1                     | 2.958e-13     | 1.62e-13            | 1.821    | 0.073 | -2.8e-14 | 6.2e-13  |
| P3                     | -99.6178      | 349.668             | -0.285   | 0.777 | -797.953 | 598.717  |
| P4                     | 515.0354      | 353.364             | 1.458    | 0.150 | -190.681 | 1220.752 |
| P5                     | 3345.1645     | 803.688             | 4.162    | 0.000 | 1740.08  | 4950.240 |
| =====                  |               |                     |          |       |          |          |
| =                      |               |                     |          |       |          |          |
| Omnibus:               | 2.388         | Durbin-Watson:      | 2.163    |       |          |          |
| Prob(Omnibus):         | 0.303         | Jarque-Bera (JB):   | 1.837    |       |          |          |
| Skew:                  | 0.105         | Prob(JB):           | 0.399    |       |          |          |
| Kurtosis:              | 3.760         | Cond. No.           | 1.71e+19 |       |          |          |
| =====                  |               |                     |          |       |          |          |

Figure E 5: OLS regression Total resources variable

**Notes:**

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.  
 Mean Absolute Percentage Error (MAPE): 0.44676808636768656  
 Root Mean Squared Error (RMSE): 1044.7677434423067

**Assumptions:**

- **Multicollinearity:**

|   | Independent Variable | VIF Factor |
|---|----------------------|------------|
| 0 | Lead time (weeks)    | 3.931327   |
| 1 | Total Resources      | 2.782487   |
| 2 | P2                   | 1.289062   |
| 3 | P3                   | 2.324124   |
| 4 | P4                   | 2.434459   |
| 5 | P5                   | 1.730702   |

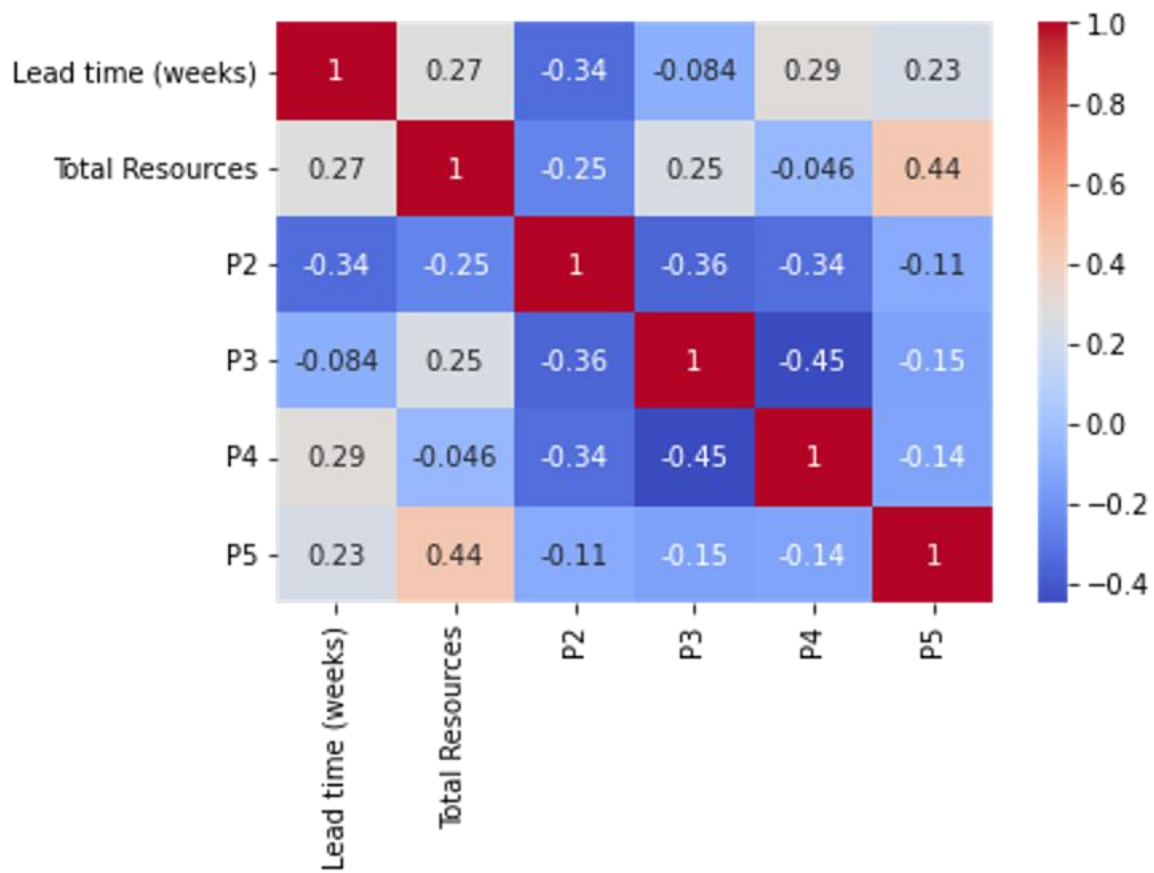


Figure E 6: Correlation Total resources variable

**Homoscedasticity:**

Jarque-Bera test statistic: 1.986

Jarque-Bera p-value: 0,370

Not significant, thus no violation

**Normality:**

Breusch-Pagan test statistic: 35.950

Breusch-Pagan p-value: 9.71e-07

Significant, thus violation

## APPENDIX F: PYTHON CODE

- **For standardization:**

```
from sklearn.preprocessing import StandardScaler

# Create a StandardScaler object
scaler = StandardScaler()

# Fit the scaler to the independent variables
scaler.fit(X)

# Apply standardization to the independent variables
X_scaled = scaler.transform(X)
```

Figure F 1: Python standardization

- For variable selection:

```
import itertools

# Create a list of all independent variable names
variable_names = X_train.columns.tolist()

# Initialize variables for storing the best model and its adjusted R-squared
best_model = None
best_adj_r_squared = float('-inf')

# Generate all possible combinations of variable names
for r in range(1, len(variable_names) + 1):
    combinations = itertools.combinations(variable_names, r)

    # Iterate through each combination
    for combo in combinations:
        # Subset the independent variables based on the combination
        X_subset = X_train[list(combo)]

        # Add a constant term to the independent variables
        X_subset = sm.add_constant(X_subset)

        # Fit the multiple linear regression model
        model = sm.OLS(y_train, X_subset).fit()

        # Get the adjusted R-squared of the model
        adj_r_squared = model.rsquared_adj

        # Check if the current model has a higher adjusted R-squared
        if adj_r_squared > best_adj_r_squared:
            best_adj_r_squared = adj_r_squared
            best_model = model

# Print the summary statistics of the best model
print(best_model.summary())

# Subset the independent variables based on the best model
X_best_subset = X_test[list(best_model.params.index[1:])]
X_best_subset = sm.add_constant(X_best_subset)

# Calculate the predictions of the best model
y_pred = best_model.predict(X_best_subset)
```

Figure F 2: Python variable selection

## APPENDIX G: INTERVIEW IMPACT WORK PROCESSES

*To ensure confidentiality, the appendix containing sensitive information has been omitted from this version of the document.*

## APPENDIX H: INTERVIEW VALIDATION PHASE

*To ensure confidentiality, the appendix containing sensitive information has been omitted from this version of the document.*

## APPENDIX I: DISCUSSION VALIDATION PHASE

*To ensure confidentiality, the appendix containing sensitive information has been omitted from this version of the document.*