ML Governance Within Banking

A study into the oversight of ML model development and compliance.

A thesis presented for the degree:
Master of Science

Author
S.B. Berends
(s1498320)

UNIVERSITY OF TWENTE.

bunq
BANK OF THE FREE

Supervisors:
Dr. B. Roorda
Dr. A. Abhishta

Supervisors:
N. Mossel
Dr. A. el Hassouni

University of Twente
Financial Engineering and Management
PO Box 217
7500 AE Enschede
The Netherlands

Friday 19th November, 2021

Summary

Financial institutions are increasingly leveraging ML to perform difficult and laborious tasks in order to save costs and/or gain competitive advantage. DNB, however, is hesitant in allowing ML as they are seen as black-box models. As part of the model governance, DNB requires financial institutions to audit their models on a regular basis. However audit requirements for ML model validation are not clearly defined. This results in an indeterminate expectation of the scope and outcomes from both sides.

In this thesis an effort is made to bridge the expectation gap by facilitating ML governance with the use of a framework for challenger banks. In order to facilitate ML governance, a framework is developed. This framework, denoted in chapter 5, 6 and 7, aids developers, data scientists and team leads by providing recommendations. These recommendations promote ML governance by recommending to take certain aspects into consideration and to documented them properly. The ML governance framework is the main deliverable of this thesis.

To gather the information that is required for the framework, three research questions are answered. The research questions are:

1. “What is, according to the literature, important in the governance of ML models?”
2. “How is Explainability in an artificial intelligence context defined?”
3. “What is according to the Dutch Central Bank important in the development and monitoring of ML systems in the financial service industry?”

To determine if bunq has ML governance issues, a fourth research question is answered in the form of a case study. In this case study, it is checked whether all recommendations from the built ML governance framework are followed. The specific research question is:

4. “Are there ML governance shortcomings in the Transaction Monitoring system?”

Firstly a literature study is done to establish important aspects of ML governance.

These aspects are split up in five categories: 'Justice and Equity', 'Use of Force', 'Safety and certification', 'Privacy and Power', 'Taxation and Displacement of Labour' and 'Other'. From this literature study can be concluded that there is no clear agreement on all aspects of ML as various papers have different interpretations on the importance of these aspects. Therefore in this thesis, all the aspects are taken into account(if within scope). There is, however, a general consensus on four core principles. An ML model should be fair, transparent, documented and accountability should be defined in case of faulty outcomes.

Secondly, the often cited term explainability is explored. Explainability is an often used term with no clear-cut definition in the literature. For example, banking regulators in Germany do not handle the same definition as in the Netherlands. Based on the definitions found in literature and from regulators, it is determined that there are two key factors that define explainability. The two factors are: 1) a good explanation and 2) an audience. The explanation must be comprehensible by the audience. Furthermore, what constitutes as a good explanation is retrieved from social sciences as the field of human explanations has been a researched far longer than explanations of ML models.

Thirdly, the publications from DNB are scrutinized for regulations, requirements and their stance on ML governance. DNB is the regulator within the dutch banking sector and is responsible for performing audit procedures such as transaction monitoring based on ML algorithms. To be compliant and aware of potential future requirements, requirements from DNB are incorporated in the framework. The most important publication is: "General principles for the use of Artificial Intelligence in the financial sector" (De Nederlandsche Bank, 2019). In this publication multiple recommendations are stated which are incorporated into the framework. The other three publications stated few to none recommendations but did supply insights into what their stance is regarding AI and thus ML.

To find gaps in bunq's ML governance, a case study is performed where the Transaction Monitoring system is vetted against the framework. It is found that 4 out of the 33 recommendations are not fulfilled based on the available documentation. Firstly, bunq's risk framework does not have a specific ML section. Secondly, it is recommended by the framework that fairness mitigations need to be approved by the risk department, which has not been done. Thirdly, the model has no specific documentation on data integrity or bias issues. Fourthly, the model is not checked for unfairness by proxy. These gaps should ideally be fixed by bunq.

The framework as most important part of this research, contains three large phases i.e. 'Development', 'Deployment' and 'Post Deployment'. The phases contain sub-

phases where chronologically, information and recommendations on ML governance aspects within the development process are described. These recommendations vary greatly and includes topics such as: what should be in the documentation, various forms of fairness, mutual entropy, unfairness by proxy and more. The developer can use these recommendations to improve ML governance and provide proof that the model is up to the standard, set by the framework. It will furthermore force developer to continuously question, document and mitigate the risks of various aspects of ML models.

The future of ML governance will become more advanced than the framework presented in this thesis, as the field of ML governance is evolving. Based on recent literature and papers, regulators and society require improved quality standards for ML models. It is therefore important that the framework goes through regular improvements based on the latest insights. The following recommendations should be considered. Firstly, it is recommended to build monitoring software for facets such as prejudice, fairness and data/concept drift to improve control. Secondly it is recommended to do research on methods to determine causal relations as it will proved a better basis for a model as well as better interpretability. Thirdly it is recommended to focus on the fundamentals of developing ML models instead of focusing on explainability as the current post-hoc models such as Shapley are not yet the end-all-be-all of being in control of a ML model. Fourthly, the use of Generative Adversarial Networks provide the ability to improve ML models but require more research on the influx of bias and other potential downsides. Lastly, I recommend society and the financial service industry to cooperate with other companies or institutions to push the field of ML governance forward, as best as possible. This can be done in multiple ways such as: participating in initiations like DNB's iForum and so called 'sandbox' environments. In such a sandbox environment, regulators pose less stringent regulations such that companies can test innovative techniques whilst regulators can learn from those tests.

Preface

I am very pleased to present you this thesis named: “ML Governance within Banking, A study into the oversight of ML model development and compliance”. This thesis, which is written as a graduation project for the master Financial Engineering & Management, focuses on the development of a framework that facilitates ML governance in an effort to diminish the indeterminate audit requirements of ML model validation. Data scientists, ML model developers and their team leads are requested to follow the recommendations from the framework such that they can provide proof that their models have the standard set by the framework.

This research has been a real challenge with changing goals and determining what the actual problem is. Fortunately, I had help from my supervisors whom I want to thank greatly. Firstly, I want to thank Berend Roorda for his guidance throughout this project as well as his lectures over the course of the master. Secondly, I want to thank Abhishta Abhishta for his fresh look on my thesis, which was very insightful.

Furthermore, I would like to thank bunq for providing me the opportunity to perform my graduation project in the risk department. I especially want to thank Nico Mossel for his guidance, insights and general companionship. I also want to thank Ali el Hassouni who is not officially my supervisor, but who has provided me with good guidance and discussions regarding the subject.

Lastly, I want to thank everyone that has supported me over the years: Femke, ’Magnaten’, my housemates, my rowing crew, my friends and especially my family. Thank you for for all the good times!

Sjors Berends

Amsterdam, November 2021

Contents

List of acronyms

AFM	Autoriteit Financiele markten
AI	Artificial Intelligence
AML	Anti-Money Laundering
API	Application Programming Interface
BaFin	Bundesanstalt für Finanzdienstleistungsaufsicht // Federal Financial Supervisory Authority
DNB	De Nederlandsche Bank
FIU	Financial Intelligence Unit
HU	University of Applied Sciences Utrecht
KL	Kullback Leibler
KPI	Key Performance Indicator
KYC	Know your Customer/client
ML	Machine Learning
NVB	Nederlandse Vereniging van Banken // Dutch Banking Association
TM	Transaction Monitoring
MSE	Mean Squared Error

List of Figures

List of Tables

CHAPTER 1

Introduction

"*Trust is a fragile thing — hard to earn, easy to lose.*"
M.J. Arlidge

In the past fifteen years, the banking sector has been the embodiment of the above mentioned saying. Since the banking crisis of 08, people have been skeptic of the sector as a whole and with good reason as banks are in the business of trust (Aerens, 2019). However, times are changing and the sector has been regulated thoroughly. But how will regulators apply rules and guidelines when new technologies emerge? Technologies that will be increasingly leveraged to perform laborious tasks, find correlations and control the allocation of money: Artificial Intelligence(AI) and its subcategory Machine Learning(ML).

1.1 Motivation

In the recent years I have noticed that the use of ML has elevated from something that was only used in far-away research projects and gimmicks such as training a video game into something that is widely available. Most of my colleagues at the University of Twente, with varying research fields, are using ML in their theses with the aid of pre-built packages such as Scikit-Learn (Scikit-learn, n.d.) and TensorFlow (TensorFlow, n.d.). With the ease of ML adoption among academics and employees, an increasingly amount of ML models will be built. However there are other factors to take into account besides having a functioning model when an ML model is deployed into a production environment at a company. Factors that pose risks to the responsible organisation, such as not adhering to fairness and regulatory aspects are extremely undesirable. I will refer to the managing of these aspects of ML models as the field of ML governance. The field of ML governance is an upcoming research field and is getting increasingly more interest from society and scientists. Therefore with this piece of research I aim to add value to the field of ML governance within banking.

1.2 bunq

This thesis is written as part of my internship at bunq. bunq with its motto: “Bank of The Free”, is the latest company in the Netherlands that has received a banking license from De Nederlandsche Bank (DNB). Since the founding of the company in 2012, the people at bunq have been trying a different approach to this age-old profession. bunq has intended to leverage tech to perform banking and payment services through an application for smart-phones.

The founding idea of bunq is based an a dislike for the conventional banking business model which encourages financial gain with large risks whilst using the clients' money as collateral. bunq's business model is primarily focused on providing financial services and keeping your money safe, not based on the money made by the interest rate spread (Huizinga, 2016). However, due to the current negative interest rate climate, bunq has revisited those principles and started to invest with a low tolerance for risk such as: investment grade bonds, mortgages and more. However, the primary revenue comes through a subscription based business model where the client's interaction with its money is the main focus i.e. easy payments, direct transfers, spending insights and more. All done through a state-of-the-art app.

1.3 Artifical Intelligence and Machine Learning

The field of AI is focused on building a non-human program that mimics the problem-solving and decision-making capabilities of the human mind (IBM Cloud Learn Hub, 2020). Through various AI techniques, the possibilities and the task that an AI model is able to perform can vary greatly. Examples of AI models are autonomous driving systems and personalised assistants e.g. Siri, Alexa and more. Although sounding very state-of-the-art, the field of Artificial Intelligence is not new and was coined in 1956 at a conference at Dartmouth College (T. Lewis, 2014). In recent years the field of AI has made giant leaps forward with the increase of computing power and labeled data. Through these innovations, it has never been easier for companies and individuals to create their own AI models.

ML is a subcategory of AI that is focused on teaching computers how to learn and act without being explicitly programmed to do so (DeepAI, n.d.). This is often done through optimizing various algorithms with training data. Using this optimized algorithm, a prediction or estimation is made on future data. Examples of such models are image classification systems and models that predict interest rates (Cornelissen, 2021).

1.4 Banking

Banking has been around since the first currencies were minted. In the ancient times bartering was the way to pay for goods and services. However having a currency increased the possibility of paying with something that was more easily exchangeable.

With the origination of banks resulting from the switch of bartering to a currency, banks have played a large role in society. Banks enabled the storage of money and enabled an accelerated economic growth throughout history due to the ability for society to take on loans to facilitate key revolutions e.g. ships, steam powered machines and the power loom (Rousseau, 2003). Services that banks currently supply are for example the processing of payments and taking out a mortgage. Without mortgages very few people would be able to buy a house at the current rate. Right up until this day the principles of banking e.g. lending money and safekeeping have largely remained the same however small changes have occurred in the form of various financial services e.g. transactions and currency exchanges.

1.4.1 Bank as a gatekeeper

Tackling money laundering has a high priority within the dutch government as it is of great importance for the effective fight against all forms of serious crime (Ministerie, 2021). As a bank controls and transfers money throughout the world, banks are deemed as gatekeepers to the financial system. Therefore banks are morally and by law (De Nederlandsche Bank, 2017) obliged to know who exactly makes use of their networks and whether people are abusing the bank's network. Among banks this is often recognized as Anti-Money Laundering operations(AML). These operations consist of: know your client (KYC) and transaction monitoring (TM). This is a large part of the operations of a bank, to illustrate: ING currently has 4000 people employed who solely focus on KYC (ING, 2021).

1.4.2 Transaction Monitoring

Transaction Monitoring is the process of analyzing the transactions that are done through the bank. Banks process an enormous amount of payments, on 2020 more than 6 billion payments have been processed in the Netherlands alone (Dutch Payments Association, 2021). Scrutinizing every single payment and its characteristics is impossible, therefore models are in place at all banks which perform initial screening based on certain rules. The hits that these rule-based models provide are checked by analysts and reported to the Financial intelligence Unit (FIU) when

there is reasonable assumption that a transaction is fraudulent, involved in money laundering or financing of terrorism. However due to the static nature of rule-based systems, they often produce a very high number of false positives or very little fraudulent transactions. Therefore bunq has produced a set of ML models to perform the initial screening. Which resulted in an increase in the accuracy and a decrease in the amount of false positives.

1.4.3 Challenger banks

Challenger banks, which are smaller and newer banks that are in a direct competition with already established banks, have come into existence due to an increase in consumers who have lost faith in the traditional financial system during the global financial crisis (CBInsights, 2021). In combination with an increase in technology and software, challenger banks were able to start streamlined retail banks that are not subdue to legacy IT and large overhead from physical branches. A bank with an IT infrastructure that has no legacy allows for certain advantages over traditional banks, with their largest advantage being able more easily leverage their technologies to solve problems at hand. These problems can vary from implementing newer payment methods, to allowing users to hold multiple currencies but also to allow users to interact with their bank account using an Application Programming Interface(API).

1.5 Research Goal

Banking is a very regulated and audited business, therefore the need for proper processes are important for both business and compliance incentives. This also holds for ML which is increasingly adopted. DNB is hesitant in allowing it without thorough audits as various ML techniques can result in a black-box model. However, as was stated to me in a personal interview with someone from EY, DNB does not exactly know how to audit these models. This results in a grey area, or expectation gap, where there is uncertainty on whether a bank adheres to regulations and expectations.

This expectation gap leads to the goal of this thesis, bunq wants to take steps to be compliant with DNB as well as to have a better control on their ML processes. Therefore the goal of this thesis is to devise a framework for challenger banks to be compliant with DNB's publications and to mitigate potential operational risks in the development and monitoring of ML models.

To put this framework into practice, the ML model which handles transaction moni-

toring will be put to the test. The transaction monitoring model and its problems will be further elaborated on in Section 1.7.1.

1.6 Scope

The establish the domain of this thesis, three items need to be defined in this scope. Firstly, bunq is a dutch challenger bank. Therefore the scope of this research is limited to ML Governance within the challenger banking sector. Secondly, bunq tries to gain an edge on the competition by leveraging tech. Therefore bunq does not want to exchange the developed models with third parties. Fourthly, as explained in Section 1.3, AI and ML are not the same. Although these terms are often used interchangeably, it is important to be aware of the difference as this thesis is focused on ML models within bunq. Since ML is a subcategory of AI, research on the governance of AI models is also taken into account. This decision will provide more information of governance aspects to consider. However, certain aspects in the literature have their name and definition rooted in AI e.g. explainable AI. To stay consistent with the literature, these names will not be altered and AI will therefore be mentioned.

1.7 ML Governance

Considering the posed goal in Section 1.5, a specific term to address the goal is desired. I have chosen to use a contraction between ML and Governance. The idea being that “Governance”, which according to the Merriam-Webster dictionary is: “*The act or process of governing or overseeing the control and direction of something*” (Merriam-Webster, n.d.), closely resembles the goal of the thesis. Moreover, ML is the specific field where bunq tries to improve its governance. Therefore a proper ML governance entails being compliant with recommendations published by DNB and having control over potential operational risks regarding the development and monitoring of ML models.

1.7.1 Example of the problem

An example of a grey area as mentioned in Section 1.5, is that of a bank and its gatekeeper function. Failing to perform it properly, meaning the transaction monitoring is not done well enough, can result in fines and even legal prosecution. However there are no specific rules in place that describe how a bank precisely should monitor transactions. DNB’s position is stated in the following quote from the guidance document on transaction monitoring by the DNB:

> *"As gatekeepers for the Dutch financial system, banks are expected to adequately and continuously monitor transactions, and to stay alert. There are statutory requirements which banks must meet in this regard, and it is our task to supervise compliance with these rules and regulations. All banks are therefore obliged to conduct transaction monitoring, although this obligation and its supervision is principle-based.*
>
> *This means that the practical interpretation of these requirements is not prescribed in detail by laws and regulations, or by the supervisory authority. It is up to you as a bank to determine how exactly you interpret this. The supervisory authority will assess the result." (De Nederlandsche Bank, 2017)*

Sole rule-based systems are deemed unsuitable as these systems are quite inaccurate. Due to the nature of rule-based systems, they either produce a very high number of false positives or very little fraudulent transactions are caught. Therefore a more clever way of transaction monitoring is by using ML models, after which people check on the hits that are generated by the system. However bunq has had extensive enquiries on the ML system by DNB. With the aid of the aforementioned framework, bunq is taking steps to deal with compliance and governance issues surrounding the ML part in transaction monitoring.

1.8 Research Question

The aim of this research is to improve ML governance within challenger banks by devising a framework. This framework will be backed up by academic research and insights from regulators to aid in the development and monitoring of ML models such that these models are compliant with the published recommendations of DNB and governed properly. To reach the aim, the following research questions will need to be answered.

RQ1: "What is, according to the literature, important in the governance of ML models?"

In order to progress the field of ML-governance within banking, this research explores what according to academic literature are good practices in the development and monitoring of ML. A list of aspects are expected from this sub-question that, according to the literature, are of great importance on ML Governance and should therefore be considered in the final framework.

RQ2: "How is Explainability in an artificial intelligence context defined?"

bunq has had extensive remarks by auditors on Explainability regarding their transaction monitoring model which leverages ML. However the field of explainable AI is quite novel and not yet matured. Therefore this sub question focuses on the exploration of a definition, as a common and agreed upon definition is not found in the literature.

RQ3:"What is according to the Dutch Central Bank important in the development and monitoring of ML systems in the financial service industry?"

The third sub question for this research goal is regarding the information and guidelines that are posed by regulatory bodies and with DNB in particular. The guidelines, regulations and other information that is provided by DNB are going to be used to ensure that a challenger banks' ML models are compliant.

With the information provided by all three sub questions, a development and monitoring framework will be devised which has a solid base in both academic literature and regulatory guidelines.

RQ4:"Are there ML governance shortcomings in the Transaction Monitoring system?"

The fourth sub question puts the framework to the test on the transaction monitoring model which is one of bunq's most important ML models. In this case study, using the framework, the expectation is to find aspects of the transaction monitoring model which are not fully coherent with proper ML governance.

1.9 Research contributions

This research provides contributions to the fields of ML and banking. This research can in some form be applied to most banks and especially challenger-banks. Therefore this research provides contributions to the field of ML governance and bunq's knowledge in the following forms:

Insights
The research will provide insights on multiple fronts that are important to bunq's operations. Firstly, the insights on regulations and requirements that are posed by DNB on the use of ML. Secondly, the insights around what according to the literature is important on ML governance when developing and monitoring ML models.

Framework
This research contributes to having a proper ML governance and compliance process around developing ML models within the banking industry in the form of a framework. A framework turns implicit steps into explicit steps. This is important

as people and experts in general tend to combine steps and combine them into larger tasks. This process of combining tasks has a negative effect on the conscious approach of the person which results in more mistakes (Agency for Healthcare Research and Quality, 2010). Having a framework enables the developer to chronologically go through the process of developing a model whilst considering the ML governance. It furthermore provides a certain standard that is set by the framework. This makes the process of developing and monitoring ML models within a challenger bank more stable and therefore less key-person dependent.

Use-Case Transaction Monitoring
The transaction monitoring is one of the key ML models that is used within bunq. This model will be vetted with the developed framework to find out whether the model has any ML governance issues which need to be addressed.

Best practice
The last contribution that this research brings forward a tool which can help in the process of improving towards a best practice environment surrounding ML. bunq is deemed a bank as well as a tech company and therefore processes around ML should become increasingly better such that bunq can use their experience and processes to gain an edge on competitors.

1.10 Outline of the thesis

The structure of the thesis is as follows:

- Chapter 2 focuses on research question 1 and provides an overview of what different aspects according to the literature should be taken into account when developing and using an ML model with proper ML governance.
- Chapter 3 delves into research question 2 and describes a definition for explainability in a machine learning context.
- Chapter 4 presents research question 3, which goes into the information that is made available by the dutch banking regulator, DNB.
- Chapter 5 describes the framework that is needed the development phase of developing an ML model.
- Chapter 6 presents the framework that should be considered when the ML model is deployed.
- Chapter 7 describe the framework that should be considered when monitoring the ML models after it has been deployed.

- Chapter 8 contains research question 4, a case study on one of bunq's ML model to test the built framework in chapters 5,6 and 7.
- Chapter 9 concludes this work with an overview of the contributions that this research makes to the field of ML governance within banking as well as the recommendations for a next research.

CHAPTER 2

Literature review into ML Governance aspects

"What is, according to the literature, important in the governance of ML models?"

To answer the posed research question, the literature surrounding Governance on ML models is explored in a literature review. During this review, academic papers on governance and ML are scrutinized for the aspects that are from a governance perspective considered important when developing and monitoring ML models. Afterwards, the aspects that do not fit the scope (defined in section 1.6) will be dropped.

To provide the review with a solid base of information, multiple databases are used to find academic works. Scopus and Web of Science are chosen as they are regarded among the top research databases (paperpile, 2019). Furthermore, see the appendix for the used key words, requirements and the elimination procedures.

2.1 Findings for a new framework

Considering the established rules, key words and elimination procedures in the appendix, 17 papers are reviewed. These 17 papers are visible in Table 2.2. From these 17 papers, the aspects regarding ML governance are explained below in several categories. For each aspect, an explanation is given what the aspect is and why it is important. Since not each aspect is clearly defined in its corresponding paper, I have chosen to use external sources to define and further elaborate the found aspects. To enable ease of reading, the aspects are subdivided into six categories which are first seen in use by Calo (2018). The categories are:

1. Justice and Equity
2. Use of Force
3. Safety and certification

AI facet
Certification
Setting safety thresholds
Validating safety thresholds

Table 2.1: Excluded ML facets

4. Privacy and Power
5. Taxation and Displacement of labor
6. Other

2.2 Excluded aspects

Considering the scope of the research, defined in section 1.6, certain aspects are exempted. The ML governance facets displayed in Table 2.1 are excluded as these facets should according to the theory (Calo, 2018) be executed by a third party. Since these models are proprietary pieces of research that are subject to both fraud sensitive information and competitive advantage bunq is not inclined to allow third parties insights into proprietary models.

№	**Title**	**Authors**	Year
1	Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI	Carla Zoe Cremer and Jess Whittlestone	2021
2	Privacy-Preserving Scoring of Tree Ensembles: A Novel Framework for AI in Healthcare	Kyle Fritchman and Keerthanaa Saminathan and Rafael Dowsley and Tyler Hughes and Martine De Cock and Anderson Nascimento and Ankur Teredesai	2018
3	Artificial intelligence policy: A primer and roadmap	R. Calo	2018
4	Artificial intelligence and the financial markets: Business as usual?	J. Schemmel	2019
5	A Layered Model for AI Governance	Urs Gasser and Virgilio A.F. Almeida	2017
6	Binary governance: Lessons from the GDPR'S approach to algorithmic accountability	M.E. Kaminski	2019
7	Privacy-Preserving Scoring of Tree Ensembles: A Novel Framework for AI in Healthcare	K. Fritchman and K. Saminathan and R. Dowsley and T. Hughes and M. De Cock and A. Nascimento and A. Teredesai	2019
8	Biobanks and Biobank-Based Artificial Intelligence (AI) Implementation Through an International Lens	Z. Kozlakidis	2020
9	Toward the agile and comprehensive international governance of AI and robotics	W. Wallach and G. Marchant	2019
10	Data Governance Technology — 数据治理技术	X.-D. Wu and B.-B. Dong and X.-Z. Du and W. Yang	2019
11	Model Governance: Reducing the anarchy of production ML	V. Sridhar and S. Subramanian and D. Arteaga and S. Sundararaman and D. Roselli and N. Talagala	2020

№	**Title**	**Authors**	Year
12	Possible extension of ISO/IEC 25000 quality models to Artificial Intelligence in the context of an international Governance	D. Natale	2020
13	Global Challenges in the Standardization of Ethics for Trustworthy AI	Dave Lewis and Linda Hogan and David Filip and P. J. Wall	2020
14	Experiences with improving the transparency of AI models and services	M. Hind and S. Houde and J. Martino and A. Mojsilovic and D. Piorkowski and J. Richards and K.R. Varshney	2020
15	AI and ML-Driving and Exponentiating Sustainable and Quantifiable Digital Transformation.	C. Naseeb	2020
16	Building the right AI governance model in Oman	Halah Al Zadjali	2020
17	Rho AI – Leveraging artificial intelligence to address climate change: Financing, implementation and ethics	I. Fischer and C. Beswick and S. Newell	2021

Table 2.2: Reviewed papers

Justice and Equity Aspects
Fairness
Accountability
Transparency
Inequality in application
Consequential Decision making
Explainability
Dependability
Responsibility
Controllability

Table 2.3: AI aspects within Justice and Equity

2.3 Justice and Equity

2.3.1 Accountability and Transparency

Accountability and transparency are the two facets that are most often mentioned in the reviewed papers. These two facets come to the core of what the common problems are within machine learning. In practice, a lot of machine learning models are deemed black-box models, which immediately poses two difficult problems. Firstly, with regards to transparency, the problem is that people cannot look into the model and therefore don't know what is happening at the core of the model. Secondly, with regards to accountability, people and companies often shelter behind an ML model as if it is an all-knowing oracle.

Accountability and Transparency are important aspects to take in mind for various reasons. Most importantly, when a model is transparent it is much more explainable, allows for thorough testing, and people can understand why particular decisions are made (Deloitte, 2019). The accountability aspect is important as people can't hide behind a model but will be held liable to face consequences from an authority such as DNB.

2.3.2 Fairness and inequality in application

According to the literature, an ML model should adhere to fairness such that the model treats each person in question fairly. Meaning that no discriminatory variables such as native descent, nationality or gender are of influence. Calo (2018) names fairness as Inequality in Application. An example of this is the unequal performance of commercial face classification services in the gender classification task where

the accuracy on dark-skinned females is significantly worse than any other group (Muthukumar et al., 2018). This issue often occurs through a bias in the dataset (Lim, 2020). However that is not the entire story as in the aforementioned example evidence is brought forward that differences in lip, eye and cheek structure across ethnicity lead to the differences. However, Inequality in application is an undesirable phenomenon and is important to handle as society strives for equality. Having an unfair model can therefore result in reputational damage.

2.3.3 Consequential decision making

Consequential decision making involves the process where systems make or help to make consequential decisions about people whilst being influenced by regulations and procedural rules (Calo, 2018). An example of such a system is an ML-enabled justice system (Završnik, 2020). Special caution should be taken in such situations as ML models can determine correlations, it cannot prove explicit causality which is much more important when decisions are to be made with a backdrop of procedural rules and regulations.

2.3.4 Explainability

Explainability of ML models is the idea that more information is provided on how a model came to a conclusion instead of the sole classification. This is very intuitive as it gives people developing, working with and overseeing these models more insight into what is happening inside the so-called black box. Having more insight into how a decision has come to be can be a valuable addition, especially if the decision is to be contested or used as an input for another process. The precise definition of explainability is to be discussed in chapter 3.

2.3.5 Responsibility

Responsibility as mentioned by Gasser and Almeida (2017) in the paper: “A Layered Model for AI Governance” is that there ought to be someone e.g. a company, person or institution that takes responsibility for the outcomes of the Ml model in question. Not to be mistaken with responsible AI which is an umbrella term for the governance of ML from an ethical and legal point of view. The definition of responsibility and accountability are often used synonymous however there is a distinction, as responsibility is an ongoing duty to handle something whereas accountability is what happens after a situation occurs (SpriggHR, 2020). Responsibility is important as it someone is in charge of maintaining a certain quality.

Use of Force Aspects
Use of Force
Human well-being
Human rights

Table 2.4: AI aspects within Use of Force

2.3.6 Controllability

Controllability is a difficult term to define in the wide range of ML. However, Yampolskiy (2020) defined it as the ability for humanity to remain safely in control while benefiting from a superior form of intelligence. This definition is defined with the ultimate form of AI (super intelligence) in mind such that humanity is safe in each form of other AI.

2.3.7 Dependability

Dependability is the quality of being able to be trusted and being very likely to do what people expect (Dictionary, n.d.). Using dependable ML is key since you need to always be able to rely on your systems to work. Although seemingly obvious, there are thinkable situations where an ML model will not be dependable for instance when a model is overfitted.

2.4 Use of Force

2.4.1 Human rights and well-being

Within the category of Use of Force, denoted in Table 2.4, human rights and well-being are the aspects to consider. The impact of ML on humans should always be taken into consideration whilst keeping in mind that ML needs to work in favour of the human and not vice-versa. There are a plethora of situations where human rights or humans' well-being are affected, but especially as a bank, the developers of the ML model should be aware of the implications a model can have on people. Banks have to show that their business is done with integrity and done in a controlled manner (Rijksoverheid, 2021). An example of a negative situation can be a model which infers the interest rate on a product differently due to a discriminatory aspect. This inequality in application is against human rights and will possibly result in large reputational damage as well as fines from DNB.

Safety and Certification Aspects
Certification
Setting safety thresholds
Validating safety thresholds
Cybersecurity

Table 2.5: AI aspects within Safety and Certification

Privacy and Power Aspects
Privacy
Pattern recognition
Data-parity problem

Table 2.6: AI aspects within Privacy and Power

2.5 Safety and Certification

The principle of Safety and Certification is that an ML model should adhere to a specific standard such as safety thresholds which would ideally be certified by an outside institution. Thus filtering sub-par ML models. Safety thresholds can be important such that it is known when action should be taken when the models' results deteriorate.

2.6 Privacy and Power

2.6.1 Privacy, pattern recognition and the data-parity problem

Privacy concerns have been growing in the last few decades as consumers are becoming more aware of how companies are using their data (Goswami, 2020). Since ML is intimately tied with the availability of data (Calo, 2018), privacy concerns will play an important role in the governance of ML.

However there are various ways that privacy ought to be taken into consideration i.e. the chance that sensitive information on people are not kept private enough as well as the problem of pattern recognition and that of data-parity. The problem with pattern recognition is that seemingly small snippets of an individuals' life can add up until it can ultimately predict patterns that are (potentially) not imaginative to the individuals' self (Hill, 2012). Data-parity is a concentration of data problem. Since the models are so reliant on data and the abundance/quality of it, larger companies have an advantage over smaller companies.

Taxation and Displacement of Labour Aspects
Taxation of labour
Displacement of labour

Table 2.7: AI aspects within Taxation and Displacement of Labour

Other Aspects
Auditability
Accuracy
Provenance/lineage
Reproducubility

Table 2.8: Other ML aspects

2.7 Taxation and Displacement of Labour

The argument can be made that the increased automation can decrease the human workforce. The displacement or decrease in work for the human workforce can be detrimental to society as people can become unemployed or need re-schooling. Aside from the direct impact on people, there is also a potential monetary impact on society. Governments are mainly funded through income taxes, if there is a shift from income through jobs to capital gains this can result in a decrease in government funding. This will in turn have impact on the the collective spending of the government (e.g. healthcare), the ability for governments to reach policy goals and the redistribution of wealth (Planbureau, 2020).

2.8 Other

2.8.1 Auditability

A not often mentioned facet related to ML governance is that of Auditability. In a banking environment, auditing processes are very common and used both internally and externally. Being able to show regulators that your systems are up to par is key, as you will be reprimanded and will not be able to use the model when it is not approved. Being auditable is being able to show the algorithms, data and design processes, but preserving the intellectual property related to the ML systems (Barredo Arrieta et al., 2020).

2.8.2 Accuracy

Accuracy is a notion of how well a model performs and can be used to evaluate classification and prediction models. A good accuracy metric is very situation dependent. Therefore some situations, need a different accuracy metric such as: Area under Curve(AUC), F1 Score or Mean Squared Error(MSE). To illustrate accuracy I have chosen one of the most fundamental metrics: classification accuracy in a binary classifier. In the case of a simple binary classifier, it is defined as shown in equation 2.1. In this equation there are four different variables. Each variable is based on whether the model has classified the data point the same as it actually is or is not. Therefore:

1. TP is the number of True Positives, the number of classifications that are classified as Positive whilst it actually is positive.
2. TN is the number of True Negatives, the number of classifications that are classified as Positive whilst it actually is negative.
3. FP is the number of False Positives, the number of classifications that are classified as Positive whilst it actually is negative.
4. FN is the number of False Negatives, the number of classifications that are classified as Negatives whilst it actually is positive.

$$\begin{aligned} Accuracy &= \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made} \\ &= \frac{TP + TN}{TP + TN + FP + FN} \end{aligned} \tag{2.1}$$

However, Accuracy as described in equation 2.1 can be an example of a metric that can give a false sense of achieving a high accuracy when it is applied the wrong situation. When your dataset is not balanced. For example when a dataset contains 90% positive and 10% negative samples, the classifier can reach a 90% accuracy by classifying each sample as positive. Therefore more metrics are defined in the literature, which will be touched upon later in this thesis.

2.8.3 Provenance/lineage

The principle of provenance is to be able to retrace the events that have occurred for a specific outcome. Which means being able to answer questions like: “On what data set is it trained?”, “What code was used?” and “What human approvals are given?”. This is important as it allows a better view of the model and less like a black-box such as described in Section 2.3.1.

2.8.4 Reproducibility

The ability to reproduce each step which is described in the provenance section and through doing that, arriving at the same prediction. This is key as when this doesn't happen, it means that there is some random aspect to each classification which is undesirable at best.

CHAPTER 3

Defining Explainability

RQ2:"How is explainability in a machine learning context defined?"

Explainability is an often used term within the guidance of DNB (2019). However an agreed-upon definition within the field of ML or AI is not clear-cut. Without a definition it is difficult for a challenger bank to use the DNB's guidelines in their operations. Therefore in this research question, research is done on how explainability is defined in a ML context.

3.1 Explainability according to regulators

Explainability in general is defined by DNB as: *"Explainability entails that an explanation can be formulated. An explanation is contextual, relevant and has the goal of addressing a stakeholder's concern or interest"* (De Nederlandsche Bank, 2021). Whereas the German banking regulator, BaFin, has defined explainability as: *"Being able to list the major factors of influence for a concrete individual decision"* (BaFin, 2018).

The difference in the definitions and the vagueness surrounding the aforementioned definitions show an example of a larger problem that Miller et al (2017) describes. Miller et al (2017) have shown that research on explainability in AI and thus ML, rarely builds on frameworks from social sciences but instead researchers use their intuition to state what a 'good' explanation is. Thereby setting themselves up for potential failure when 20 years of research into how people generate explanations and evaluate their quality is ignored.

3.2 Interview with DNB

To further the research into what explainability is, a personal interview has been held with the 'senior policy advisor Artifical Intelligence' from DNB. In this interview was stated that DNB struggles with defining explainability within the field of artificial

intelligence. Therefore DNB in cooperation with the AFM, the university of applied sciences Utrecht and three major banks did an exploratory study on explainable AI. In that study they have decided on a definition of explainable AI(XAI), which is:

> "A set of capabilities that produces an explanation, in the form of details, reasons, or underlying causes, to make the functioning and/or results of an AI solution sufficiently clear so that it is understandable and addresses stakeholders' concerns." (De Nederlandsche Bank, 2021)

This is a very large definition where various types of explanations can adhere to. And thus 'explanations' range from how a particular outcome has come about, to the people that are accountable for development and use of the AI solution (De Nederlandsche Bank, 2021). This interview has made it clear that DNB likes to use explainability as an umbrella term where the true question is actually whether banks develop, use and monitor AI models in a valid way. What that valid way is, is decided by DNB on a case-by-case basis.

In my opinion, umbrella terms such as how XAI is coined by DNB are undesirable. Using a single term to address multiple issues creates uncertainty and misunderstanding. In the same publication where the explainable AI is coined (De Nederlandsche Bank, 2021), it is stated by the DNB that: *"There appears to be a disparity between views of the supervisory authorities and the participating banks, regarding the desired scope of explainability required for AI solutions in banking."* Therefore in this research question the scope of the term explainability will be focused on the outcomes of ML models and not on e.g. who is accountable(this is handled separately in the aspect: accountability).

3.3 Explainability and its key factors

In the scientific literature a lot of (slightly) different definitions of explainability in AI are found. The most clear-cut definition I could find is:

> "Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand." (Barredo Arrieta et al., 2020)

Based on this and the aforementioned definitions in section 3.1, two key factors are distilled that define explainability in a machine learning context. namely an explanation and a target audience.

Miller has written multiple articles with various co-authors on explanations within AI and has shown that social sciences is a sound research field to start when look-

ing into what an explanation entails. The definition that Miller puts forward is: *"An explanation is an answer to a why-question"* (2019). Miller systematically surveyed over 250 social science papers on how people explain to each other and found that explanations are contrastive. Instead of answering the question: *"Why P?"*, people actually explain the question: *"Why P rather than Q?"*. Since explanations on ML models are interpreted by humans, it is important to consider how people ask for and provide explanations.

The contrastive event Q is often implicit from the context but a contrasting explanation is important for two reasons. First, people ask contrasting questions when they are surprised by an event and expect something different. This provides a 'window' into the questioner's mental model, identifying what they do not know (D. K. Lewis, 1986). The second reason is that the explainer does not have to reason about, or know about, all causes of the fact that are relative to the contrast case (Miller, 2020). This corresponds with Lewis' (D. K. Lewis, 1986) statement that the use of a contrastive explanation is simpler, more feasible and cognitively less demanding for both the explainer and the explainee. An example of a good explanation in a ML context would be an image classifier with the following explanation: "The creature is classified as a spider as opposed to an insect because an insect has 6 legs where spiders have 8."

The target audience that is to be kept in mind can vary per ML model. However within challenger banks, the three most important audiences are: Developers, Users of the model and regulators such as DNB. These audiences may require a different explanation because they have a different goal when asking an explanation.

To provide an example, consider a ML model predicts whether a customer is eligible for a loan. A developer will be focused on maximizing its accuracy and is therefore looking at minute details or edge cases. A user wants to know what the reasons are for being refused a loan which are most often done by the variables with the most distinctive power. Whereas a regulator might want to know if a classification is abiding by the regulations.

Therefore to answer the research question. Explainability in a machine learning context is defined by two key factors, a good explanation for a specific audience. Where a good explanation is able to answer a contrastive question.

CHAPTER 4

Interpreting publications from DNB

RQ3:"What is according to the Dutch Central Bank important in the development and monitoring of ML systems in the financial service industry?"

There are multiple pieces of information that provide information on how DNB views AI. In this chapter, the publications from DNB are scrutinized for aspects that are important in the development and monitoring of ML systems. Since ML is a subcategory of AI as explained in Section 1.3, DNB's publications on AI are considered applicable on ML. As DNB is the main auditor of dutch banks, dutch banks should comply to their regulations. However, DNB has not (yet) imposed formal regulations but they have published articles where guidelines are provided and their stance is implicitly stated. Banks can use these guidelines to leverage on performed research whilst simultaneously comply with the regulator. Furthermore, through being aware of DNB's stance on AI, it allows for a progressive insight on the direction of ML governance from the regulators perspective.

The publications from DNB are handled one-by-one where aspects that suit ML governance are stated below. The scrutinized publications are:

1. "General principles for the use of Artificial Intelligence in the financial sector" (De Nederlandsche Bank, 2019).
2. "DNB Position Paper 'Wettelijk kader en toezicht'" (De Nederlandsche Bank, 2020b).
3. "Guideline on the Anti-Money Laundering and Anti-Terrorist Financing Act and the Sanctions Act" (De Nederlandsche Bank, 2020a).
4. "Perspectives on Explainable AI in The Financial Sector" (De Nederlandsche Bank, 2021).

4.1 General principles for the use of Artificial Intelligence in the financial sector

DNB states in its first publication, general principles for the use of AI, that financial institutions should adhere to the 'SAFEST" principles (De Nederlandsche Bank, 2019). Where 'SAFEST' is an acronym for: 'Soundness', 'Accountability', 'Fairness', 'Ethics', 'Skills' and 'Transparency'.

From these principles, Soundness is the principle that DNB is most caring about. Soundness is an aggregate of multiple facets such as: reliability, accuracy, predictability and operating within regulatory boundaries. All of these facets contribute to a model that, simply put, works and behaves as expected.

The opinion of DNB around accountability is mostly focused on educating stakeholders that AI applications can be complex and might not work as intended. Despite these principles being guidelines instead of regulations, there are situations described which are condemned. In the case of accountability, model complexity and third party reliance are stated as arguments that can never be used for limiting the organisation's accountability.

The definition of fairness is explained in Section 2.3.2. Moreover DNB states that fairness is essential for society's trust in the financial sector and therefore an institutions' concept of fairness ought to be definable and that their AI application behaves accordingly.

The Ethics principle is very similar to the previously described 'consequential decision-making' i.e. that one must be critical of decisions that are made and that humans or institutions are not mistreated.

The Skills principle is based around the premise that with an increasing reliance on AI, (senior) management, risk management and compliance functions should have adequate expertise.

The last principle is that of transparency where financial firms should be able to explain how they use AI in their business process and how these applications function. Adhering to this principle enables adequate risk management and internal audits. If this principle is executed properly, it allows for further optimisation of the applications.

4.2 DNB Position Paper 'Wettelijk kader en toezicht'

In the second published item, a position paper of DNB, the question is asked how the oversight on AI will look like in ten years time. DNB states that the speed at which technological developments evolve will continue to rise whereas regulations are generally slow and might take years to develop. "*Through this, the risk arises where the regulations underestimate new risks whilst bona fide innovation will be slowed unintentionally*" (De Nederlandsche Bank, 2020b). DNB states that through this conviction, legislators and regulators should aim for a future resilient technology neutral policy.

4.3 Guideline on the Anti-Money Laundering and Anti-Terrorist Financing Act and the Sanctions Act

The third publication that contains a mention of AI is the publication by DNB which guides banks on how to interpret and deal with the WWFT. This publication has information on how DNB views the use of AI in operations, specifically Transaction Monitoring. DNB states:

> *"When an institution deems the use of highly advanced systems, for example Artificial Intelligence. The quality and effectiveness of the system must be demonstrable. An institution can adequately assess the quality and effectiveness of its transaction monitoring by arranging a model validation or audit" (De Nederlandsche Bank, 2020a)*

This shows that DNB requires banks to be able to show that AI models work effectively and that they are up to the set standards.

4.4 Perspectives on Explainable AI in The Financial Sector

In the fourth publication, the iForum publication(an initiative from DNB), DNB explores the field of explainable AI in cooperation with the AFM, the University of Applied Sciences Utrecht, the dutch banking association and the representatives from three major dutch banks. According to the report,

> *"DNB's iForum aims to create more room for technological innovation within the financial system and does so by developing joint experiments*

> *in the areas where technology and supervision meet. With this study, we have therefore outlined the perspectives from the banks as well as the two involved supervisory authorities, on where explainable AI meets supervision." (De Nederlandsche Bank, 2021)*

The quote above illustrates rather well what the report is about. No true guidelines or expectations of the financial service industry are stated and it is emphasized that this is an exploratory study and it does not formulate new supervisory policies. Furthermore this publication shows that DNB is searching for ways to improve AI governance whilst simultaneously finding support within the industry.

CHAPTER 5

Framework: Development Phase

In the following three chapters the results from the first three subquestion come together to form a framework on ML governance. This framework is based on the published recommendations from DNB and should be used by a developer/programmer to get a ML model which is coherent to the published standards by DNB and the aspects from academic literature.

The product is set up in a chronological order. The entire modelling process is split up into three (common) large phases, development, deployment and post deployment. Within these phases, there are multiple sub phases which are based on Oracle's Lifecycle of Machine Learning Models eBook (2020). At each stage the relevant aspects will be noted that need attention in that specific time frame. What a certain aspect is and why it is important is explained in each section. If each recommendation is checked off, one could consider that their model is sufficiently well-grounded in ML governance. The considerations are denoted with a letter for the corresponding large phase i.e. A, B & C followed with a numbering per section and per check.

In the upcoming chapter the development phase will be discussed. During the development phase the bulk of the process is done and will be an inherently iterative process until a model with proper documentation is devised.

5.1 Development

The first phase, development, spans across the entirety from defining a goal to having an initial working model. This phase generally encompasses the following sub phases:

1. Data gathering / Cleaning
2. Feature Engineering
3. Data Splitting

4. Model Selection
5. Model Training
6. Model Validation

Some aspects found in the literature should be considered the entirety of the development phase. These facets are displayed in Table 5.1. The aspects of transparency, auditability and reproducibility all find their core in the choices and the documentation thereof. This documentation is an ongoing process which should be on the mind of the developer at all times such that no choices are forgotten to document. The specific choices can vary greatly from why specific features are used, to why data splitting for training is done in a certain way. To ensure proper ML governance, the document should be readable by someone that is not necessarily an ML expert. Meaning that the choices are explained and that the reader can read it as a standalone document

Accountability is very focused around being aware of the limitations, risks and the (unintended) impact that ML can have. This is also emphasized by DNB which states:

> *"Especially when AI applications become more material, financial firms should demonstrate unequivocally that they understand their responsibility for AI applications and that they have operationalised accountability for these applications throughout their organisation." (De Nederlandsche Bank, 2019)*

In an actionable way this translates to two requirements from DNB. Firstly, operational accountability is explicitly assigned at all relevant levels of the organisation where final accountability for ML applications and their outcomes (both for the organisation and their customers) is assigned to one (or more) board member(s). Secondly, The adoption and use of ML is integrated in the organisation's risk management framework where clear roles and responsibilities are assigned throughout the organisation to ensure the responsible use, management and auditability of ML applications.

A.1.1 Are sufficient efforts made to produce an as transparent, auditable and reproducible as possible model, through proper documentation?

A.1.2 Does the developer and the organisation recognize that machine learning models can produce undesired/faulty results for which the organization is responsible?

Name of feature
Accountability
Transparency
Auditability
Reproducibility

Table 5.1: Features to consider in the entire development process.

A.1.3 Is operational accountability assigned at all relevant levels of the organization?

A.1.4 Is the adoption of ML integrated in the organization's risk management framework?

5.2 Data gathering and cleaning

Due to the nature of ML, the data on which it learns ought to be of a high quality. Facets that influence the quality of data are for example bias, missing/faulty labels and other inputs that are not expected. All of these facets will inhibit the quality of the actual model and will curb its accuracy. A model trained on low quality data will always result in a low quality model. Raw data can be messy, duplicated or inaccurate therefore an important part of this process is to explore the available data, then cleanse the data by identifying corrupt, inaccurate, and incomplete data and replacing or deleting it (Oracle, 2020).

Besides data quality there is also the facet of privacy. During the data gathering process, all the data should be anonymised as banking data is highly confidential.

DNB (2019) describes six best practice requirements for handling data:

1. A minimal requirement regarding data quality is defined.
2. Efforts are made on a continuous basis to ensure that data are correct, complete and representative.
3. Special attention is paid to missing or incorrect data-points, potential sources of bias in data, features and inference results(such as selection and survival bias).
4. Procedures and safeguards are in place to maintain and improve data integrity and security during the process of data collection, data preparation and data management.

5. Issues with data integrity and bias, both in development and production are evaluated and documented in a structural manner for future reference.
6. Original datasets used to (re)train and (re)calibrate models are systematically archived.

These six best practices for handling data are solid best practices and lead to the following considerations:

A.2.1 Is a minimal requirement regarding data quality defined?

A.2.2 Is the data anonymised?

A.2.3 Are continuous efforts made to ensure correct, complete and representative data?

A.2.4 Is special attention paid to missing or incorrect data-points, potential sources of bias in data, features and inference results(such as selection and survival bias)

A.2.5 Are procedures and safeguards in place to maintain and improve data integrity and security during the process of data collection, data preparation and data management?

A.2.6 Are issues with data integrity and bias, both in development and production evaluated and documented in a structural manner for future reference?

A.2.7 Are the original datasets used to (re)train and (re)calibrate models systematically archived?

5.3 Feature Engineering

A dataset is a collection of labeled examples which can be written as: $(x_i, y_i)_{i=1}^{N}$. In this collection, each element x_i among N is called a feature vector. A feature vector is a vector in which each dimension $j = 1, ..., D$ contains a value that describes the example. That value is called a feature and is denoted as $x_i^{(j)}$. Each feature contains some information. The specific information will depend on what the model tries to predict and what is available to the developer (Burkov, 2019).

To solve the identified problem, the ML model requires features to predict the required outcomes/classifications. Building these features is called feature engineering, where the modeller will create highly informative features from raw data. These informative features should have high predictive power in translating inputs to actual classifications/labels.

During feature engineering, the modeller should question the features that are created and check the validity of such choices. Do you build features that can inhibit fairness by determining race, sex, or other sensitive features? A modeler should always aim for fair features. However this raises questions such as: “What is fair?” and “Can we leave this features out?”

There are multiple views on how to define fairness, however there are two dominant worldviews that will be explained. The first worldview called: “We’re All Equal”, is under the opinion that all groups have similar abilities with respect to the task, even if we cannot explicitly observe it. To test whether this worldview is violated one can test statistical or demographic parity. The second worldview is that of: “What You See Is What You Get”, which holds that the observations reflect ability with respect to the task. To test this worldview one can apply the equalized odds test (EY, 2020).

5.3.1 Statistical Parity

The statistical parity test is a test that verifies, if a sensitive subset of a group has the same classification probabilities as the group as a whole. This is done by subtracting the probability of the group as a whole from probability of the specific subset. If the absolute value of the remaining number is smaller than the bias of the whole classifier, the classifier has statistical parity with respect to the sensitive subset.

Suppose we have an entire population denoted as set X where there is a sensitive subset $S \subset X$ and a distribution D over X. The distribution D represents the probability of success. Furthermore we have a classifier $h : X \to \{0, 1\}$ which labels X, where 0 is a failure and 1 is a success. The formula denotes the bias will then be:

$$bias(X, S) = Pr[h(x) = 1 | x \in S] - Pr[h(x) = 1 | x \in S^C] \tag{5.1}$$

When S is an unfavourable subset, the left leg of the equation will be smaller and the bias will be negative. Whereas if being part of a subset S, has a favourable influence on the classification than the bias will be positive. Being aware of the sign, provides insights to whether the sensitive subset is better or worse of. Therefore, classifier $h(x)$ is said to have statistical parity on D with respect to S up to bias if: $|bias(X, S)| < \epsilon$ (Kun, 2015). Where ϵ is the bias of the whole classifier.

There are some caveats to using statistical parity in the debate on fairness. If there is statistical parity, there would be no statistical difference in the two distributions based on a specific attribute. However if one of the two distributions is better qualified, e.g. a sensitive subset has a better credit score when applying for loans, this will not be taken into account. This is further explained in Section 5.3.2. Furthermore it can become a self-fulfilling prophecy when malicious people or companies will actively

skew their results. This can for example be done by providing loans to people who can't afford it to afterwards, when they are unable to repay them, point to these people that they are justified in discriminating these people (Kun, 2015). However, it can be appropriate to use as a test to check the model for bias when on features that should not exhibit a difference in their distribution.

5.3.2 Equalized Odds

Equalized odds is a statistical notion of fairness that ensures that classification algorithms do not discriminate against protected groups (Gölz, Kahng, & Procaccia, 2019). Equality of Odds is different to statistical parity in that it obtains a sense of fairness without disregarding a protected attribute. Instead, it enforces an equal True Positive Rate and a True Negative Rate of all groups. Therefore not exempting differences in protected groups. Suppose there are two groups of people applying for a loan. These groups are different in that on average group A has a better credit score than group B and therefore more people from group A are qualified for the loan. Equalized odds is satisfied when someone irrespective of group A or B, if they are qualified are equally as likely to get a loan. Moreover if they are not qualified, they are equally as likely to not get a loan (Google Machine Learning Glossary, n.d.). This can be formally noted like Equation 5.2 as proposed by Hardt, Price, and Srebro (2016). In this equation, $\hat{Y}$ denotes a predictor of the target variable Y where the variable A denotes whether a person is qualified or not.

$$Pr[\hat{Y} = 1|A = 0, Y = y] = Pr[\hat{Y} = 1|A = 1, Y = y], \quad y \in [0, 1] \tag{5.2}$$

A variety on equalized odds is called Equality of Opportunity. Equality of Opportunity is a more relaxed version of equalized odds where the only metric to adhere to is the True Positive Rate. The concept is, for most cases there is a preferred classification label and the rest is not of importance. Considering the aforementioned loan example, equality of Opportunity is satisfied when the same percentage of qualified people are granted a loan irrespective of which group they belong to.

Equalized odds and equality of opportunity are interesting notions of fairness however there is still room for debate. Fairness is used to close the gap between two or more groups. However, equalized odds might not be of help to close it. A group might have a privileged position such that it has more qualified people. When they are granted the same percentage of loans, given that they are qualified, as the on average less qualified group will not close the gap. Actually, the gap between group A and group B will tend to enlarge over time (Zhong, 2018).

5.3.3 Unfairness by proxy

Given the aforementioned views on how to handle fairness raises the question whether it is possible to exempt models from using sensitive features entirely. Although a straightforward thought, this will not solve the problem of fairness due to the (at least) three sources that cause unfairness. These three causes are: prejudice, underestimation and negative legacy (Kamishima, Akaho, & Sakuma, 2011).

The first cause of unfairness is prejudice which involves the statistical dependency between a sensitive variable S, the target variable Y and a non-sensitive variable X. The way these dependencies can manifest are again in three ways, direct prejudice, indirect prejudice and latent prejudice. Direct prejudice is the use of a sensitive variable in a model, resulting in direct discrimination. Removing direct prejudice can be done by removing the sensitive variable. Indirect prejudice, however, occurs when the target variable Y still portrays the influence of S. This correlation can still be present as S had an influence on Y during data collection. However when there is no correlation between the sensitive variable and the labels, there is no scope for bias and thus no indirect discrimination. The indirect prejudice is defined as the mutual information between Y and S, which can be calculated as seen in Equation 5.3 (Kamishima et al., 2011). The letters y, s and x denote the different datapoints within the variables Y, S and X.

$$IP = \sum_{y,s} Pr[y,s] ln \left[\frac{Pr(y,s)}{Pr(y)Pr(s)} \right] \tag{5.3}$$

Latent prejudice occurs when there is a correlation between a non-sensitive variable and the sensitive variable. The indirect prejudice is defined as the mutual information between X and S and can therefore be calculated using Equation 5.3 when swapping out the target variable Y with a non-sensitive variable X. If a correlation would be present, the non-sensitive variable would, to the ML, function as the sensitive variable by proxy.

The second cause of unfairness is due to underestimation. When underestimation occurs the learned model is not fully converged due to the finiteness of the training dataset. Given a learning algorithm without an indirect prejudice, it will make a fair determination if infinite training examples are available (Kamishima et al., 2011). The third cause is that of negative legacy, meaning that the sampling or labelling of the training data has been unfair. This can occur due to biased sampling such as selection bias. However detecting and correcting negative legacy is very difficult unless there is different information such as a smaller-sized fairly labeled dataset to exploit (Kamishima et al., 2011).

Example Indirect and Latent Prejudice

To show how indirect and latent prejudice works, the different prejudice are calculated in a fictive dataset of an ML model that classifies fraudulent transactions. The dataset is based on two variables and one target variable Y. The first predictive variable, which is deemed sensitive, is gender. The second variable denotes, whether the person making the transaction has previously made transactions at a nail salon. Both gender(denoted as: S) and nail salon purchase(X) are binomial variables and their distributions are denoted in Table 5.2. Where $S = 1$ refers to the male gender and $S = 0$ to the female gender. Applying the formulas for indirect and latent prejudice(equation 5.3), provides $IP = 0.22$ and $LP = 0.11$. This means that gender has an impact on Y, the actual classification. However the Latent prejudice between X and S is quite low. Therefore, the ML model cannot use the variable of nail salon purchases as a proxy for gender.

Probability	
$Pr(s)$	0.70
$Pr(x)$	0.40
$Pr(y)$	0.60
$Pr(y = 0, s = 0)$	0.40
$Pr(y = 0, s = 1)$	0.10
$Pr(y = 1, s = 0)$	0.20
$Pr(y = 1, s = 1)$	0.30
$Pr(s = 0, x = 0)$	0.20
$Pr(s = 0, x = 1)$	0.10
$Pr(s = 1, x = 0)$	0.60
$Pr(s = 1, x = 1)$	0.10

Table 5.2: Fictive data distributions

Mutual information and Entropy

The formula's on indirect and latent prejudice are based on the mutual information between two variables. Mutual information in turn is based on Entropy from the field of information theory. The formula of Entropy, which is denoted as H(X), can be explicitly written as shown in Equation 5.4. Within information theory, entropy is a measure of information and uncertainty (Shannon, 1948). Using a binary classifier as an example, most information is provided when the probability of a classification is 0.5. When the probability deviates from 0.5, each time a classification is done, one classification is more likely than the other. Therefore there is less uncertainty

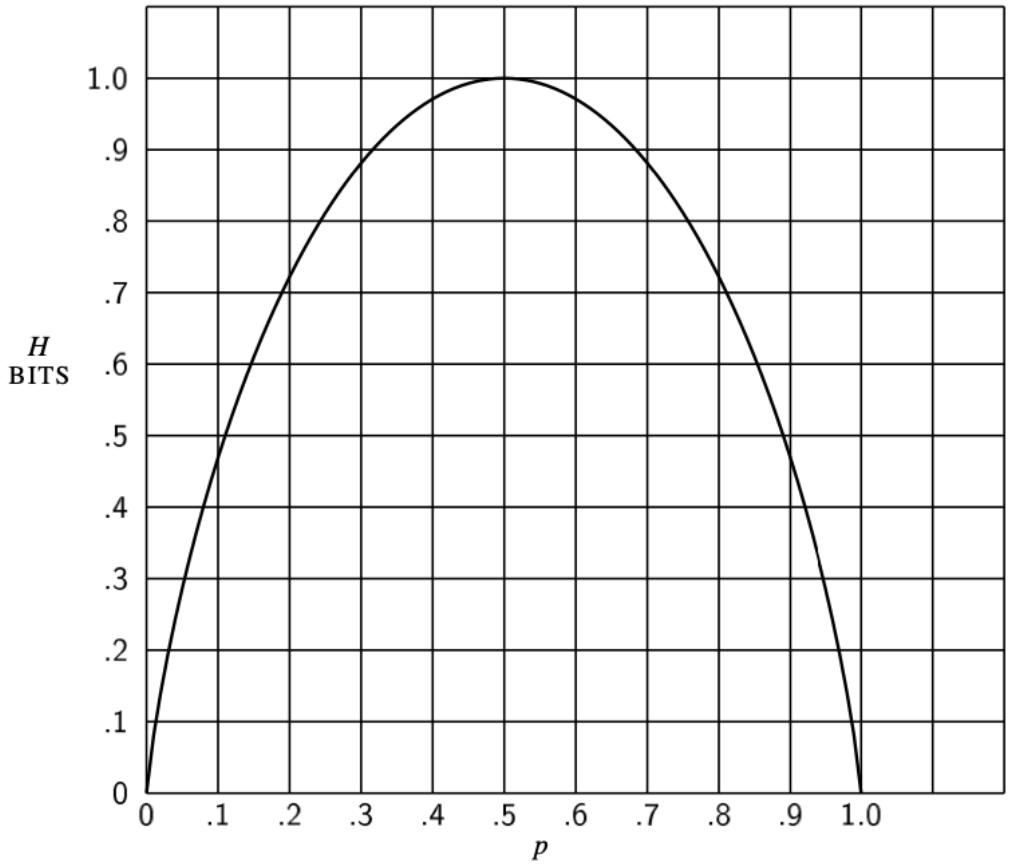

Figure 5.1: Entropy in the case of a binary classifier. (Shannon, 1948)

and by definition, less information is provided by the classifier. This can be illustrated by Figure 5.1, where any deviation from 0.5 results in a lower Entropy(Denoted with H). The Indirect- and Latent prejudice formula's use the concept of adding/removing information through checking whether the union of two distributions is equal to the multiplication of the two distributions. Suppose there is a target variable Y and a feature variable S. If $Pr(Y, S)$ is equal to $Pr(Y) * Pr(S)$, then by definition the two distributions are independent. Therefore when the distributions are independent, the logarithmic term in Equation 5.3 and thus the entire formula will evaluate as 0. As it is 0, there is no prejudice between the evaluated variables. When the distributions are not independent, there is a change in entropy and therefore an advantage to a specific class.

$$H(X) = -\sum_{i=1}^{n} Pr[x_i] ln\big(Pr(x_i)\big) \tag{5.4}$$

The entropy of a single variable in the case of an ML model is not necessarily helpful as the goal is to predict a target variable from a dataset. Therefore relative entropy is

Kullback-Liebler Divergence	
$D(Y\|\|S)$	0.023
$D(S\|\|Y)$	0.022
$D(S\|\|X)$	0.18
$D(X\|\|S)$	0.19

Table 5.3: Kullback-Liebler Divergence metrics, based on the example in section 5.3.3

a concept that is of more use. Relative Entropy is coined by Solomon Kullback and Richard Leibler and therefore more often referred to in the literature as the Kullback-Leibler(KL) divergence (Kullback & Leibler, 1951). KL divergence can be used to calculated how one distribution is different from another. KL divergence, however, is an asymmetric measure. Therefore it does not qualify as a metric of spread and the order of the two distributions needs to be taken into account. In Equation 5.5 the divergence between two distributions, P and Q, is calculated. In the notation: $D_{KL}(P||Q)$, distribution Q is used to approximate distribution P. The larger the value on KL's domain: $[0, \infty)$, the larger the separation between the two distributions.

In the example used in Section 5.3.3, the KL divergence can be calculated to get a sense of the difference between the distributions of being a fraudster(Y), gender(S) and Nail salon purchases (X). The KL divergences that resemble indirect and latent prejudice are displayed in Table 5.3. Important to note is that the order of the divergence levels coincide with the levels calculated for the various forms of prejudice. KL divergence is a good metric to determine order between how different one distributions is with respect to another. However, it is difficult to use a metric to determine how different two distributions are.

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) log_b \left(\frac{P(x)}{Q(x)} \right) \tag{5.5}$$

5.3.4 Adjusting for fairness

From a governance perspective, it is of utmost importance for the ML models to treat people fairly. Since there are different mathematical notions of fairness, e.g. demographic parity versus equalized odds, modelling decisions regarding fairness should be made on a case-by-case basis. The decisions should documented and ideally approved by the risk department.

There are multiple ways of handling the found imbalances such as manipulating labels, reweighting samples, a regularization method proposed by Kamishima et

al (2011) training with an adversarial machine and more. This choice is up to the developer. However, the type of fairness metric that is used and the used technique to solve it should be documented appropriately.

A.3.1 Does the data contain discriminatory features based on demographic parity or equalized odds?

A.3.2 Does the dataset contain unfair features by proxy i.e. due to prejudice, underestimation or negative legacy?

A.3.3 Is the chosen fairness mitigation approved by the risk department?

5.4 Dataset Splitting

After having gathered, cleaned, and enriched the dataset using feature engineering, an actual model can be run. To do this, the dataset should be split into three different sets: a training dataset, a test set and a hold-out dataset (Burkov, 2019). There is no optimal split for these datasets. However, generally the training dataset is considerably larger where the test and hold-out dataset are similar in size. The proportion can range from 70% training, and 15% for both the test and hold-out set to 95% training to 2.5% for the test and hold-out set if the dataset is large enough(millions of examples). However, when splitting data to train, test and validate, the developer should keep in mind that the data distributions of the various sets of data might not be the same due to data sparsity and how the split is done.

Two main types of dataset shift that can occur due to the aforementioned causes, Covariate Shift and Prior Probability Shift. Covariate shift occurs when the input distribution of the two sets are different whilst their is no change in the underlying relationship between input and output (Stewart, 2019). It can be formally noted as can be seen in Equation 5.6. In Equation 5.6 and 5.7, the training and test probability functions($P_{training}$ and P_{test} respectively) are dependent on the dataset that has been used for training. Furthermore, x is an input element with the predicted value y.

$$P_{training}(y|x) = P_{test}(y|x) \quad \text{and} \quad P_{training}(x) \neq P_{test}(x) \tag{5.6}$$

Prior Probability shift is essentially the reverse of the Covariate Shift as it does not focus on a change of the input distribution but a change in the output distribution. Prior Probability shift can be formally noted as in Equation 5.7. However, to get a sense of how this would manifest in a model, consider a spam filter which trains on data where 50% of its examples are spam. If in reality(or the test set) 90% of the mails are considered spam, the prior probability will have changed considerably.

Name of feature
Bias Reduction
Fairness/Inequality in application
Explainability
Accuracy

Table 5.4: Features to consider during model selection and hyper parameter tuning.

Therefore inhibiting the models functionality.

$$P_{training}(x|y) = P_{test}(x|y) \quad \text{and} \quad P_{training}(y) \neq P_{test}(y) \tag{5.7}$$

A.4.1 Is your data split suitable for the dataset's size?

A.4.2 Are the distributions in the training, test and validation similar or has a dataset shift occurred?

5.5 Model Selection and Hyper parameter tuning

During the model selection process it's important to come back to the problem at hand and determine which model is suitable for it. From a ML governance aspect, it is up to the developer to determine which model will be best suited for the problem at hand. During the selection phase, it is wise to consider multiple modelling techniques after which those are narrowed down. The process of narrowing down should be documented and it is important to describe what selection criteria will be used during e.g. a Model performance Assessment. In a Model Performance Assessment one could test multiple models on the validation set and choose the model based on the predefined selection criteria. These predefined selection criteria will be situation specific and should be decided upon with the relevant stakeholders.

In a model performance assessment, there is a distinction between a regression model and a classification model. To test whether a model is better than another, it is possible to test the models on the hold-out data and review those against the outcomes in the test set. If those models predict well on both the training and the test data, it means that the model generalizes well and is in other words a good model. This can be done by using a loss function, like the MSE. When the MSE value is considerably higher than on the test data, it can be a sign of overfitting and therefore facets of your model should be adjusted such as: the hyperparameters, regularization or a different form of regression. Note, 'considerably higher' is not defined in the literature and is defined by the developer, is situation specific and should be documented.

For classification algorithms the process is less straightforward than with a regression as there are more facets to consider. According to Burkov (Burkov, 2019) the most widely used metrics, also named Key Performance Indicators (KPI's), and tools to assess classification models are:

1. Confusion matrix,
2. Accuracy
3. F-Score
4. Cost-sensitive accuracy
5. Precision/recall
6. Area under the ROC curve

Using the confusion matrix it is possible to tune your model to what is required by the situation. Distinguishing between minimizing a type I(False Positive) or the type II error(False negative) is one of these decisions. These errors will decrease when the model generalizes well. However at some point the decrease in one will result in an increase in the other. Therefore distinguishing which is most important to the problem should be defined and taken into account. In order to objectively state whether the results are suitable, it is good practice to have predefined KPI's and their required levels.

To find the most suitable hyperparameters, e.g. learning rate and kernel type, multiple models with different hyperparameters should be trained and tested. Consider the fact that models in a bank are often run on enormous datasets such as transactional data. For such situations, hyperparameter tuning can be a costly operation. Therefore it is required to think about how the parameters are tuned, for example using random search or Bayesian Hyperparameter optimization. For future reference and retraining it is good practice to document how the hyperparameters are optimised.

From an oversight perspective it is important to have insights into why a model makes the decisions that it makes. Explainability in AI(XAI) can be of help in this process. As described in Chapter 3, a good explanation answers a contrastive question about the prediction that is tailored to a specific audience. Such explanations will help in determining the functioning of the model and the outcomes. However there is a significant discrepancy between the vision of explainability and how it is being incorporated in practice (Newman, 2021). XAI is a rapidly evolving field of research but it is not yet ready to give sensible recommendations for the framework. It is however recommended to closely monitor the developments in this field.

To mitigate the accuracy and inequality in application the developer should analyse the model they intend to use thoroughly. Question the model thoroughly, does this model represent the problem domain? Is the model balanced regarding facets such as race, gender or other aspects?" To aid the answering of questions about the model it is possible to use fairness tests as described in Section 5.3.1, 5.3.2 and 5.3.3.

A.5.1 Are the Key Performance Indicators and their required levels predefined?

A.5.2 On what criteria/metrics is the final model selected?

A.5.3 In comparable situations, is the best explainable model chosen?

A.5.4 What are the hyperparameters that should be optimised?

A.5.5 What method is used to optimise the hyperparameters?

5.6 Validation

In the validation stage the trained model is tested on the hold-out dataset to see how well it performs on new data whilst having optimised hyperparameters. The metrics derived through this test are the metrics that should be reported to DNB and auditors.

A.6.1 Are the final metrics reported?

CHAPTER 6

Framework: Deployment Phase

During deployment, the built model in the development phase needs to be adjusted and monitored to get it in production and to actually provide value. This deployment section is split up in three phases, namely: preliminary testing in a copy of the production environment, operating the model in a shadow environment and operating the model in the production environment. Following these steps are important for ML governance as it provides the bank time to monitor the operations and to abort or failover when needed.

6.1 Pre-Production

The initial thing that should be thought about in the deployment phase is that of how the predictions of the built model should be interacted with. Important facets to think about in this section are speed, when the prediction is done and how it's returned. To illustrate this, consider the transaction monitoring system, the classification should be done afterwards to monitor behaviour and initiate investigations when it's out of the ordinary. Since it is done afterwards, the model has time to predict the outcome. If it is classified as fraudulent, the compliance team is alerted. However, if the model was a transaction filtering model instead of transaction monitoring, the model needs to stop the payment when it is classified as fraudulent. The classification model should work very fast or near instant as it will otherwise slow down the payment process which is undesirable. Thus the deployment of a different model will require a different environment to interact with.

B.1.1 How will the model interact with the environment and what is needed to ensure that it works?

6.2 Testing & Shadow Mode

Having figured out how the model will interact with the production environment it is a sound idea to test the functionality on a test environment. This test environment is ideally a copy of the production environment. When a model works in this environment one can then reasonably assume that it will also work in the production environment.

After having tested the model in a testing environment it is time to deploy the model into the production environment. However, prior to the model taking over work from previous models, it is good practice to deploy the model in a 'shadow mode'. In shadow mode, the model is deployed in the production environment where the data is simultaneously run through the regular system and the new model. The responses from the old system are used to serve responses and predictions where the responses from the new model are captured and stored for analysis (Samiullah, 2019).

B.2.1 Does the model work as intended in a test environment?

B.2.2 Does the model work as intended with production data whilst in shadow mode?

6.3 Transition to production

When the model has proven itself in shadow mode it is time to flip the system such that the new model predictions/responses are used. When a model is put to work in a production environment, the model infers a prediction based on the training and tuning it has had in the development phase. Therefore, in the process of ML inference, it is still possible to see the model stray from the accuracy found during validation. This is due to the fact that the model is trained on historic data and the inference is made on future data. Expecting to get the same results would therefore be an assumption that often will not hold (Patruno, 2019). Therefore it is imperative to keep the old system available such that it is possible to transition back to the proven and tested system.

B.3.1 Does the model work as intended whilst being deployed in production?

CHAPTER 7

Framework: Post Deployment

An often overlooked and underrated phase in the world of ML is post deployment. The Post Deployment phase is important due to the nature of machine learning i.e. a model which is trained on historical data whilst inferring future outcomes assumes a stationary process. However, rarely are processes stationary. Therefore it's possible for an ML model to diminish in effectiveness when the history does not perfectly translate to the future.

7.1 Monitor the Key Performance Indicators

In Section 5.5 the KPI's and their required levels are defined. During the post deployment phase it is important to monitor the KPI's and to take action when they approach the defined minimum values. Possible culprits are data- and concept drift which are elaborated on in Section 7.2 and 7.3.

C.1.1 Are the KPI's still within the required range?

7.2 Data Drift

There are multiple types of drift that can occur in an ML model. The first drift concept, is data drift. During data drift, something unforeseen has happened to the data pipeline. When the data fed to the system changes due to e.g. an alteration in the data collection process, the efficiency of the model will probably decay as it is not trained for that situation. To combat data drift, it's important that the data pipeline is routinely checked for alterations. Besides routine checks, it's important to check the incoming data when an extra source is added to the model. For instance in the case of transaction monitoring, it is necessary to check whether the data from a new form of payment has the same distribution as the data on which the model is trained. If this is not the case, adjust the data, the pipeline or the model to decrease the model decay.

Addendum 4 by DNB (2019) focuses on the need for emphasis on the correctness of the data pipeline and thus on data drift. This addendum is also cited in Section 5.2 where the focus is on ensuring that correct data is used in the starting phase. In Recommendation A.2.1 the minimal data quality is defined. When the data quality does not adhere to the definitions, one can assume the occurrence of data drift and action should be taken. Besides reactive checking whether the data quality is up to par, proactive and continuous efforts should be made to ensure that the data is correct, complete and representative. Finally, issues regarding data quality and bias, both in development and production, should be documented in a structured manner for future reference.

C.2.1 Routinely check if alterations have been made in the data pipeline.

C.2.2 Proactive measures should be taken to ensure correct, complete and representative data.

7.3 Concept Drift

During concept drift a model decay is present just like with the data drift, however there is difference in the root cause of it. In concept drift, the statistical properties of the target variable which the model tries to predict change over time (Widmer & Kubat, 1996). Essentially the real world is changing and the model is not trained for it. Therefore the model should be retrained for the real world changes. In a model such as Transaction monitoring, the goal is to detect money laundering. Therefore, monetary gain is to be had for people who can circumvent the transaction monitoring system. Due to the monetary incentive to game the system, there is a high likelihood of concept drift.

DNB (2019) points out, in addendum 2, that ML models should be periodically retrained, recalibrated and assessed, especially in the event of significant changes in the input data, relevant external factors and/or in the legal or economic environment. The aforementioned situations are all examples of events that can cause a separation between the real world and the concept of the model. Criteria for when these changes are significant as well as other fail criteria should therefore be documented for each ML model.

To check whether there is concept drift occurring, broadly the four stage process in Figure 7.1 should be followed. Firstly the data should be split in two batches which will be compared. Secondly, the data should be abstracted to retrieve the key features that impact the system most when they drift. This stage is optional as it mainly concerns dimensionality reduction, or sample size reduction, to meet storage

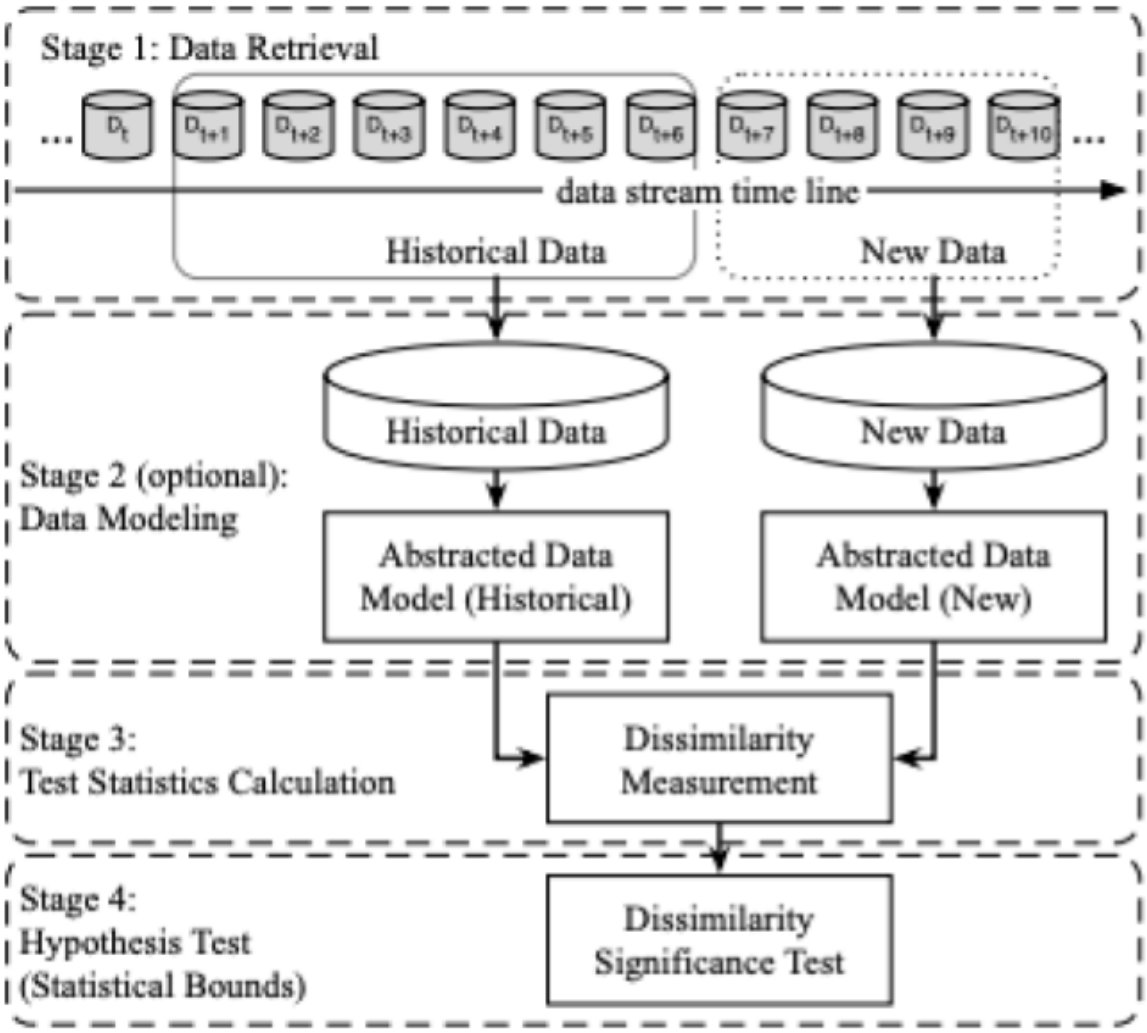

Figure 7.1: Four common stages to check for concept drift (Lu et al., 2018).

and online speed requirements (Liu, Song, Zhang, & Lu, 2017). In the third stage, the dissimilarity should be measured in a test statistic, where in the fourth stage a hypothesis test should be done to check whether the dissimilarity is significant (Lu et al., 2018). Test statistics to consider are the basic variance and mean but also divergence and distance tests such as: Kullback-Leibler divergence, Kolmogorov-Smirnov statistics, Population Stability Index(PSI), Hellinger distance and so on. To test categorical variables it's possible to use the chi-squared test or entropy (Oladele, 2021).

C.3.1 Routinely check if there are features that display significant concept drift.

C.3.2 All criteria for significant changes as well as other fail criteria should be documented to assess whether retraining is necessary.

7.4 Fall Back plans

If the models fail to work as intended it is critical to have fall-back plans to revert to. A common way to do this is by reverting to a rule-based system or to a previous version of the model that has proven to work. To revert to a previous version, having a proper version history in place where aspects like changes, reasons, used training data and models are documented is key.

C.4.1 Are back up plans in place?

C.4.2 Is the version history properly documented?

Phase	Label	Recommendation
Development		
	A.1.1	Produce an as transparent, auditable, and reproducible as possible model through proper documentation.
	A.1.2	Realize that machine learing models can produce undesired/faulty results for which the organization is responsible.
	A.1.3	Assign operational accountability at all relevan levels of the organisation.
	A.1.4	Integrate AI in the organization's risk management framework
	A.2.1	Define a minimal required data quality
	A.2.2	Anonymise the data.
	A.2.3	Make continuous efforts to ensure correct, complete and representative data.
	A.2.4	Pay special attention to missing or incorrect data-points, potential sources of bias in data, features and inference results (such as selection and survival bias)
	A.2.5	Procedures and safeguards are in place to maintain and improve data integrity and security during the process of data collection, data preparation and data management.
	A.2.6	Issues with data integrity and bias are both in development and production evalutated and documented in a structural manner for future reference.
	A.2.7	Systematically archive the original datasets which are used to (re)train and (re)calibrate models
	A.3.1	Determine whether the data contain discriminatory features based on demographic parity or equalized odds.
	A.3.2	Determine whether there are unfair features by proxy e.g. due to prejudice, underestimation or negative legacy.
	A.3.3	Get approval from the risk department on the chosen fairness mitigation.
	A.4.1	Determine whether the data split is suitable for the set's size.

Phase	Label	Recommendation
	A.4.2	Determine if the training, test and validation set have similar distributions or whether a data shift has occurred.
	A.5.1	Predefine the Key Performance Indicators and their required levels.
	A.5.2	Predefine the criteria and metrics on which the final model will be selected.
	A.5.3	Choose the best explainable model in comparable situations
	A.5.4	Determine what hyperparameters should be optimised.
	A.5.5	Determine what method to optimise the hyperparameters is most suitable.
	A.6.1	Report the final metrics.
Deployment		
	B.1.1	Determine how the model will interact with the outside environment and what is needed to ensure that it works with the outside environment.
	B.2.1	Determine whether the model works as inteded in the test environment.
	B.2.2	Determine whether the model works as inteded with production data whilst in shadow mode.
	B.3.1	Determine whether the model works as inteded whilst being deployed in production.
Post Deployment		
	C.1.1	Monitor whether the Key Performance Indicators are still within range.
	C.2.1	Routinely check if alterations have been made in the data pipeline.
	C.2.2	Proactive measures should be taken to ensure correct, complete and representative data.
	C.3.1	Routinely check if there are features that display significant concept drift.
	C.3.2	All criteria for significant changes as well as other fail criteria should be documented to assess whether retraining is necessary.
	C.4.1	Have back-up plans in place.
	C.4.2	Properly document the version history.

CHAPTER 8

Transaction Monitoring Case Study

RQ4:"Are there ML governance shortcomings in the Transaction Monitoring system?"

In this chapter the framework originating from Chapter 5, 6 and 7 is applied to the transaction monitoring model to establish ML governance shortcomings.

8.1 Background

Bunq is obliged by law to perform its gatekeeper function where it should report fraudulent, money laundering and the financing of terrorism to the FIU (De Nederlandsche Bank, 2020a). Failing to catch criminal flows of money can result in fines and even legal prosecution if the transaction monitoring is not done well enough. One of the main ways that banks use to catch unwanted flows of money is monitoring of transactions that flow through the bank. A simple rule-based system can be used to detect fraudulent transactions. However the initial simplicity comes at a cost. Due to the ingenuity of fraudsters, banks have to constantly improve their fraud detection systems. Consistently improving rule-based systems is labor intensive, brings high numbers of false positives and increases complexity over time (el Hassouni, 2016). Therefore bunq has developed a system of ML models to determine fraudulent behaviour based on previous made transactions.

Each transaction has certain features that are available during the occurrence of a transaction such as: payment amount, payee, GPS location of the payer, the clients' age etcetera. Besides features generated during the occurrence of a transaction, more features can be made available by updating or calculation. Examples of these calculated features are the number of outgoing transactions in the last 24 hours, the number of cross-border payments in the last 4 days or the number of bitcoin related transactions in the last 12 hours. Using the aforementioned features, a model determines whether the transaction is either common or unusual after which a compliance agent will look into it.

8.2 Findings

A whole host of information on the transaction monitoring system can be found on an internal web-based workspace. The information available provides the done research, why decisions have been made and how to interact with it. Applying the framework to the available information presented some interesting findings and also shows that a lot of work has been done to establish proper ML governance. The framework items which are not found in the documentation are visible in Table 8.1, the full framework and the results are visible in Appendix 10.4.

Check nr.	**Recommendation**
A.1.4	Integrate AI in the organization's risk management framework
A.2.6	Issues with data integrity and bias are both in development and production evalutated and documented in a structural manner for future reference.
A.3.2	Determine whether there are unfair features by proxy e.g. due to prejudice, underestimation or negative legacy.
A.3.3	Get approval from the risk department on the chosen fairness mitigation.

Table 8.1: Not done recommendations

From the 33 items in the framework, 4 are not found. The first recommendation that is not found, is A.1.4. ML is not incorporated as a separate section in bunq's risk framework which poses a problem to bunq's ML governance. The second and third not available framework items have to do with how bias is handled. The transaction monitoring system is supposed to handle bias by not using sensitive variables in the model. However in Section 5.3.3, it is described that bias can still be in a model even though the sensitive variable is removed. Therefore, issues with bias are not properly evaluated and documented in a structural manner. Furthermore no checks are done to rule out the presence of unfairness by proxy. The fourth item that is not available is regarding the approval and the documentation of the fairness mitigation policy. At the time of building the TM model, this item was not a standard policy and therefore is not considered. Fairness is, however, considered in the building of the TM process. All sensitive variables are removed from the dataset which is an important step. However as is found out in Section 5.3.3, solely removing the sensitive variables is not always sufficient.

8.3 Conclusion

Overall a lot aspects of the transaction monitoring is documented and proper ML governance is taken into account. However there are some aspects that need attention. Firstly, in terms of handling data integrity, bias and the documentation thereof. A lot is to be gained when a process and policy on how to handle it would be made. Secondly, an item on how ML is handled within the risk management framework is necessary. Thirdly, approval of the risk department on how fairness is mitigated should be done. Fourthly, no check has been done to determine whether the model contains unfair features by proxy. Lastly, the documentation consists of too many different files or are not documented while the problem has been addressed. Therefore, one master file with the most important aspects, numbers and references would do the ML governance more justice.

8.4 Recommendations

For proper ML governance, some items need to be handled. First and foremost, ML should be taken into account in the risk management framework as it is imposed by DNB. Secondly, a policy should be set up on how potential bias issues are handled and documented. This provides knowledge on potential downsides of the model and more importantly provides grounds when being audited as it shows that it is taken into consideration. Thirdly, the way fairness has been handled in the past should be retroactively checked to determine whether omitting sensitive variables has been a sufficient mitigation. Fourthly, a better structured master document. In such a master document, important aspects numbers and references to specific documentation should be made.

CHAPTER 9

Conclusions and recommendations

9.1 Conclusions

In this section the research will be concluded. The general outline of this section will be as followed. First, an introduction explains which research questions provide information for the built framework. This will be followed with a paragraph per research question, discussing how the research question has come about and what can be concluded. The penultimate section is on the framework and discusses what the impact of it is and how it will evolve in the future. The final section considers the limitations of this research.

Unnoticed, ML is becoming an important part of everyone's everyday lives and has become instrumental in the financial markets. Therefore challenger banks employing ML techniques in their operations have a duty to society to handle these models with care and skepticism.

The goal of this thesis was to develop a framework to facilitate ML governance with the purpose to provide guidance for banks to become compliant and mature the internal control environment of their ML processes, this framework can be used to self assess ML governance.

In order to create the framework, information needs to be gathered on ML governance within the banking sector. Three research questions are answered to gather information that is needed for the framework. The conclusions on these research questions are provided below. A separate fourth research question is specific to bunq, where research is done to find out whether there are shortcomings in the ML governance of the transaction monitoring system.

In the first research question, an answer is sought to the question: *"What is, according to the literature, important in the governance of ML models?"* The results from this question are used as input for the framework. It can be concluded that the ML field is very extensive, and there is a discrepancy between authors on the impor-

tance of certain aspects. All authors, however, agree on four core principles, an ML model should be fair, transparent, documented and accountability should be defined in case of faulty outcomes. Furthermore, a list of aspects is created and explained. These aspects form the basis of the framework.

The second research question is focused on providing more clarity around the term explainability with the research question: *"How is explainability in a machine learning context defined?"* Explainability is an often used term in the publications of DNB, however an agreed-upon definition within the field of ML is not clear-cut. This makes it difficult for a bank to follow DNB's publications. To answer the research question, literature research has been performed and an interview with the 'senior policy advisor Artificial Intelligence' of DNB has been held. When analysing the information based on the definitions, we can conclude that there are two key factors that define explainability in a machine learning context. These factors are: 1) a good explanation and 2) a specific audience. Where a good explanation is able to explain a contrastive question, as it is found that people generally ask contrastive questions in the form: *"Why P rather than Q?"* (D. K. Lewis, 1986) Lastly, the audience is important as different people want to know different aspects of a model. For example, a developer might want to know minute details about the answer of a ML model whereas a regulator might want to know if the predictions abide by the rules.

To conclude the third research question, the following question is answered: *"What is according to the Dutch Central Bank important in the development and monitoring of ML systems in the financial service industry?"* In this section, all of the publications are analysed one-by-one to determine the ML aspects that DNB finds important. From the publications that DNB has done, the 'General principles for the use of Artificial Intelligence in the financial sector' (De Nederlandsche Bank, 2019) is the most important publication to consider when aiming for compliant operations. In this publication their SAFEST principle is coined, which is an acronym for: 'Soundness', 'Accountability', 'Fairness','Ethics', 'Skills' and 'Transparency'. Based on this acronym DNB state their stance on AI. Based on their second and third publication, it can be concluded that DNB is uncertain about their role as they do not want to inhibit technical progress. However they do want all financial institutions to be able to show that they are aware of their responsibilities and that modelling choices and uncertainties should be documented. The fourth publication provides no clear stance for DNB as it is an exploratory study where they try to find support for better AI control from within the banking sector.

From the fourth research question, *"Are there ML governance shortcomings in the Transaction Monitoring system?"*, can be concluded that there are shortcomings. When vetting the TM system against the framework, four shortcomings are found.

Firstly, there is no ML section present in the organization's risk framework. Secondly, bias is handled by removing sensitive variables. However it is shown that the technique might not suffice and that indirect and latent prejudice should be checked. Thirdly, there is no documentation on data integrity and bias issues. Fourthly, it is recommended by the framework that fairness mitigations need to be approved by the risk department, which has not been done. Lastly, although a lot of documentation is present the information is scattered across different documents which is unwanted for the governance of ML models.

To conclude the research, a framework to facilitate a better ML governance is developed. This framework can be used by challenger banks as a steppingstone to improve ML governance. ML governance is a very important and highly watched field as society demands improved control over the functioning of models, especially machine learning algorithms. The framework(Chapter 5, 6 and 7), which is the most important part of this research contains three large phases i.e. 'Development', 'Deployment' and 'Post Deployment'. Within these phases you can find subphases which describe, information and recommendations on ML governance aspects within the development process on chronological basis. These recommendations vary greatly and includes topics such as: what should be in the documentation, various forms of fairness, mutual entropy, unfairness by proxy and more. By using this framework bunq can improve their ML governance as the framework will enable the use of self-assessments, the outcomes will help bunq by setting and maintaining a ML modelling standard.

The framework provides a generic approach in its application, but will always need tailoring to the specific environment where it will be used. The important advantage in using a framework is the structured approach which is beneficial for challenger banks because of the following reasons. Firstly, the structured approach can provide the guideline during audits and move the discussion based on individual cases to a broader discussion about the generic framework (objectives vs. mitigations). Secondly, a framework will fundamentally improve the process of model development and the change management process. This will lower the dependency on specific people during the model development process. Lastly, the framework guide the model developer to continuously follow up on questions, enforce writing documentation and mitigate the risks of various aspects of ML models in a structured manner. However, this framework will be subject to continuous improvements as new information will be published and technology will evolve.

The field of ML governance is getting traction and will not stop evolving anytime soon. There is however an important limitation in this research thesis. Due to the wide application of ML models there are a lot of different aspects that could be

considered, which in turn hinders the research depth. However, since this is the first version of such a framework within bunq, it is chosen to focus on a broad scope. It would for example be possible to develop an entire framework on just explainability, let alone all other aspects. The limitation can be resolved in a following research project when DNB has provided their expectations and/or findings during an audit. Based on their remarks and expectations it will become possible for a new student to specify the focus, thereby improving and tailoring the framework.

9.2 Recommendations

The field of ML governance is quite novel based on the recent publication dates of the reviewed papers. However, a large amount of the research is still in the phase of figuring out what is wrong. The expectation is that in the upcoming years more regulations, frameworks and techniques will be developed to further structure how companies and society deal with ML. In my opinion, to build further on this research, it would be good to translate the theoretical items such as prejudice, fairness and data/concept drift into monitoring software that can be applied to different forms of ML systems. Using such software would enable better control, faster interventions and possibly quantification of certain aspects.

ML techniques aim to find correlations in large datasets to infer various outcomes such as classifications or predictions. However this generally accepted practice might be a large blind spot for developers, as correlation does not imply causation. Therefore ML developers need to be aware and should put extra effort into proving relationships between variables, as there is a chance that a third variable is the actual reason why two variables are correlated. Therefore I recommend to perform more research into methods that provide insights on causation. Having more insights on causal relationships will in turn also provide a model which can explain better, why a certain prediction has been made.

Explainability in a machine learning context is not well-defined due to the variance of all possible explanations. In my opinion the field of Explainability is not mature enough to provide a post-hoc model that works with the vision of explainability. Techniques such as Shapley and LIME are great first steps but these techniques are not good enough to provide true explainability. Therefore I recommend to focus on getting the basis i.e. ML governance right whilst keeping an eye out for big changes in the field of explainability when it matures.

Recent advances within the field of ML has enabled the use of a ML technique called: Generative Adversarial Networks. Within the field of Generative Adversarial Networks it is possible to generate new data which has the same distribution as the

original dataset. This allows developers to create more data to further train their ML models, essentially pushing the ML model to a desired state. However this poses ML governance issues as bias from an external model can be imposed in the new data. Therefore more research on bias and other potential downsides of synthetic generated data using Generated Adversarial Networks needs to be performed.

DNB has started a cooperation with the banking industry in the form of: “iForum”. I, however, recommend to improve cooperation and information sharing between banks so learnings and roadblocks on this relatively new topic are shared without regulatory consequences. This could be established by experimentation. There are examples of regulators who provide sandbox test environments for their members to experiment with innovative technology. The expectation is that such environments will fuel ML progress and the ML governance thereof.

References

Aerens, P. (2019, 9). *Trust is the only product banks have left.* Retrieved from https://medium.com/swlh/trust-is-the-only-product-banks-have-left-d31a7ba2ea5c

Agency for Healthcare Research and Quality. (2010). *What makes a good checklist?* Retrieved from https://psnet.ahrq.gov/perspective/what-makes-good-checklist

BaFin, F. F. S. A. (2018). *Big data meets artificial intelligence.* Retrieved from https://www.bafin.de/SharedDocs/Downloads/EN/dl_bdai_studie_en.html

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82-115. Retrieved from https://www.sciencedirect.com/science/article/pii/S1566253519308103 doi: https://doi.org/10.1016/j.inffus.2019.12.012

Burkov, A. (2019). *The hundred-page machine learning book* (Vol. 1). Retrieved from http://ema.cri-info.cm/wp-content/uploads/2019/07/2019BurkovTheHundred-pageMachineLearning.pdf doi: 10.1080/15228053.2020.1766224

Calo, R. (2018). Artificial intelligence policy: A primer and roadmap. *University of Bologna Law Review*, *3*, 180-218.

CBInsights. (2021). *The challenger bank playbook: How 6 digital banking upstarts are taking on retail banking.* Retrieved from https://www.cbinsights.com/research/report/challenger-bank-playbook/

Cornelissen, J. (2021, June). *A study on forecasting sofr with a recurrent neural network using long short-term memory cells.* Retrieved from http://essay.utwente.nl/86418/

De Nederlandsche Bank. (2017). *Post-event transaction monitoring process for banks.* Retrieved from https://www.dnb.nl/media/xzhnz40r/guidance-document-transactiemonitoring-banken.pdf

De Nederlandsche Bank. (2019). *General principles for the use of artificial intelligence in the financial sector.* Retrieved from https://www.dnb.nl/media/voffsric/general-principles-for-the-use-of-artificial-intelligence-in-the-financial-sector.pdf

De Nederlandsche Bank. (2020a). *Guideline on the anti-money laundering and anti-terrorist financing act and the sanctions act.*

De Nederlandsche Bank. (2020b). *Position paper dnb t.b.v. hoorzitting/ ron-*

detafelgesprek 'wettelijk kader en toezicht' d.d.

De Nederlandsche Bank. (2021, July). *Perspectives on explainable ai in the financial sector.* Retrieved from https://www.dnb.nl/media/jifijieq/perspectives-on-explainable-ai-in-the-financial-sector.pdf

DeepAI. (n.d.). *What is machine learning.* Retrieved from https://deepai.org/machine-learning-glossary-and-terms/machine-learning

Deloitte. (2019). *Transparency and responsibility in artificial intelligenc.* Retrieved from https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/innovatie/deloitte-nl-innovation-bringing-transparency-and-ethics-into-ai.pdf

Dictionary, C. (n.d.). *Meaning of dependability.* Retrieved from https://dictionary.cambridge.org/dictionary/english/dependability

Dutch Payments Association. (2021). *Facts and figures on the dutch payment system in 2020.* Retrieved from https://factsheet.betaalvereniging.nl/en/

el Hassouni, A. (2016). Fraud detection using machine learning methods.

EY. (2020). *Ml in the banking landscape - trustworthy, responsible ai in financial services.* Retrieved from https://www.firm.fm/wp-content/uploads/2020/12/2020-12-FIRM-Roundtable-EY-on-Trusted-AI-in-Financial-Services.pdf

Gasser, U., & Almeida, V. A. (2017, 11). A layered model for ai governance. *IEEE Internet Computing*, *21*, 58-62. Retrieved from https://dash.harvard.edu/bitstream/handle/1/34390353/w6gov-18-LATEX.pdf?sequence=1 doi: 10.1109/MIC.2017.4180835

Google Machine Learning Glossary. (n.d.). *Machine learning glossary: Fairness.* Retrieved 2021-09-15, from https://developers.google.com/machine-learning/glossary/fairness

Goswami, S. (2020, 12). *The rising concern around consumer data and privacy.* Retrieved from https://www.forbes.com/sites/forbestechcouncil/2020/12/14/the-rising-concern-around-consumer-data-and-privacy/

Gölz, P., Kahng, A., & Procaccia, A. D. (2019). *Paradoxes in fair machine learning.* Retrieved from https://proceedings.neurips.cc/paper/2019/file/bbc92a647199b832ec90d7cf57074e9e-Paper.pdf

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *CoRR*, *abs/1610.02413*. Retrieved from http://arxiv.org/abs/1610.02413

Hill, K. (2012). *How target figured out a teen girl was pregnant before her father did.* Retrieved from https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=4466d4166668

Huizinga, S. (2016). *Breken met banken* (T. van de Put, Ed.). Kompas.

IBM Cloud Learn Hub. (2020). *Artificial intelligence (ai).* Retrieved 2021-09-10, from https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

ING. (2021, 2). *Know your customer and anti-money laundering measures.* Retrieved from https://www.ing.com/About-us/Compliance/KYC-and-anti-money-laundering-measures.htm

Kamishima, T., Akaho, S., & Sakuma, J. (2011). Fairness-aware learning through regularization approach. *2011 11th IEEE International Conference on Data Mining Workshops*. Retrieved from https://www.kamishima.net/archive/2011-ws-icdm_padm.pdf

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. Retrieved from http://www.jstor.org/stable/2236703

Kun, J. (2015, 10). *One definition of algorithmic fairness: statistical parity.* Retrieved from https://jeremykun.com/2015/10/19/one-definition-of-algorithmic-fairness-statistical-parity/

Lewis, D. K. (1986). *Causal explanation.* Oxford University Press.

Lewis, T. (2014, 12). *A brief history of artificial intelligence.* Retrieved from https://www.livescience.com/49007-history-of-artificial-intelligence.html

Lim, H. (2020). *7 types of data bias in machine learning.* Retrieved from https://lionbridge.ai/articles/7-types-of-data-bias-in-machine-learning/

Liu, A., Song, Y., Zhang, G., & Lu, J. (2017, 8). Regional concept drift detection and density synchronized drift adaptation. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2017/317

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*. doi: 10.1109/TKDE.2018.2876857

Merriam-Webster. (n.d.). *Definition of governance.* Retrieved from https://www.merriam-webster.com/dictionary/governance

Miller, T. (2019, 2). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*. doi: 10.1016/j.artint.2018.07.007

Miller, T. (2020). *Contrastive explanation: A structural-model approach.* Retrieved from https://arxiv.org/abs/1811.03163

Miller, T., Howe, P., & Sonenberg, L. (2017, 12). Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *CoRR*, *abs/1712.0*. Retrieved from http://arxiv.org/abs/1712.00547

Ministerie, O. (2021). *Abn amro betaalt 480 miljoen euro vanwege ernstige tekortkomingen bij het bestrijden van witwassen.* Retrieved from https://www

.om.nl/actueel/nieuws/2021/04/19/abn-amro-betaalt-480-miljoen-euro-vanwege-ernstige-tekortkomingen-bij-het-bestrijden-van-witwassen

Muthukumar, V., Pedapati, T., Ratha, N. K., Sattigeri, P., Wu, C.-W., Kingsbury, B., ... Varshney, K. R. (2018). Understanding unequal gender classification accuracy from face images. *CoRR*, *abs/1812.0*. Retrieved from http://arxiv.org/abs/1812.00099

Newman, J. (2021, 5). *Explainability won't save ai.* Retrieved from https://www.brookings.edu/techstream/explainability-wont-save-ai/

Oladele, S. (2021, 6). *A comprehensive guide on how to monitor your models in production.* Retrieved from https://neptune.ai/blog/how-to-monitor-your-models-in-production-guide

Oracle. (2020). *Lifecycle of machine learning models*. Author. Retrieved from https://www.oracle.com/a/ocom/docs/data-science-lifecycle-ebook.pdf

paperpile. (2019). *The best academic research databases.* Retrieved from https://paperpile.com/g/academic-research-databases/?fbclid=IwAR1CMIiVvvKQp9BBXjPAef5yJlL2VMae14WrrxFd91a12KALYHxgjoDjr6Q

Patruno, L. (2019, 6). *The ultimate guide to model retraining.* Retrieved from https://mlinproduction.com/model-retraining/

Planbureau, C. (2020). *Kansrijk belastingbeleid.* Retrieved from https://www.cpb.nl/sites/default/files/omnidownload/CPB-Kansrijk-belastingbeleid-2020.pdf

Rijksoverheid. (2021). *Wet op het financieel toezicht.* Retrieved from https://wetten.overheid.nl/BWBR0020368/2021-07-01

Rousseau, P. L. (2003). Historical perspectives on financial development and economic growth. *Review*, *85*. doi: 10.20955/r.85.81-106

Samiullah, C. (2019, 3). *Deploying machine learning models in shadow mode.* Retrieved from https://christophergs.com/machine%20learning/2019/03/30/deploying-machine-learning-applications-in-shadow-mode/

Scikit-learn. (n.d.). *Machine learning in python.* Retrieved from https://scikit-learn.org/stable/

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379-423. Retrieved from https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

SpriggHR. (2020). *Responsibility vs accountability - wht;s the difference?* Retrieved from https://sprigghr.com/blog/hr-professionals/responsibility-vs-accountability-whats-the-difference/

Stewart, M. (2019, 12). *Understanding dataset shift.* Retrieved from https://towardsdatascience.com/understanding-dataset-shift-f2a5a262a766

TensorFlow. (n.d.). *An end-to-end open source machine learning platform.* Retrieved from https://www.tensorflow.org/

Widmer, G., & Kubat, M. (1996, 4). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, *23*. doi: 10.1007/BF00116900

Yampolskiy, R. V. (2020). On controllability of AI. *CoRR*, *abs/2008.04071*. Retrieved from https://arxiv.org/abs/2008.04071

Završnik, A. (2020). Criminal justice, artificial intelligence systems, and human rights. *ERA Forum*, *20*, 567-583. Retrieved from https://doi.org/10.1007/s12027-020-00602-0 doi: 10.1007/s12027-020-00602-0

Zhong, Z. (2018). *A tutorial on fairness in machine learning.* Retrieved from https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

CHAPTER 10

Appendix

10.1 Literature review

The literature review is split up in a number of parts. Firstly, the goals of this literature review are determined such that we know what we will look for. To actually find the information, the goals section has to be divided into two subsections which are focused on the key words and the sources. The key words will in some form resemble the determined goals such that the desired information is found. The choosing of sources is mentioned to aid reproducibility. Secondly, the elimination requirements are set. Elimination requirements are used as it will help to weed out the pieces of information that are of no use to the research. Thirdly, an oversight is created such that the research is reproducible and transparent. Fourthly, insight into the results are presented.

10.1.1 Goal

The goal of the literature review is to gather information on the governance of ML models to be able to determine a new AI development and monitoring framework for the banking sector. Through using this framework, challenger banks can proactively improve their ML operations with a focus on developing and monitoring ML models that are compliant with the regulations.

Key words
Governance
AI
ML
Models

Table 10.1: Key words

	Scopus (# of hits)	Web of Science (# of hits)	Duplicates
"Governance" AND ("AI" OR "ML") AND "Models"	125	31	22

Table 10.2: Search Queries

10.1.2 Key words

The words or set of words that are chosen to fulfill the requested information of the aforementioned goals are displayed in Table 10.1.

10.1.3 Sources

To provide the research with a solid base of information, multiple databases for academic works are used. Scopus and Web of Science are chosen as they are highly regarded (paperpile, 2019) and information is accessible to the researcher through his University of Twente credentials. Besides these databases for academic research a large part of the research will be done on the position of dutch regulatory bodies in the financial industry regarding ML models. Therefore the retrieval of those publications will be done manually. The regulatory bodies that will be scrutinized is DNB.

10.1.4 Elimination requirements and procedures

Since the searches will most likely generate a wealth of information, the elimination requirements and procedures displayed in algorithm 1.

Algorithm 1: Requirements and elimination procedures

```
Data: Found Papers
Result: To be analysed papers
initialization;
foreach Paper in the set do
    if Language is not [English or Dutch] then
        Discard paper
    end
    case Title of the paper is not applicable do
        Discard paper
    case Abstract of the paper is not applicable do
        Discard paper
    otherwise do
        Analyse paper
    end
end
```

10.1.5 Found AI aspects

All the different aspects which are found in literature are displayed in Table 10.3. All the facets that are in scope are discussed in chapter 2

10.2 Results Transaction monitoring Case Study

In this section, all checks are denoted in Table 10.4 and whether it is done or not done.

AI facet
Accountability
Transparency
Privacy
Bias reduction
Explainability
Fairness
Human rights
Human well-being
Inequality in application/Discrimination
Auditability
Certification
Cybersecurity
Data- parity problem
Displacement of labour
Pattern recognition
Setting safety thresholds
Taxation of labour
Use of force
Validating safety thresholds
Accuracy
Consequential Decision making
Dependability
Fairness
Investment and procurement
Leverage (Ability to reuse work)
Provenance/lineage
Reproducibility
Responsibility
Sustainable development

Table 10.3: AI facets found in the literature

Phase	Check nr.	Recommendation	Done/Not Done
Development			
	A.1.1	Produce an as transparent, auditable, and reproducible as possible model through proper documentation.	Done
	A.1.2	Realize that machine learing models can produce undesired/faulty results for which the organization is responsible.	Done
	A.1.3	Assign operational accountability at all relevan levels of the organisation.	Done
	A.1.4	Integrate AI in the organization's risk management framework	Not Done
	A.2.1	Define a minimal required data quality	Done
	A.2.2	Anonymise the data.	Done
	A.2.3	Make continuous efforts to ensure correct, complete and representative data.	Done
	A.2.4	Pay special attention to missing or incorrect data-points, potential sources of bias in data, features and inference results (such as selection and survival bias)	Done
	A.2.5	Procedures and safeguards are in place to maintain and improve data integrity and security during the process of data collection, data preparation and data management.	Done
	A.2.6	Issues with data integrity and bias are both in development and production evalutated and documented in a structural manner for future reference.	Not Done
	A.2.7	Systematically archive the original datasets which are used to (re)train and (re)calibrate models	Done
	A.3.1	Determine whether the data contain discriminatory features based on demographic parity or equalized odds.	Done

Phase	Check nr.	Recommendation	Done/Not Done
	A.3.2	Determine whether there are unfair features by proxy e.g. due to prejudice, underestimation or negative legacy.	Not Done
	A.3.3	Get approval from the risk department on the chosen fairness mitigation.	Not Done
	A.4.1	Determine whether the data split is suitable for the set's size.	Done
	A.4.2	Determine if the training, test and validation set have similar distributions or whether a data shift has occurred.	Done
	A.5.1	Predefine the Key Performance Indicators and their required levels.	Done
	A.5.2	Predefine the criteria and metrics on which the final model will be selected.	Done
	A.5.3	Choose the best explainable model in comparable situations	Done
	A.5.4	Determine what hyperparameters should be optimised.	Done
	A.5.5	Determine what method to optimise the hyperparameters is most suitable.	Done
	A.6.1	Report the final metrics.	Done
Deployment			
	B.1.1	Determine how the model will interact with the outside environment and what is needed to ensure that it works with the outside environment.	Done
	B.2.1	Determine whether the model works as inteded in the test environment.	Done
	B.2.2	Determine whether the model works as inteded with production data whilst in shadow mode.	Done
	B.3.1	Determine whether the model works as inteded whilst being deployed in production.	Done
Post Deployment			
	C.1.1	Monitor whether the Key Performance Indicators are still within range.	Done

Phase	Check nr.	Recommendation	Done/Not Done
	C.2.1	Routinely check if alterations have been made in the data pipeline.	Done
	C.2.2	Proactive measures should be taken to ensure correct, complete and representative data.	Done
	C.3.1	Routinely check if there are features that display significant concept drift.	Done
	C.3.2	All criteria for significant changes as well as other fail criteria should be documented to assess whether retraining is necessary.	Done
	C.4.1	Have back-up plans in place.	Done
	C.4.2	Properly document the version history.	Done

Table 10.4: Transaction Monitoring results