The Impact of Behavioural Bias in Credit Risk Decision-Making

by
Lesley Munyaneza, ANR: 929202
MSc. Tilburg University

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Quantitative Finance and Actuarial Science

Tilburg School of Economics and Management
Tilburg University

Supervisors

prof. dr. B. Melenberg
dr. A. Balter

Date: April 27, 2025

Contents

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisor prof. Bertrand Melenberg for his excellent guidance; for our bi-weekly meetings, his suggestions for new ideas and improvements and also for his continuous support throughout the process of writing this thesis. I would also like to thank my friend Ella for her emotional support and help in reviewing my thesis.

Management summary

Credit risk is a major risk for all financial institutions because it is their main source of revenue. It is important to manage this risk to mitigate the impact of large unexpected losses. The decisions made regarding credit risk determine the effectiveness of this risk management. Behavioural finance is a branch of finance, that deviates from the assumption of people being rational in their decision-making. One of the main assumptions of behavioural finance is that people are influenced by biases which are systematic errors that occur when people interpret information and make decisions. These biases can lead to incorrect interpretations, which can affect financial decision-making.
This thesis explores the impact of one behavioural bias, overconfidence, displayed by loan officers on the decisions made for a microfinance institution's credit risk. Literature shows that overconfidence behaviour can lead to excessive risk-taking, as risks are underestimated or people's ability to gauge risk is overestimated. This can be detrimental to financial institutions, such as microfinance institutions. In addition, most of a microfinance institution's loans are unsecured (that is, no collateral), so there is a higher risk of default. Thus, managing credit risk is crucial for microfinance institutions.
Background on behavioural finance and credit risk is provided for better understanding and to give context. In addition, overconfidence bias is described and also how credit risk occurs for a microfinance institution. In order to measure the impact of overconfidence, we used two quantitative models, linear regression and random forest. Both methods are extensively described, with their formulations, assumptions and an overall evaluation. Additionally, we discuss how they quantify the impact of overconfidence. For linear regression, we used least square estimated coefficients, standardised coefficients and probability values (p-values). For random forest, we used permutation importance values, accompanied by approximated p-values, and partial dependence plots.
Due to the limitations of this research, we used three proxies that indicate overconfidence, which were constructed by researchers using relevant balance sheet data. The dataset used for analysis was obtained from the Mix Market database and World Development Indicators database. Due to a substantial amount of missing data, we implemented an imputation method (miss Forest algorithm) to replace those missing values. Evaluating the impact of this imputation method was limited, so results should be interpreted carefully. The two models were applied on the final dataset containing financial and country specific data on 1,312 microfinance institutions from 100 countries over time period $2005 - 2015$. To measure a microfinance institution's credit risk, we used measures of a loan portfolio's quality: Portfolio at Risk greater than 30 days and Portfolio at Risk greater than 90 days. These two measures were chosen as response variables for both models. Besides the three proxies for overconfidence, the models included other predictors such as bank characteristics and country specific characteristics.
The results are first shown for linear regression and random forest separately, for both response variables, then, the results are compared for all models. For two out of the three proxies, all the models agreed on their impact on credit risk. Next, we discuss the implications of the results. Although not all proxies were significant, we showed that there is an association between loan officers' overconfidence and an MFI's credit risk for one of the proxies.

1 Introduction

Financial institutions are exposed to numerous potential threats that can affect their financial health and performance. Such threats can take many forms like liquidity risk, credit risk, market risk, operational risk, cyber risk, reputational risk, etc. Hence, it is crucial for financial institutions to employ effective risk management to mitigate such risks, that is, to minimise the impact those risks could have on an institution's stability and profitability. The key components in risk management are identifying (all) possible threats, assessing the damage they may cause, designing strategies to mitigate those threats and evaluating the implementation of the strategies (Bouteillé & Coogan-Pushner, 2013).

One of the most common types of risk faced by financial institutions is credit risk. In particular, credit risk is a major source of risk for banks as it is their main source of revenue. Credit risk refers to the likelihood that a borrower (obligor) will not meet future debt obligations in accordance with the terms agreed when credit was provided by a lender (Doumpos et al., 2019). If many loans are not paid back on time, or at all, then this will reduce a bank's ability to provide loans and increase its expenses as they have to chase after those borrowers. Additionally, some loans may never be fully recovered, that is, only a proportion of the loan and accumulated interest is repaid, leading to losses for the bank. If such losses are substantial, the bank could become insolvent, that is, the bank will have insufficient funds to pay back its own debts and may even file for bankruptcy. Thus, credit risk management is important to lower the risk of nonpayment and cut down on bank losses, which is essential for the long-term survival of banks (Olorunsola et al., 2023). To effectively manage credit risk, banks need to thoroughly screen (future) clients, analyse their exposure to credit risk and assess the decisions made to mitigate them (Churchill & Coster, 2001).

Research on the determinants of loan repayment defaults and repayment performance of banks typically focuses on banks' clients and the types of services offered (Bandyopadhyay & Saxena, 2023; Chaibi & Ftiti, 2015; Foos et al., 2010; Hayati & Ariff, 2007). These studies link default rates to moral hazard and information asymmetries, hence the focus is on the borrowers' side. However, others claim that the reason for poor repayment performance depends rather on the lender's side, in their decision-making process (Fersi & Boujelbène, 2022).

Behavioural finance is a branch of finance that explores the influence of psychological factors on financial decision-making (Almansour et al., 2023). This is a relatively new school of thought that deviates from traditional finance as it assumes that an agent, or a decision maker, is not always rational, has a limit to their self-control and is influenced by biases (Barberis & Thaler, 2002). Behavioural experts have identified the role of behavioural biases like self-attribution, overconfidence, and herd behaviour in fuelling market anomalies (Kapoor & Prosad, 2017). Hence, behavioural finance is an extremely relevant topic in today's times.

According to the assumptions of behavioural finance, lenders' biases may affect the credit risk methods used to assess customers and the amount of credit they are given. For instance, an overconfident lender may underestimate the risks of future cash flows, by reducing the provisions for loss or granting more loans with higher risk. Thus, overconfidence can lead to higher risk-taking, which in turn exposes a bank to the detrimental effects of credit risk (Baklouti & Baccar, 2013; Fersi & Boujelbène, 2022). Therefore, behavioural biases may influence a bank's credit risk management and its financial performance.

This study will explore the influence of behavioural bias overconfidence on the credit risk

decisions made for a microfinance institution, using two quantitative models. A microfinance institution (MFI) is a bank that offers financial services, including credit, savings and insurance, to low-income individuals and communities who are excluded from the formal banking sector (Olorunsola et al., 2023). MFIs typically have dual objectives of social impact and financial performance. They aim to alleviate poverty, empower marginalised groups and foster entrepreneurship whilst covering their operational costs and maintaining adequate reserves for future growth and expansion (Parihar et al., 2024). By balancing these two objectives, they can achieve financial sustainability and operational viability.

MFIs play a crucial role in promoting financial inclusion and economic development in under-served communities and developing markets (Ibtissem & Bouri, 2013). Thus, managing the credit risk of MFIs is not only important for their survival, but also for the communities they serve and the countries they operate in. However, in the last few years, MFIs have experienced an increasing number of defaulted loans, which led to high inefficiency levels in the microfinance sector and costs associated with the MFI operations (Fersi & Boujelbène, 2022). According to Agier and Szafarz (2013) and Churchill and Coster (2001), credit risk is one of the risks challenging the sustainability of MFIs and their efficiency. One of the reasons for this is that most microlending is unsecured (that is, traditional collateral is not used to secure microloans). This increases the risk of non-repayment for an MFI as if a loan defaults, there is no collateral that can be seized and sold to pay for the entire, or partial, loan repayment.

1.1 Literature Review

Research on the impact of behavioural biases on financial decision-making has mainly been focused on investors, showing how they can sometimes make irrational investment decisions, contradicting traditional finance theory. However, the impact of behavioural bias on lenders' credit risk decisions and its effect on the performance of financial institutions has scarcely been explored, especially for microfinance institutions.

There is a great deal of research on behavioural finance and its influence on financial decision-making. Madaan and Singh (2019) investigated how an investor's behavioural biases affect investment decision-making in the National Stock Exchange. The study showed the existence of four biases (overconfidence, herding bias, anchoring, disposition effect) in individual investment decisions. Using linear regression, they observed that investors with overconfidence and herding bias have a significant positive impact on investment decision. Bihari et al. (2023) studied how biases on information based on knowledge affect decisions about investments. Through research and consultation with specialists, they determined four relevant biases: overconfidence, loss aversion, regret aversion and the Barnum effect. They applied multiple linear regression and artificial neural networks to analyse the data of 337 retail investors. The findings showed that the investment choice was heavily impacted by regret aversion, followed by loss aversion, overconfidence, and the Barnum effect.

In the last two decades, research on behavioural finance in the banking context has been growing. Nkamnebe and Idemobi (2011) and Olomola (2002) showed that repayment default performance is significantly affected by both the borrower's and lender's characteristics. Addae-Korankye (2014) conducted a survey on the causes of loan repayment default and the different means for future loss control. The author found that poor loan portfolio quality is strongly associated with the managers' poor evaluation and selection of borrowers along with high interest rates and inadequate loan size. In light of such findings, Baklouti (2015) explored the relationship between loan officers' psychological traits and

the accuracy of their predictions for possible default events within MFIs in Tunisia. The main results of the study revealed that loan officers' information evaluations are largely influenced by behavioural biases that are linked to judgmental errors in the granting process. Similar studies were conducted, with the focus on Tunisian MFIs, and further showed that default risk can be linked to the decision-making of loan officers and their behavioural biases (Baklouti & Baccar, 2013; Mighri, 2014; Mighri et al., 2018).

One example of a behavioural bias is overconfidence. This bias results in a person's overestimation of his or her abilities, chances of success, the probability of obtaining positive results or the accuracy of his or her knowledge and gathered information (Fersi & Boujelbène, 2022). This behavioural bias has been widely debated and proven in the context of financial markets and was typically related to the investor's behaviour but also the managerial behaviour in companies and financial institutions (Abbes, 2012; Ben Salah Mahdi & Boujelbène Abbes, 2018; Daniel et al., 2001; Malmendier & Tate, 2008). However, studies in the context of MFIs are rather scarce (Baklouti, 2015; Mighri, 2014).

Fersi and Boujelbène (2022) investigated how loan officers' overconfidence affects risk-taking decisions and solvency performance of Islamic and conventional microfinance institutions (MFI). The dataset covers 326 conventional MFIs and 57 Islamic MFIs in six different regions over the period of 2005–2015. Proxies for overconfidence were used such as loan growth and net interest margin. They were treated as explanatory variables in a generalised least squares regression amongst others with the dependent variable being the portfolio at risk. The study showed that overconfidence has a significant positive impact on risk-taking decisions, which lowers the loan portfolio quality and harms the solvency performance of the institution. An article by Chen and Chen (2015) also investigated the impact of overconfidence on bank risk-taking, but uses press data instead. The sample consists of the financial institutions in G20 [1] and Taiwan over the period of 2005-2012. It showed that overconfidence increases credit risk and insolvency risk, especially in recession periods.

1.2 Problem Statement

Behavioural studies are of great interest since they allow us to understand the decision-makers' attitudes towards risk-taking, however they have not been fully researched in the context of MFIs. This thesis adds to the existing literature as it deals with behavioural finance and introduces other determinants of the risk-taking of MFIs on a global level.

The aim of the thesis is to explore the impact of loan officers' overconfidence on the credit risk decision-making of an MFI, using two quantitative models, linear regression and random forest, which will measure and quantify the potential influence of overconfidence. The data is based on MFIs in several countries in Asia, Africa, Eastern Europe and Latin America.

Thus, the main research question is :

Does loan officers' overconfidence influence the credit risk decision-making of an MFI?

To answer the main research question, the following sub-questions will be answered.

i) *Is there a relationship between overconfidence and an MFI's credit risk?*

[1]The G20 comprises 19 countries including: Argentina, Australia, Brazil, Canada, China, France, Germany, India, Indonesia, Italy, Japan, Republic of Korea, Mexico, Russia, Saudi Arabia, South Africa, Türkiye, United Kingdom, and United States and two regional bodies, namely the European Union and the African Union (G20, 2023).

This sub-question addresses whether there exists a relationship between loan officers' overconfidence and credit risk decision-making of an MFI. We will use two quantitative models to determine whether there is an association between loan officers' overconfidence and an MFI's credit risk decision-making.

ii) *Will the two quantitative methods, linear regression and random forest, agree on the (potential) effect of overconfidence on an MFI's credit risk?*

This sub-question addresses whether the two different quantitative models, linear regression and random forest, will agree on how they describe the association between loan officers' overconfidence and an MFI's credit risk decision-making.

1.3 Outline

This thesis is structured as follows. Chapter 2 provides the theoretical background on behavioural finance, touching on behavioural bias overconfidence, and credit risk, with a focus on microfinance institutions. In Chapter 3, we describe the methodology of two models, and how the relationship between variables is measured and quantified. Chapter 4 provides a description of the data, the different variables and their statistics. In addition, this chapter discusses how missing data is handled. In Chapter 5, we present the results for the data and analyse them. Finally, Chapter 6 discusses the implications of the results and how these address the research sub-questions, and, thus, the main research question. This chapter concludes with a summary of the thesis, discusses recommendations for future research, and the limitations of this research.

For the transparency of the use of AI, Microsoft Edge copilot was used for finding some of the literature relevant to behavioural finance, credit risk, and random forest. So, in chapters 2 and 3.

2 Theoretical Background

Before we examine how a behavioural bias affects a microfinance institution's credit risk, let us understand what behavioural bias is in the context of finance, what is the credit risk of an MFI and how these two concepts link together. In this chapter, we introduce the two main topics: behavioural finance and credit risk. We first describe behavioural finance, how this approach to finance came to be and its current status. We also briefly introduce behavioural biases and elaborate on one bias, overconfidence. We then move on to describing credit risk and what it looks like for a microfinance institution. Lastly, we discuss behavioural finance in the context of credit risk management.

2.1 Behavioural Finance

2.1.1 Brief Introduction to Traditional Finance

In 1844, John Stuart Mill introduced the concept of the rational economic man or in Latin *homo economicus*, whose aim is to maximise his or her satisfaction (or utility) given the constraints he or she faces (Mill, 2011). Utility is a measure of the satisfaction of an individual by the consumption of a good or service (Kapoor & Prosad, 2017). There are three underlying assumptions for this agent (that is, a rational man), which are perfect rationality, perfect self-interest and perfect information (Kapoor & Prosad, 2017). According to Barberis and Thaler (2002), rationality has two implications. First, when agents receive new information, they correctly update their current beliefs, according to Bayes' law [2]. Second, agents make choices that maximise their satisfaction, given their beliefs. This concept of the rational economic man became very popular among economists for two main reasons (Pompian, 2006). Firstly, *homo economicus* makes economic analysis moderately simple. Secondly, it allows economists to quantify their findings, make their work more elegant and easier to understand. The concept of the rational economic man became the foundation of the traditional finance paradigm, which seeks to understand financial markets using models with rational agents (Barberis & Thaler, 2002). This paradigm led to the development of several traditional finance theories such as the Expected Utility Theory (Bernoulli, 1954; Neumann & Morgenstern, 2004), Markowitz Portfolio Theory (Markowitz, 1952) and Efficient Market Hypothesis (Fama, 1970).

2.1.2 Problems in Traditional Finance

Although traditional finance theories were well constructed to make calculated financial decisions, they were unable to explain disruptions or anomalies in the stock markets such as market bubbles, overreactions and underreactions to new information (Kapoor & Prosad, 2017). A market bubble arises when the price of an asset exceeds the asset's fundamental value, that is, the present value of future dividend payments (Werner, 2012).

A famous example of a market bubble is the Dutch Tulip Mania (Mackay, 1852), which occurred during the Dutch Golden Age (that is, 1588 to 1672). Imported from Constantinople (currently known as Istanbul, Türkiye), this beautiful flower became a consumer sensation once introduced in Holland (currently known as the Netherlands). As

[2] Bayes' law or Bayes' Theorem relates the direct probability of a hypothesis H conditional on given data E, $\mathbb{P}(H|E)$, to the inverse probability of the data conditional on the hypothesis, $\mathbb{P}(E|H)$ (Joyce, 2021). This is expressed as follows:

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(H) \cdot \mathbb{P}(E|H)}{\mathbb{P}(E)}$$

this flower was difficult to obtain, it became a status symbol for the Dutch elite. Soon, the tulip mania caught on all over Holland, with people of all classes in a frenzy to obtain this flower. The growth in the demand for tulips increased so much that people began to invest in tulip stocks. Naturally, the price for a tulip soared and at its peak, it held the same value as expensive, indispensable, durable goods. Eventually, the Dutch stock market crashed when the investors decided that they had spent a substantial amount on a commodity having very low utility such as the tulip flower. This realisation resulted in a severe drop in tulip prices, leading to heavy losses. Events such as the Tulip Mania make one question the rationality of investors and hence theories based on rational agents.

A big drawback of traditional theories is that they are based on rules about how agents should behave, but not on principles describing how they actually behave. The concept of the rational economic man oversimplifies reality and disregards inner conflicts that real people face, which can lead to irrational behaviour, according to research (Pompian, 2006). For example, this concept does not take into account that people have difficulty prioritising short-term versus long-term (like spending versus saving) or reconciling inconsistencies between individual goals and societal values. When challenging the rational economic man, the three underlying assumptions received most of the criticisms. Let us discuss some of those criticisms described by Pompian (2006).

The assumption of *perfect rationality* implies that rational humans have the ability to reason and make beneficial judgments. Nevertheless, rationality is not the only driver of human behaviour. In reality, it may not even be the primary driver since many psychologists believe that the human intellect is actually submissive to human emotion. In other words, logic is not the dominant driver of human behaviour as this behaviour is largely influenced by subjective impulses like fear, love, hate, pain and pleasure. Research has shown that humans only use their intellect to achieve or to avoid these emotional outcomes.

Numerous studies have shown that humans are not *perfectly self-interested*. If they were, then religions that value selflessness, sacrifice and kindness to strangers would be unlikely to prevail as they have over the centuries. In particular, perfect self-interest would result in people focusing only on themselves and things that benefit them only. This would prevent people from performing unselfish deeds such as volunteering, helping the needy or serving in the military, as there is no (sufficient) gain for them. It would also rule out self-destructive behaviour like suicide, alcoholism and substance abuse.

In conclusion, humans are neither perfectly rational nor perfectly irrational. They possess diverse combinations of rational and irrational characteristics and benefit from different degrees of enlightenment with respect to different issues (Pompian, 2006). Hence, the assumption of rational economic man needs to be replaced by more realistic assumptions on human behaviour.

2.1.3 New Approach to Finance: Behavioural Finance

Behavioural finance is a new approach to financial markets that has emerged, in part, in response to the difficulties faced by the traditional theories (Barberis & Thaler, 2002). In general, it argues that some agents are not fully rational. More specifically, it analyses when we relax one or both of the two implications of rationality. On the one hand, some behavioural finance models have agents that fail to update their beliefs correctly. On the other hand, there are models where agents apply Bayes' law correctly but make choices that are questionable. In order to build such models, one needs to specify the form of agents'

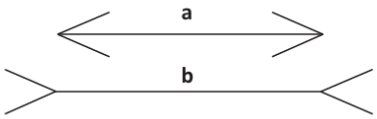

Figure 1: An adapted version of the Muller-Lyer Illusion (Howe & Purves, 2005), which illustrates visual illusions. When asked which segment is longer, a typical answer is to pick segment **b**. However, this is incorrect as both segments have the same length, but the arrows give the impression (or illusion) that segment **b** is longer than segment **a**.

irrationality. For guidance on this, behavioural economists typically turn to cognitive psychologists. More specifically, to the extensive experimental evidence compiled by those psychologists on the biases that arise when people form beliefs and on people's preferences, or on how they make decisions, given their beliefs (Pompian, 2006). These are generally described as two disciplines within behavioural finance: cognitive psychology and limit to arbitrage. In the following sub-section, we introduce cognitive psychology, but more information on limit to arbitrage can be found in Hens and Rieger (2010), Pompian (2006), and Tversky and Kahneman (1974).

2.1.4 Cognitive Psychology

Cognitive psychology is the scientific study of cognition, or the mental processes that are believed to drive human behaviour (Pompian, 2006). Research in cognitive psychology investigates a variety of subjects, such as memory, attention, perception, knowledge representation, reasoning, creativity, and problem solving. Cognitive psychology emerged in the late 1950s and early 1960s, and is considered as a relatively recent development in the history of psychological research (Pompian, 2006). Though the term "cognitive psychology" was coined by Ulrich Neisser in 1967, the cognitive approach was brought to prominence by Donald Broadbent in 1958 (Pompian, 2006). Broadbent's approach treats mental processes as software running on a computer, that is, the brain. Cognitive psychology typically describes human thought in terms of input, representation, computation or processing, and output.

As a result of experiments on cognition, it became clear that people's judgments and decisions are often far from rational (Pompian, 2006). In other words, they are affected by seemingly irrelevant factors or fail to take into account crucial information. Additionally, these departures from the rational norm are typically systematic because people fail consistently in the same type of problem, making the same mistake (Blanco, 2017). Namely, people seem to be irrational in a predictable way (Ariely, 2008). In order to develop a theory that aims to model human judgment and decision-making, this theory must be able to explain these instances of consistent irrationality or behavioural biases. A behavioural (or cognitive) bias is a systematic (that is, non-random and, thus, predictable) deviation from rationality in judgment or decision-making (Blanco, 2017).

To illustrate cognitive bias, let us consider a similar type of phenomenon, visual illusions. Figure 1 shows a famous object configuration, the *Muller-Lyer Illusion* (Howe & Purves, 2005). The idea is that one must observe the picture and decide which of the two horizontal segments, **a** or **b**, is longer.

According to Blanco (2017), most people in Western societies would agree that segment **b** looks slightly longer than segment **a**. Despite this common impression, both segments **a** and **b** have exactly the same length. These segments only differ in the ori-

entation of the arrows at the edges of the segments (that is, arrows pointing inward or outward). Hence, this difference creates the illusion that they differ in length (Howe & Purves, 2005). This simple example of visual illusion shows how people's deductions and reasoning can be deceived by irrelevant information like, for example, the arrows, which led to a systematic error in perception. Similar situations can occur across a variety of domains and tasks. This reveals the presence of behavioural biases not only in perception but also in judgment, decision-making, memory and so on (Blanco, 2017).

A typical consequence of behavioural bias is the predictable form of irrational behaviour. Behavioural biases have been proposed to underlie many beliefs and behaviours that are dangerous or problematic to individuals such as superstitions, pseudoscience, prejudice, poor consumer choices, etc. (Blanco, 2017). Moreover, they become very dangerous at a group-level because these mistakes are systematic rather than random (thus, one individual's mistake is not cancelled out by another one's), leading to disastrous collective decisions such as those observed in the 2007-2008 economic crisis (Ariely, 2009).

Though the main focus of research on behavioural biases has been to elaborate on experimentally documented biases, there has been some research on offering coherent conceptual frameworks to understand where these biases originate from. Some of these approaches are limited cognitive resources, influence of motivation and emotion, social influence and heuristics. More details on such approaches can be found in Blanco (2017), Pompian (2006), and Tversky and Kahneman (1974).

There are numerous behavioural biases such as overconfidence, optimism/pessimism, loss aversion bias, framing bias, anchoring bias, overreaction/underreaction, etc. The following literature provides further details (Barberis & Thaler, 2002; Baron, 2023; Gilovich et al., 2002; Pompian, 2006). Behavioural biases allow for the specification of how agents form expectations, and hence make decisions, which is a crucial component of any model of financial markets (Barberis & Thaler, 2002). In this project, we consider behavioural bias overconfidence and how this may impact the credit risk decisions made by a microfinance institution.

2.1.5 Behavioural Bias: Overconfidence

Canikli and Aren (2019) define overconfidence as a cognitive bias where a person tends to overvalue their knowledge and abilities. Namely, overconfidence is the difference between the value that a person attributes to their knowledge and ability, and the real situation. The higher this difference is, the more overconfident a person is. The concept of overconfidence derives from a large body of cognitive psychological experiments and surveys in which participants overestimate their own predictive abilities and the precision of the information given (Pompian, 2006). In reality, people are poorly calibrated in estimated probabilities, that is, events that they think are certain to happen are often far less than 100 percent certain to occur. In other words, people think they are smarter and have better information than they actually do.

A consequence of overconfidence is to ignore the uncertainty that is inherent in the financial markets (Aren & Nayman Hamamci, 2021). This is mainly due to information overload, which overwhelms and confuses individuals. Nowadays, the real problem is the multitude of information, not its scarcity as it was decades ago. A common belief is that if an individual has a lot of information, they can make better decisions (Aren & Nayman Hamamci, 2021). However, the increase of information channels leads to several consequences. One of them is the difficulty of evaluating a large amount of information. According to Aren and Nayman Hamamci (2021), the human brain cannot evaluate more

than eight pieces of information at the same time. They further claim that an individual distinguishes the information as relevant or irrelevant based on their knowledge and experience, and attributes importance to the information they consider important. Overconfidence is activated at this stage and the evaluation is made within this framework (Aren & Nayman Hamamci, 2021). Overconfidence does not only lead to incorrect decisions, it can be a source of correct decisions (Aren & Nayman Hamamci, 2021). It can be said that successful people have overconfidence but overconfidence is not the determinant of success.

There are various ways to strengthen overconfidence, such as easy decision-making, experience, level of knowledge and past achievements (Aren & Nayman Hamamci, 2021). For instance, successful managers show a tendency to overestimate their analytical capabilities when they make many successful decisions in the past, hence they can have overconfidence bias. However, a crucial point successful managers miss is that uncertainty dominates the markets which have no memory (Pompian, 2006). Past successful decisions are not the sole determinants of future outcomes, as those past decisions may have been successful due to variables that cannot be controlled (and could even be defined as luck). Nonetheless, human psychology is more prone to link successful decisions to their abilities (Aren & Nayman Hamamci, 2021). Therefore, past successes can strengthen overconfidence.

Overall, there are several possible ways that result in managers having overconfidence bias and hence, being prone to systematically making correct or incorrect decisions. Thus, this can impact the decisions made for any business and its performance.

2.2 Credit Risk

2.2.1 What is Credit Risk?

Credit risk is the possibility of losing money due to the inability, unwillingness or untimeliness of a counterparty to honour a financial obligation (Bouteillé & Coogan-Pushner, 2013). Namely, whenever there is a chance that a counterparty will not pay an amount of money owed (whether full or partial), live up to a financial commitment or honour a claim, there is credit risk. The counterparties having the responsibility of fulfilling an obligation are called obligors. The obligations themselves typically represent a legal liability in the form of a written or oral (in some countries) contract between counterparties to pay or perform.

According to Bouteillé and Coogan-Pushner (2013), there are three concepts associated with the inability to pay. First is insolvency, which defines the financial state of an obligor whose liabilities exceed its assets. Second is default, which is the failure to meet a contractual obligation such as through nonpayment. This is usually due to insolvency, but not always. Third is bankruptcy, which is a legal process allowing individuals or companies who are unable to repay their debts to seek relief through court-supervised reorganisation or liquidation of assets. Though it is common to use bankruptcy as a synonym for insolvency, they are different events.

Generally, losses from credit risk involve an obligor's inability to pay a financial obligation. Common scenarios include companies whose products or services become obsolete or whose revenues simply no longer cover operating and financing costs. When the scheduled payment becomes due and the company does not have sufficient funds available, it defaults and may generate a credit loss for the lenders and all other counterparties. Credit losses can also arise from an obligor's unwillingness to pay. Although less common, it can result in the same consequences for the creditors. Typical cases are commercial disputes over the validity of a contract. In the cases where unwillingness is the issue, if the dispute leads to litigation and the lender prevails, there is recovery of the amount owed and ultimate losses

are reduced, or even entirely avoided, since the borrower has the ability to pay. There are various other ways in which credit losses stem from, such as in the form of timing or other risks faced by a company like political risk. One common feature of all credit exposure is that the longer the term of a contract, the riskier the contract is, because every additional day increases the likelihood of an obligor's failing to make good on a financial obligation. Hence, time is also a risk affecting credit risk.

All businesses and individuals are exposed to credit risk, whether willingly or unwillingly. Nevertheless, not all credit risk exposure is inherently detrimental since banks and hedge funds exist and even profit from their ability to create and manage credit risk (Bouteillé & Coogan-Pushner, 2013). Financial institutions bear the greatest exposure to loaned money, which is reasonable as they either invest in debt instruments or assume credit risk as a primary business endeavour. Such institutions include banks, asset managers, hedge funds, insurance companies and pension funds. To read more on the credit exposure of these institutions, see Bouteillé and Coogan-Pushner (2013). In the next sub-section, we describe the credit exposure faced by a microfinance institution, but first we explain why credit risk should be managed.

An essential aspect of credit risk is that it is controllable (Bouteillé & Coogan-Pushner, 2013). In other words, credit exposure does not fall from the sky, but can actually be managed if credit risk is understood in terms of its fundamental sources and can be anticipated. In addition, credit risk is the product of human behaviour, that is, of people making decisions (Bouteillé & Coogan-Pushner, 2013). Obligors may find themselves in precarious financial situations due to decisions that the company's managers have made. These decisions are the consequences of their incentives and the incentives of the shareholders whom they represent. Thus, understanding the motivation of shareholders and managers is an important aspect of a counterparty's credit risk profile.

To put it briefly, weak management of a credit risk can be costly and even lead to bankruptcy (Bouteillé & Coogan-Pushner, 2013). The exposure to credit risk is capital intensive [3].To survive large and unexpected losses, a large equity base must be built. For businesses dealing with credit portfolios[4], a large number of small losses are expected and manageable. However, there is also a small chance of facing large losses, which can lead to insolvency or even bankruptcy. Therefore, all companies dealing with credit portfolios should dedicate significant attention and resources to credit risk management for their own survival, profitability and return on equity (Bouteillé & Coogan-Pushner, 2013).

For financial institutions, credit risk is a primary concern for their survival because colossal losses can lead to insolvency or bankruptcy. Yet, even a nonfinancial company can have credit losses resulting in insolvency or bankruptcy. Though a trivial statement, it is clear to see that the less money one loses, the more money one makes and hence this summarises the importance of credit risk management to a business' profitability. Lastly, companies cannot run their businesses at a sufficient return equity if they hold too much equity capital. However, holding a substantial amount of debt capital is also not the solution as debt does not absorb losses and can introduce more risk into the equation. Therefore, the key to long-term survival is an adequately high amount of equity capital complemented by prudent risk management.

[3]Capital intensive refers to business processes or industries that require large amounts of investment to produce a good or service (Scott, 2003).

[4]Credit Portfolio is any collection of credit exposures that is formed as part of financial intermediation activities (e.g., regular Lending products or derivative contracts) or as an investment in Credit Risk sensitive securities (such as corporate bonds) (Open Risk Manual, 2019).

In summary, actively managing a credit portfolio can help increase a company's return on equity in any economic environment and for many types of company. Thus, it is important to employ and maintain an effective credit risk management.

2.2.2 Microfinance Institution

Microfinance refers to the financial services provided to low-income individuals or groups who are typically excluded from traditional banking (Pinz & Helmig, 2014). Microfinance institutions (MFIs) tend to offer credit in the form of small working capital loans (also known as microcredit or microloans), insurance, money transfers and savings accounts. For instance, small loans enable entrepreneurs to start or expand small and medium businesses, and savings help families build assets to finance school fees, improve homes and achieve goals. Microfinance aims to make financial services more accessible to marginalised groups, to address their needs and promote self-sufficiency, by fostering financial inclusion (Pinz & Helmig, 2014). Financial inclusion is defined as the availability of and the ease of access to basic formal financial services to all members of the population (Ozili, 2020). This inclusion means that all individuals and businesses have access to useful and affordable formal financial services - transactions, payments, credit, insurance and savings - that meet their needs in a responsible and sustainable way.

Microfinance programs have proven to have a great contribution in reducing poverty (Ibtissem & Bouri, 2013). They enable poor entrepreneurs to initiate their own business, teaching them how to protect the capital they have, to deal with risk and to expand the circle of their economic activities. The greater the access to microcredit schemes, the more the number of small enterprises will increase. This in turn creates employment opportunities for the poorest and therefore stimulates economic development and social inclusion (Ibtissem & Bouri, 2013).

Besides their social goals, MFIs need to perform well financially to survive in increasingly competitive markets (Pinz & Helmig, 2014). Hence, many MFIs have dual objectives, on a social and financial level, though there are some differences in their relative focus. For instance, nonprofit MFIs tend to put more weight on their social performance, whereas for-profit MFIs mainly focus on the achievement of clearly stated financial goals (Pinz & Helmig, 2014). On a financial level, MFIs aim to achieve self-sustainability, that is, to be able to cover all their current costs and make profits on services that they offer (Ibtissem & Bouri, 2013). This requires applying high interest rates, largely above market rates. This can be at the expense of social aims, since high interest rates can exclude low-income people especially those living in rural or marginal areas.

This dual mission in delivering financial services to low-income people whilst achieving self-sustainability (or even profit) represents one of the most widely discussed dilemmas in the field of microfinance (Ibtissem & Bouri, 2013). In order to face such a dilemma, it's crucial for an MFI's stability to employ the best practice, improve the efficiency of their portfolio risk management and apply accurate pricing policies. This will allow them to find an adequate equilibrium between sustainability and outreach. Further information on how MFIs can achieve this can be found in the following literature Churchill and Coster (2001), Ibtissem and Bouri (2013), and Pinz and Helmig (2014). Next, we concentrate on the risks faced by MFIs, with a special focus on credit risk.

2.2.3 Credit Risk of a Microfinance Institution

Microfinance institutions face numerous risks that threaten their financial viability and long-term sustainability. For example, some of the most serious risks come from the external environment in which an MFI operates in, such as the risk of a natural disaster,

economic crisis or war. Since an MFI cannot directly control these risks, there are several ways in which an MFI can minimise the potential negative impact. In order to employ an effective risk management, an MFI must first identify, understand and assess all the risks that can have a severe impact on the institution and their likelihood of occurrence. Once those risks are identified, an MFI can design strategies and control mechanisms to deal with them and assign responsibility to key individuals and teams to address them (Steinwand, 2000).

There are many risks that are common to all financial institutions, from banks to unregulated MFIs. These risks include credit risk, liquidity risk, market risk, fraud risk, legal and compliance risk, and governance risk. According to Steinwand (2000), financial institution managers and regulators evaluate these risks by considering i) the institution's potential exposure to loss, ii) the quality of internal risk management and information systems, and iii) the adequacy of capital and cash to absorb both identified and unidentified potential losses. In simple terms, management determines whether the risk can be appropriately measured and managed, considers the size of the potential loss, and assesses the institution's ability to bear such a loss. We now focus on credit risk, but for more details on other types of significant risks, how they interact and current challenges faced by MFIs, see Churchill and Coster (2001) and Steinwand (2000).

The most common and often the most serious vulnerability in a microfinance institution is credit risk, which is the deterioration in loan portfolio quality that leads to loan losses and high delinquency management costs (Churchill & Coster, 2001). This risk is also known as default risk and it relates to a client failing to meet the terms of a loan contract.

This is a serious concern for MFIs because most microlending is unsecured, that is, traditional collateral is not used to secure microloans. This means that if a loan defaults, an MFI will not have a collateral to seize and sell off to pay off the loan, thus increasing the risk of greater loss if a default occurs. This type of microlending allows MFIs to reach more low-income people, thus fostering financial inclusion, but this means that MFIs have higher credit risks compared to traditional banks (Uddin et al., 2024). To mitigate this risk, MFIs incur more overhead expenses for intensive screening and close monitoring of unsecured loans. Overall, microfinance business requires MFIs to maintain extended social networks, remote offices and a large pool of field staff. Hence, this leads to higher operational costs per loan as compared to banks and, as a result, MFIs charge higher interest rates than banks (Uddin et al., 2024).

One microloan does not pose a serious credit risk because it is such a small percentage of the total portfolio. Since most microloans are unsecured, delinquency [5] can quickly spread from a few loans to a significant portion of the portfolio. This contagious effect is magnified by the fact that microfinance portfolios have a high concentration in certain business sectors (Churchill & Coster, 2001). As a result, a large number of clients may be exposed to the same external threat, such as a crackdown on street vending or a livestock disease. These factors create volatility in microloan portfolio quality, thus emphasising the importance of managing credit risk.

2.3 Behavioural Finance in Credit Risk Management

Generally, financial institutions are more exposed to credit risk as compared to other risks. Hence, it is essential for these institutions to manage this risk to mitigate heavy losses. A

[5] Delinquency is a term used to characterise the status of contractual lending relationships with respect to repayments of interest and/or principal (Majaski, 2015).

key aspect of credit risk management is the decisions made to mitigate credit risk, such as client selection, loan approval and monitoring of loans. In the context of an MFI, such decisions would be handled by loan officers and their managers. According to behavioural finance, biases may influence the decisions made by loan officers and/or their managers, which can affect how effectively an MFI's credit risk is managed. Thus, this implies that rate of repayment for an MFI is not only affected by the borrowers' characteristics, but also the lenders' characteristics.

There are several studies which support this notion of lenders influencing the credit risk of an MFI. A study by Addae-Korankye (2014) found that poor loan portfolio quality is strongly linked to the managerial poor evaluation and selection of borrowers. As previously discussed, poor loan portfolio quality increases the likelihood of a default and hence increases an MFI's credit risk. In addition, a study by Baklouti (2015) revealed that loan officers' information evaluations are largely influenced by behavioural biases that are linked to judgmental errors in the granting process. Furthermore, Mighri (2014) conducted a study based on the assumption that credit risk can be linked to the decision-making process of loan officers and their behavioural biases. This study found that loan officers (in the sample) are generally considered as risk-takers and responsible for the lack of reimbursement at an advanced stage of the loan granting decision-making process.

Thus, behavioural biases may influence decision-making in the context of MFIs. Let us consider behavioural bias, overconfidence. According to Fersi and Boujelbène (2022), an overconfident manager or loan officer believes he or she has superior decision-making abilities, leading him or her to underestimate the risks and the exogenous noisy signals, on the one hand, and overestimating future returns and his or her problem solving capabilities, on the other. Possible consequences are reducing the provisions for future losses and approval of more loans with higher risks. These could be detrimental to an MFI as unexpected heavy losses can occur if, for example, many of the high risk loans default, which will be hard to manage with reduced provision. In the worst case scenario, an MFI becomes insolvent. Thus, it is of interest if such a bias can have a meaningful impact on the credit risk of an MFI.

In the next section, we describe the two quantitative methods that will measure and test the potential impact of overconfidence.

3 Research Methods

In this chapter, we discuss the two quantitative models, linear regression and random forest, used to examine the relationship between overconfidence and an MFI's credit risk. We describe how the models are formulated, how they measure the impact of a set of variables on one particular variable and evaluate the use of these models.

3.1 Linear Regression

In statistics and econometrics, linear regression is one of the most fundamental tools and it is widely used for modelling relationships between variables (Nguyen, 2020). This simple method typically provides adequate and interpretable descriptions of how a set of variables affects one particular variable (Hastie et al., 2009). Linear regression serves as a foundation for much of applied data analysis in fields ranging from economics to healthcare studies because of its simplicity and versatility (Nguyen, 2020).

There are two main uses for linear regression: understanding patterns in data and prediction. Linear regression provides a framework to summarise and explore relationships between variables, which allows for users to identify patterns such as trends or associations, further guiding analysis or decision-making. Additionally, linear regression is used for making predictions. For example, given historical data, a regression model can be used to predict future outcomes such as sales, prices and demand.

Nevertheless, it is crucial to understand that while linear regression can provide insights into potential causal relationships, it does not establish causation. For example, linear regression may show a strong positive association between advertising and sales, however this does not prove that the advertising directly causes sales to increase. There are other factors such as seasonality, market trends or unobserved variables that could also influence the results. Establishing causality requires additional steps like controlled experiments, instrumental variables techniques or careful observational study designs (Nguyen, 2020).

3.1.1 Model Formulation

So, why is it called 'linear'? This term refers to the structure of the model, where the response variable (or dependent variable) is modeled as a linear combination of one or more predictors (or independent variables) (Nguyen, 2020). For instance, simple linear regression is expressed as

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where Y is the response variable, X is the predictor variable, β_0, β_1 are parameters (or coefficients) to estimate and ϵ is the error term capturing random or unobserved factors that X cannot explain. We elaborate on the distribution of the error terms in the next sub-section.

Multiple linear regression extends simple linear regression by including more predictors. Suppose we have a dataset with N observations, then multiple linear regression has the form

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p} + \epsilon_i, \tag{1}$$

where Y_i is the response variable for the i-th observation, $X_{i,j}$ is the j-th predictor value for the i-th observation, β_j quantifies the association between that predictor $X_{i,j}$ and the response variable Y_i, and ϵ_i is the error term for the i-th observation. We interpret coefficient β_j as the average effect on Y_i of a one unit increase in $X_{i,j}$, while holding all other predictors fixed (James et al., 2021). Equation (1) holds for all $i \in \{1, 2, \ldots, N\}$.

3.1.2 Least Square Coefficients

The coefficients β_j's describe how each predictor $X_{i,j}$ relates to the response variable Y_i. However, these coefficients are unknown, so they need to be estimated. One of the most popular estimation methods is least squares. This method aims to find the coefficients that best describe the relationship between response variable and set of predictors, by choosing the coefficients that minimise the residual sum of squares (RSS) (James et al., 2021), given that certain assumptions hold.

Let us consider these assumptions. Let vector $\mathbf{Y} = Y_1, \ldots, Y_N$ denote the response variable and let matrix $\mathbf{X} = X_1, \ldots, X_p$ denote the set of p predictors, where X_j is a $N \times 1$ vector representing the j-th predictor. There are five classical assumptions for least squares and they are as follows:

1. Linearity: There is a linear relationship between the response variable and the predictors, which can be expressed as in equation (1).

2. No Perfect Multicollinearity: The columns of $\mathbf{X}$ must be linearly independent, that is, no predictor can be written as a linear combination of the other predictors.

3. Strict Exogeneity of Predictors: The predictors carry no information about the error term, which can be expressed as

   $$\mathbb{E}[\epsilon_i | X_{i,1}, \ldots, X_{i,p}] = 0$$

   for all $i \in \{1, \ldots, N\}$. By the Law of Iterated Expectations, this implies that

   $$\mathbb{E}[\epsilon_i] = \mathbb{E}\left[\mathbb{E}[\epsilon_i | X_{i,1}, \ldots, X_{i,p}]\right] = \mathbb{E}[0] = 0$$

   for all $i \in \{1, \ldots, N\}$.

4. Homoscedasticity: The variance of the error term is the same for all observations, regardless of the values of the predictors. This is formulated as

   $$\mathrm{Var}(\epsilon_i | \mathbf{X}) = \mathrm{Var}(\epsilon_i) = \sigma^2$$

   for some positive constant σ and for all $i \in \{1, \ldots, N\}$.

5. No Serial Correlation of Error terms: All error terms are uncorrelated with one another, which is expressed as

   $$\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$$

   for all $i, j \in \{1, \ldots, N\}$ such that $i \neq j$.

These assumptions are considered to be strong, as they are quite strict, but they are essential for least squares to produce the best estimates. There exists other, weaker assumptions for which least squares still provides reliable estimates.

For instance, a weak form of the third assumption (also known as weak exogeneity) is that the predictors are uncorrelated with the error term. This is expressed as $\mathbb{E}[\mathbf{X}_i^T \epsilon_i] = 0$ for all $i \in \{1, \ldots, N\}$. Note that strict exogeneity implies weak exogeniety, but not the other way around. Moreover, weak exogeneity sustains statistical inference about parameters of interest without loss of information while strict exogeneity sustains statistical forecasting (that is, prediction) (Hendry, 1995). As we are mainly interested in inference rather than prediction, we can replace the third assumption with weak exogeneity.

An additional assumption about the distribution of the error terms is that they are normally distributed, that is, $\epsilon_i \sim N(0, \sigma^2)$ for all $i \in \{1, \ldots, N\}$(Nguyen, 2020). This

is necessary for reliable and accurate statistical inference, such as hypothesis tests, confidence intervals and also predictions. This assumption is not required for least squares to produce the best estimates.

Let all the above assumptions hold and reconsider RSS. The residuals $e_1, \ldots, e_N$ are the difference between the observed values of $\mathbf{Y}$, denoted as $y_1, \ldots, y_N$, and the expected outcome predicted by the linear regression model conditioned on the set of predictors, denoted as $\mathbb{E}[Y_i|X_{i,1}, \ldots, X_{i,p}]$. Let $x_1, \ldots, x_p$ denote the observed values of predictors $X_1, \ldots, X_p$, then the residual is expressed as

$$\begin{aligned}
e_i &= y_i - \mathbb{E}[Y_i|X_{i,1} = x_{i,1}, \ldots, X_{i,p} = x_{i,p}] \\
&= y_i - \mathbb{E}\left[\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j + \epsilon_i | X_{i,1} = x_{i,1}, \ldots, X_{i,p} = x_{i,p}\right] \\
&= y_i - \mathbb{E}[\beta_0|X_{i,1} = x_{i,1}, \ldots, X_{i,p} = x_{i,p}] - \mathbb{E}\left[\sum_{j=1}^{p} X_{i,j}\beta_j | X_{i,1} = x_{i,1}, \ldots, X_{i,p} = x_{i,p}\right] \\
&- \mathbb{E}[\epsilon_i|X_{i,1} = x_{i,1}, \ldots, X_{i,p} = x_{i,p}] \\
&= y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j,
\end{aligned}$$

where we used that $\mathbb{E}[\epsilon_i|X_{i,1}, \ldots, X_{i,p}] = 0$ for all i in the second last equality. The residual is also known as the deviation of an observed value from its expected value, based on the regression model, hence it represents the error in the prediction of the observed value. It is clear to see that the residuals e_i estimate the error terms ϵ_i.
The RSS is defined as

$$\text{RSS} = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j\right)^2.$$

We note that the error terms are always unknown because we do not know the true values of the coefficients (Nguyen, 2020). Given the RSS, the least squares estimated coefficients $\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)$ are defined as

$$\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right) = \underset{(\beta_0, \beta_1, \ldots, \beta_p)}{\arg\min} \quad \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j\right)^2.$$

The least squares estimators are commonly used as they have three desirable properties (James et al., 2021; Nguyen, 2020). First, a least square estimator is unbiased, meaning that the expected value of the estimator is equal to the true value of the parameter. Second, this estimator is consistent, which means that the estimator converges to the true value of the parameter as the sample size increases. This property relies on the Law of Large Numbers, as it guarantees that larger samples reduce random fluctuations and lead to more precise estimates. Third, this estimator is efficient because it is the Best Linear Unbiased Estimator under the Gauss-Markov Theorem. An efficient estimator has the smallest variance amongst all unbiased estimators. These three properties show that least squares estimators are reliable, precise and robust, so using this method to estimate coefficients leads to accurate inferences.

Once we estimate the coefficients, we can quantify the relationship between the predictors and the response variable. Additionally, we can make more inferences using the coefficient estimates, such as standard errors, confidence intervals, hypothesis tests, p-values, etc.

A coefficient's standard error measures the uncertainty around the estimate of the coefficient for each predictor (James et al., 2021). In other words, it indicates how much the estimate deviates from the actual value. This shows how accurate the coefficient estimates are. Moreover, we can use standard errors to perform hypothesis tests on the coefficients.

A hypothesis test is a statistical tool that evaluates claims or hypotheses about a population based on data (Hastie et al., 2009). This test examines whether observed data provide sufficient evidence to reject a null hypothesis in favor of an alternative hypothesis. To perform this test, first define the hypotheses, then select a test statistic that qualifies evidence against the null hypothesis and then define the decision rule, where we reject the null hypothesis if the test statistic exceeds a critical value or if the p-value is below a confidence level. Otherwise, we fail to reject the null hypothesis, as there is insufficient evidence to rule it out. The p-value is the probability of obtaining results as extreme, or more extreme than, the observed results under the assumption that the null hypothesis is true (James et al., 2021).

The aim of this thesis is to establish whether there is a relationship between overconfidence and credit risk. Hence, our null hypothesis H_0 states that there is no relationship between overconfidence and credit risk and the alternative hypothesis H_1 states the converse. This is equivalent to $H_0 : \beta_o = 0$ and $H_1 : \beta_o \neq 0$, where β_o is the coefficient of overconfidence. For linear regression, the t-test statistic typically is used as the test statistic and it is defined as

$$t = \frac{\hat{\beta}_o - \beta_o}{\mathrm{SE}(\hat{\beta}_o)},$$

where $\hat{\beta}_o$ is the estimated coefficient of overconfidence, β_o is the true coefficient of overconfidence and $\mathrm{SE}(\hat{\beta}_o)$ is the standard error of the estimated coefficient. This test statistic measures the number of standard deviations $\hat{\beta}_0$ is away from β_o.

Under the null hypothesis $\beta_o = 0$, this test statistic has a t-distribution with d degrees of freedom. This allows for the computation of the probability of observing a specific testing statistic and we obtain the p-value. If the p-value is smaller than a set confidence level, then there is sufficient evidence to reject the null hypothesis and hence, there exists a relationship between overconfidence and credit risk. We set the confidence level to 0.05, which is a typical level when using hypothesis tests. If the p-value of an estimated coefficient is below the confidence level, then we call it a significant p-value.
For completeness, we apply this hypothesis test to all predictors of the models and present the results in Chapter 5.

3.1.3 Standardised Coefficients

Comparing the estimated coefficients can be challenging as predictors are measured in different units or have different variances, which means that the coefficients may vary in value. Thus, we consider the method of standardised regression coefficients. This method involves standardising each predictor and response variable, that is subtracting the mean of each variable from the original variable values and then dividing by the variables standard deviation, then performing linear regression, which results in standardised regression coefficients (Nimon & Oswald, 2013). This means that all predictors and response variable are standardised to have mean 0 and variance 1. This method allows for the direct comparison of the importance of each predictor in influencing the response variable. The

higher the standardised coefficient (ignoring the sign, as this shows the direction of the relationship between predictor and response variable), the more influential the predictor is on the standard deviation of the response variable.

3.1.4 Benefits and Limitations

Since linear regression is a widely used statistical method in various fields, let us consider the benefits and limitations of linear regression (Hastie et al., 2009; James et al., 2021).

Linear regression models are straightforward as a simple equation portrays the relationship between the outcome and the predictors. In addition, the impact of the predictors on the outcome and the significance of this impact can be quantified. Furthermore, linear regression can be used for prediction, which allows organisations and researchers to forecast future trends based on historical data.

Nevertheless, linear regression has several limitations. One of them is the violation of one or more assumptions, which reduces the quality of the estimates. Given that real world data tends to exhibit more complex, non-linear relationships, linear regression may fail to capture the true nature of the relationship and result in incorrect inferences.

However, there are ways to solve the violations of the assumptions. For example, if the relationship is non-linear, then apply transformations such as logarithm and square root, as they may help linearise the relationship between the predictors and the response variable. In addition, higher-order terms can be included in the regression equation, which can capture the non-linear relationships. If the residuals are not normally distributed and the sample size is large, linear regression can still provide reliable results thanks to the Central Limit Theorem [6]. More specifically, inference on the least square estimates is robust to moderate departures from normality, due to Central Limit Theorem (Nguyen, 2020).

Thus, linear regression is flexible, despite its assumptions. However, transformations and including higher-order terms makes the model more complex and may limit the interpretation of the coefficients.

[6]The Central Limit Theorem states that for a sufficiently large sample size ($n \geq 30$), the sampling distribution of the sample mean approaches a normal distribution, regardless of the population's original distribution (Nguyen, 2020).

3.2 Random Forest

Random forest is a machine learning method based on decision trees. Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on developing algorithms and models capable of learning from data (Nguyen, 2020). ML algorithms are repeatedly trained, that is, refined many times, until they accumulate a comprehensive list of instructions that allow them to function correctly. Sufficiently trained ML algorithms can learn patterns and relationships from data, and they can perform specific tasks such as sorting images, predicting housing prices, or making chess moves. Although the general principles underlying machine learning are relatively straightforward, the models that are produced at the end of the process can be very elaborate and complex (Molnar, 2025).

In recent years, there has been a surge in the use of ML methods for data analysis, as they rely on fewer assumptions on the underlying data distributions, can handle high dimensional data and make accurate predictions. Machine learning includes various techniques such as supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, algorithms are trained using a labeled training dataset to understand the relationships between features (or inputs) and an outcome (or output) (James et al., 2021). A labeled dataset means that the input data have corresponding labels, which are known as the output data. This type of learning is often used for creating ML models for predicting and classifying purposes. Unsupervised learning uses unlabeled datasets to train algorithms, that is, datasets with only features. This is typically used for identifying patterns within large, unlabeled datasets quickly and efficiently.

In general, supervised learning works by first training the model's algorithm, that is, the algorithm processes large datasets to explore potential relationships between features and outcome. Then, the model's performance is evaluated with test data (new data, that is not used for training) to find out whether the model was trained successfully using cross-validation. Cross-validation is the process of testing a model using a different portion of the dataset (James et al., 2021). Tree-based regression methods are one of the many supervised learning approaches in ML.

3.2.1 Tree-based Methods

A decision tree is a machine learning method for constructing prediction models from a dataset (Loh, 2011). This involves partitioning the set of features (or predictor), then for every observation in each partition, predict the observation by using the mean or mode of the outcome (or response variable) for the training observations in the region to which it belongs (James et al., 2021). Hence, fitting a single prediction model for each partition. using a tree, the splitting rules used to partition the set of features can be summarised (James et al., 2021). Hence, this approach is known as a decision tree method. Decision trees can be used for regression, classification, survival modelling, and more. They are very easy to interpret and explain to others as they can be displayed graphically.

Let us consider an example of a decision tree, shown in Figure 2, and explain its structure (GeeksforGeeks, 2022). A root node is the first or topmost node in a tree. Every tree always has one root node. Splitting is dividing the node into decision nodes, based on a specific feature. For example, if the root node is the number of home runs in baseball, then split this node into high batting scores and low batting scores. A decision node, also known as a child node, is a node which is the descendant (that is, it follows another node) of any node in a tree. A leaf node, or terminal node, is a node in a tree that does not have any child nodes. That is, this node is located at the edge of the tree, signifying the end of the branch. A sub-tree is a group of connected nodes with at least one child node within a tree.

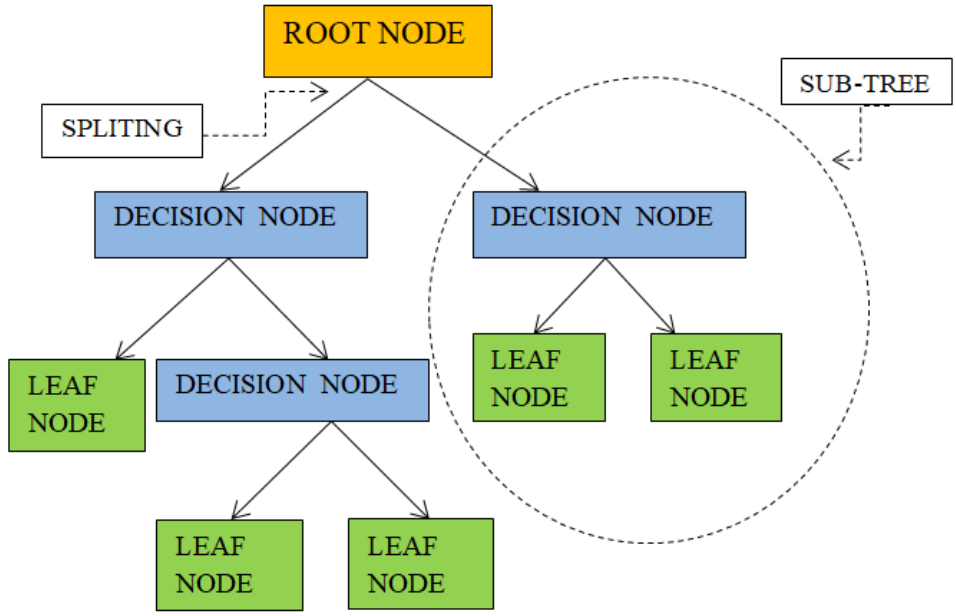

Figure 2: Graph showing a decision tree, with its different nodes, and a sub-tree (Bishnoi & Hooda, 2022).

Although decision trees are intuitive to understand and easy to interpret, they have a poor level of prediction accuracy, high variance and are very non-robust, so a small change in the data can lead to a large change in the final predicted tree (James et al., 2021).

Nevertheless, there are other tree-based methods such as bagging and random forest that aggregate many decision trees and this substantially improves the predictive performance of the trees and also their robustness (James et al., 2021). Such tree-based methods are also known as ensemble learning methods. These improvements come at the expense of some loss in interpretation, which we elaborate in the following section.

An ensemble method is an approach that combines many simple building block models to obtain a single and possibly very powerful model (James et al., 2021). These building block models are often known as weak learners, because on their own they may result in mediocre predictions. For tree-based ensemble methods, the simple building block is a decision tree. From here onwards, we focus on the context of regression and hence when we refer to decision trees we mean decision regression trees. For further reading on tree-based methods see Hastie et al. (2009), James et al. (2021), Liaw and Wiener (2002), and Loh (2011) for classification trees, Ishwaran et al. (2008) for survival trees and Meinshausen (2006) for quantile trees.

Bagging or bootstrap aggregation is a general-purpose technique for reducing the variance of a statistical learning method (James et al., 2021). Bagging is known to work well for decision trees as they can capture complex interaction structures in the data, have high variance and if grown sufficiently deep, have relatively low bias (Hastie et al., 2009). For regression, we repeatedly fit the same regression tree to bootstrap samples from the training dataset, that is, sampling from the training dataset with replacement, and then average the resulting predictions. Though each individual tree has high variance, taking the average reduces the variance. However, if the trees are highly correlated due to, for example, a very strong feature in the dataset, then most or all trees will use this feature in the top split. As a result, all of the bagged trees will be highly correlated and averaging

them will not reduce the variance as much as when the trees are uncorrelated. In this case, bagging will not lead to a substantial reduction in variance compared to a single decision tree (James et al., 2021).

3.2.2 Random Forest: Model Formulation

To overcome this, random forest is proposed by Breiman (2001) as a modified version of bagging and it has the aim to improve the variance reduction of bagging by decorrelating trees (that is, reducing the correlation between trees), without increasing the variance too much (Hastie et al., 2009). This is achieved by incorporating an additional layer of randomness when partitioning (Breiman, 2001).

In standard decision trees, each node is split using the best split among all features, where this split maximises the reduction in the sum of squared residuals. The residual is the difference between the observed outcomes and the predicted outcomes. In a random forest, before splitting a node, first take a random sample of m features from the set of p features and then find the best split among the features in the random sample, where typically $m = p/3$. Referring to the case where there is a very strong feature, this is a clever approach as it forces each split to consider only a subset of the features and thus, on average $(p-m)/p$ of the splits will not even consider the strong feature and other features will get a chance (James et al., 2021). This process can be described as decorrelating the trees and hence taking the average substantially reduces the variance and maintains the low bias of each tree (Breiman, 2001). Note that bagging can be thought of as a special case of random forest, when $m = p$. Algorithm 1 summarises how random forest works.

Algorithm 1 Random Forest

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}$ of size N from the training data.
   (b) Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m features at random from the p features.
      ii. Pick the best split-point among the m features.
      iii. Split the node into two child nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.
3. To make a prediction at a new point x: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$.

Based on the training data, an estimate of the error rate can be obtained at each bootstrap iteration, by predicting the data not in the bootstrap sample (which is also known as "out-of-bag" or OOB data) using the tree grown with the bootstrap sample (Liaw & Wiener, 2002). This is possible as a bootstrap sample is only a subset of the training data. Given an observation from the training data, predict it using all trees that excluded this observation in the bootstrap sample, and these predicted values are known as OOB predictions. Then take the average of the OOB predictions and calculate the error rate by comparing the average OOB predictions to the observed outcome. This error is known as the OOB estimate of the error rate. This estimate is quite accurate, given that a sufficient number of trees are grown (Breiman, 2001). The OOB error serves as an efficient, unbiased estimate of the test error without the need for cross-validation (Nguyen, 2020).

Hyperparameters

An important part of an ML's training process is the hyperparameters. Unlike parameters which allow the model to learn the rules from the data, hyperparameters control how the model is trained (Molnar, 2025). Hyperparameters are manually set in advance, that is, before the training process. While parameters can be estimated by fitting the given data to a model, hyperparameters cannot be learned this way. Instead, many potential combinations of hyperparameters are tested and the best performing one is selected. The choice of hyperparameters determines the efficiency of the ML algorithm, hence finding the optimal hyperparameter values can considerably improve the performance of the model (Nguyen, 2020).

Random forest involves many hyperparameters such as the structure of each individual tree (e.g., the minimal size or number of observations n_{min} a node should have to be to split), the structure and size of the forest (e.g., the number of trees n_{tree} to grow), and the level of randomness (e.g., the number m of features considered as candidate splitting features at each split or the sampling scheme used to generate the datasets on which the trees are built) (Probst et al., 2019). Since the hyperparameters control how the random forest trains, this influences the random forest's performance and hence, the accuracy of its predictions.

Although random forests perform well with default hyperparameters, tuning these hyperparameters can improve the performance of random forest (Probst et al., 2019). Tuning involves testing multiple values for a hyperparameter and finding the value which leads to the lowest OOB error rate.

When applying random forest, we will compare the results of the default and tuned hyperparameters. We use the R function 'ranger' to perform random forest (Wright, 2017b), which is a fast implementation of random forests by Breiman (2001) for high dimensional data. To tune the hyperparameters, we use R function 'tuneRanger' (Probst & Klau, 2024), an automatic tuning method for random forests of the 'ranger' function. This tuning function uses a tuning strategy called sequential model based optimisation (SMBO) and the following three hyperparameters are tuned: the number of features in random subsample m, the minimum node size n_{min} and the fraction of the sample size s_f to be used. OOB predictions are used for evaluation, which makes it faster than other packages and tuning strategies that use for example K-fold cross validation (Breiman, 2001). SMBO is a very successful tuning strategy that iteratively tries to find the best hyperparameter settings based on evaluations of the previously done hyperparameters. For more details on this tuning strategy, see Probst et al. (2019).

3.2.3 Variable Importance

Random forests offer measures of variable importance, which rank features according to their predictive abilities. Variable importance is a measure reflecting the importance of a variable (or feature) in a random forest prediction rule or split-point (Probst et al., 2019). This measure can also capture non-linear and interaction effects. Examples of variable importance include impurity importance (or mean decrease of impurity) and permutation importance (or mean decrease of accuracy). In the impurity variable importance, a feature is important if it leads to a large decrease in impurity when used for splitting. Impurity is a measure used in decision algorithms to determine the best feature to split the data at each node (Nguyen, 2020). Impurity measures depend on the type of decision trees used (Nembrini et al., 2018).

For regression, the impurity measure is the residual sum of squares (RSS). Thus, a fea-

ture is considered to be important if it causes a large decrease in RSS when splitting. For permutation importance, a feature is important if it has a positive effect on the prediction performance, estimated by the OOB prediction error (Nembrini et al., 2018). In both cases, the larger the importance value, the higher the influence on the outcome. Though both are commonly used, we choose to use permutation importance as the impurity importance is known to be biased to features with many possible split points (Nembrini et al., 2018). This is because for such features, it may have a higher probability to produce a split with a larger impurity reduction (that is, a reduction in RSS) than a feature with fewer possible splits, even if both features are unassociated with the outcome (Nembrini et al., 2018).

As the impurity importance is calculated based on the features selected for splitting, it is also biased. For example, a continuous feature will obtain higher impurity importance values than dichotomous features, even if both are uninformative. In our dataset, most features are continuous, but the same issue could arise due to the difference in the number of distinct numerical values between the features, therefore, we use permutation importance.

Approximating p-values

Based on variable importance measures, there are a few methods that allow the approximation of p-values for the importance of the features. One such method is proposed by Altmann et al. (2010), which uses the permutation importance as the importance measure and is known as the PIMP (permutation importance) Algorithm. The idea behind this algorithm is to assess the relevance of a set of features with respect to an outcome.

The PIMP Algorithm permutes the outcome s many times, but does not change the features, which preserves the dependence between the features. For each permutation of the outcome, the relevance for all features is assessed. This leads to a vector of s importance measures for each feature, which is called the null importances. The algorithm then fits a probability distribution to the population of null importances, by first applying the Kolmogorov-Smirnov tests to identify the most appropriate probability distribution between Gaussian, lognormal and gamma. Given the chosen distribution, maximum likelihood estimators of the parameters are computed and then the probability of observing a relevance of v or higher using the true outcome can be computed, that is, the PIMP p-values. With the same interpretation as statistical p-values, permutation importance values with a p-value below a set confidence level, are considered to have significant p-values. Thus, features with significant p-values are informative features. We use the R function 'importancepvalues' (Wright, 2017a) to obtain the permutation importance and p-values for every feature.

3.2.4 Partial Dependence Plot

Although variable importance measures help identify informative features, they do not quantify the association between features and outcome. However, we can visualise this association using a partial dependence plot. This plot shows the average marginal effect of one or two features on the predicted outcome of a machine learning model (Friedman, 2001). This plot can show whether the relationship between the feature and predicted outcome is linear, monotonic or non-linear. To measure the partial dependence, marginalise the machine learning model over the distribution of the set of features (excluding the feature of interest) (Molnar, 2025). This results in a function showing the relationship between the feature of interest and predicted outcome. In other words, marginalising over all other features leads to a function that depends only on the feature of interest, while

still including the interactions with the other predictors. The partial dependence function for regression is defined as

$$f_S^{pdp}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, \mathbf{x}_C^{(i)}),$$

where x_S is some value for the feature of interest, $n \times (p-1)$ matrix $\mathbf{x}_C$ are the other features used in the machine learning model $\hat{f}$ and n is the number of observations.

A partial dependence plot provides a clear interpretation of how changing the j-th feature changes the average prediction in the dataset. In addition, it is intuitive and easy to compute. However, the maximum number of features in a partial dependence function that can be meaningfully visualised is two. This is because it is difficult to imagine more than three dimensions. Moreover, partial dependence assumes all features to be independent. When this assumption is violated, the plot creates new data points in areas of the feature distribution where the actual probability is very low (Molnar, 2025).

3.2.5 Benefits and Limitations

Since we have a better understanding of how random forest works, let us consider the benefits and limitations of using this method (Breiman, 2001; Hastie et al., 2009; James et al., 2021; Nguyen, 2020).

One of the benefits of random forest is its high accuracy in predictions. This can be attributed to the decorrelation of regression trees, as the variation associated with individual trees decreases, and the aggregation of the regressions, which both lead to more accurate predictions. This result is shown in several studies that compared random forest with other models, such as decision trees, linear regression, logistic regression, etc. Furthermore, random forest makes no assumptions about the data distribution or correlation between the predictors and the response variable. This method can also handle outliers. With variable importance, random forest can help with identifying features that are most influential, and hence, that are most relevant for prediction.

However, there are some limitations to using random forest. Given the ensemble approach, it is hard to interpret the model itself and to quantify the direct impact of each feature. This is why such methods are known as 'black-box' models. It is computationally expensive to fit multiple trees or to train a random forest model on a large dataset, which can result in greater memory utilisation and longer training times. In addition, prediction times are greater as to get a final forecast each observation must navigate through several decision trees in the forest. Moreover, random forest can suffer from overfitting when the model captures noise in the training data, resulting in poor generalisation on new data. This is because of the algorithm's ability to fit the training data too closely.

4 Data Analysis

In order to test the impact of overconfidence on the credit risk of a microfinance institution, we apply quantitative modelling to data obtained from The World Bank Group. We explore the data by describing the variables and their characteristics. Due to numerous missing observations, we will apply an imputation method and evaluate its effect on the final data. We then describe the final dataset, on which we perform quantitative analysis. All coding is done in R.

4.1 Data

We obtain the data from the Mix Market (Microfinance Information Exchange Market) database (The World Bank Group, 2019) and World Development Indicators database (The World Bank Group, 2023), both accessed from The World Bank Group.

The Mix Market database contains financial and balance sheet performances of numerous microfinance institutions (MFIs) across the world. This database has been reported by financial services providers targeting people who do not have access to traditional banks in developing markets around the globe (The World Bank Group, 2019). The data points include data on financial statements (income statement, balance sheet), operations, financial products, end clients, social performance, etc. They were collected and reported in line with broadly recognized reporting standards within microfinance and inclusive finance. This data reports 68,520 observations of 2,855 MFIs over the period June 1999 to September 2019. Those MFIs operated in countries in Africa, Eastern Europe, the Pacific, Latin America, the Caribbean, and Asia. Many studies have used this database to compare and analyse the performance of financial services providers and MFIs in developing markets.

To involve macroeconomic aspects, we also include data from the World Development Indicators. The World Development Indicators database is the primary World Bank collection of development indicators, compiled from officially recognised international sources (The World Bank Group, 2023). It provides the most up to date and accurate global development data available. We only consider development data indicators involving all countries included the Mix Market database.

We choose 12 variables for the analysis of behavioural bias and an MFI's credit risk, based on past research showing these variables as relevant (Assefa et al., 2013; Banto & Monsia, 2021; Fersi & Boujelbène, 2022; Leite et al., 2018). Due to the limited time and resources for this thesis, we were unable to undergo a behavioural study, by interviewing loan officers/managers at various MFIs and analysing how their own biases (specifically, overconfidence) affect their risk decisions. Thus, we rely on constructed proxies for overconfidence, based on financial data.

4.1.1 Variables

We describe the variables, obtained from the Mix Market data and World Development Indicators data, that will be included in the analysis of behavioural bias and credit risk. We first consider the two response variables, then indicators of an MFI's performance, including the proxies of overconfidence, and lastly, macroeconomic variables.

Response Variables

In order to measure the credit risk of an MFI, we need to monitor the quality of their loan portfolio. The quality of an MFI's portfolio refers to the quality of its clients' repayment and thus, a credit is qualified as healthy when it is repaid on its due date (Fersi

& Boujelbène, 2022). The most accepted measure of portfolio quality is the Portfolio at Risk (PAR) ratio (Churchill & Coster, 2001; de Oliveira Leite et al., 2020). This is the outstanding amount of all loans that have one or more installments of principal past due by a certain number of days, such as 30, 90 or 180 days (Christen & Lyman, 2003). The most commonly used ratio for assessing an MFI's loan portfolio is the PAR 30 days (Assefa et al., 2013; D'Espallier et al., 2011; Fersi & Boujelbène, 2022). The formula for PAR 30 days is defined as

$$\begin{aligned}&\text{Portfolio at Risk} > 30 \text{ days} \\ &\qquad = \frac{\text{Portfolio Overdue at least 30 days} + \text{Renegotiated Portfolio}}{\text{Average Gross Loan Portfolio}}.\end{aligned}$$

In banking literature, PAR 90 days is also commonly used to convey a bank's credit risk as it represents non-performing loans (NPLs) (Arko, 2012; Fofack, 2005). A non-performing loan generally refers to a loan that does not generate income for a relatively long period of time, that is, the principal and/or interest on these loans has been left unpaid for at least 90 days (Alton & Hazen, 2001; Naili & Lahrichi, 2022b). Hence, NPLs indicate more severe delinquency. Additionally, a large amount of NPLs can lower an MFI's ability to grant more credit because the loanable funds tend to deplete when repayment of loans is delayed or fails to come (Berger & DeYoung, 1997).

We consider both PAR 30 days and PAR 90 days as indicators of loan portfolio quality and hence, the credit risk of an MFI. We select the *portfolio at risk of more than 30 days* and *portfolio at risk of more than 90 days* as response variables in the credit risk models.

Proxies for Overconfidence

We describe the three proxies that indicate the overconfidence of a loan officer/manager at an MFI.

The *loan growth (LG)* indicator is the level at which the loan portfolio grows and is used in literature as an indicator of credit risk (Foos et al., 2010). An aggressive loan growth is linked with loans granted to more risky borrowers (Naili & Lahrichi, 2022a). This is a good proxy for behavioural bias overconfidence, as overconfidence leads to lenders (that is, loan officer/manager) overestimating the repayment capacities of borrowers and hence, being more willing to grant loans (Fersi & Boujelbène, 2022). Additionally, these overconfident lenders tend to underestimate the risk profile of their borrowers and thus, tend to lend more to high-risk borrowers, further increasing the MFI's exposure to credit risk (Fersi & Boujelbène, 2022). Therefore, the two consequences of overconfidence lead to declining lending standards and granting more credits, which increases credit risk. Using the net loan portfolio variable, which is calculated in US dollars as the value of loan portfolio net of impairment loss allowance and unearned income and discount (The World Bank Group, 2019), we calculate loan growth as follows:

$$\text{Loan Growth}_t = \frac{\text{Net Loan Portfolio}_t - \text{Net Loan Portfolio}_{t-1}}{\text{Net Loan Portfolio}_{t-1}} \times 100,$$

where t refers to the year [7].

[7]Note that when the value of the Net Loan Portfolio at time t is 0 or not available (NA), we set Loan growth$_{i,t}$ to NA.

Loan loss provision (LLP) is an expense set aside as an allowance for uncollected loans and loan payments (Banto & Monsia, 2021). This is expressed mathematically as follows (Fersi & Boujelbène, 2022; The World Bank Group, 2019):

$$\text{Loan Loss Provision} = \frac{\text{Impairment Losses on Loans}}{\text{Average Total Assets}} \times 100$$

This is to address the inherent risks prevalent in the loan portfolio and can be considered as a measure of the quality of a financial institution's assets (Banto & Monsia, 2021). The higher this ratio is, the worse the quality of bank assets is. Additionally, this provision reflects the prospects of lenders for possible future risk and thus, the misestimation of future risk would be reflected by smaller provisions against borrowed assets (Fersi & Boujelbène, 2022). Moreover, this provision is expected to increase simultaneously with the growth of the loan portfolio. Therefore, this can also serve as a proxy for overconfidence.

The *net profit margin (NPM)* measures the percentage of operating revenue remaining after all financial, loan loss provision, and operating expenses are paid (Christen & Lyman, 2003). This indicator helps to measure the commercial performance of MFIs and it depends on an MFI's ability to generate a causal link between operating revenue and operating expenses (Banto & Monsia, 2021). Thus, to ensure their profitability, operating expenses should not grow faster than operating revenue. According to Fersi and Boujelbène (2022), net profit margin is negatively linked to overconfidence, where low margins reflect higher overconfidence. Foos et al. (2010) illustrated that an excessive risk-taking is detected through low-interest loans or offering high-interest deposits, which negatively affects the risk levels in the long run. Since the demand for microfinance services is growing, an MFI is more likely to decrease the NPM if the demand for saving services increases relative to the demand for loans (Fersi & Boujelbène, 2022). Hence, the net profit margin is an appropriate proxy for overconfidence and it is formulated as follows:

$$\text{Net Profit Margin} = \frac{\text{Net Operating Income}}{\text{Operating Revenue}} \times 100.$$

Other indicators of MFI's performance

Risk coverage ratio shows how much of the portfolio at risk is covered by an MFI's loan loss provision (Christen & Lyman, 2003) and is expressed mathematically as follows:

$$\text{Risk Coverage} = \frac{\text{Loan Loss Provision}}{\text{Portfolio at Risk of more than 30 days}} \times 100$$

This ratio shows how prepared an institution is to absorb loan losses in the worst-case scenario. MFIs should provision according to the age of their portfolio at risk, that is, the older the delinquent loan, the higher the loan loss provision (Christen & Lyman, 2003). For example, a ratio for PAR $>$ 180 days may be close to 100%, whereas the ratio PAR $>$ 30 days is likely to be significantly less. Thus, a risk coverage ratio of 100% is not necessarily optimal. We include this ratio as several studies have shown that risk management is achieved through the provisions for losses (Fersi & Boujelbène, 2022).

In the banking literature, *bank size* is reported to have an impact on credit risk (Chaibi & Ftiti, 2015; Fersi & Boujelbène, 2022; Naili & Lahrichi, 2022a). Though MFIs tend to have

lower total assets than banks, hence are smaller in size, bank size may also influence risk-taking decisions and thus, also the credit risk of MFIs. The size of banks determines the effectiveness of smaller and larger banks and their ability to manage risk (Bandyopadhyay & Saxena, 2023). Bank size is estimated by taking the natural logarithm of a bank's total assets [8] (Chaibi & Ftiti, 2015; Naili & Lahrichi, 2022b), that is,

$$\text{Bank Size} = \ln\left(\text{Total Assets}\right).$$

Total assets include all asset accounts net of all contra-asset accounts, such as the loan loss provision and accumulated depreciation (Christen & Lyman, 2003). Since the variable total assets is larger, in size, than other variables, the natural logarithm expression for total assets can make the variable more comparable in scale to other variables (Akash, 2022).

Another indicator of an MFI's performance is the *number of active borrowers*. Several studies show that MFIs with more borrowers can effectively achieve economies of scale and become more efficient in loan recovery (Abusharbeh, 2023; Ngonyani & Mapesa, 2019). Additionally, they revealed that growth in the number of borrowers has a positive influence on the quality of loans in MFIs. Given that the number of active borrowers is quite skewed, we take the natural logarithm and hence, define the number of active borrowers as

$$\text{Active Borrowers} = \ln\left(\text{Number of Active Borrowers}\right).$$

Return on assets (ROA) reflects how well banks and other commercial institutions use their total assets to generate returns (Banto & Monsia, 2021). Moreover, return on equity is a common indicator that measures the attractiveness of an MFI. However, the two types of MFIs (i.e. for-profit and non-profit) are both included in the data. Hence, ROA is a suitable indicator for both types of MFIs (Banto & Monsia, 2021; Fersi & Boujelbène, 2022). This is formulated as follows:

$$\text{Return on Assets} = \frac{\text{Net Operating Income} - \text{Taxes}}{\text{Average Total Assets}} \times 100.$$

Country-level variables

We include three country-level variables to control for the macro-environment: GDP growth, inflation rate and interest rate.

The *Gross Domestic Product (GDP) growth* is defined as the yearly percentage change in the natural logarithm of real GDP in each country (Naili & Lahrichi, 2022a). During economic growth, individuals and firms face sufficient stream of revenues to repay their financial obligations (Naili & Lahrichi, 2022a). During challenging times, firms and households are more likely to default on their loans due to the decrease in asset values that act as collateral. This results in an increase in delinquent loans and hence, increases credit risk. Many studies have shown that GDP growth is negatively related to a bank's credit risk (Chaibi & Ftiti, 2015; Hayati & Ariff, 2007; Naili & Lahrichi, 2022a; Naili & Lahrichi, 2022b). The World Bank Group (2023) defines GDP as the sum of gross value added by

[8] Note that when total asset is zero or not available (NA), we set bank size to NA.

all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products.

The *inflation rate* indicates the broad increase in the prices of goods and services over time (Eurosystem, n.d.). In particular, inflation reduces the value of the currency over time. Although there is ample evidence that there exists a causal relationship between a bank's credit risk and inflation rate (Chaibi & Ftiti, 2015; Naili & Lahrichi, 2022a; Nkusu, 2011), the direction of this relationship is ambiguous. On the one hand, higher inflation could reduce the real value of outstanding loans. On the other hand, it can cause the debt-servicing capacity of borrowers to deteriorate by reducing their real income. Consequently, the relationship between inflation rate and credit risk can be either positive or negative. To capture inflation rate, we use the annual inflation rate in a country as measured by the consumer price index which reflects the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly (Naili & Lahrichi, 2022a; Nkusu, 2011; The World Bank Group, 2023).

Another factor that can influence the credit risk of MFIs is the *lending interest rate*. The lending interest rate is the bank rate that usually meets the short- and medium-term financing needs of the private sector (The World Bank Group, 2023). Prior studies show that high interest rates impact the interest rate of banks, which results in increased levels of NPLs (Naili & Lahrichi, 2022b). Furthermore, increasing interest rates lead to higher loan losses as soaring interest payments deteriorate borrowers' repayment capacity (Naili & Lahrichi, 2022b). The International Monetary Fund (IMF) collects the lending interest rates, which act as representative interest rates offered by banks to resident customers (The World Bank Group, 2023).

4.2 Handling Missing Values

Due to a substantial amount of missing values in the dataset, we only keep the observations in time periods with at most 75% of the data missing, so the dataset is reduced from $68,520$ to $30,998$ observations [9] over the time period 2005 to 2015. Table 1 provides a summary of statistics of the current dataset.

Despite reducing the number of time periods, we still observe a large amount of missing values. Missing data is a common problem in data science, and it impacts the quality and reliability of inferences derived from datasets. There are several ways to handle missing data. But first, let us visualise the missing data.

A heatmap is a visualisation tool that uses colours to graphically represent data values (Nguyen, 2020). This tool highlights where the missingness occurs in a dataset. Figure 3 shows a heatmap of our dataset, which indicates in black the location of missing values. We notice that most of the missingness is in the Mix Market data rather than the World Development Indicator data. Nevertheless, no pattern of missingness is visible due to the substantial amount of missing data.
A straightforward approach to deal with missing data is to delete all observations with missing values, and perform our analysis on the complete observations. However, this may lead to the loss of valuable information as there is a high fraction of missing values and

[9] Note that some observations were removed whilst computing loan growth, in the case where net loan portfolio at a certain time was zero. This was to avoid loan growth equal to infinity. Hence, there are less than half the original number of observations.

Variables	Min.	Median	Mean	Max.	N
Active Borrowers	0.000	9.047	8.966	15.92	12,288
Bank Size	0.000	15.461	15.551	24.47	12,774
Loan Growth (%)	-4,214.092	22.329	113.367	145,914.92	9,898
Loan Loss Provision (%)	-61.060	1.020	1.868	125.51	10,264
Net Profit Margin (%)	-3,549,562.460	10.400	214.715	7,209,813.12	11,832
Portfolio at Risk 30 Days (%)	0.000	3.810	7.202	711.43	1,0128
Portfolio at Risk 90 Days (%)	-1.660	2.300	5.095	512.58	9,329
Return on Assets (%)	-746.370	2.000	0.635	97.31	10,455
Risk Coverage (%)	-107.69	86.600	46,635	412,968,200	9,275
GDP Growth (%)	-46.082	5.961	5.702	53.38	30,908
Inflation Rate (%)	-10.067	5.924	6.770	62.17	30,241
Interest Rate (%)	3.454	12.936	14.277	82.33	26,109

Table 1: Summary of statistics for dataset over the period $2005 - 2015$, showing the minimum, median, mean, and maximum values, and the number of observations N.

hence this will further reduce the final dataset. An alternative approach is imputation, which is to replace the missing values with new values (James et al., 2021). Missing values can be replaced using various methods, such as by the mean of column of the missing variable or by taking the median of the column. Another method is to estimate a predictive model for each variable given the other variables if the variables have at least some moderate degree of dependence, and then impute each missing value by its prediction from the model (Hastie et al., 2009). To preserve the information within the data, we choose to implement an imputation method.

4.2.1 Imputation Method

There are several imputation methods for missing data, such as KNN Imputation and Random Forest (missForest). The K-Nearest Neighbor (KNN) Imputation approach identifies the k nearest observations based on a distance metric and imputes missing values using a weighted average (for continuous variables) or the mode (for categorical variables) (Nguyen, 2020). This is a simple non-parametric method [10] and is easy to interpret. However, it has a high run-time for imputation (especially for high-dimensional samples) and is sensitive to values of k.

The Random Forest Imputation (also known as MissForest algorithm) is an approach that uses an iterative process where a random forest model predicts missing values for one variable at a time, treating other variables as predictors (Nguyen, 2020; Stekhoven & Bühlmann, 2011). This process continues until the average out-of-bag (OOB) error of the forests stops improving (that is, convergence of OOB error). The missing values are filled by the OOB predictions of the best iteration. This imputation method captures complex interactions and non-linearities and handles mixed data types very well, but it is computationally expensive for large datasets and sensitive to the quality of data relationships.

[10] A non-parametric method is a flexible statistical approach that makes no (or minimal) assumptions about the underlying distribution of the data (Bagdonavicius et al., 2010).

Figure 3: A heatmap of the dataset, showing the missingness patterns for each variable.

However, there is an alternative method to the MissForest algorithm which is called missRanger. This is a a multivariate imputation algorithm that is faster than MissForest, as it uses the R package 'ranger' to fit random forests (Wright & Ziegler, 2017). Wright and Ziegler show in their 2017 paper that this package is much faster than MissForest and other methods. The missRanger algorithm runs similarly to the MissForest algorithm, as it iterates until the average OOB error of the random forests stops improving and fills the missing values by the OOB predictions of the best iteration.

One difference between these two algorithms is that missRanger has the optional step of using predictive mean matching (PMM) after finding the OOB predictions of the best iteration. PMM operates by finding observed values with similar predictive characteristics to the missing entries (Nguyen, 2020). The missing values are then imputed, thus preserving the distribution and variance of the original data more effectively than simpler methods, such as mean imputation. In addition, this step avoids imputing values not present in the original data such as a value 0.3334 in a $0-1$ coded variable. Hence, we decide to use missRanger as the imputation method.

Before we impute the data, we first remove outliers for three variables, Loan Growth, Net Profit Margin, and Risk Coverage, as they are heavily affected by outliers. This can be seen in Table 1, as those variables have minimum and maximum values that are very different and far from the mean, especially compared to the other variables. These outliers can distort the distribution of those variables and will affect the imputation method, for example by imputing values that are within this wide range of values. This will also affect the results of the statistical and machine learning methods, potentially rendering them inaccurate.

Hence, we apply the interquartile range (IQR) method to remove the outliers. The IQR measures the spread of the middle 50% of the data values, and an observation is an outlier if it has a value 1.5 times greater than the IQR or 1.5 times less than the IQR (James et al., 2021). This reduces the number of observations to 14,432, with 1,312 MFIs

across 100 countries[11]. Table 2 shows the resulting minimum, median, mean and maximum values for the three variables after applying the IQR method. We observe more reasonable values, and they are in line with previous studies (Fersi & Boujelbène, 2022; Foos et al., 2010; Leite et al., 2018; Naili & Lahrichi, 2022b).

Variables	**Min.**	**Median**	**Mean**	**Max.**
Loan Growth (%)	-62.76	18.006	20.684	113.798
Net Profit Margin (%)	-34.70	13.280	14.124	56.100
Risk Coverage (%)	-69.79	69.850	74.475	271.500

Table 2: Summary of statistics for three variables after removing outliers using the interquartile range method, showing the minimum, median, mean, maximum values.

We use the R function 'missRanger' (R Documentation, 2017) to impute the missing data and combining it with the observed values, this yields the new complete dataset, which is summarised in Table 3. In the next section, we evaluate the effect of the imputation method missRanger.

Variables	**Min.**	**Median**	**Mean**	**Max.**	**Std. Dev.**
Active Borrowers	0.693	9.154	9.107	15.787	2.038
Bank Size	4.779	15.117	15.205	24.468	2.179
Loan Growth (%)	-62.760	18.049	20.792	113.798	30.253
Loan Loss Provision (%)	-29.910	1.170	2.001	54.350	3.592
Net Profit Margin (%)	-34.700	13.090	13.958	56.100	14.883
Portfolio at Risk 30 Days (%)	0.000	5.625	9.584	373.180	14.911
Portfolio at Risk 90 Days (%)	-1.660	2.410	5.525	512.580	14.998
Return on Assets (%)	-246.680	1.710	0.136	97.310	13.449
Risk Coverage (%)	-69.790	66.550	72.459	271.500	50.768
GDP Growth (%)	-46.082	6.014	5.713	53.382	3.629
Inflation Rate (%)	-10.067	5.924	6.734	59.220	4.989
Interest Rate (%)	3.552	13.021	14.342	65.418	7.882

Table 3: Summary of statistics for imputed dataset over the period $2005 - 2015$, showing the minimum, median, mean, and maximum values, and standard deviation.

4.2.2 Impact of Imputation Method

By imputing missing data, this will clearly have an influence on the dataset and hence the quality of the statistical results. For instance, an imputation approach can introduce bias if it is poorly executed, which is when the imputed data fails to preserve the integrity of the relationships among the variables (Nguyen, 2020). Therefore, it is important to evaluate the impact of the chosen imputation method.

[11]The following 15 countries were removed: Angola, Central African Republic, Croatia, Fiji, Gabon, Gambia, Guinea-Bissau, Guyana, Hungary, Liberia, North Macedonia, Serbia, Solomon Islands, Turkiye, and Venezuela.

Let us have a look at the summary statistics of the imputed data and compare them to the original data. Table 4 shows the median, mean and standard deviation of the original data and the imputed data. We exclude the minimum and maximum values because they are identical for both datasets. We see that for most variables, the values are close for the original data and imputed data. There are a few cases where the difference is quite large, such as for the standard deviation of Portfolio at Risk 90 Days and the median and mean of Risk Coverage. Overall, we have that the imputation method preserves the median, mean and standard deviation of the variables reasonably well.

	Original Data			*Imputed Data*		
Variables	**Median**	**Mean**	**St.Dev**	**Median**	**Mean**	**St.Dev**
Active Borrowers	9.139	9.108	2.044	9.154	9.107	2.038
Bank Size	15.157	15.272	2.239	15.117	15.205	2.179
Loan Growth (%)	18.006	20.684	30.108	18.049	20.792	30.253
Loan Loss Provision (%)	1.190	2.011	3.420	1.170	2.001	3.592
Net Profit Margin (%)	13.280	14.124	14.875	13.090	13.958	14.883
Portfolio at Risk 30 Days (%)	5.335	8.928	13.645	5.625	9.584	14.911
Portfolio at Risk 90 Days (%)	2.320	5.062	12.150	2.410	5.525	14.998
Return on Assets (%)	1.720	0.472	12.418	1.710	0.136	13.449
Risk Coverage (%)	69.850	74.475	50.587	66.550	72.459	50.768
GDP Growth (%)	6.014	5.714	3.631	6.014	5.713	3.629
Inflation Rate (%)	5.924	6.736	4.996	5.924	6.734	4.989
Interest Rate (%)	13.021	14.370	8.163	13.021	14.342	7.882

Table 4: Comparison of the summary of statistics for the original and imputed dataset over the period $2005 - 2015$, showing the median, mean and standard deviation for each variable.

Let us also consider the correlation between the variables, which is shown in Figures 4 and 5. We see that some correlations are higher in the imputed data than in the original and also lower in the imputed data than in the original data. The most significant difference is for the correlation between Loan Loss Provision and Return on Assets, which is -0.31 original data but increased to -0.23 in the imputed data. Hence, the correlation between these two variables is weaker. We also notice that the correlation between Portfolio at Risk 30 days with Return on Assets has strengthened, whilst with Risk Coverage it has weakened. However, on an overall level the correlations have not changed considerably and all directions are maintained, that is, the correlations have the same signs for both datasets.

Although there are various other ways to evaluate the impact of the imputation method, we are unable to do many of them as there are too few complete observations (that is, an MFI with no missing value in a specific year) over consecutive years. Hence, this limits the extent to which we can examine the effect of using an imputation method. We decide to continue the rest of the analysis using the imputed data, however we should be careful in the interpretations of the results as we cannot guarantee that the imputed data accurately represents the original data.

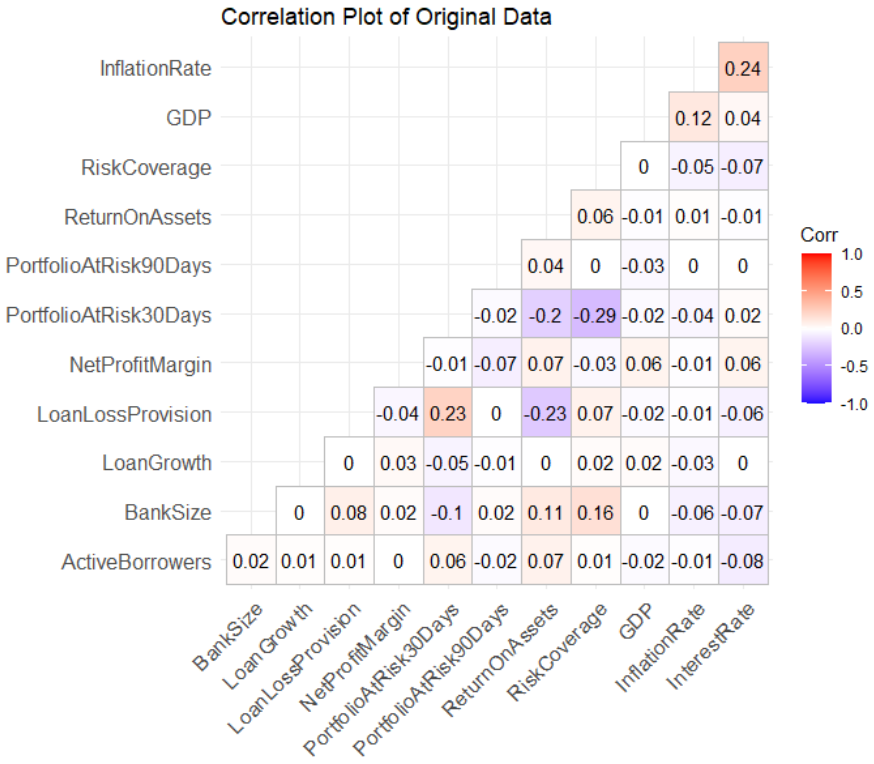

Figure 4: Plot showing the correlation between the variables from the original dataset (with missing data present).

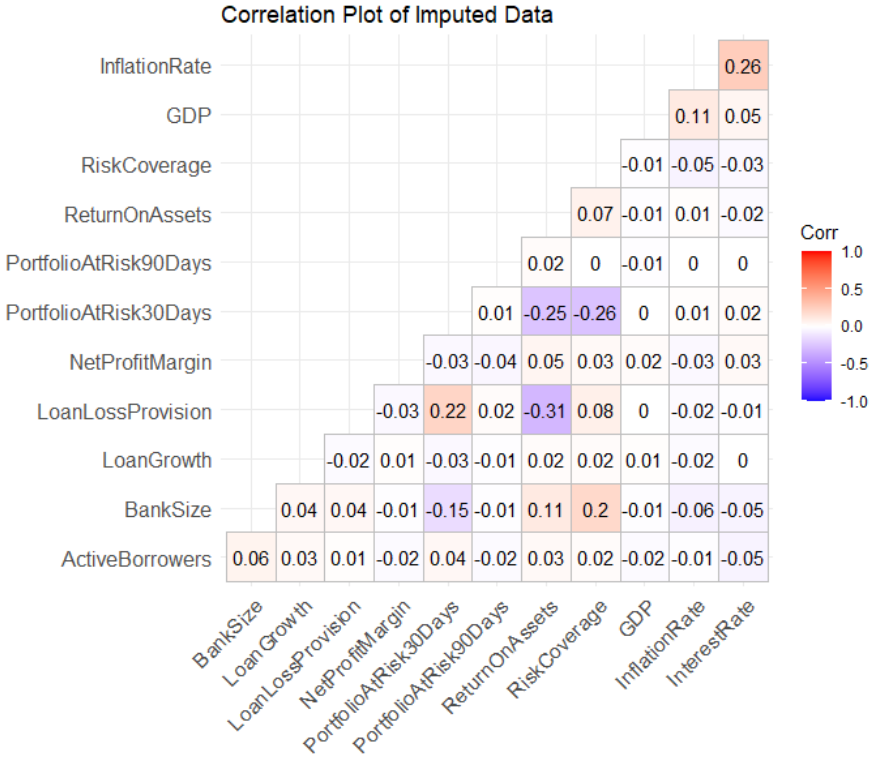

Figure 5: Plot showing the correlation between the variables from the imputed dataset, where missing values were replaced with imputed values.

4.3 Descriptive Statistics

Let us explore our final dataset by considering some summary statistics, given in Table 5, and the correlations between the variables, as given in Figure 6. Given that the final dataset contains observations of 1,312 MFIs over a time period $2005 - 2015$, we decide to include *Year* as a predictor in both linear regression and random forest models to account for time.

Table 5 shows a summary of the variables, with the following statistics: minimum, median, mean, and maximum values, and standard deviation. Considering the response variables, we see that they vary in their statistics, though their minimum and maximum values are close. In terms of the predictors, we observe various ranges, with some variables having negative minimum values.

Variables	**Min.**	**Median**	**Mean**	**Max.**	**Std. Dev.**
Year	2005	2010	2010	2015	3.162
Active Borrowers	0.693	9.154	9.107	15.787	2.038
Bank Size	4.779	15.117	15.205	24.468	2.179
Loan Growth (%)	-62.760	18.049	20.792	113.798	30.253
Loan Loss Provision (%)	-29.910	1.170	2.001	54.350	3.592
Net Profit Margin (%)	-34.700	13.090	13.958	56.100	14.883
Portfolio at Risk 30 Days (%)	0.000	5.625	9.584	373.180	14.911
Portfolio at Risk 90 Days (%)	-1.660	2.410	5.525	512.580	14.998
Return on Assets (%)	-246.680	1.710	0.136	97.310	13.449
Risk Coverage (%)	-69.790	66.550	72.459	271.500	50.768
GDP Growth (%)	-46.082	6.014	5.713	53.382	3.629
Inflation Rate (%)	-10.067	5.924	6.734	59.220	4.989
Interest Rate (%)	3.552	13.021	14.342	65.418	7.882

Table 5: Summary of statistics for the final dataset over the period $2005 - 2015$, showing the minimum, median, mean, and maximum values, and standard deviation.

Let us consider the correlation plot. Figure 6 shows that Return on Assets and Risk Coverage are negatively correlated with PAR 30 days, hence this may suggest a negative association between these variables. Looking at the proxies for overconfidence, we see that Loan Loss Provision is strongly correlated to PAR 30 days compared to Net Profit Margin and Loan Growth. For PAR 90 Days, there seems to be no strong correlations with any of the proxies.

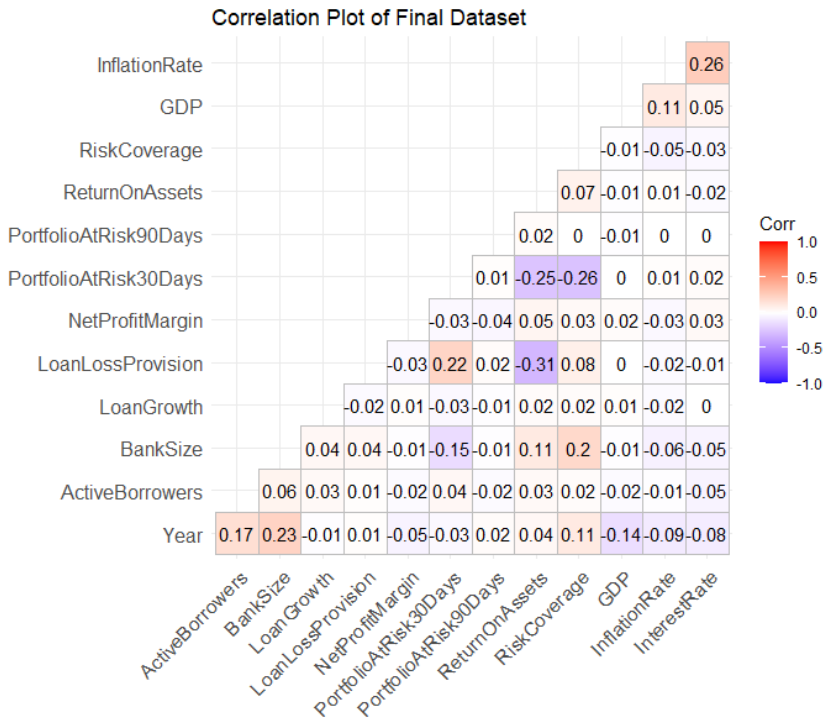

Figure 6: Plot showing the correlation between the variables from the final dataset.

Lastly, Figure 7 shows all the countries are included in our final dataset. There are in total 100 countries. More details, such as the names of the countries and the number of MFIs in each country, can be found in the Appendix.

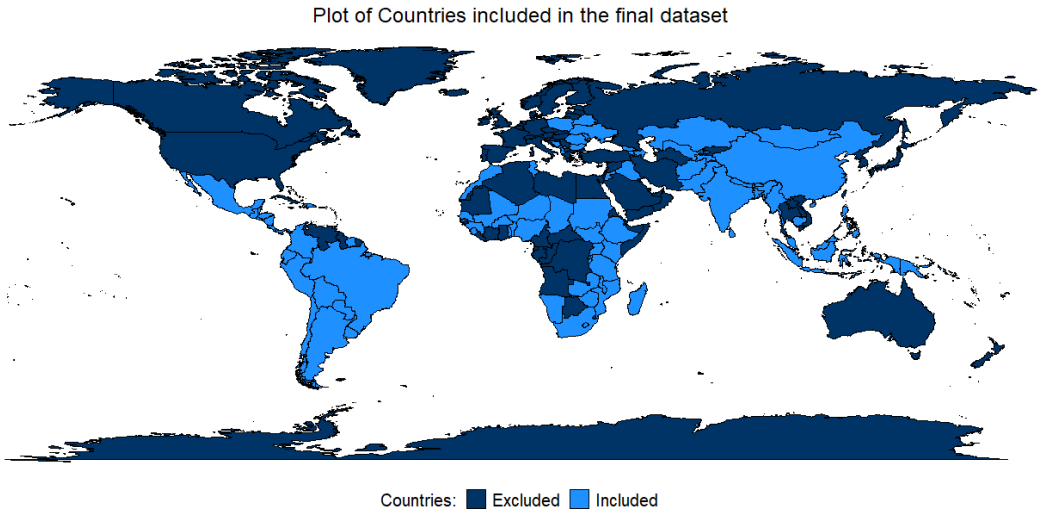

Figure 7: This is a world map, highlighting the countries included in our final dataset.

5 Results

In this section, we present the results of applying linear regression and random forest to our dataset for both response variables (outcomes). For linear regression, we show the estimated least square coefficients and also the standardised coefficients. For random forest, we show the permutation importance values for each feature along with the approximated p-values. We also plot partial dependence plots for some features. In the last part, we compare both models in terms of the predictors (features) they found to have a relationship with one (or both) of the response variables.

5.1 Linear Regression

We present the results of linear regression for both the full model with unstandardised (that is, data that has not been standardised) and standardised data when the response variable is Portfolio at Risk 30 days and Portfolio at Risk 90 days. We show the coefficient estimates, their standard errors and the p-values for each predictor. We set the confidence level at 0.05.

5.1.1 Portfolio at Risk 30 days

Table 6 shows the linear regression result for the full model with unstandardised and standardised data when the response variable is Portfolio at Risk 30 days. we see that for the full model all predictors have estimated coefficients that are in the range $[-1, 1]$. Loan Loss Provision has the largest coefficient (ignoring the sign, as this just shows the direction of the relationship), followed by Active Borrowers and Return on Assets. Large coefficients imply that the marginal effect of these predictors on the response variable is greater than the rest of the predictors.

Predictors	Full Untand. Model		Full Stand. Model	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Year	0.075	0.038*	0.016	0.038*
Active Borrowers	0.332	0.057***	0.045	0.057***
Bank Size	−0.617	0.055***	− 0.09	0.055***
Loan Growth (%)	−0.009	0.004*	−0.018	0.004*
Loan Loss Provision (%)	0.770	0.034***	0.186	0.034***
Net Profit Margin (%)	−0.005	0.008	−0.005	0.008
Return on Assets (%)	−0.188	0.009***	−0.17	0.009***
Risk Coverage (%)	−0.071	0.002***	−0.243	0.002***
GDP Growth (%)	0.008	0.032	0.002	0.032
Inflation Rate (%)	−0.008	0.024	−0.003	0.024
Interest Rate (%)	0.013	0.015	0.007	0.015
Note:	*p<0.05; **p<0.01; ***p<0.001			

Table 6: Table containing the linear regression results for our dataset with response variable Portfolio at Risk 30 days. For the full model with unstandardised and standardised data, the table shows the estimated coefficients, standard errors and p-values (using asterisks *).

Out of the 11 predictors, 7 of them have statistically significant coefficients, that is, the p-value is below the confidence level, which is 0.05 (as indicated by the asterisk(s)). Hence, we reject the null hypothesis which claims that those coefficients are equal to zero. This suggests that the coefficients are different from zero. More specifically, this suggests that there is an association between these predictors and the response variable, Portfolio at Risk 30 days.

Out of the three proxies for overconfidence, only Loan Loss Provision has a large coefficient and also a statistically significant coefficient. Hence, there is a strong association between Loan Loss Provision and Portfolio at Risk 30 days. Looking at the full model, this predictor has a coefficient equal to 0.770. This positive coefficient implies that increasing Loan Loss Provision increases Portfolio at Risk 30 days. More precisely, a 1% increase in Loan Loss Provision leads to an average expected increase of 0.770% in Portfolio at Risk 30 days.

The other two proxies have negative coefficients, meaning that a decrease in those predictors leads to an increase in Portfolio at Risk 30 days. Though Loan Growth has a statistically significant coefficient in the full model with unstandardised data, its coefficient is quite small (-0.009), as in it is close to zero, so varying this predictor has a small impact on Portfolio at Risk 30 days. This suggests that there is weak association between Loan Growth and Portfolio at Risk 30 days.

Like Loan Growth, Net Profit Margin has a small coefficient but unlike Loan Growth, its coefficient is not statistically significant. This implies there is no association between Net Profit Margin relationship and Portfolio at Risk 30 days.

Let us consider the full standardised linear regression model in Table 6. We observe that the standard errors are the same as in the full unstandardised model and the statistical significance is the same. Thus, we can arrive at the same conclusions as in the full unstandardised model case, in terms of statistical significance.

Let us also consider the economic significance, by looking at the standardised coefficients (in the full standardised linear regression model). From Table 6, we see that the estimated coefficients differ from the full linear model before standardising. For instance, now the predictor with the largest coefficient is Risk Coverage, with Loan Loss Provision coming in second and Return on Assets in third. The rest of the predictors have coefficient values less than 0.1 (ignoring the sign). Risk Coverage has a positive coefficient, which means that increasing this predictor decreases Portfolio at Risk 30 days. More precisely, one standard deviation increase in Risk Coverage is associated with a 0.243 standard deviation decrease in Portfolio at Risk 30 days, assuming all other predictors are held constant (that is, unchanged). Compared to the other predictors, those three predictors have a much larger effect on Portfolio at Risk 30 days. Hence, the most influential predictors on response variable Portfolio at Risk 30 days are Risk Coverage, Loan Loss Provision and Return on Assets.

Given the standardised coefficients, we can directly compare the impact of each proxy for overconfidence. Loan Loss Provision has the highest coefficient among the three proxies, and hence has the largest impact on Portfolio at Risk 30 days.

5.1.2 Portfolio at Risk 90 days

Table 7 shows the linear regression results for the full model with both unstandardised and standardised data when the response variable is Portfolio at Risk 90 days. In the full linear model with unstandardised data, all predictors have estimated coefficients in the range $[-1, 1]$. Predictors Year, Active Borrowers, Bank Size and Loan Loss Provision

Predictors	Full Unstand. Model		Full Stand. Model	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Year	0.128	0.042**	0.027	0.042**
Active Borrowers	−0.179	0.062**	−0.024	0.062**
Bank Size	−0.117	0.061	−0.017	0.061
Loan Growth (%)	−0.003	0.004	−0.007	0.004
Loan Loss Provision (%)	0.106	0.037**	0.025	0.037**
Net Profit Margin (%)	−0.045	0.008***	−0.044	0.008***
Portfolio at Risk 30 Days (%)	0.014	0.009	0.014	0.009
Return on Assets (%)	0.039	0.010***	0.035	0.010***
Risk Coverage (%)	0.001	0.003	0.004	0.003
GDP Growth (%)	−0.008	0.035	−0.002	0.035
Inflation Rate (%)	0.014	0.026	0.005	0.026
Interest Rate (%)	0.006	0.016	0.003	0.016
Note:	*p<0.05; **p<0.01; ***p<0.001			

Table 7: Table containing the linear regression results for our dataset with response variable Portfolio at Risk 90 days. For the full model with unstandardised and standardised data, the table shows the estimated coefficients, standard errors and p-values (using asterisks *).

have large estimated coefficients (ignoring the sign). However, not all of them have a significant coefficient. Predictors Year, Active Borrowers, Loan Loss Provision, Net Profit Margin and Return on Assets have significant coefficients. This suggests that there is an association between those predictors and the response variable, Portfolio at Risk 90 days.

Out of the three proxies for overconfidence, Loan Loss Provision and Net Profit Margin have statistically significant coefficients in the full unstandardised model. Hence, there is an association between these two proxies and Portfolio at Risk 90 days. Loan Loss Provision has a coefficient equal to 0.106. This positive coefficient implies that increasing Loan Loss Provision increases Portfolio at Risk 90 days. More precisely, a 1% increase in Loan Loss Provision leads to an average expected increase of 0.106% in Portfolio at Risk 90 days. Net Profit Margin has a negative coefficient, thus a 1% increase in Net Profit Margin results in an average expected decrease of 0.045% in Portfolio at Risk 90 days. Considering both coefficients, Loan Loss Provision has a stronger association with Portfolio at Risk 90 days than Net Profit Margin.

Loan Growth also has a negative coefficient, meaning that a decrease in this predictor leads to an increase in Portfolio at Risk 90 days. However, Loan Growth does not have a statistically significant coefficient.

Looking at the full linear model with standardised data, we observe that the standard errors are consistent with those from the full model with unstandardised data and the statistical significance stayed the same. This means that, in terms of statistical significance, the results are the same.

Let us consider the economic significance. Standardising the data has decreased the estimated coefficients for most predictors such that the overall range of coefficient values is

$[-0.1, 0.1]$. We note that Net Profit Margin has the highest coefficient, followed by Return on Assets, Year, Loan Loss Provision and Active Borrowers. Thus, these predictors have a larger effect on Portfolio at Risk 90 days than the rest of the predictors.

Considering the standardised coefficients of the proxies, Net Profit Margin has the largest absolute coefficient. More precisely, one standard deviation decrease in Net Profit Margin is associated with a 0.045 standard deviation increase in Portfolio at Risk 90 days, assuming all other predictors are held constant. While Loan Loss Provision follows closely behind, with a coefficient value of 0.025, Loan Growth is further behind, with a value of -0.007. This suggests that there is a stronger association between Net Profit Margin and Portfolio at Risk 90 days, than with the other two proxies. In other words, Net Profit Margin has a larger effect on Portfolio at Risk 90 days than the other two proxies.

To check the validity of the linear regression results, we tested the assumptions. The graphs and tests used can be found in the Appendix (8.2). We found that three assumptions are violated for both response variables. Though this reduces the quality and reliability of the linear regression results, we proceed with the analysis and later on, we provide suggestions on how to deal with these violations.

5.2 Random Forest

We show the results of applying random forest to our dataset with both default and tuned hyperparameters, for both outcomes. More specifically, we show the permutation importance values and approximated p-values for each feature. Additionally, we plot the partial dependence for some of the proxies of overconfidence. We show these plots for one proxy per response variable, but the rest of the plots are in the Appendix (8.3). For the approximated p-values, we set the confidence level at 0.05. Thus, a permutation importance for a feature is significant if its p-value is below this level.

5.2.1 Portfolio at Risk 30 days

Let the outcome be Portfolio at Risk 30 days. We set the number of trees to be grown to 1000. For classical random forest, the default hyperparameters are $m = 3$, $n_{min} = 5$ and $s_f = 1$ (that is, use the entire dataset). On the other hand, the tuned hyperparameters are $m = 2$, $n_{min} = 15$ and $s_f = 0.6077014$. Table 8 shows the random forest results.

On the one hand, we observe that for the random forest with default hyperparameters, Risk Coverage and Loan Loss Provision have the highest permutation importance values. This shows that they have a greater influence on Portfolio at Risk 30 days than the rest of the features. Additionally, their importance values have approximated p-values below 0.05. Thus, this suggests there is a significant relationship between these features and Portfolio at Risk 30 days.

On the other hand, in the random forest with tuned hyperparameters, there are four predictors with statistically significant approximated p-values, Bank Size, Loan Loss Provision, Return on Assets and Risk Coverage. We also note that the permutation importance values changed after tuning the hyperparameters.

Since permutation importance is used to indicate the importance of the features, let us rank those values and visualise them in a plot. Figure 8 shows the permutation importance for random forest with both the default and tuned hyperparameters. In the default case, Risk Coverage and Loan Loss Provision have far greater permutation importance values than the rest of the features. Though in the tuned case, the difference between those two features and the other features is lessened. We now have that Return on Assets and Bank

Predictors	Default Hyp. Par.		Tuned Hyp. Par.	
	Permutation Importance	Approximated p-value	Permutation Importance	Approximated p-value
Year	5.981	0.5248	4.897	0.3366
Active Borrowers	0.743	0.6931	1.906	0.1782
Bank Size	8.393	0.1485	8.839	0.0198*
Loan Growth (%)	0.182	0.5644	0.916	0.2178
Loan Loss Provision (%)	28.431	0.0099**	22.837	0.0099**
Net Profit Margin (%)	0.684	0.5644	1.437	0.1584
Return on Assets (%)	9.236	0.0792	13.996	0.0099**
Risk Coverage (%)	42.814	0.0099**	37.436	0.0099**
GDP Growth (%)	0.413	1	1.932	0.8812
Inflation Rate (%)	8.674	0.6139	4.468	0.8218
Interest Rate (%)	6.414	0.8218	4.695	0.7921
Note:	*p<0.05; **p<0.01; ***p<0.001			

Table 8: Table containing the random forest results for our dataset with outcome Portfolio at Risk 30 days. It shows the results of random forest with default and tuned hyperparameters: the permutation importance values and the approximated p-values for each feature (using asterisks *).

Size have the third and fourth largest permutation importance.

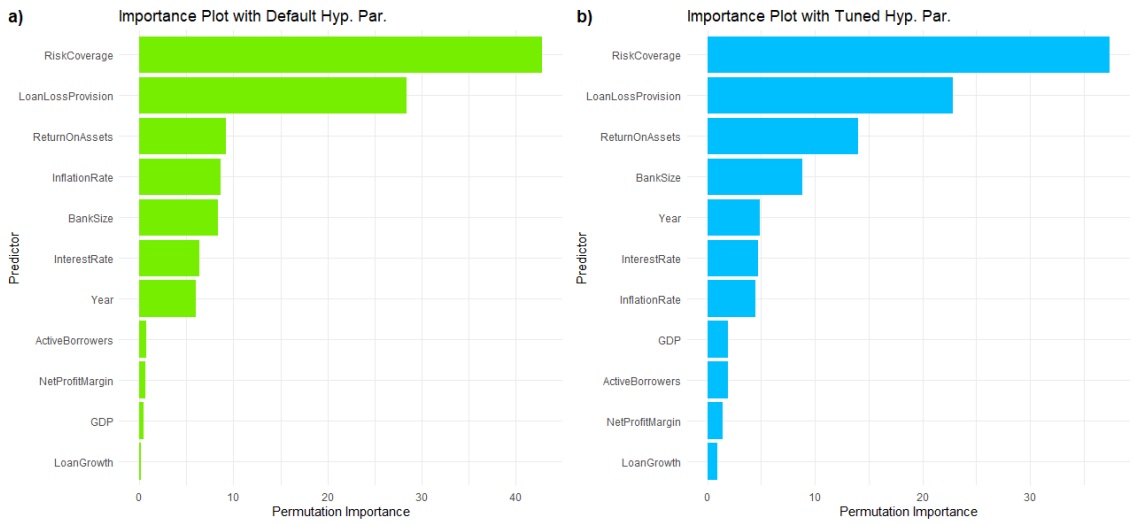

Figure 8: Plot showing the ranked permutation importance values for random forest with outcome Portfolio at Risk 30 days and for both default and tuned hyperparameters.

Let us consider the proxies for overconfidence. As noted earlier, Loan Loss Provision has a high permutation importance value with a statistically significant p-value for random forest with default and tuned hyperparameters. Thus, this implies that Loan Loss Provision has an association with Portfolio at Risk 30 days. The other two proxies have small permutation importance values and their p-values are above 0.05. Hence, this suggests

they do not have an association with Portfolio at Risk 30 days.

Since Loan Loss Provision is the most influential proxy for Portfolio at Risk 30 days, let us observe how various values for this feature affect the outcome. Figure 9 shows the partial dependence plot of Loan Loss Provision with outcome Portfolio at Risk 30 days. We see that as Loan Loss Provision increases, Portfolio at Risk 30 days decreases (when Loan Loss Provision is below 0) and increases (when Loan Loss Provision is above 0). This suggests that this relationship is non-linear. This holds for random forest with default and tuned hyperparameters.

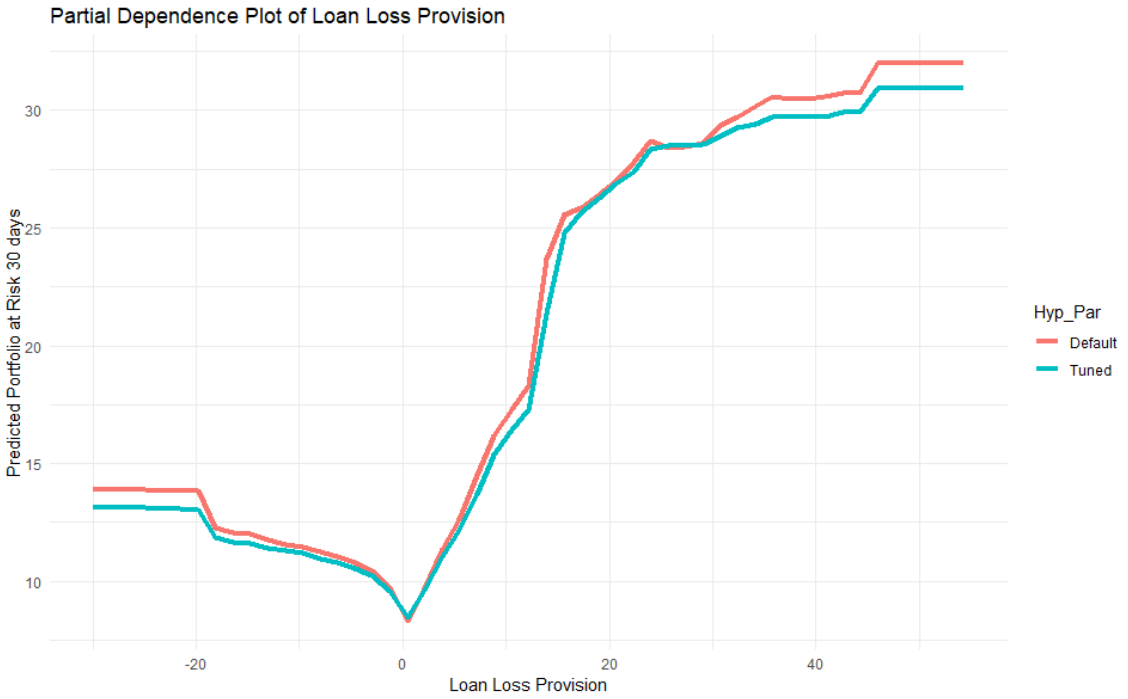

Figure 9: Plot showing the partial dependence plot of Loan Loss Provision with response variable Portfolio at Risk 30 days and for both default and tuned hyperparameters.

Out of all the proxies for overconfidence, only Loan Loss Provision has an importance value with a significant p-value. The partial dependence plot shown in Figure 9 indicates that positive values of Loan Loss Provision increase Portfolio at Risk 30 days while negative values decrease it.

5.2.2 Portfolio at Risk 90 days

Let the outcome be Portfolio at Risk 90 days. We set the number of trees to be grown to 1000. For classical random forest, the default hyperparameters are $m = 3$, $n_{min} = 5$ and $s_f = 1$ (that is, use the entire dataset). On the other hand, the tuned hyperparameters are $m = 1$, $n_{min} = 172$ and $s_f = 0.7169451$. Table 9 shows the random forest results.

For random forest with default hyperparameters, Risk Coverage, Net Profit Margin and Portfolio at Risk 30 days have the highest permutation importance values. However, only Net Profit Margin's importance value has a significant p-value. For the case with tuned hyperparameters, we see that all the permutation importance values decreased and are now within the range $(-1, 2)$. Though Year, Risk Coverage, and Net Profit Margin have

Predictors	Default Hyp. Par.		Tuned Hyp. Par.	
	Permutation Importance	Approximated p-value	Permutation Importance	Approximated p-value
Year	4	0.7823	1.542	0.198
Active Borrowers	6.700	0.1188	-0.504	0.9406
Bank Size	5.441	0.604	1.034	0.2277
Loan Growth (%)	4.927	0.0792	0.222	0.3465
Loan Loss Provision (%)	8.933	0.2772	1.337	0.0594
Net Profit Margin (%)	10.593	0.0099**	1.726	0.0099 **
Portfolio at Risk 30 days (%)	16.206	0.2475	1.024	0.3465
Return on Assets (%)	5.651	0.4158	0.557	0.4653
Risk Coverage (%)	17.896	0.2079	1.5223	0.099
GDP Growth (%)	1.02	0.8614	0.085	0.8416
Inflation Rate (%)	7.366	0.5941	0.646	0.6436
Interest Rate (%)	8.106	0.6337	0.729	0.6931
Note:	*p<0.05; **p<0.01; ***p<0.001			

Table 9: Table containing the random forest results for our dataset with outcome Portfolio at Risk 90 days. It shows the results of random forest with default and tuned hyperparameters: the permutation importance values and the approximated p-values for each feature (using asterisks *).

the highest permutation importance values, only Net Profit Margin's importance value has a significant p-value. Hence, this suggests that there exists a significant relationship between this feature and Portfolio at Risk 90 days.

Let us consider Figure 10, which shows the ranked permutation importance values for random forest with default and tuned hyperparameters. In the default case, Risk Coverage and Portfolio at Risk 30 days are far higher importance values than the rest of the features. However, in the tuned case, Net Profit Margin has the highest importance value, but is followed closely by the Year, Risk Coverage and Loan Loss Provision.

Considering the three proxies for overconfidence, only Net Profit Margin has a high permutation importance value with a significant p-value in random forest with default and tuned hyperparameters. Thus, Net Profit Margin has an association with Portfolio at Risk 90 days. Though Loan Loss Provision is close in importance value to Net Profit Margin, its p-value is not significant, though it is close to 0.05 in the case of tuned hyperparameter. Similarly, Loan Growth has a p-value close to 0.05 in the case of default hyperparameters. However, it has small importance values in both cases. These two proxies do not have an association with Portfolio at Risk 90 days.

Since Net Profit Margin has an importance value with a significant p-value, we plot a partial dependence plot to visualise the relationship between Net Profit Margin and Portfolio at Risk 90 days. Figure 11 shows the partial dependence plot, for random forest with default and tuned hyperparameters. We see that the predicted Portfolio at Risk 90 days is quite volatile as the Net Profit Margin changes in value, when using default hyperparam-

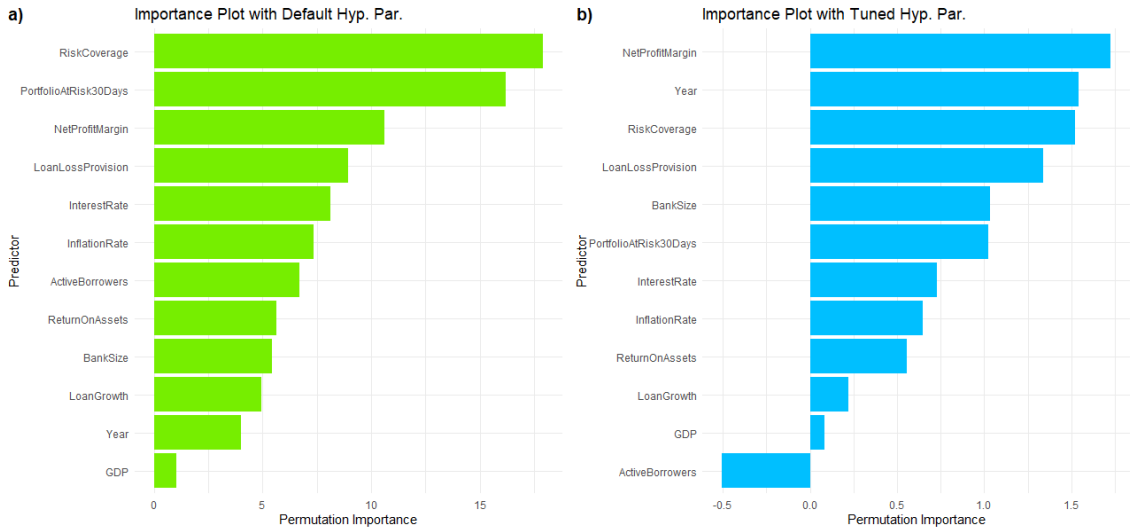

Figure 10: Plot showing the ranked permutation importance values for random forest with response variable Portfolio at Risk 90 days and for both default and tuned hyperparameters.

eters. However, we notice a downward trend as Net Profit Margin decreases, then a sharp increase followed by a sharp decrease. The predicted outcome stabilises for a while, then begins to increase. The case of tuned hyperparameters, we see less volatility in predicted Portfolio at Risk 90 days. More precisely, it decreases as Net Profit Margin increases to zero, stabilisies once Net Profit Margin is positive then slowly increases. Overall, we observe a non-linear relationship between Net Profit Margin and Portfolio at Risk 90 days.

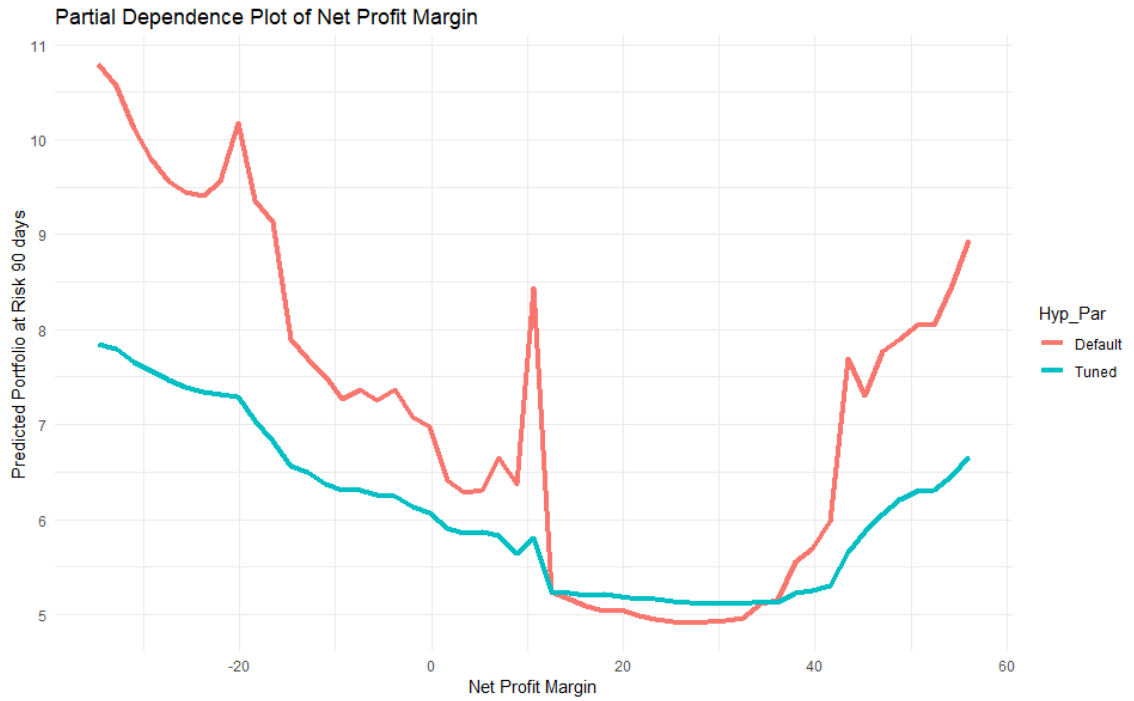

Figure 11: Plot showing the partial dependence plot of Net Profit Margin with response variable Portfolio at Risk 90 days and for both default and tuned hyperparameters.

5.3 Comparison of Linear Regression and Random Forest

Having analysed the results of all models, let us compare their them. Figure 12 shows bubble charts of the linear regression and random forest results when the response variable (or outcome) is Portfolio at Risk 30 days. The charts summarise the estimated coefficients or permutation importance values and their significance (that is, whether p-value is below 0.05).

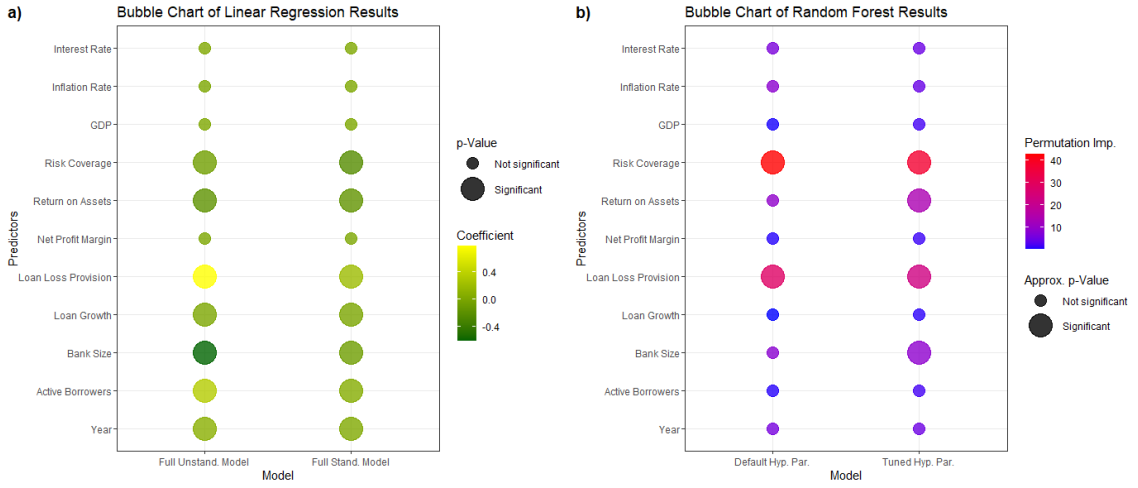

Figure 12: Bubble charts of model results for Portfolio at Risk 30 days, showing the significance of the coefficients or permutation importance (indicated by the size of the bubbles) and the coefficient or permutation importance value (indicated by the colour of the bubble). A coefficient or or permutation importance value has a significant p-value if its p-value is below 0.05.

Figure 12a) shows the full linear regression model results with unstandardised and standardised data. In terms of the estimated coefficient values, both models have similar values for most predictors, except for Bank Size and Loan Loss Provision. We note that standardising the data decreased the coefficient of Loan Loss Provision, but increased the coefficient of Bank Size. Regarding the significance of the estimated coefficients, both models agree on all of the predictors.

Figure 12b) shows the random forest results with default and tuned hyperparameters. When looking at the permutation importance values, they are all quite similar before and after tuning. In regards to the significance of the permutation importance values, we note that after tuning Return on Assets and Bank Size have significant importance values, along with Risk Coverage and Loan Loss Provision which have significant importance values both before and after tuning the hyperparameters.

Comparing linear regression to random forest, we observe that Loan Loss Provision and Risk Coverage have significant coefficient/permutation importance values in all models. In terms of the coefficient/permutation importance values, they are both among the highest values in all models. Thus, all models indicate that these two predictors/features have a strong association with Portfolio at Risk 30 days. Overall, it seems like more predictors/features have significant coefficient/permutation importance values in the linear models than in the random forest models.

In terms of the proxies for overconfidence, all models agree that Loan Loss Provision has a significant coefficient/permutation importance, while Net Profit Margin does not. Only the linear models agree that Loan Growth has a significant coefficient. The random forest models show that Loan Growth does not have a significant permutation importance value.

Let us look at the bubble charts when the response variable (or outcome) is Portfolio at Risk 90 days, shown in Figure 13. In Figure 13a), both models have similar coefficient values, except for three predictors. Both Loan Loss Provision and Year have a lower coefficient value after standardising the data, while Active Borrowers has a higher coefficient value. In addition, standardising did not affect the significance of the coefficients. Looking at the random forest results in Figure 13b), we see that the permutation importance values decreased considerably after tuning the hyperparameters. Nevertheless, Net Profit Margin is the only feature with significant permutation importance value before and after tuning.

Figure 13: Bubble charts of model results for Portfolio at Risk 90 days, showing the significance of the coefficients or permutation importance (indicated by the size of the bubbles) and the coefficient or permutation importance value (indicated by the colour of the bubble). A coefficient or or permutation importance value has a significant p-value if its p-value is below 0.05.

Considering all the models, only Net Profit Margin has significant coefficient/permutation importance values. As previously noted, linear models have more predictors/features with significant coefficient/permutation importance values than the random forest models.

In terms of the proxies, the linear models only agree with the random forest models for Net Profit Margin (significant) and Loan Growth (not significant). The models disagree on Loan Loss Provision, as the linear models show that its estimated coefficient is significant, while the random forest models show that its permutation importance value is not significant.

6 Discussions and Conclusion

The purpose of this thesis was to examine the potential impact of loan officers' overconfidence behaviour on an MFI's credit risk management. In the previous section, we presented the results of applying two quantitative models and in this chapter we discuss these findings and the implications of the results in more depth. Lastly, we conclude this thesis with a brief summary and describe some of the limitations of this research.

6.1 Discussion

Before looking at the main research question, let us consider the sub-questions. The first sub-question,

i) Is there a relationship between overconfidence and an MFI's credit risk ?

addresses the existence of a relationship between loan officers' overconfidence behaviour and an MFI's credit risk. We used three proxies for overconfidence, which acted as indicators of an MFI's loan officers' overconfidence. Based on the results of the two quantitative models, we observed statistical significance of some of the proxies with an MFI's credit risk (that is, coefficient/permutation importance value has a significant p-value). On the one hand, proxy Loan Loss Provision has a strong association (statistically significant) with response variable Portfolio at Risk 30 days, while Loan Growth has a weak association with the same response variable. On the other hand, proxies Net Profit Margin and Loan Loss Provision have a strong association with response variable Portfolio at Risk 90 days.

However, the economic significance is not as straightforward. For instance, Loan Loss Provision has a positive coefficient, which implies that increasing this proxy increases Portfolio at Risk 30 days. According to Fersi and Boujelbène (2022), overconfidence behaviour leads to loan officers' misestimating future risk by underestimating the likelihood of defaults. This is not in line with our results (positive coefficient and partial dependence plot Figure 9), which show that holding too much capital leads to higher credit risk. Instead of overconfidence bias, this may be underconfidence bias as future risk of default is overestimated, which results in holding too much capital for provisions for loss, leading to lower quality in loan portfolio and higher credit risk.

Moreover, Loan Growth has a negative coefficient. Thus, increasing Loan Growth decreases Portfolio at Risk 30 days. This means that aggressively increasing the number of approved loans will considerably improve the loan portfolio quality and hence reduce the credit risk of an MFI. This too is not in line with expectations of overconfidence as past studies show that aggressive loan growth leads to greater risk-taking, which harms an MFI's credit risk (Ben Salah Mahdi & Boujelbène Abbes, 2018; Fersi & Boujelbène, 2022).

When the response variable is Portfolio at Risk 90 days, the economic significance is still complicated. Loan Loss Provision has a positive coefficient, that is also significant. As previously explained, this is not consistent with the expectations of the impact of overconfidence. That is, overconfidence would lead loan officers to underestimate future risks by reducing provisions of loss and this hence would impact credit risk. Moreover, proxy Net Profit Margin has a significant coefficient which is negative. This implies that as this proxy decreases, the loan portfolio at risk increases (that is, loans are more likely to default). However, overconfidence behaviour should imply that high profit margins lead to loan officers overestimating their abilities to manage credit risk and hence approve of more loans, even if clients have a high risk profile. This, in turn, would increase an MFI's credit risk. This result is shown in the studies by Ben Salah Mahdi and Boujelbène Abbes (2018)

and Fersi and Boujelbène (2022). Nevertheless, if we consider the partial dependence plot in Figure 11 when the response variable is Portfolio at Risk 90 days, we do observe this result. As Net Profit Margin increases above 40%, the response variable begins to increase (Figure 11).

Therefore, statistically, we have shown that there is an association between at least one proxy of overconfidence and an MFI's credit risk (as represented by Portfolio at Risk 30 days and Portfolio at Risk 90 days). However, in terms of economic significance, we only showed that overconfidence impacts credit risk, with proxy Net Profit Margin for Portfolio at Risk 90 days. More precisely, when the model is random forest, higher profit margins lead to greater risk-taking which lowers loan portfolio quality (indicated by high credit risk measure, Portfolio at Risk 90 days).

The second sub-question,

ii) Will the two quantitative methods, linear regression and random forest, agree on the effect of overconfidence on an MFI's credit risk?

addresses whether the two methods show the same results for the effect of overconfidence on an MFI's credit risk. When the response variable is Portfolio at Risk 30 days, all models show that proxy Loan Loss Provision has a significant coefficient/permutation importance, while Net Profit Margin does not a have significant coefficient/permutation importance. However, they do not agree on Loan Growth. More specifically, the linear models show that this proxy has a significant coefficient, but the random forest models do not.

We observe the same issue when the response variable is Portfolio at Risk 90 days. Net Profit Margin has a significant coefficient/permutation importance, while Loan Growth does not. However, linear models show that Loan Loss Provision has a significant coefficient, but the random forest models show that this proxy does not have a significant permutation importance.

Although two out of the three proxies the linear and random forest models agree, they still disagree for at least one proxy for either response variable. Hence, the two quantitative methods do not agree, on all proxies, on the effect of overconfidence on an MFI's credit risk.

Recall that the main research question is

Does loan officers' overconfidence influence the credit risk decision-making of an MFI?

The answer to the first sub-question confirms that an MFI's loan officers' overconfidence behaviour has an impact on the credit risk decisions made, at least when the measure of loan portfolio quality is Portfolio at Risk 90 days. With proxy Net Profit Margin, we showed that increasing this proxy deteriorates loan portfolio quality, which increases credit risk. The answer to the second sub-question indicates that the models do not all agree on the effect of overconfidence on credit risk for all proxies, but only agree for two of them. In particular, all the models agree that there is a strong association between Net Profit Margin and Portfolio at Risk 90 days. This answers the main research question as Net Profit Margin represents loan officers' overconfidence and Portfolio at Risk 90 days is a measure of an MFI's credit risk.

However, the use of an imputation method limits the accuracy of this result. Thus, we should be careful in the interpretation of the results and further research with different datasets or a different imputation method used should be considered to further analyse the impact of overconfidence on an MFI's credit risk.

6.2 Conclusion

Credit risk is one of the major risks faced by financial institutions because it is their main source of revenue. Effective management of this risk is important to minimise (or even eliminate) the risk of loans defaulting. This is because large unexpected losses caused by defaults can lead to financial institutions becoming insolvent, if insufficient provisions for loss are held. For microfinance institutions, managing credit risk is crucial for their survival. MFIs tend to offer small loans with no collateral, hence there is no security if a loan defaults. This increases the risk of defaults for MFIs, and thus affects their financial performance.

Behavioural finance brings a new aspect on how decisions are made, specifically with its assumption of people being irrational (or at least not always rational). This irrationality is caused by behavioural biases, as research has shown that people's judgments and decisions are often far from rational. Since biases can influence financial decision-making, this means that it can lead to incorrect decisions made in regard to an MFI's credit risk. We looked at behavioural bias overconfidence, which occurs when people overestimate their knowledge and abilities. In the context of MFIs, an overconfident loan officer may understimate future risk of defaults by, for example, holding lower provisions for loss. This could be detrimental to an MFI's survival if heavy, unexpected losses occur.

This thesis aimed to examine the potential impact of overconfidence on an MFI's credit risk. Due to limited resources, three proxies were used to indicate loan officers overconfidence. With a focus on MFIs, the dataset used in the analysis contained financial and country specific data on MFIs in many countries across the globe over a period of 10 years. We used two quantitative models, linear regression and random forest, to measure the potential impact of overconfidence on an MFI's credit risk. In terms of statistical significance, all models agreed on the significance of two out of the three proxies. More specifically, they showed that there is an association between loan officers' overconfidence (proxy Net Profit Margin) and an MFI's credit risk (indicator Portfolio at Risk 90 days). Using linear regression, the economic significance shows that lower profit margins lead to higher credit risk (as loan portfolio quality is worsened). However, this does not show overconfidence behaviour in loan officers. Using random forest, we observed that higher profit margins lead to increased risk of defaults. A possible explanation is that loan officers' overestimate their ability to manage credit risk as current high profit margins indicate their success and hence they approve more loans, even if clients have a high risk profile (that they would in normal circumstances not approve).

However, having employed an imputation method to deal with the large amount of missing data limits the reliability of the results and hence our interpretations. Further research is recommended to have a better analysis of this potential relationship.

6.3 Limitations and Suggestions for Further Research

In this section, we discuss the limitations of this research and recommend suggestions for further research for this topic.

There are a few limitations. For instance, the final data sample used for the application of the quantitative models. The data obtained by the Mix Market database contained a substantial amount of missing data, and deleting observations with missing values would have greatly reduced the sample size. Thus, we employed an imputation method to replace the missing data. Though this method preserved the statistical measures, such as median, mean and correlation, we were unable to analyse the full impact of the imputation

method, for example, by performing a sensitivity analysis. This constraint was caused by the substantial amount of missing data, which severely limited the amount of complete observations. This meant that comparing the two datasets after applying a quantitative model was not feasible. This limited the accuracy of the results obtained in this thesis and hence the conclusions derived.

In addition, there are several other imputation methods (such as K-Nearest Neighbor, Multivariate Imputation by Chained Equations, and Regression Imputation) which could have been used instead of the random forest (missForest) algorithm. If different methods were used, then obtained results might have differed and so would the interpretations. Thus, further research could include applying different imputation methods and comparing the results obtained from applying one or more quantitative methods.

When applying linear regression, the results were less accurate as some of the assumptions of least squares were not met. For example, the variance of the residuals was not constant. Hence, the standard errors could be biased, which also affects the p-values obtained. Further research could include dealing with the assumption violations and comparing the results, that is, to check whether the assumption violations considerably distorted the results.

In addition, the dataset has many observations of many different MFIs over a period of time, so a more suitable regression model than standard linear regression is panel regression as such data is considered panel data. Panel regression allows to control dependencies of unobserved predictors on the response variable, which can lead to biased estimators in traditional regression models (Nguyen, 2020). Moreover, with this type of regression model, we can use lagged variables, that is, the value of a variable at a time point before the current time point. This can be useful in examining the potential impact of the current credit risk caused by past values of the proxies for overconfidence. Although there is little literature on random forest applied to panel data, this could be an interesting subject for further research.

This research topic can be extended to include other biases, such as loss aversion, risk aversion, optimism and pessimism. This can examine how other biases influence an MFI's credit risk. In addition, this can also investigate how several biases interact with one another. This may require more proxies or, if the necessary resources are available, conducting a behavioural study on loan officers' from many MFIS and analyse their credit risk decision-making. A possible proxy for optimism and pessimism is an MFI's loan loss provisions. This will require comparing actual provisions for loss with an estimate of provisions for loss based on current bank characteristics and the financial environment. More specifically, if an MFI underestimates future risk (that is, optimistic about future risk so hold less capital), the provisions for loss are lower than this estimated provisions for loss and the reverse for the case when an MFI overestimates future risk (that is, pessimistic about future risk so hold more capital).

7 References

Abbes, M. B. (2012). Does overconfidence bias explain volatility during the global financial crisis? *Transition Studies Review*, *19*(3), 291–312. https://doi.org/10.1007/s11300-012-0234-6

Abusharbeh, M. T. (2023). Modeling the factors of portfolio at risk for microfinance institutions in palestine. *Cogent Economics amp; Finance*, *11*(1). https://doi.org/10.1080/23322039.2023.2186042

Addae-Korankye, A. (2014). Causes and control of loan Default/Delinquency in microfinance institutions in ghana. *American International Journal of Contemporary Research*, *4*, 36–45.

Agier, I., & Szafarz, A. (2013). Microfinance and gender: Is there a glass ceiling on loan size? *World Development*, *42*, 165–181. https://doi.org/https://doi.org/10.1016/j.worlddev.2012.06.016

Akash, M. (2022, January). *Determinants of capital adequacy of banks: A comparative study on bank asia ltd. and ab bank ltd.* [Doctoral dissertation, University of Dhaka]. https://doi.org/10.13140/RG.2.2.20728.72963/2

Almansour, B. Y., Elkrghli, S., & Almansour, A. Y. (2023). Behavioral finance factors and investment decisions: A mediating role of risk perception. *Cogent Economics amp; Finance*, *11*(2). https://doi.org/10.1080/23322039.2023.2239032

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Alton, R. G., & Hazen, J. H. (2001). As economy flounders, do we see a rise in problem loans? *Federal Reserve Bank of St. Louis*, *11*(4), 45–65.

Aren, S., & Nayman Hamamci, H. (2021). Biases in managerial decision making: Overconfidence, status quo, anchoring, hindsight, availability. *Journal of Business Strategy, Finance and Management*, *3*(1-2), 08–23.

Ariely, D. (2008). *Predictably irrational: The hidden forces that shape our decisions.* HarperCollins Publishers.

Ariely, D. (2009). The end of rational economics. *Hardvard Business Review*, *87*(7), 78–84.

Arko, S. K. (2012). *Determining the causes and impact of Non-Performing loans on the operations of microfinance institutions: A case of sinapi aba trust.* [Doctoral dissertation, Kwame Nkrumah University of Science and Technology].

Assefa, E., Hermes, N., & Meesters, A. (2013). Competition and the performance of microfinance institutions. *Appl. Fin. Econ.*, *23*(9), 767–782.

Bagdonavicius, V., Kruopis, J., & Nikulin, M. S. (2010). *Nonparametric tests for complete data.* ISTE Ltd; John Wiley & Sons.

Baklouti, I. (2015). On the role of loan officers' psychological traits in predicting microcredit default accuracy. *Qualitative Research in Financial Markets*, *7*(3), 264–289. https://doi.org/10.1108/qrfm-04-2014-0011

Baklouti, I., & Baccar, A. (2013). Evaluating the predictive accuracy of microloan officers' subjective judgment. *International Journal of Research Studies in Management*, *2*(2). https://doi.org/10.5861/ijrsm.2013.343

Bandyopadhyay, A., & Saxena, M. (2023). Interaction between credit risk, liquidity risk, and bank solvency performance: A panel study of indian banks. *Indian Econ. Rev.*, *58*(2), 311–328.

Banto, J. M., & Monsia, A. F. (2021). Microfinance institutions, banking, growth and transmission channel: A GMM panel data analysis from developing countries. *Q. Rev. Econ. Finance*, *79*, 126–150.

Barberis, N., & Thaler, R. (2002). A survey of behavioral finance. *National Bureau of Economic Research Working Paper Series*, *No. 9222*.

Baron, J. (2023). *Thinking and deciding* (5th ed.). Cambridge University Press.

Ben Salah Mahdi, I., & Boujelbène Abbes, M. (2018). Risk and inefficiency: Behavioral explanation through overconfidence in islamic and conventional banks. *Managerial Finance*, *44*(6), 688–703. https://doi.org/10.1108/mf-04-2017-0130

Berger, A. N., & DeYoung, R. (1997). Problem loans and cost efficiency in commercial banks. *J. Bank. Financ.*, *21*(6), 849–870.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, *22*(1), 23.

Bihari, A., Dash, M., Muduli, K., Kumar, A., Mulat-Weldemeskel, E., & Luthra, S. (2023). Does cognitive biased knowledge influence investor decisions? an empirical investigation using machine learning and artificial neural network. *VINE Journal of Information and Knowledge Management Systems*, *55*(2), 445–469. https://doi.org/10.1108/vjikms-08-2022-0253

Bishnoi, S., & Hooda, B. K. (2022). Decision tree algorithms and their applicability in agriculture for classification. *Journal of Experimental Agriculture International*, 20–27. https://doi.org/10.9734/jeai/2022/v44i730833

Blanco, F. (2017). Cognitive bias. In J. Vonk & T. Shackelford (Eds.), *Encyclopedia of animal cognition and behavior* (pp. 1–7). Springer International Publishing. https://doi.org/10.1007/978-3-319-47829-6_1244-1

Bouteillé, S., & Coogan-Pushner, D. (2013). *The handbook of credit risk management*. John Wiley & Sons.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Canikli, S., & Aren, S. (2019). Typology of behavioral biases and heuristics. *The European Proceedings of Social and Behavioural Sciences*.

Chaibi, H., & Ftiti, Z. (2015). Credit risk determinants: Evidence from a cross-country study. *Research in International Business and Finance*, *33*, 1–16. https://doi.org/https://doi.org/10.1016/j.ribaf.2014.06.001

Chen, H. J., & Chen, C. H. (2015). Managerial overconfidence and bank risk taking: A cross-country analysis. *European Financial Management Association 2015 Annual Meetings*, *1*, 1–23.

Christen, R. P., & Lyman, R., Timothy R.and Rosenberg. (2003). Microfinance consensus guidelines : Guiding principles on regulation and supervision of microfinance. *CGAP and The World Bank Group*.

Churchill, C., & Coster, D. (2001). *Microfinance risk management handbook*. Pact Publications.

Daniel, K. D., Hirshleifer, D., & Subrahmanyam, A. (2001). Overconfidence, arbitrage, and equilibrium asset pricing. *The Journal of Finance*, *56*(3), 921–965. https://doi.org/10.1111/0022-1082.00350

de Oliveira Leite, R., dos Santos Mendes, L., & de Lacerda Moreira, R. (2020). Profit status of microfinance institutions and incentives for earnings management. *Research in International Business and Finance*, *54*, 101255. https://doi.org/https://doi.org/10.1016/j.ribaf.2020.101255

D'Espallier, B., Guérin, I., & Mersland, R. (2011). Women and repayment in microfinance: A global analysis. *World Development*, *39*(5), 758–772. https://doi.org/https://doi.org/10.1016/j.worlddev.2010.10.008

Doumpos, M., Lemonakis, C., Niklis, D., & Zopounidis, C. (2019). *Analytical techniques in the assessment of credit risk: An overview of methodologies and applications*. Springer International Publishing. https://doi.org/10.1007/978-3-319-99411-6

Eurosystem. (n.d.). *What is inflation?* European Cental Bank. https://www.ecb.europa.eu/ecb-and-you/explainers/tell-me-more/html/what_is_inflation.en.html

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417.

Fersi, M., & Boujelbène, M. (2022). Overconfidence and credit risk-taking in microfinance institutions: A cross-regional analysis. *Int. J. Organ. Anal.*, *30*(6), 1672–1693.

Fofack, H. (2005). Non performing loans in sub-saharan africa : Causal analysis and macroeconomic implications. *World Bank Policy Research Working Paper no. WP3769*.

Foos, D., Norden, L., & Weber, M. (2010). Loan growth and riskiness of banks [International Financial Integration]. *Journal of Banking Finance*, *34*(12), 2929–2940. https://doi.org/https://doi.org/10.1016/j.jbankfin.2010.06.007

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, (5), 1189–1232.

G20. (2023). The group of 20 members. https://g20.org/about-g20/g20-members/

GeeksforGeeks. (2022). What are the different types of nodes in a tree? https://www.geeksforgeeks.org/what-are-the-different-types-of-nodes-in-a-tree/

Gilovich, T., Griffin, D., & Kahneman, D. (2002). Heuristics and biases: The psychology of intuitive judgment. https://doi.org/10.1017/CBO9780511808098

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Hayati, N., & Ariff, M. (2007). Multi-country study of bank credit risk determinants. *The International Journal of Banking and Finance*, *5*(1), 135–152. https://doi.org/10.32890/ijbf2008.5.1.8362

Hendry, D. F. (1995). 156exogeneity and causality. In *Dynamic econometrics*. Oxford University Press. https://doi.org/10.1093/0198283164.003.0005

Hens, T., & Rieger, M. O. (2010). *Financial economics* (2010th ed.). Springer.

Howe, C., & Purves, D. (2005). The müller-lyer illusion explained by the statistics of image-source relationships. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 1234–9. https://doi.org/10.1073/pnas.0409314102

Ibtissem, B., & Bouri, A. (2013). Credit risk management in microfinance: The conceptual framework. *ACRN Journal of Finance and Risk Perspectives*, *2*, 9–24.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3). https://doi.org/10.1214/08-aoas169

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.

Joyce, J. (2021). Bayes' theorem. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University.

Kapoor, S., & Prosad, J. M. (2017). Behavioural finance: A review [5th International Conference on Information Technology and Quantitative Management, ITQM 2017]. *Procedia Computer Science*, *122*, 50–54. https://doi.org/https://doi.org/10.1016/j.procs.2017.11.340

Leite, R., Mendes, L., & Moreira, R. (2018). Profit status of microfinance institutions and incentives for earnings management. *SSRN Electron. J.*

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*, 18–22. https://journal.r-project.org/articles/RN-2002-022/

Loh, W.-Y. (2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/10.1002/widm.8

Mackay, C. (1852). *Memoirs of extraordinary popular delusions and the madness of crowds* (2nd ed.). London : Office of the National Illustrated Library.

Madaan, G., & Singh, S. (2019). An analysis of behavioral biases in investment decision-making. *International Journal of Financial Research*, *10*(4), 55. https://doi.org/10.5430/ijfr.v10n4p55

Majaski, C. (2015, June). Delinquency vs. default: What's the difference? https://www.investopedia.com/ask/answers/062315/what-are-differences-between-delinquency-and-default.asp

Malmendier, U., & Tate, G. (2008). Who makes acquisitions? ceo overconfidence and the market's reaction. *Journal of Financial Economics*, *89*(1), 20–43. https://doi.org/https://doi.org/10.1016/j.jfineco.2007.07.002

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91. http://www.jstor.org/stable/2975974

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*(35), 983–999. http://jmlr.org/papers/v7/meinshausen06a.html

Mighri, Z. (2014). The behavioral finance and the lack of reimbursement: Empirical application in the case of tunisian institutions ofmicrofinance. *Journal of Business and Finance*, (2), 73–83.

Mighri, Z., Bel Hadj, T., Kheireddine, H., & Jarboui, A. (2018). The contribution of behavioral finance in the decision of the microcredit granting: Empirical application to the tunisian AMC case. *International Journal of Information, Business and Management*, *10*(3), 197–226.

Mill, J. S. (2011). *Essays on some unsettled questions of political economy*. Andrews UK.

Molnar, C. (2025). *Interpretable machine learning: A guide for making black box models explainable* (3rd ed.). https://christophm.github.io/interpretable-ml-book

Naili, M., & Lahrichi, Y. (2022a). Banks' credit risk, systematic determinants and specific factors: Recent evidence from emerging markets. *Heliyon*, *8*(2), e08960. https://doi.org/https://doi.org/10.1016/j.heliyon.2022.e08960

Naili, M., & Lahrichi, Y. (2022b). The determinants of banks' credit risk: Review of the literature and future research agenda. *Int. J. Finance Econ.*, *27*(1), 334–360.

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the gini importance? (A. Valencia, Ed.). *Bioinformatics*, *34*(21), 3711–3718. https://doi.org/10.1093/bioinformatics/bty373

Neumann, J. v., & Morgenstern, O. (2004). *Theory of games and economic behavior* (60th anniversary ed.). Princeton University Press.

Ngonyani, D. B., & Mapesa, H. J. (2019). The effects of credit collection policy on portfolio at risk of microfinance institutions in tanzania. *Journal of Accounting, Finance and Auditing Studies*, *5*(1), 241–253. https://doi.org/10.32602/jafas.2019.12

Nguyen, M. (2020). *A guide on data analysis*. Bookdown. https://bookdown.org/mike/data_analysis/

Nimon, K. F., & Oswald, F. L. (2013). Understanding the results of multiple linear regression: Beyond standardized regression coefficients. *Organizational Research Methods*, *16*(4), 650–674. https://doi.org/10.1177/1094428113493929

Nkamnebe, A. D., & Idemobi, E. I. (2011). Recovering of micro credit in Nigeria: Implications for enterprise development and poverty alleviation. *Management Research Review*, *34*(2), 236–247.

Nkusu, M. (2011). Non-performing loans and macrofinancial vulnerabilities in advanced economies. *IMF Working Paper*, *11*(161). https://doi.org/10.2139/ssrn.1888904

Olomola, A. (2002). Social capital, microfinance group performance and poverty implications in nigeria. *Nigerian Institute of Social and Economic Research, Centre for the Study of African Economies, Oxford University*.

Olorunsola, G. I., Olalekan, T. G., & Dele-Oladejo, O. O. (2023). Assessment of credit risk management and financial performance of microfinance banks. *International Journal of Research and Innovation in Social Science*, *7*(3).

Open Risk Manual. (2019). Credit Portfolio. https://www.openriskmanual.org/wiki/Credit_Portfolio

Ozili, P. K. (2020). Theories of financial inclusion. *SSRN Electron. J.*

Parihar, A., Parihar, B., & Parihar, V. (2024). Examine the role of microfinance institutions in reducing poverty and promoting economic development. *Journal of Xidian University*, *10*, 236–246. https://doi.org/10.5281/Zenodo.11488827

Pinz, A., & Helmig, B. (2014). Success factors of microfinance institutions: State of the art and research agenda. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, *26*(2), 488–509. https://doi.org/10.1007/s11266-014-9445-2

Pompian, M. M. (2006). *Behavioral finance and wealth management: How to build optimal portfolios that account for investor biases*. John Wiley & Sons, Ltd.

Probst, P., & Klau, S. (2024). *tuneRanger: tuneRanger*. Rdocumentation. https://www.rdocumentation.org/packages/tuneRanger/versions/0.7/topics/tuneRanger

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, *9*(3). https://doi.org/10.1002/widm.1301

R Documentation. (2017). MissRanger function. https://www.rdocumentation.org/packages/missRanger/versions/2.6.1/topics/missRanger

Scott, G. (2003, November). Capital intensive: Definition, examples, and measurement. https://www.investopedia.com/terms/c/capitalintensive.asp

Steinwand, D. (2000). *A Risk Management Framework for Microfinance Institutions*. MicroFinance Network, Shorebank Advisory Services.

Stekhoven, D. J., & Bühlmann, P. (2011). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

The World Bank Group. (2019). Mix Market Data. https://datacatalog.worldbank.org/search/dataset/0038647/MIX-Market

The World Bank Group. (2023). World Development Indicators Data. https://datacatalog.worldbank.org/search/dataset/0037712/world_development_indicators

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Uddin, M. H., Akter, S., Mahi, M. A., & Mollah, S. (2024). Why do microfinance institutions charge higher interest rates than banks? the role of operating costs. *Finance Research Letters*, *70*, 106319. https://doi.org/https://doi.org/10.1016/j.frl.2024.106319

Werner, J. (2012). Rational asset pricing bubbles revisited. *2012 Meeting Papers, Society for Economic Dynamics*.

Wright, M. N. (2017a). *ranger variable importance p-values*. Rdocumentation. https://search.r-project.org/CRAN/refmans/ranger/html/importance_pvalues.html

Wright, M. N. (2017b). *ranger: Ranger*. Rdocumentation. https://www.rdocumentation.org/packages/ranger/versions/0.16.0/topics/ranger

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01

8 Appendix

8.1 Countries Included in Final Dataset

In this section, we present the countries included the final dataset (after removing outliers and imputing missing values). More specifically, Tables 10 and 11 show the number of MFIs per country.

Table 10: Number of MFIs per Country

Country	Number of MFIs
Afghanistan	3
Albania	1
Argentina	1
Armenia	3
Azerbaijan	20
Bangladesh	40
Belarus	1
Belize	1
Benin	14
Bhutan	1
Bolivia	18
Bosnia and Herzegovina	5
Brazil	22
Bulgaria	15
Burkina Faso	7
Burundi	5
Cambodia	40
Chad	2
Chile	2
China	26
Colombia	14
Congo, Dem. Rep.	6
Congo, Rep.	3
Costa Rica	3
Cote d'Ivoire	12
Dominican Republic	7
Ecuador	45
Egypt, Arab Rep.	7
El Salvador	9
Ethiopia	13
Georgia	7
Grenada	1
Guatemala	14

Country	Number of MFIs
Guinea	2
Haiti	3
Honduras	18
India	123
Indonesia	46
Iraq	3
Jamaica	4
Jordan	3
Kazakhstan	20
Kenya	14
Kosovo	5
Kyrgyz Republic	18
Lao PDR	14
Lebanon	1
Madagascar	9
Malawi	6
Malaysia	1
Mali	6
Mexico	58
Moldova	6
Mongolia	8
Montenegro	3
Morocco	5
Mozambique	7
Myanmar	6
Namibia	1
Nepal	32
Nicaragua	13
Niger	9
Nigeria	45
Pakistan	19
Panama	4
Papua New Guinea	3

Table 11: Number of MFIs per Country

Country	Number of MFIs
Paraguay	3
Peru	57
Philippines	68
Poland	3
Romania	6
Rwanda	28
Samoa	1
Senegal	33
Sierra Leone	7
Slovak Republic	1
South Africa	6
South Sudan	3
Sri Lanka	20
St. Lucia	2
Sudan	1
Suriname	2
Syrian Arab Republic	2
Tajikistan	41
Tanzania	15
Timor-Leste	2
Togo	22
Tonga	1
Trinidad and Tobago	4
Tunisia	1
Uganda	32
Ukraine	2
Uruguay	2
Uzbekistan	17
Vanuatu	1
Viet Nam	28
West Bank and Gaza	6
Yemen, Rep.	3
Zambia	5
Zimbabwe	4

8.2 Checking the Assumptions for Linear Regression

Since linear regression had many assumptions, let us check them to ensure they hold. Let us first recall these assumtions.

1. Linearity: There is a linear relationship between the response variable and the predictors, which can be expressed as

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p} + \epsilon_i, \tag{2}$$

   where Y_i is the response variable for the i-th observation, $X_{i,j}$ is the j-th predictor value for the i-th observation, β_j quantifies the association between that predictor $X_{i,j}$ and the response variable Y_i, and ϵ_i is the error term for the i-th observation.

2. No Perfect Multicollinearity: The columns of $\mathbf{X}$ must be linearly independent, that is, no predictor can be written as a linear combination of the other predictors.

3. Strict exogeneity of Predictors: The predictors carry no information about the error term, which can be expressed as $\mathbb{E}[\epsilon_i | X_{i,1}, \ldots, X_{i,p}] = 0$ for all $i \in \{1, \ldots, N\}$. By the Law of Iterated Expectations, this implies that

$$\mathbb{E}[\epsilon_i] = \mathbb{E}\left[\mathbb{E}[\epsilon_i | X_{i,1}, \ldots, X_{i,p}]\right] = \mathbb{E}[0] = 0$$

   for all $i \in \{1, \ldots, N\}$.

4. Homoscedasticity: The variance of the error term is the same for all observations, regardless of the values of the predictors. This is formulated as $\mathrm{Var}(\epsilon_i | \mathbf{X}) = \mathrm{Var}(\epsilon_i) = \sigma^2$ for some positive constant σ and for all $i \in \{1, \ldots, N\}$.

5. No Serial Correlation of Error terms: All error terms are uncorrelated with one another, which is expressed as $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for all $i, j \in \{1, \ldots, N\}$ such that $i \neq j$.

6. $\epsilon_i | \mathbf{X} \sim N(0, \sigma^2)$ for all $i \in \{1, \ldots, N\}$.

We note that we already checked the 2nd assumption as we plotted the correlations between the variables and we found that there is weak correlation between the variables, with the highest being 0.31 (ignoring the sign). Hence, we only have 5 assumptions to check. We shall check them separately for each response variable, Portfolio at Risk 30 days and Portfolio at Risk 90 days.

Portfolio at Risk 30 days

Figure 14 shows three plots for full linear regression (with unstandardised data) for response variable Portfolio at Risk 30 days. The Residuals vs Fitted plot shows the variability of the residual values with predictors. This is to check the first assumption, that is, linearity. A relationship is linear if the red line is horizontal, without distinct patterns. We see that the red line is almost horizontal, though the tails curve upwards. Additionally, the residuals seem to be concentrated at one region and there are several outliers. Hence, this may suggest a non-linear relationship. Transforming some of the variables may help linearise the relationship.

The second plot shows a Q-Q plot (that is, a normal probability plot), which displays the distribution of the residuals across the model. This plot checks the 6th assumption. If the residuals are normally distributed, then all the points should fall approximately along the reference line. Though the majority of the points fall on this reference line, a lot of

the other points deviate from this line. So, the residuals are roughly normally distributed, but with heavier-tailed than the theoretical distribution.

The third plot, Scale-Location plot, is used to assess whether the variance of the

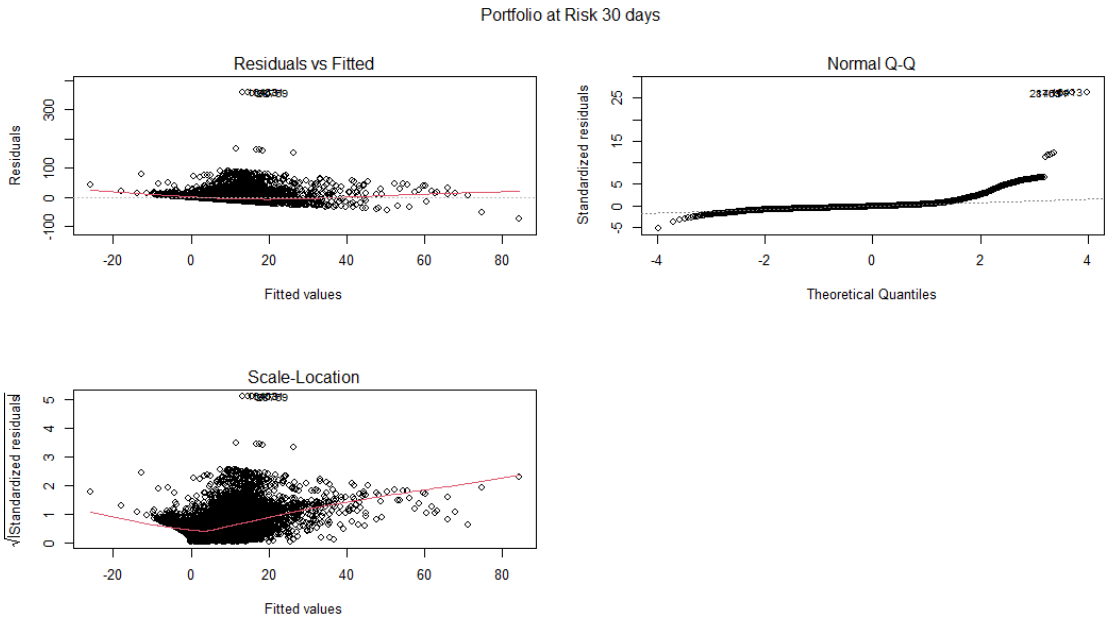

Figure 14: We plot the Residuals vs Fitted, a Q-Q plot and Scale-Location for full linear regression (with unstandardised data) for response variable Portfolio at Risk 30 days.

residuals is constant (the 4th assumption). This plot shows if the residuals are spread equally along the ranges of the predictors. If the red line is horizontal with equally spread points, then the variance of the residuals is constant. In our case, we do not observe that. The red line is curved and the points are mostly concentrated in one region. Thus, this assumption also does not hold.

To check if there is no correlation between the residuals (the 5th assumption), we can use Durbin-Watson test's. This hypothesis test detects the presence of autocorrelation in the residuals of a linear regression. The null hypothesis is that there is no correlation among the residuals and the alternative hypothesis is the reverse. In R, we use this test and obtain a p-Value of 0.0002. This means that we reject the null hypothesis, and there is autocorrelation between the residuals. We also obtained a value for the autocorrelation, which is 0.05386211. For positive serial correlation, we could consider adding lags of the response variable or predictors to this model. This can be applied for further research. Unfortunately, there is no direct test to check for strict exogeneity (the 3rd assumption).

Portfolio at Risk 90 days

Figure 15 shows three plots for full linear regression (with unstandardised data) for response variable Portfolio at Risk 90 days. The first plot shows the Residuals vs Fitted, and we see that the red line is approximately horizontal. However, the points are not randomly spread around the red line. This suggests that the relationship may not be linear.

The second plot shows the Q-Q plot, which assesses the distribution of the residuals. We observe that most of the points fall on the line, however there are some points that deviate when the theoretical quantiles are positive. This suggests that residuals have a right-skewed normal distribution. This is not a direct violation, however this skewness may lead to linear regression overfitting. The third plot, Scale-Location plot, checks whether

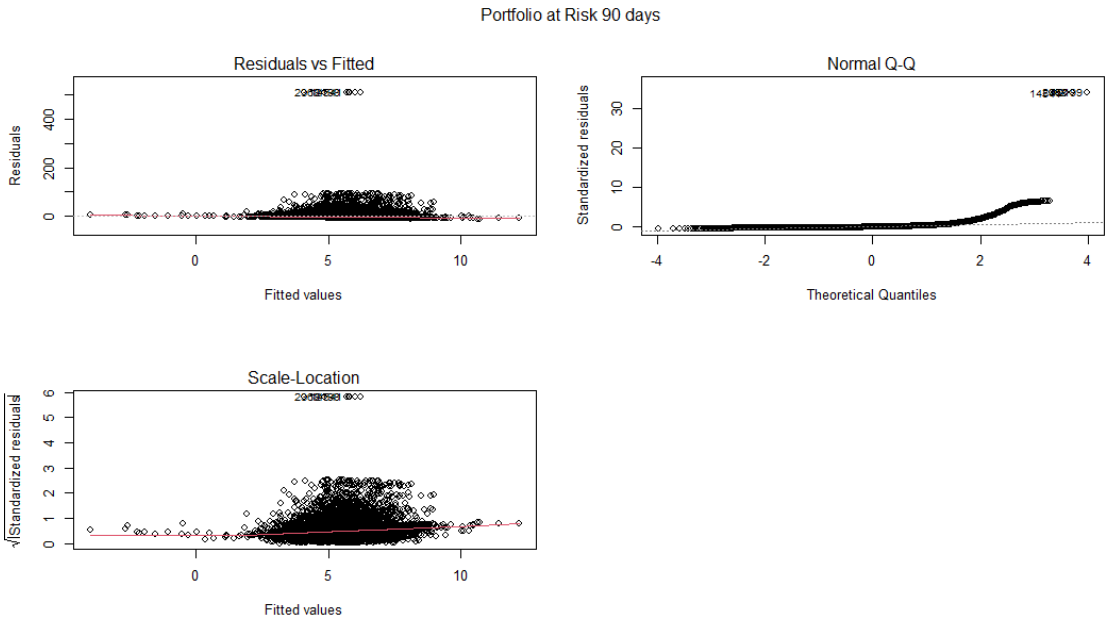

Figure 15: We plot the residuals against the fitted values, a Q-Q plot and Scale-Location for full linear regression (with unstandardised data) for response variable Portfolio at Risk 90 days.

the variance of the residuals is constant. The red line is slightly curved and above 0, and the points are concentrated in one region, thus the variance of the residuals is not constant.

We also apply the Durbon-Watson Test and obtain a p-value of 0.01, with autocorrelation 0.03247181. Hence, we reject the null hypothesis and find that there is some correlation between the residuals.

8.3 Partial Dependence Plots

We show the partial dependence plots for the two response variable, Portfolio at Risk 30 days and Portfolio at Risk 90 days. We show the plots for the proxies that are not shown in the main body of the thesis, as they did not have statistically significant permutation importance values.

Portfolio at Risk 30 days

Let the response variable be Portfolio at Risk 30 days. Figure 16 shows the partial dependence plot for Loan Growth. We see that for negative values for Loan Growth, the predicted response variable has a downward trend. Once Loan Growth becomes positive, the predicted value stabilisies then begins to increase quickly. We observe this for both hyperparameters, and hence this shows that the relationship between Loan Growth and the predicted Portfolio at Risk 30 days is non-linear.

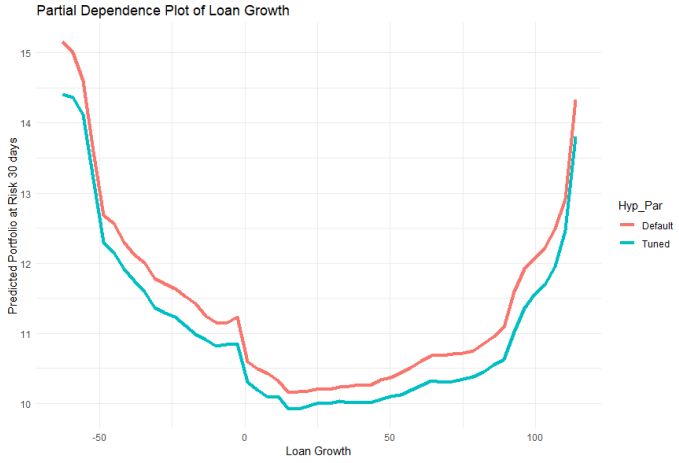

Figure 16: Plot showing the partial dependence plot of Loan Growth with response variable Portfolio at Risk 30 days and for both default and tuned hyperparameters.

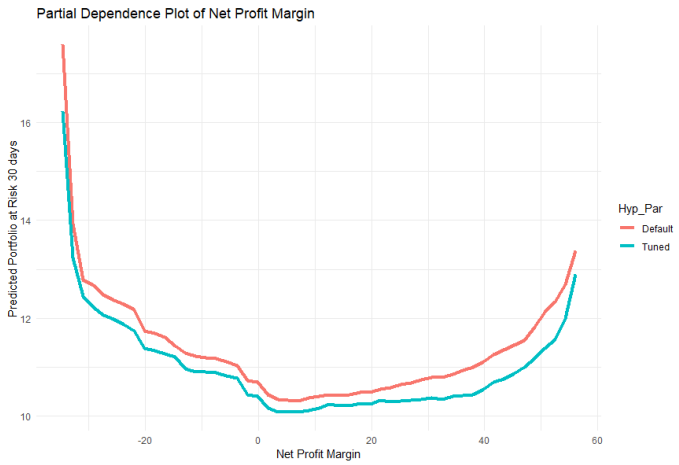

Figure 17: Plot showing the partial dependence plot of Net Profit Margin with response variable Portfolio at Risk 30 days and for both default and tuned hyperparameters.

Figure 17 shows the partial dependence plot for Net Profit Margin. For negative values of Net Profit Margin, the predicted response variable decreases, stabilises once Net Profit

Margin is positive then slowly starts to increase. Once more, this holds for both hyperparameters, and so this shows that the relationship between Net Profit Margin and the predicted Portfolio at Risk 30 days is non-linear.

Portfolio at Risk 90 days

Let the response variable be Portfolio at Risk 30 days. Figure 18 shows the partial dependence plot for Loan Growth. For large negative or positive values of Loan Growth, the predicted response variable is high, but in between this variable roughly stabilises. This shows that the relationship between Loan Growth and the predicted Portfolio at Risk 90 days is non-linear. There is a gap between the random forest with default and tuned hyperpearmeters. More specifically, tuning minimised the effect of large negative or positive values of Loan Growth and overall tuning led to lower predicted response variable.

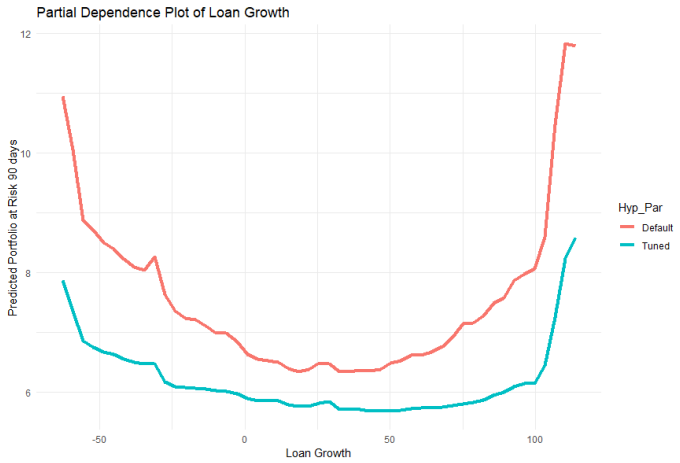

Figure 18: Plot showing the partial dependence plot of Loan Growth with response variable Portfolio at Risk 90 days and for both default and tuned hyperparameters.

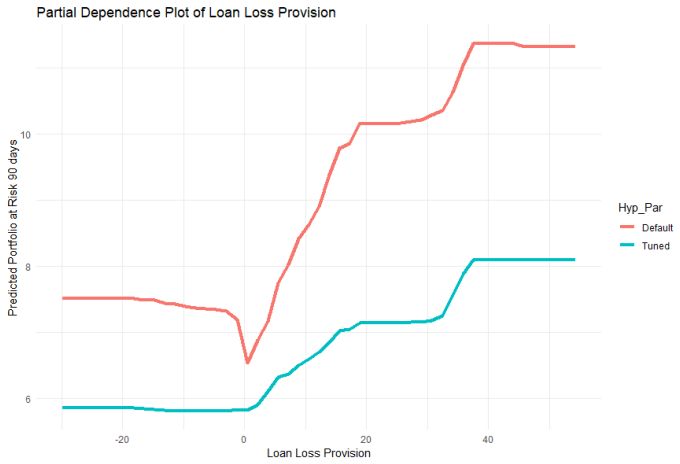

Figure 19: Plot showing the partial dependence plot of Loan Loss Provision with response variable Portfolio at Risk 90 days and for both default and tuned hyperparameters.

Figure 19 shows the partial dependence plot for Loan Loss Provision. We see that for random forest with default parameters, the predicted response variable has an upward trend, though there is a dip when Loan Loss Provision is zero. The case with tuned

hyperparameters, we see a less pronounced upward trend which is lower than the case with default hyperparameters. Since tuning improves the performance of random forest and hence its results, we have that there is approximately a non-decreasing monotonic relationship between Loan Loss Provision and predicted Portfolio at Risk 90 days.