UNIVERSITY OF THE AEGEAN

DEPARTMENT OF STATISTICS AND ACTUARIAL - FINANCIAL MATHEMATICS

STATISTICS AND DATA ANALYSIS

A MACHINE LEARNING APPROACH FOR MICRO-CREDIT SCORING AND LIMIT OPTIMIZATION

MASTER THESIS

TSELEKIDOU EFTYCHIA

SAMOS 2023

UNIVERSITY OF THE AEGEAN

DEPARTMENT OF STATISTICS AND ACTUARIAL - FINANCIAL MATHEMATICS

STATISTICS AND DATA ANALYSIS

A MACHINE LEARNING APPROACH FOR MICRO-CREDIT SCORING AND LIMIT OPTIMIZATION

MASTER THESIS

TSELEKIDOU EFTYCHIA
sasm21004
JUNE, 2023

EVALUATION COMMITTEE

LAPPAS PANTELIS
(SUPERVISOR)

KARAGRIGORIOU ALEXANDROS

XANTHOPOULOS STYLIANOS

To my family.

Contents

Acknowledgements

I would like to express my deepest appreciation to the individuals and organizations who have provided invaluable support and made significant contributions throughout the completion of this thesis.

First and foremost, I am profoundly grateful to my supervisor, Pantelis Lappas, for his invaluable guidance, expertise, and unwavering support. His insightful feedback and constructive criticism have been instrumental in shaping the development of this research.

I would like to extend my sincere thanks to Alex Karagrigoriou and Stelios Xanthopoulos, my thesis committee members, for their valuable input, suggestions, and dedicated time spent reviewing and evaluating my thesis.

I am truly thankful to the University of the Aegean for providing me with access to the necessary resources and information that have made this research possible.

My heartfelt appreciation goes to my colleagues and friends who have provided continuous encouragement and assistance. I am incredibly grateful to have had the opportunity to meet and interact with them as they have become an integral part of my life throughout these years.

Furthermore, I am deeply indebted to my family for their unwavering love, support, and understanding, particularly during challenging times. Their constant encouragement and belief in my abilities have played a pivotal role in overcoming obstacles and maintaining my motivation throughout the entire thesis process.

In the end, I am deeply grateful to the person who first introduced me to the field of statistics, which made me really love and value this subject. I am very thankful for his priceless help and support during this journey.

Although it is impossible to mention everyone who has contributed to this thesis, please accept my sincere gratitude for your support and encouragement.

Abstract

This thesis, conducted within the postgraduate program ”Statistics and Data Analysis” at the University of the Aegean in Department of Statistics and Actuarial - Financial Mathematics, aims to enhance credit risk assessment in the context of micro-loans by analyzing optimal boundary results for both new and existing clients. The research objective is to contribute to improved lending practices and financial inclusion through the development of a more accurate and effective credit risk assessment framework. A multi-method approach is employed to achieve this objective.The research begins by exploring the theoretical foundations of credit risk and examining various credit rating methods used in real-life scenarios. This comprehensive analysis provides a solid understanding of the current landscape of credit risk assessment in micro-finance.

Furthermore, a case study analysis is conducted using real micro-loan data, focusing on determining optimal thresholds for both new and existing clients. For the total dataset, statistical techniques are applied for data cleansing and feature selection. Principal Component Analysis (PCA) is utilized for dimensionality reduction, identifying key factors contributing to credit risk. Logistic Regression is further employed to develop a predictive model that assesses creditworthiness, considering relevant variables and estimating probabilities of repayment. Additionally, for evaluating and comparing reasons, the Random Forest algorithm was employed in conjunction with the aforementioned methods. The integration of Logistic Regression and Random Forest facilitates a comprehensive and meticulous analysis of credit risk factors, thereby augmenting the precision of risk evaluation and decision-making procedures in the domain of micro-lending. To optimize constraints and enhance decision-making, the simplex algorithm is employed. This examination enables the acquisition of optimal quantities of micro-loans that can be disbursed and determines the number of loans to be approved, ensuring efficient and effective loan allocation. By integrating these methodologies, this research advances the field of credit risk analysis in micro-loans. The findings provide valuable insights into improving risk assessment accuracy and decision-making processes.

Key words: credit scoring, micro-finance, machine learning, logistic regression, random forest, limit optimization, simplex method.

Περίληψη

Η παρούσα διπλωματική εργασία, η οποία διεξήχθη στο πλαίσιο του μεταπτυχιακού προγράμματος ¨Στατιστική και Ανάλυση Δεδομένων¨ του Πανεπιστημίου Αιγαίου στο Τμήμα Στατιστικής και Αναλογιστικών - Χρηματοοικονομικών Μαθηματικών, αποσκοπεί στην ενίσχυση της αξιολόγησης του πιστωτικού κινδύνου στο πλαίσιο των μικροδανείων, αναλύοντας τα βέλτιστα αποτελέσματα των ορίων τόσο για νέους όσο και για υφιστάμενους πελάτες. Ο ερευνητικός στόχος είναι να συμβάλει στη βελτίωση των πρακτικών δανειοδότησης και της χρηματοπιστωτικής ένταξης μέσω της ανάπτυξης ενός πιο ακριβούς και αποτελεσματικού πλαισίου αξιολόγησης του πιστωτικού κινδύνου. Για την επίτευξη αυτού του στόχου χρησιμοποιείται μια προσέγγιση πολλαπλών μεθόδων. Η έρευνα ξεκινά με τη διερεύνηση των θεωρητικών βάσεων του πιστωτικού κινδύνου και την εξέταση διαφόρων μεθόδων αξιολόγησης της πιστοληπτικής ικανότητας που χρησιμοποιούνται σε πραγματικά σενάρια. Αυτή η ολοκληρωμένη ανάλυση παρέχει μια σταθερή κατανόηση του σημερινού τοπίου της αξιολόγησης του πιστωτικού κινδύνου στη μικροχρηματοδότηση.

Στη συνέχεια, πραγματοποιείται ανάλυση μελέτης περίπτωσης με τη χρήση πραγματικών δεδομένων μικροδανείων, εστιάζοντας στον προσδιορισμό των βέλτιστων ορίων μικροδανείων τόσο για νέους όσο και για υφιστάμενους πελάτες. Για το σύνολο των δεδομένων, εφαρμόζονται στατιστικές τεχνικές για τον καθαρισμό των δεδομένων και την επιλογή χαρακτηριστικών. Η Ανάλυση Κύριων Συνιστωσών χρησιμοποιείται για τη μείωση της διαστατικότητας, προσδιορίζοντας τους βασικούς παράγοντες που συμβάλλουν στον πιστωτικό κίνδυνο. Η Λογιστική Παλινδρόμηση χρησιμοποιείται περαιτέρω για την ανάπτυξη ενός προγνωστικού μοντέλου που αξιολογεί την πιστοληπτική ικανότητα, λαμβάνοντας υπόψη τις σχετικές μεταβλητές και εκτιμώντας τις πιθανότητες αποπληρωμής. Επιπλέον, για λόγους αξιολόγησης και σύγκρισης, ο αλγόριθμος Random Forest χρησιμοποιήθηκε σε συνδυασμό με τις προαναφερθείσες μεθόδους. Η ενσωμάτωση της Λογιστικής Παλινδρόμησης και του Random Forest διευκολύνει την ολοκληρωμένη και ενδελεχή εξέταση των παραγόντων πιστωτικού κινδύνου, αυξάνοντας έτσι την ακρίβεια των διαδικασιών αξιολόγησης του κινδύνου και λήψης αποφάσεων στον τομέα των μικροδανείων. Για τη βελτιστοποίηση των περιορισμών και την ενίσχυση της λήψης αποφάσεων, χρησιμοποιείται ο αλγόριθμος εν Σιμπλεξ. Η μέθοδος αυτή επιτρέπει την απόκτηση των βέλτιστων ποσοτήτων μικρών δανείων που μπορούν να εκταμιευθούν και καθορίζει τον αριθμό των προς έγκριση δανείων, διασφαλίζοντας την αποδοτική και αποτελεσματική κατανομή τους. Με την ενσωμάτωση αυτών των μεθοδολογιών, η παρούσα έρευνα προάγει τον τομέα της ανάλυσης του πιστωτικού κινδύνου στα μικροδάνεια. Τα αποτελέσματα παρέχουν πολύτιμες πληροφορίες για τη βελτίωση της ακρίβειας της αξιολόγησης του κινδύνου και των διαδικασιών λήψης αποφάσεων.

Λέξεις-κλειδιά: πιστωτική βαθμολόγηση, μικρο-χρηματοδότηση, μηχανική μάθηση, λογιστική παλινδρόμηση, random forest, **βελτιστοποίηση ορίων, μέθοδος** simplex.

1 Introduction

The provision of Micro-loans has surfaced as a crucial means of financial sustenance for both individuals and small-scale enterprises experiencing constraints in accessing customary banking services. Precisely evaluating credit risk with regards to micro-loans poses a noteworthy difficulty. The distinctive attributes of micro-loans, characterized by a substantial number of modest loan amounts, restricted availability of past financial information, and absence of collateral, render it challenging to establish the creditworthiness of borrowers with complete assurance. Consequently, financial institutions encounter elevated unpredictability in establishing optimal loan limits for both novel and current customers. The resolution of these obstacles and the delivery of discernment concerning the most fitting boundaries for credit risks can significantly enhance the potency and longevity of micro-finance schemes, bringing benefits to both creditors and debtors.

The primary objective of this dissertation is to effectively tackle the associated challenges and furnish significant perspectives concerning the analysis of credit risk within the micro-loan industry. The investigation endeavors to surmount the deficiencies of traditional credit risk assessment methodologies through the implementation of inventive paradigms and the utilization of advanced statistical and machine learning techniques.

The present study is aimed at uncovering the most effective threshold outcomes concerning both newly acquired andpreexisting client bases in the sphere of micro-finance. By incorporating comprehensive credit scoring models, which encompass both traditional credit risk factors and non-conventional data sources, financial institutions can enhance their ability to make well-informed decisions regarding loan approval and borrowing limits. The goal of this investigation is to enhance the effectiveness and endurance of micro-loan initiatives by rectifying the disparities between established procedures for assessing credit risks and the obstacles encountered within micro-finance.

This study aims to illuminate the fundamental factors affecting credit risk in the context of micro-loan disbursement. Through the identification and comprehensive analysis of various determinants that significantly impact creditworthiness, stakeholders such as policymakers, regulators, and microfinance institutions can acquire vital perspectives to facilitate the creation of personalized interventions and policies aimed at promoting ethical lending practices. The principal aim of the study is to improve the financial inclusion among disadvantaged communities, encourage the implementation of sustainable lending methodologies, and instigate economic progress at a grassroots level.

In summary, the present study aims to provide motivation, innovation, and commitment to the domain of micro-loan credit risk assessment. Through the presentation of innovative approaches, utilization of advanced methodologies, and integration of alternative data sources, this research endeavors to address the obstacles associated with assessing credit risk in the microloan sector. The findings and insights of this study present a significant potential for transforming the practices of credit risk assessment, enhancing decision-making procedures for financial institutions, and promoting economic opportunities and growth for disadvantaged populations.

The study is structured into four separate chapters that aim to examine various crucial aspects pertaining to the management and assessment of credit risk. Chapter 2 establishes the theoretical framework for credit risk and explores the diverse credit rating techniques implemented in real-world settings. Furthermore, this chapter presents an analysis of the various industries in which credit rating practices are frequently utilized, furnishing a comprehensive overview of the extant credit risk evaluation assessment. Chapter 3 explores statistical methodologies utilized in the micro-credit approach. This chapter focuses into the utilization of statistical models, such as logistic regression, for the purpose of evaluating creditworthiness and scrutinizes their efficacy within the domain of micro-loans. Afterwards, Chapter 4 discusses techniques for optimizing problems, with a specific emphasis on the well-established Simplex method for solving linear programming problems. Upon conclusion, Chapter 5 offers a thorough investigation of micro-loans utilizing a comprehensive approach based on a real case study. This chapter is dedicated to the examination of data pre-processing and analysis, accomplished through the use of advanced analytical techniques.

2 Credit Risk and Credit Scoring

2.1 Introduction

Credit risk represents a tremendously significant commodity within the context of banking establishments and financial organizations. There exists a constant demand from clients seeking financial assistance in the form of a loan. Due to the inherent risks associated with loans, it is crucial to discern viable loan candidates and establish measures to distinguish favorable applicants from unfavorable ones. In order to mitigate the aforementioned challenge, commercial entities such as banks initiated the process of creating credit scores. By utilizing the credit scores of prospective borrowers, financial institutions are able to determine the level of risk associated with loan applications. The process of computing an individual's credit score is utilized by lenders in order to determine the suitability of extending credit. Factors such as the potential borrower's ability to meet their loan obligations and the percentage of credit or loan that they are eligible to receive are taken into account when making this determination [111].

Lending institutions, commonly employs "historical" information acquired from clientele in order to construct the scorecard for prospective borrowers. The aforementioned task was initialy executed through the collection of pertinent information regarding prospective candidates, comprising their income, employment sector, present occupation, remaining obligations, financial holdings, tenure with the financial institution, credit record, and any past history of defaults or tardiness in payments. In contemporary times, the practice of credit scoring gained significant prevalence subsequent to the 80s, in accordance with scholarly research conducted by Lyn et al. [111]. Historically, credit scoring was a practice exclusively employed by financial institutions, specifically banks. However, its application has since been broadened to encompass the issuance of credit cards, which serve as an alternative type of loan. Presently, credit scoring is utilized across multiple industries including credit cards, membership cards, mobile network providers, insurance companies, and various government agencies.

An essential premise that underlies the development of a credit scoring model is the assumption that past patterns can serve as reliable indicators of future trends. Through the examination of historical repayment patterns exhibited by former clients, it is feasible to glean insight into the prospective default risks posed by future patrons. In particular, with regard to the prior clientele, knowledge is gained regarding those who have demonstrated favorable financial reliability and those who have exhibited unfavorable financial reliability. The aforementioned binary target variable, Y, denoting a positive or negative status, is presently established as the primary focus of our analysis. The main objective is the examination of its correlation with all relevant information accessible at the time of scoring regarding the obligors in question. The aim of credit scoring nowadays, is to accurately measure this linkage with the purpose of aiding credit assessments, monitoring, and administration. Financial institutions evaluate the creditworthiness of borrowers both at the time of loan application and periodically throughout the duration of the financial agreement, typically encompassing loans, loan commitments, and guarantees. This entails a scoring system designed to assess the borrower's risk, allowing financial institutions to make informed lending decisions.

2.2 Credit Risk

When financial institutions extend loan agreements to borrowers in the form of mortgages, credit cards or other types of credit instruments, there exists a certain level of inherent risk regarding the possibility of non-repayment by the borrower. In the event that a business extends credit to its clientele, there exists a possibility that said clientele may default on their payment obligations.

One of the crucial determinants of the value and return rate of financial endeavors is credit risk. Whilst the study and evaluation of credit risk's impact on bond value has been extensively explored over time, the emergence of analytical models for the examination and quantification of such effects has only recently surfaced.

Credit risk is assessed by evaluating the borrower's comprehensive capability to reimburse a loan in compliance with its initial conditions. In the process of evaluating the credit risk associated with a consumer's loan, lenders commonly examine the five key factors, commonly referred to as the five Cs of credit (see subsection 2.4), including the borrower's credit history, ability to repay the loan, financial capital, loan terms and conditions, and any collateral held in association with the loan agreement.

The assessment and quantification of credit risk is considered to be a complex and extensively examined dimension of market risk. The conventional approach towards studying the subject matter has been to employ the actuarial techniques of risk evaluation which are rooted in the analysis of past data. The accelerated expansion of derivative activity within the financial market, specifically within the domain of over-the-counter and credit derivatives, along with the elevated level of complexity exhibited by particular financial instruments, has exposed the insufficiency of conventional approaches in accurately assessing real-world risks.

In essence, credit risk can be characterized as the possibility that a contractual counterpart fails to fulfill its monetary responsibilities, resulting in a detrimental outcome for the loaning entity. Nonetheless, this description solely considers the most extreme scenario in which the debtor is rendered insolvent. A decline in the creditworthiness of the debtor may also result from a deterioration in their financial condition, without necessarily leading to insolvency. A more comprehensive explication of the concept of ”credit risk” entails the impact that an unanticipated fluctuation in a debtor’s level of creditworthiness may have on the credit value. The analytical assessments furnished by rating agencies, such as Standard & Poor’s and Moody’s, represent an appraisal of the creditworthiness of corporations and nations.

Consequently, there exist two distinct interpretations of the term ”credit risk” that serve to differentiate between credit loss incurred subsequent to a debtor’s insolvency, herein referred to as the default-mode paradigm, and variations in exposure value resulting from the depreciation of the debtor’s creditworthiness, with insolvency being an infrequent occurrence. This alternative interpretation is recognized as the Mark-to-Market or Mark-to-Model paradigm.

2.3 Credit Risk Measure

Undoubtedly, **Value at Risk (VaR)** is the most commonly employed metric in financial risk management. This particular risk measurement methodology though may not be optimally suited for quantifying counterparty risk given a myriad of underlying factors. The notion of credit exposure necessitates a comprehensive assessment across multiple temporal horizons in order to illuminate the impact of the passage of time and overarching trends. Counterparty risk must be evaluated from both pricing and risk management perspectives, necessitating the use of multiple metrics. In order to conduct a comprehensive assessment of counterparty risk within a portfolio, it is imperative to attain a nuanced understanding of the effective exposure that exists in relation to each and every individual counterparty.

Although VaR is the most well-known risk measure, there exist several alternative ones:

- In the context of risk management, institutions demonstrate a preference for positions exhibiting positive Mark to Market values as they are uniquely positioned to generate exposure, a factor previously noted given the asymmetrical nature of counterparty risk. The anticipated level of exposure at time t is solely based on positive Mark to Market evaluations. The term **"mean future exposure"** is operationally defined as the calculated arithmetic average of the distribution of exposure at a predetermined future date.

- Regarding **potential future exposure (PFE)**, it can be posited that at a predetermined time interval (t) and accounting for the most unfavorable hypothetical situation, it denotes the future exposure value to a specific level of certainty. Due to the uncertainty of future Mark-to-Market values at a specified date, the PFE is best characterized as a probability distribution.

- The **anticipated favorable exposure** within a designated period [0,T], represents a methodological appraisal of potential exposure for transactions that are susceptible to counterparty risks. This evaluation is conducted through a weighted average computation, spanning a duration that is determined by the anticipated values of these exposures. The notion of a weighted average is established based on the correlation between the temporal scope of individual anticipated exposures and the overall duration of the relevant time frame under consideration.

2.4 The 5 Cs of Credit

The quintessential attributes of financial credit, also known as the "**Five Cs of Credit**", serve as an evaluative framework employed by lending institutions to assess the reliability and solvency of prospective borrowers. The theoretical framework underpinning the lending system entails an assessment of five intrinsic attributes of the prospective borrower alongside the terms and conditions of the loan. The goal of this rigorous evaluation exercise is to ascertain the likelihood of default and potential exposure to financial loss borne by the lender. The quintet of credit evaluation parameters referred to as the "Five Cs of Credit" encompasses character, capacity, capital, collateral, and conditions.

Hence, borrower assession method involves the employment of both qualitative and quantitative criteria. When assessing a borrower's financial eligibility, lenders usually evaluate a range of pertinent documentation, including, but not limited to, credit reports, credit scores, and income statements. Additionally, they take into account information pertaining to the loan itself. However, every lender possesses a unique approach in assessing the creditworthiness of a prospective borrower.

2.4.1 Character

The term character, refers to the credit history of a borrower, encompassing his standing or past performance in terms of repayment of debts. The aforementioned data is documented on the credit reports of the recipient and is created by the three prominent credit bureaus, namely Equifax, Experian, and TransUnion. Credit reports include comprehensive data concerning an individual's previous borrowing activities, such as the amount borrowed, and his compliance with loan repayment schedules.

Moreover, the aforementioned reports comprise data related to collection accounts and bankruptcies, with the majority being preserved for a period of seven to ten years. The data presented in these reports facilitate the assessment of the borrower's creditworthiness by financial institutions. A prominent instance of this can be witnessed in the workings of FICO, which utilizes the data available in a consumer's credit report to generate a credit score. This score serves as a mechanism for lenders to obtain a swift overview of the individual's capacity to honor their financial commitments, preceding a more extensive appraisal of their credit reports.

A specific credit score threshold must be met by potential borrowers in order to secure approval for a fresh loan from a considerable number of lenders. The minimum credit score criteria typically exhibit inconsistencies between lenders as well as divergences amongst distinct loan offerings. The positive correlation between a borrower's credit score and the probability of loan approval is a pervasive norm in lending practices.

Lenders frequently rely upon credit scores to determine the interest rates and conditions of loans. The outcome frequently manifests in more desirable loan propositions for individuals possessing high-quality credit scores. Considering the paramount importance of possessing a commendable credit score and an accurate credit report to obtain loan approvals, it would be prudent to contemplate availing the benefits offered by the premier credit monitoring services to ensure the safeguarding of this valuable data.

2.4.2 Capacity

The capacity of a borrower to repay a loan, is gauged by means of evaluating his income in relation to their recurring debts, and by scrutinizing their debt-to-income (DTI) ratio. Loan providers determine the DTI ratio by aggregating a borrower's overall monthly indebtedness and dividing it by the gross monthly earnings of the borrower. The propensity for a potential borrower to receive approval for a new loan is directly correlated with a diminished DTI.

It is noteworthy that every lending organization has its individual set of criteria. However, a substantial number of lenders tend to favor an applicant's DTI to be at or below 36% before sanctioning a new financing application. It is noteworthy that on certain occasions, lenders may be restricted from extending loans to customers who possess elevated DTI ratios [47].

According to the Consumer Financial Protection Bureau (CFPB), meeting the criteria for a new mortgage generally necessitates a DTI of 43% or less, which is intended to guarantee that the borrower can conveniently afford the monthly payments for the new loan [27].

2.4.3 Capital

In addition to the aforementioned factors, financial lenders also take into account the capital contributions made by the borrower in the context of a potential investment. The probability of default is inversely related to the magnitude of the borrower's capital contribution.

Borrowers who are able to provide a down payment for their prospective home, tend to encounter less difficulty when seeking to secure a mortgage including specialized mortgages intended to expand access to home ownership for a wider range of individuals. As an illustration, loans offered by the Federal Housing Administration (FHA) which are backed with an assurance may necessitate a minimum down payment of 3.5% or greater, while 90% loans underwritten by the United States are also subject to similar conditions ([64] and [65]). The Department of Veterans Affairs (VA) does not necessitate a down payment whatsoever. The measure of capital contributions serves to convey a borrower's degree of investment, thereby potentially increasing lender's level of assurance in granting credit.

The magnitude of the initial payment may also exert an influence on the interest rates and conditions associated with the credit facility granted to a debtor. Typically, a higher initial deposit or augmented capital contributions tend to yield superior rates and contractual conditions. In the realm of mortgage finance, it is widely recognized that providing a down payment of at least 20% can effectively circumvent any obligation on the part of the borrower to procure supplementary private mortgage insurance (PMI) [78].

2.4.4 Collateral

Collateral can serve as a means by which a borrower can secure a loan. This statement posits that providing collateral enables the lending institution to obtain a form of guarantee that in the event of loan default by the borrower, the lender can recuperate the collateral by means of repossession. The security collateral serves as the underlying asset against which a borrower secures monetary lending, for instance, an auto loan is pledged against an automobile and mortgages are pledged against residential properties.

Due to their inherent nature, loans that are supported by collateral are known as secured loans or secured debt. It is commonly perceived that these instruments entail lower risk for the issuers. Consequently, financing arrangements that are guaranteed by a form of collateral are frequently provided with more favorable lending conditions, including reduced interest rates, in contrast to alternative unsecured financing options.

2.4.5 Conditions

Influence on the lender's willingness to finance the borrower is exerted by the loan conditions, encompassing key aspects such as the interest rate and principal amount. The term conditions can pertain to the purpose for which a borrower intends to utilize the funds obtained. The conditions of business loans designed to enhance prospective cash flow may be superior to those of house renovation loans during a period of depressed housing market with regards to borrowers who have no plans of selling their property.

Moreover, lenders may take into account factors beyond the purview of the borrower, such as the prevailing economic climate, shifts in industry patterns, or impending regulatory modifications. For businesses endeavoring to acquire a loan, imponderable factors encompassing the future financial stability of crucial suppliers or clientele could serve as potential hurdles.

Figure 1, illustrates the aforedescribed 5 Cs of credit risk.

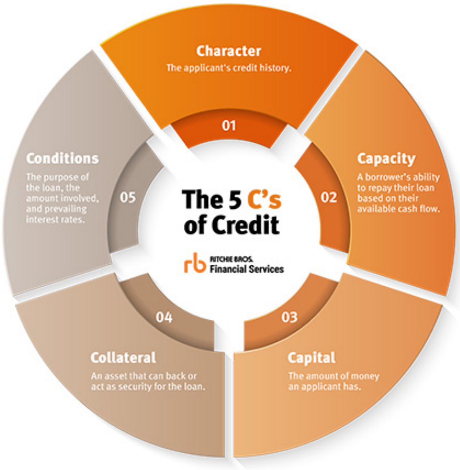

Character
The applicant's credit history.
01
Capacity
A borrower's ability to repay their loan based on their available cash flow.
02
The 5 C's of Credit
RITCHIE BROS. Financial Services
03
Capital
The amount of money an applicant has.
04
Collateral
An asset that can back or act as security for the loan.
05
Conditions
The purpose of the loan, the amount involved, and prevailing interest rates.

Figure 1: The 5 Cs of Credit

2.4.6 Importance of 5 Cs

Lending institutions utilize a standardized set of guidelines, colloquially referred to as the ”five Cs,” to evaluate the financial credibility of prospective loan seekers, as well as to determine the appropriate interest rates and credit limitations. Financial institutions make use of risk assessment methodologies to determine the level of risk involved in extending credit to a specific borrower, as well as the likelihood of timely and effective repayment of the principal and interest of the loan.

In contrast, the criteria employed by lenders are commonly referred to as the ”four Cs”, according to certain perspectives. Due to the possibility of uniformity in circumstances among debtors, there may be instances wherein certain criteria that can be regulated by the debtor are given preferential emphasis while others are excluded.

However, the inquiry remains as to which out of the five Cs holds utmost significance?

Each of the quintuple Cs possess inherent value and ought to be deemed significant. Certain lenders may assign greater significance to particular categories in comparison to others depending on the current prevailing circumstances.

The unwavering ascertainment of an applicant’s character and capacity holds significant prominence in the decision-making process of a lender when contemplating the extension of credit. Financial institutions often employ various evaluative metrics, such as DTI ratios, credit score minimums, or household income limits, when assessing loan applicants. These criteria generally assess two major categories of financial standing. Although a sizable down payment or collateral can positively impact loan conditions, these factors are typically not the foremost consideration in a lender’s decision to extend credit.

2.5 Basel Regulations

The Basel Regulations, also known as the Basel Accords, refer to a set of international banking supervision standards that were established to provide a unified framework for assessing banking institutions capital adequacy and risk management practices. These regulations were first introduced by the Basel Committee on Banking Supervision in 1988, and have since been revised and updated several times, with the most recent version being known as Basel III. The primary objective of the Basel Regulations is to ensure the stability and soundness of the global banking system, by promoting prudent risk management practices and strengthening the resilience of individual banks to potential shocks or crises.

In the upcoming section, we will endeavor to scrutinize the Capital Accords of Basel I, Basel II, and Basel III. The regulatory guidelines delineated herein aimed to facilitate financial institutions in the effective assessment and determination of their provisions and capital buffers, in order to mitigate the impact of diverse risk exposures. Credit risk is identified as a crucial category of risk that requires considerable attention. This segment undertakes a discussion focused on the effect of these accords on the growth of probability of default (PD), loss given default (LGD), and exposure at default (EAD) credit risk models. The Basel Accords serve as a foundational framework for various components of credit risk analysis and have been proposed by the Basel Committee on Banking Supervision in the context of international financial regulation. The genesis of this committee traces back to 1974, when it was established by the consortium of G10 central banks whereas today, the total number of members stands at 27. Regular meetings are held at the Bank for International Settlements (BIS) located in Basel, Switzerland.

Banks obtain monetary inflow from diverse origins. The primary pertinent sources comprise banking deposits inclusive of savings accounts, term accounts, and related financial instruments. As a reciprocal agreement, depositors are provided with a predetermined or fluctuating rate of interest. An additional source of funding for the bank is derived from shareholders or investors who acquire shares, thereby gaining a degree of ownership in the institution. In the event of the corporation realizing a positive net income, a proportional amount may be disbursed to its shareholders as dividend payments. The acquisition of funds by a bank is predicated on the inclusion of savings money and shareholder capital as integral components of its funding structure. On the asset side, a financial institution typically deploys the funds secured through various channels to make diverse investment decisions. Lending, as a primary investment and integral aspect of essential banking operations, occupies a pivotal position in financial institutions. Financial institutions extend credit to borrowers, enabling them to fund acquisitions of residential properties, vehicles, academic pursuits, or leisurely expeditions. Alternative investment opportunities may include the acquisition of diverse market securities, such as bonds or equities.

It is imperative to acknowledge that, investments of any nature inherently carry a degree of risk. There is a possibility that obligors may potentially default on their loans, leading to non-repayment of principal and interest. Additionally, there exists a risk of market collapse, which may subsequently result in a decrease in the valuation of securities. Due to the significant impact that banks hold within an economic system, it is imperative that they are safeguarded against the array of risks to which they are vulnerable. The mitigation of bank insolvency or failure, is imperative and necessitates addressing the potential hazards associated with bank's asset-side pursuits through proportional liabilities, thereby protecting depositors. The individuals in question ought to be assured of receiving their savings funds on demand at any given time. Therefore, it is imperative for a financial institution to possess a sufficient amount of shareholder capital to act as a safeguard against potential financial setbacks. Indeed, an amplification of analysis would entail the incorporation of retained earnings and reserves, thereby considering equity or capital as a superior metric. To clarify, a financial institution that holds a robust level of capital possesses an adequate quantity of equity in order to safeguard itself from a diverse range of hazards. Therefore, it is imperative that a positive correlation exists between risk and equity within the context of financial matters.

Typically, the quantification of this relationship occurs in a two-step process. Initially, the quantification of risk present on the asset side is accomplished through a numerical representation denoting a designated level of risk. The numerical value is subsequently inserted into an equation that accurately computes the corresponding equity and, therefore, the necessary capital. There exist two divergent perspectives regarding the determination of the risk metric and the corresponding mathematical expression to be employed.

The initial perspective pertains to a regulatory approach, which involves the establishment of regulations such as Basel I, II, and III to accurately delimit the manner in which risk number calculations ought to be conducted and the formula that must be employed. Regulatory capital refers to the aggregate amount of capital that a financial institution must maintain in accordance with stipulated regulations. In the absence of regulatory frameworks, banks would remain aware of the essentiality of equity capital as a safeguard mechanism. In this particular instance, the pertinent parties would employ proprietary risk modeling techniques to ascertain a quantitative risk metric and rely upon their individualized calculation protocols to derive the buffer capital. The aforementioned phenomenon defines the notion of economic capital, which refers to the quantum of capital that a financial institution possesses, ascertained through its internal modeling approach and protocol. The present capital denotes the quantifiable sum of capital possessed by a bank, which comprises both the economic capital and the regulatory capital, with the former being comparatively greater. As an illustrative instance, the Bank of America disclosed a proportion of overall capital to risk-weighted assets employing advanced methodologies, amounting to 13.2% at the conclusion of 2015, alongside the present regulatory minimum capital of 8%, with the latter figure anticipated to rise as the Basel III framework attains complete implementation. Consequently, the existing value of the capital buffer presently stands at 5.2%.

Furthermore, it is noteworthy that diverse forms of capital exist, which are differentiated based on their ability to absorb losses. Tier 1 capital typically comprises the combination of common stock, preferred stock, and retained earnings. Tier 2 capital is deemed to possess a moderately inferior quality as it comprises of subordinated loans, revaluation reserves, undisclosed reserves, and general provisions. The Basel II Capital Accord incorporated Tier 3 capital into its regulatory framework, which encompasses short-term subordinated debt.

2.5.1 Basel I

In 1988, the initial accord unveiled was the Basel I Capital Accord, which aimed to establish minimal regulatory capital prerequisites to guarantee that banks have the capacity to always reimburse depositors funds. The primary emphasis of the Basel I Accord, was centered on credit risk, and it introduced the novel concept of the Capital or Cooke ratio. This ratio is a measure of the available buffer capital in relation to the risk-weighted assets. The minimum threshold for said ratio has been established at 8%, signifying that the magnitude of the capital reserves must exceed 8% of the total risk-weighted assets. Altering the capital requirement by a marginal percentage poses arduous difficulties for prominent banking entities and entails a prolonged period of time. The capital may comprise of both Tier 1 and Tier 2 capital, as per the relevant regulatory guidelines.

Concerning credit risk, the Basel I Capital Accord introduced predetermined risk weights that are contingent upon the exposure class. In terms of cash exposures, a risk weight of 0% was applied, while a risk weight of 50% was employed for mortgage-based exposures, and a risk weight of 100% was utilized for alternative commercial exposures.

While the Basel I Accord marked a considerable stride towards enhancing risk management, it encountered a number of significant limitations. Initially, it must be noted that the debtor's solvency was inadequately considered, as the risk weights relied solely on the exposure class, without taking into consideration the obligor or product characteristics. Moreover, insufficient acknowledgment was rendered towards the contribution of collateral guarantees in mitigating credit risk. The aforementioned situation presented an array of possibilities for regulatory arbitrage whereby entities could employ strategic exploitation of regulatory loopholes for the purpose of limiting their required capital. Therefore, the analysis solely encompassed the evaluation of credit risk, while disregarding operational and market risk factors.

2.5.2 Basel II

In response to the limitations of the Basel I Capital Accord, the Basel II Capital Accord was implemented. The structure of this framework is comprised of three central components, namely, Pillar 1 which pertains to the basic capital requirement, Pillar 2 that concerns the supervisory review procedure, and Pillar 3 that pertains to market discipline and transparency in disclosures.

Within the framework of Pillar 1, three distinct categories of risk are encompassed. Particularly, credit risk is an inherent risk faced by lenders when extending loans or credit facilities to borrowers or obligors, where there exists a probability that the obligor may not be able to fulfill their repayment obligations in accordance with the agreed terms and conditions. The concept of operational risk refers to the risk associated with the prospect of experiencing either direct or indirect losses that may arise from insufficient or malfunctioning internal mechanisms, workforce, and technologies, or even from extraneous contingencies. Well-known paradigmatic situations include instances of deceit, impairment to tangible resources, and malfunctions within systems. The market risk is referred as the potential exposure to unfavorable fluctuations in the market that may affect a financial institution's market position in terms of cash or derivative products. Commonly cited examples include the inherent risks associated with equity, currency, commodity, and interest rate fluctuations. The Basel II Capital Accord encompasses three methods for credit risk modeling, namely the standard approach, the foundation internal ratings based approach, and the advanced internal ratings based approach.

The assessment and examination of all quantitative models constructed under Pillar 1 is mandated to be performed by supervisory authorities and this topic is deliberated within the framework of the second pillar. The proposed activities for implementation entail the incorporation of robust procedures aimed at appraising risks, including the internal capital adequacy assessment process (ICAAP), and regulatory oversight. In due course, upon the sanctioning of all quantitative risk models, market disclosure of the same stands permissible. The subject matter in question falls under the purview of Pillar 3. The bank shall periodically reveal its risk profile, encompassing both qualitative and quantitative information about its risk management processes and strategies to the financial market. The aim of this communication is to convey pertinent information to potential investors and furnish them with cogent reasons to believe that the banking institution in question has implemented a robust and sustainable risk management plan. Consequently, it is anticipated that such proficiency will lead to a positive rating, thereby facilitating the attainment of funds at more advantageous rates.

2.5.3 Basel III

The introduction of the Basel III Capital Accord was a direct consequence of the Global Financial Crisis (GFC). The present initiative endeavors to augment the foundational principles laid down in the Basel II Accord by reinforcing worldwide standards of capital. A crucial aspect deserving of particular consideration is a heightened emphasis on tangible equity capital, as it possesses the most substantial capacity for absorbing losses. The implementation of this measure results in a decreased dependence on proprietary models generated by the bank itself and evaluations procured from external rating agencies. Furthermore, this approach emphasizes stress testing to a greater extent than others.

The Basel III Accord incorporates a liquidity coverage and net stable funding ratio as measures to fulfill liquidity prerequisites. Its framework, was first introduced on January 1st, 2013, taken its final form on January 2019. In contrast to Basel II guidelines, Basel III Accord exhibits minimal influence on the credit risk models and introduces supplementary capital buffers.

As per the Basel II Capital Accord, the Tier 1 Capital Ratio was determined to be 4% of the Risk-Weighted Assets (RWA). The percentage rate was raised to 6% in adherence to the updated regulatory framework of Basel III. The Tier 1 capital ratio, a commonly used measure of bank solvency, is comprised of common equity, namely common stock and retained earnings, and excludes preferred stock. In the Basel II regulatory framework, the aforementioned ratio was 2% of the risk-weighted assets. However, in Basel III, this ratio has been elevated to 4.5% of risk-weighted assets. A novel capital conservation buffer has been implemented, which mandates the allocation of 2.5% of the risk-weighted assets to be secured by common equity. Furthermore, an augmenting countercyclical capital buffer has been introduced, which spans over a continuum of 0% to 2.5% of the assets weighed by risk.

In conclusion, the assessment of **credit risk** can be modeled in accordance with the Capital Accords established by **Basel II** and **III**. This can be achieved through the implementation of three distinct approaches, namely the standardized approach, the foundation internal ratings based approach, and the advanced internal ratings based approach. The aforesaid strategies exhibit variation with regard to their refinement and degree of adaptability in connection with the utilization of risk numbers that have been estimated internally.

2.6 Credit Scoring

Credit scoring is a set of decision models and underlying techniques that help lenders provide credit to consumers. These techniques determine who gets credit, how much credit needs to be taken, and operational strategies that improve borrower profitability relative to lenders.

Credit scoring techniques evaluate the risk of lending to specific consumers. Furthermore, credit scoring is occasionally used as a measure for consumer's creditworthiness. Creditworthiness is not a physical characteristic like height, weight, or money. It is a lender's appraisal of a borrower that reflects both parties conditions as well as the lender's projection of possible future economic scenarios. As a result, some lenders will consider an individual to be creditworthy while others will not. One of the longer-term risks of credit scoring is that this will no longer be the case, and there will be those who can receive credit from all lenders and those who cannot. Since it is somewhat offensive to label someone as ”untrustworthy” it is preferable for the lender to convey the truth, that is financing to this consumer constitutes a risk that the lender is unwilling to incur.

Thus, lenders have to make two kinds of decisions. First, whether to extend credit to new applicants and second, how to treat existing applicants, including whether to increase credit limits. Techniques that support the first type of decision are called credit scoring, while techniques that support the second type, are called behavioral scoring.

Regardless of the technique used, the key point in both cases is that a very large sample of previous customers is available with application details and resulting credit histories. All techniques use samples to identify the association between consumer characteristics and the ”good” or ”bad” of the broader story.

Numerous techniques result in the construction of a scorecard, wherein the attributes are assigned respective scores, and the cumulative summation of these scores serves as an indicator of the extent to which the risk associated with a consumer's likelihood to default on payments may be deemed unacceptable. Furthermore, alternative methods exist which do not entail the utilization of scorecards. These methods serve to ascertain the probability that a given consumer displays positive characteristics, thereby permitting a determination as to whether accepting the account is a advisable course of action. Despite the absence of a scorecard, these methodologies are commonly referred to as credit and behavioral scoring approaches.

Apart from the utilization of scoring in credit applications, it has been applied in various other contexts throughout the past decade. Conspicuously, it is demonstrating great efficacy in directing clientele through means such as direct mailing and various other marketing approaches. The emergence of data warehousing has enabled numerous organizations, particularly those within the financial and retail industries, to acquire the necessary data to facilitate the utilization of scoring techniques.

Various approaches may be utilized to accomplish the desired outcome. These methodologies or methods may be referred to as techniques, and are often informed by research and experimentation to determine their efficacy. Implementation of such techniques typically involves employing a systematic and structured approach, with careful consideration given to factors such as feasibility, applicability, and potential unforeseen consequences. The use of techniques is a fundamental aspect of many academic disciplines, including the sciences, engineering, and social sciences. Likewise, data mining, a highly publicized advancement in the realm of information systems, has demonstrated remarkable efficacy in the domain of response classification. Essentially, the act of scoring is being applied within a distinct contextual framework.

But how did the concept of credit scoring come about?

From the dawn of human communication, it is evident that the exchange of borrowing and repayment practices ensued. The earliest documented example of credit originates from the civilization of ancient Babylon. During the era of the Greek and Roman empires, significant progress was made in the development of banking and credit institutions. However, it is plausible to assert that the promotion of a credit card that offers XVIII.IX% APR may have been encountered with difficulties. Subsequent to the turn of the first millennium, commonly referred to as the ”Dark Ages” in European history, scant progression in matters pertaining to credit was observed. However, with the advent of the Crusades in the 13th century, the emergence of pawn shops transpired. At the outset, these were philanthropic organizations that imposed no interest; however, merchants expeditiously perceived the potential of such establishments, thereby ushering the emergence of commercial pawn shops that levied interest by the year 1350.

This can be observed in various regions across Europe. Pawn shops, which provide loans on a wide range of items placed as collateral, and are symbolized by the three-ball sign, are still prevalent in various European and South American nations. Throughout the Middle Ages, a longstanding discourse concerning the ethical implications of levying interest on loans persisted, an issue that persists in contemporary Islamic nations.

The European debate yielded a verdict which declared that lenders were within their rights to impose minor fees as interest payments, a convention that was deemed legitimate. Conversely, collections of exorbitant charges were labeled as practices of usury, which were deemed unacceptable within the sphere of financial and economic transactions. During this era, monarchs and influential leaders were compelled to engage in borrowing in order to facilitate the financing of their military campaigns and other expenditures. The act of lending at this particular level was associated more with the realm of politics rather than business, and the potential negative ramifications resulting from substandard lending practices could prove to be substantial for those in impoverished circumstances.

During the 19th century, the emergence of the middle classes prompted the establishment of various private financial institutions that showed willingness to extend bank overdrafts for the purpose of financing commercial ventures and daily expenditures. Nonetheless, the inception of consumer credit was limited to a nominal segment of the populace.

During the latter half of the 20th century, there has been a substantial surge in consumer lending. The sector of consumer credit has experienced a remarkable escalation in growth, surpassing that of numerous other business sectors. The emergence of credit cards during the 60s serves as a conspicuous indication of the expansion of financial systems, and currently, it is arduous to operate within societal norms without one. In numerous instances, procurement transactions primarily require the utilization of a credit card, particularly in scenarios such as online or telephonic purchases. It is crucial to note that the utilization of credit cards constitutes a relatively insignificant proportion of consumer credit, amounting to less than 15%. A significantly greater magnitude of debt is incurred through the acquisition of personal loans, hire purchases, overdrafts, and especially mortgage loans.

Although the concept of **credit** has a legacy as far back as five millennia, **credit scoring** only emerged quite recently, a mere half-century ago. The employment of credit scoring is fundamentally a means of distinguishing discrete clusters within a populace in circumstances where the defining attributes of such clusters are not perceptible, but rather display a correlation with related attributes. The initial method for addressing the challenge of identifying clusters within a populace was initially proposed within the field of statistics by Fisher in 1936 [44]. The researcher endeavored to distinguish between two distinct cultivars of iris based on their physical dimensions, as well as differentiate between the provenances of skulls utilizing physical metrics. In 1941, Durand [40] made a pioneering observation regarding the potential application of identical techniques for distinguishing favorable and unfavorable loans. This study constituted a research endeavor conducted under the auspices of the United States. The dataset in question was sourced from the reputable National Bureau of Economic Research and was deemed unsuitable for implementation in a predictive capacity.

At the onset of World War II, financial institutions and mail-order enterprises encountered challenges in managing credit operations. Credit analysts were being conscripted into military service, leading to a significant dearth of individuals possessing such specialized knowledge and skills. Thus, the corporations directed the analysts to document the heuristics that guided their loan allocation decisions [66]. There existed a number of scoring systems of a numerical nature that had been previously introduced, alongside additional collections of prerequisites that required fulfillment.

Subsequently, individuals lacking expertise in the field of credit assessment applied these regulations to facilitate their credit decisions. Following the interuption due to the war, quickly emerged an inclination among some individuals to link credit decision automation with the classification procedures being developed in statistics, recognizing the advantages of employing statistically derived models in lending determinations [120]. During the early 50s, San Francisco witnessed the establishment of the pioneering consultancy company by Bill Fair and Earl Isaac. The aforementioned consultancy primarily catered to clients hailing from the finance, retail, and mail-order industries.

During the late 60s, the advent of credit cards prompted banks and other credit card issuers to recognize the significance of credit scoring as a crucial tool. Due to the significant volume of credit card applications received on a daily basis, it was both practically and financially unfeasible to undertake a manual lending decision process, thus necessitating the automation of said process. The increase in computational capacity facilitated the achievement of this objective.

These entities have discovered that credit scoring is a significantly more effective forecaster in comparison to all other judgement-based models, resulting in a reduction of default rates by 50% or more. The work of Myers and Forgy [90] provides an early account of this accomplishment, whilst the research conducted by Churchill et al. further highlights this phenomenon.

During the 80s, credit scoring rose to prominence in the credit card industry, prompting financial institutions to apply this methodology to additional financial products, namely personal loans. Later, the utilization of scoring models has extended to the realm of home loans and small business loans. During the 90s, the expansion of direct marketing prompted the utilization of scorecards as a means to enhance the efficacy of advertising campaigns by increasing the rate of response. In point of fact, scoring was employed in the 50s as one of the earliest applications, whereby Sears implemented it as a mechanism to make informed decisions on the recipients of its catalogues [79]. Significant advancements in computer technology have facilitated the exploration of alternative methods for constructing scorecards. During the 80s, logistic regression and linear programming, which are presently considered to be foundational pillars of card building, were introduced. In contemporary times, there has been a surge in experimentation with artificial intelligence methodologies, such as expert systems and neural networks.

Currently, there is a shift in focus from the objective of reducing the likelihood of customer default on a specific product towards exploring ways in which the organization can optimize its revenue generation from said customer. Additionally, the initial concept of risk assessment regarding loan default has been expanded to encompass the use of scorecards, which facilitate the evaluation of an individual's likelihood of responding to a direct mailing promoting a new product, utilizing a particular product, maintaining product usage beyond the introductory offer period, switching to an alternative lender, handling debt in the event of delinquency, and the potential for fraudulent activity in connection with a loan application.

2.7 Aspects of Credit Scoring

The fundamental goals of credit scoring are practicality and experimentation. The point of credit scoring and behavioral scoring is to anticipate hazard, not to clarify it. Over the past five decades, considerable emphasis has been placed on predicting the potential for a borrower to default on their credit transactions. The credit scoring system maintains a high level of quality due to its sound strategy and reliance on scientifically derived information.

As a result, credit scoring systems are formulated based on the historical conduct of individuals with characteristics akin to those individuals who will be subjected to such scoring systems. In order to obtain reliable data on the subsequent performance history of customers, it is customary to seek a sample of previous customers who recently applied for the product. If it is not feasible to construct a system based on comprehensive data due to the novelty of the product or limited customer usage, alternative approaches may involve developing systems on smaller or similar product samples. However, it should be noted that the predictive capability of such systems will not be as effective as those built on historical data from previous customers specific to that product.

The pragmatic and empirical nature of credit scoring necessitates the inclusion of all relevant factors pertaining to the customer or consumer and their environment in order to improve predictive accuracy within the scoring system. The variables employed, exhibit a discernible correlation with default risk. Several factors are used to assess consumer stability and financial status. These include the duration of the consumer's residence and employment, possession of current or checking accounts and credit cards, length of time with their current bank, and the consumer's available resources such as residential status and employment, as well as the employment of their spouse. Additionally, the number of children and dependents can also provide insights into the consumer's potential outgoings. However, it is not obligatory to provide justification for any given variable. The application of this technique is recommended if it contributes to enhancing predictive capabilities.

Furthermore, more emphasis should be placed on the illegality of incorporating certain demographical attributes, such as race, religion, and gender, into credit scoring mechanisms. Several studies [30] have demonstrated that the inclusion of gender identification leads to a higher number of credit allocations for women compared to present circumstances. The aforementioned phenomenon can be attributed to the fact that certain variables such as low income and part-time employment, have been found to serve as predictors for favorable repayment patterns among the female demographic, while yielding unfavorable outcomes within the broader population. The utilization of gender has been deemed discriminatory against women by legislators, thus rendering it impermissible.

Although not prohibited by law, certain characteristics are not employed in determining default risk as they are considered culturally inappropriate. The presence of a subpar health history or a significant number of driving violations have been identified as potential indicators of higher default risk. However, financial institutions refrain from utilizing this information due to apprehension regarding societal critique. The act of verifying if a consumer has secured insurance coverage for credit card debts in case of job loss is employed by specific loan providers and has been observed to exhibit a constructive correlation with the probability of default. The selection of characteristics is a highly subjective matter. The early 80s witnessed a fervent discourse regarding the ethical implications surrounding credit scoring. Specifically, proponents of this practice, such as Nevin and Churchill [92] and Saunders [104], espoused their support for credit scoring's use. Conversely, Capon [28] was among those who criticized the underlying philosophy and operationalization of credit scoring, arguing that it was inferior to subjective judgmental systems that relied on the opinions of credit analysts and underwriters.

Detractors of credit scoring questioned the underlying philosophy and rigor of its methodology. The critique was leveled against the absence of any elucidation regarding the relationships between the salient attributes identified by the study and the ensuing credit performance. It is noteworthy that contemporary advancements in credit scoring, such as graphical modeling, are endeavoring to simulate such interrelated sequences. The statistical methodology's soundness was subject to criticism due to the potential bias in the sample utilized, as it failed to account for those who had been previously excluded. The adequacy of sample size and the issue of overriding system decisions were both subject to scrutiny. Additional concerns that were raised included the issue of collinearity amongst the variables, as well as the introduction of discontinuities that comes with the use of course classification on continuous variables. Eisenbeis [42] also acknowledged the existence of these criticisms. The credit-scoring industry has extensively acknowledged these criticisms and has either developed methods to mitigate their shortcomings or incorporates them into its decision-making practices.

2.8 Credit Scoring in Real Life

Credit scoring provides mutual benefits to both lenders and customers. From the standpoint of the financial institution, the utilization of credit scores is a valuable tool for appraising prospective clientele and establishing an appropriate credit cap. This practice enables financial institutions to mitigate the likelihood of encountering credit risk. The application of credit scoring represents a more expeditious means by which to evaluate the reliability of prospective borrowers, when contrasted with conventional practices that tend to involve protracted procedures. From a client's perspective, the continuous enhancement of credit score and expansion of credit limit can be achieved. Credit scoring has the potential to mitigate unjustifiable credit risk to both the creditor and debtor.

One of the primary benefits of credit scoring is the expeditious evaluation of each individual client. Moreover, the utilization of an automated system results in substantial cost savings for the lenders. The process of applying for credit can be simplified for customers by providing solely the information that is utilized in the scoring system. Moreover, this practice enables lenders to uniformly apply the same set of criteria for credit assessment to all potential borrowers, without any consideration of their demographic or social characteristics such as gender, race, or other pertinent factors. Hence, this procedure exhibits greater objectivity towards clientele while precluding the manifestation of discriminatory practices.

2.9 Types of Credit Scoring

A plethora of credit score models are presently utilized, and each possesses distinct attributes:

- The **FICO score**, developed by the Fair Isaac Corporation, is presently regarded as the preeminent credit scoring model in circulation. The score employs a scale ranging from 300 to 850 points. Although, the companies that disseminate credit scores to their clients are Experian, TransUnion, and Equifax, the direct provision of FICO scores to clients is unavailable. The credit agencies are responsible for maintaining the credit history and files of their clients. The determination of credit score is contingent upon the data existing in the customer's account at the specific moment of evaluation.

- The **PLUS Score** is a credit scoring model, created by Experian, that aims to enhance a layperson's comprehension of their creditworthiness. Ranging from 330 to 830, this system provides customers with scores reflecting how creditors appraise their ability to pay debts. Increased scores are indicative of an augmented probability of customers repaying their debts, thereby being perceived by lenders as possessing diminished credit risk. Over the course of time, there is a likelihood that alterations may occur to the client's details. Moreover, their credit score may exhibit variability over time.

- The **Vantage Score**, a new credit scoring model developed jointly by Experian, TransUnion, and Equifax, serves to promote uniformity and precision in credit scoring methodologies. The aforementioned score offers to lenders a comparable evaluation of risk across the three major credit reporting agencies. The Vantage scoring system encompasses a spectrum of values ranging from 501 to 990.

To summarize, irrespective of the scoring systems adopted by financial institutions, maintaining a favorable credit score is advantageous for individuals since, a higher score facilitates approval for lower interest rates on lending products.

2.10 Credit Scoring Applications

The utilization of credit scoring methodologies and techniques in providing financial services and assessing creditworthiness of individuals and businesses can be referred to as credit scoring applications.

The concept of credit scoring, along with complementary concepts of behavioural and profit scoring, has gained significant attention from scholars and practitioners in the fields of finance and risk management. The utilization of credit scoring techniques has become increasingly popular in the assessment of borrower creditworthiness and in the development of predictive models for credit risk management. Alongside credit scoring, behavioural and profit scoring techniques have also emerged as valuable tools for analyzing and predicting consumer behavior and profitability. Specifically, the key principles behind credit scoring and its related scoring methodologies are discussed with a focus on the use of statistical models and data analysis techniques. Additionally, the potential applications and limitations of these scoring techniques are explored, and suggestions for future research in this field are provided. The works of Fritz and Hosemann [49], Sarlija et al. [103], and Banasik and Crook [19] pertain to relatively contemporary subjects, when juxtaposed against established business constructs such as credit. Notwithstanding, the utilization of credit scoring has been extensively employed across various domains, encompassing a juxtaposition of diverse statistical methodologies implemented for predictive and classificatory objectives. The categorization of various applications is evidenced in scholarly literature. Altman et al. [7], and Landajo et al. [71]

contend that applications can be categorized into accounting and finance. The work of Kumar et al. [69], and Chiang et al. [33], identify marketing as another classification. Smith and Mason [106] and Dvir et al. [41] contend that applications can be classified into engineering and manufacturing. Furthermore, Warner and Misra [118] and Behrman et al. [20] classify some applications under health and medicine. Finally, Hardgrave et al. [57] and Nikolopoulos et al. [93] identify general applications.

Undoubtedly, in the realm of accounting and finance, the utilization of credit scoring applications has been employed for a variety of aims, most notably in light of the accelerated expansion experienced by this domain. Over the past few decades, there has been a sharp increase in the number of applications of predictive modeling, encompassing several domains including: bankruptcy prediction [117], bankruptcy classification ([88] and [76]), scoring applications [36], as well as classification problems ([94], [114] and [22]). Other applications have also emerged, such as financial distress [62], financial decisions, and financial returns ([121] and [124]).

Nevertheless, banking institutions have exhibited significant growth in the utilization of credit scoring applications over the past several decades ([56] and [19]). This proliferation can be attributed to the surge in credit applications for several banking products, which has facilitated the emergence of a plethora of novel product channels that these banks can capitalize on. Various banking applications have been developed, of which consumer loans are deemed to be crucial and widely used ([95], [83] and [73]). Credit-card-scoring applications, which were implemented early on in the banking sector, are also noteworthy ([53] and [72]). Additionally, small businesses have emerged as an important factor in the banking industry ([110] and [116]). Mortgages are another variety of banking products that have gained significant popularity amongst lenders recently [107].

3 Statistical Techniques for Credit Scoring

3.1 Data Mining in Credit Scoring

Upon examination of the primary methodologies utilized in the field of data mining, it is evident that their efficacy in credit scoring applications serves as no surprise. Jost [67] recognized this observation in his scrutiny of data mining. The fundamental techniques employed in the field of data mining encompass data summary, variable reduction, observation clustering, prediction and explanation. The conventional technique of data analysis involves the utilization of standard descriptive statistics, namely frequency analyses, measures of central tendency such as means, dispersal statistics such as variances, and the creation of cross-tabulations to facilitate a comprehensive compilation of data summation. Moreover, it can be helpful to classify continuous variables by binning them into distinct categories. The aforementioned approach of coarse classification has been demonstrated to be highly beneficial in the realm of credit scoring. Identifying the most crucial variables with the aim of diminishing the quantity of variables to be taken into account is a common practice in numerous statistical applications. The techniques employed in credit scoring have exhibited efficacy in other fields of data-mining application as well. One of the data-mining tools is the segmentation of customers into distinct groups based on their purchasing behavior and preferences, which facilitates targeted marketing efforts and product offerings. Credit scoring is a process that entails grouping consumers into distinct clusters based on their behavior, and then developing discrete scorecards for every group.

According to Jost [67], explanation analysis represents a critical function of data mining. However, Jost also argued that achieving satisfactory explanations through this process is rare. Instead, the approach of data mining employs the method of segmentation analysis for the purpose of identifying the most probable segment that would manifest a certain kind of behavior. The present context presents persistent philosophical challenges akin to those expounded by Capon [28] concerning credit scoring, with the added complexity of explicating the behavior of the aforementioned segment. The intricacies of human behavior are multifaceted, posing a challenge to identifying clear causations. A more viable approach entails elucidating the relationships between various customer segments and their corresponding purchase and repayment patterns. It is advised to refrain from endeavoring to establish a causal relationship between the aforementioned entities.

Data mining is characterized by its utilization of credit scoring techniques and methodologies, which serve as its fundamental basis, albeit applied in a broader context. It is imperative for proponents and practitioners of the data mining technique to examine the accomplishments and progressions of credit scoring, as it serve as a valuable reference point to elude the drawbacks and embrace the concepts that exhibit effectiveness in similar areas of application.

3.2 Statistical Techniques

A diverse array of statistical methodologies is employed in constructing scoring models. The majority of statistical models, including some nonlinear options, possess the capability to construct a proficient credit scoring system for anticipatory purposes with regard to both its efficiency and effectiveness. Various analytical methodologies, including the weight-of-evidence approach, have been utilized in practical applications.

The techniques commonly utilized in constructing credit scoring models by credit analysts, researchers, lenders, and computer software developers and providers include mainly measure, regression analysis, discriminant analysis, probit analysis, logistic regression, linear programming, Cox's proportional hazard model, support vector machines, decision trees, neural networks, K-Nearest Neighbour (K-NN), genetic algorithms, and genetic programming.

The contrast between advanced statistical methods and traditional statistical methods though, is a common topic of discussion in scientific community.

Indeed, ***advanced statistical methodologies***, including neural networks and genetic programming, propose a distinct approach to ***traditional statistical methodologies***, namely discriminant analysis, probit analysis and logistic regression. The purpose of employing advanced methods, such as neural networks, resides in their capacity to model intricate functions, which diverges from classical linear techniques, like linear regression and linear discriminant analysis. Probabilistic neural networks have been shown to exhibit faster training times when presented with input cases compared to multilayer feed-forward neural networks, while also demonstrating equivalent or superior classification accuracy.

Although multilayer feed-forward nets have demonstrated exceptional classification abilities in previous studies [98], it is noteworthy that alternative approaches have exhibited superior performance. On the contrary, a diverse array of advanced algorithms have been developed for training neural networks, rendering them a compelling alternative to traditional methods.

Various methodologies have become accessible as corroborated by Masters [84] and Palisade Corporation [98]. In contemporary times, genetic programming has emerged as one of the foremost efficacious substitutes for conventional methods within the respective domain. The utilization of genetic programming enables the automatic determination of both the adequate discriminant functions and the relevant features in a concurrent manner. It has been suggested that while dissimilar neural networks may solely be suited for extensive datasets, genetic programming may exhibit favorable performance even when confronted with limited datasets [91]. It is pertinent to engage in a discourse regarding the credit scoring modeling approaches that have been previously alluded to.

3.2.1 Linear Regression

Linear regression techniques have become a crucial constituent of data analysis endeavors that aim to explicate the interdependence between a dependent variable and one or multiple independent variables. The formula for simple linear regression can be expressed as:

$$Y = a + bX + \varepsilon,$$

where Y represents the response (dependent) variable, X is the predictor (independent) variable, a represents the intercept (the value of Y when X = 0), b denotes the slope, and finally ε is the regression error.

Linear regression has been employed in the context of credit scoring, given that the two-class problem may be effectively represented through the utilization of a dummy variable. An alternative approach to handle situations where customers make partial repayments of varying amounts would be to employ a Poisson regression model. Consequently, the proportional repayments possess the potential to be reformulated as Poisson counts. Credit analysts have the capability to closely examine factors including customer's payment patterns, guarantees, and punctual default rates utilizing linear

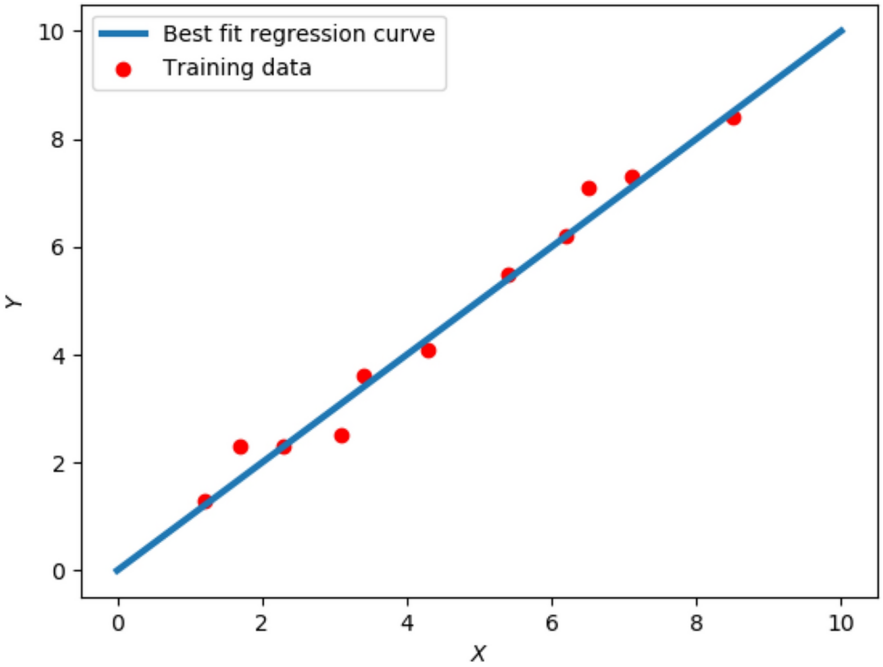


Figure 2: Simple Linear Regression

regression, to ultimately establish a score for each aforementioned factor. This score is then side-by-side against the bank's designated cut-off score as a means of comparison. If a prospective customer attains a score surpassing the bank's creditworthiness assessment criterion, credit shall be extended to the individual.

In 1970, Orgler [95] conducted an analysis in which regression techniques were employed to examine commercial loans. The outcome of this analysis was restricted to the assessment of extant loans, making it a viable tool for the scrutiny and review of loans. Subsequently, Orgler [95] employed a regression methodology to assess consumer loans that remained unpaid. The researcher arrived at a conclusion suggesting that undisclosed information possessed more significant prognostic potential than the information already mentioned in the initial application, while evaluating the potential for future loan quality. The application of regression analysis has been expanded in various fields and several scholars ([58], [55] and [113]) have conducted research in this area.

3.2.2 Binary Logistic Regression

Binary logistic regression, stands as a commonly employed statistical method within the discipline. One distinct characteristic that sets a logistic regression model apart from a linear regression model, is that the dependent variable in logistic regression is characterized by dichotomous outcomes, with only two possible values of ”0” or ”1”. The disparity between logistic and linear regression is manifested in both the selection of a parametric model and the underlying assumptions. Upon consideration of this distinction, the techniques utilized in a logistic regression analysis adhere to the analogous fundamental principles as those employed in a linear regression analysis [61].

The binary logistic regression model may be readily expanded to encompass two or more independent variables. Undoubtedly, the degree of difficulty in acquiring multiple observations at all levels of diverse variables increases proportionally with the number of variables. Consequently, the majority of logistic regression analyses featuring more than one independent variable are executed utilizing the maximum likelihood approach [48]. From a theoretical perspective, logistic regression would appear to be a more appropriate statistical method in comparison to linear regression, specifically in light of the two categorizations of credit as either ”good” or ”bad”, as described by Hand and Henley [55]. The application of logistic regression in credit scoring has been widely adopted and extensively studied from Lenard et al. [75]; Desai et al. [37]; Lee and Jung [74]; Baesens et al. [12]; Crook et al. [36].

Linear regression is able to investigate the probable relationship between one (or more) independent variables X and a continuous dependent variable Y. There are, however, several scenarios in which the effect of one (or more) independent variables on a discrete dependent variable needs to be investigated. If the discrete dependent variable has two possible outcomes (say, 1: ”success” and 0: ”failure”), binary logistic regression, which is one of the most popular special cases of Generalized Linear Models, is the most appropriate technique for studying the relationship between the independent(s) and the dependent variable. Some common examples of bivariate outcomes include: ethnicity (”Greek” or ”Not Greek”), gender (”Male” or ”Female”), loan request (”Accepted” or ”Not Accepted”), and so on.

Let us, consider Bernoulli random variable Y so that:

$$Y = \mathbb{I}^{(Y)}_{\text{"success"}} = \begin{cases} 1, & \text{"success"} \\ 0, & \text{"failure"} \end{cases},$$

with

$$E(Y) = p, \;\; Var(Y) = p(1-p), \;\; p \epsilon (0,1),$$

where p is the probability of success.

Furthermore, consider the multiple linear model with:

$$E(\underset{\sim}{y}) = p_i = X_i' \underset{\sim}{\beta} = \beta_0 + \sum_{i=1}^{p-1} \beta_i X_i. \tag{1}$$

The relationship implied in the two hands of equation (1), raises a critical issue. In particular, the right hand ranges in $\mathbb{R}$, whereas the left hand, as a probability, ranges in $(0,1)$. In order to have the two hands share the same range, a "cunning" solution must be enlisted. First, an analogous for the probability of success p has to be created. Let us divide the probability of success p by the probability of failure q. The outcome:

$$\frac{p}{q} = \frac{p}{1-p},$$

is called odds and is similar to the concept of p. The main difference of odds is that it ranges in $(0,+\infty)$.

The first step to correct the problem of different ranges is achieved, and hence:

$$\frac{p_i}{1-p_i} = X_i' \underset{\sim}{\beta}. \tag{2}$$

The final step is to bring the range of the left handside from $(0,+\infty)$ to $\mathbb{R}$. This can be achieved by using the logarithm as follows:

$$log\left(\frac{p_i}{1-p_i}\right) = X_i' \underset{\sim}{\beta} = g(\mu_i).$$

The quantity:

$$log\left(\frac{p_i}{1-p_i}\right),$$

is called the logit - transformation and constitutes the link function of the mean p with the independent variables of the model.

To determine the probability of success, all you need to do is solve for p_i:

$$E(\underset{\sim}{y}) = p_i = \frac{exp(X_i'\underset{\sim}{\beta})}{1+exp(X_i'\underset{\sim}{\beta})}.$$

The model parameters will be estimated using the maximum likelihood method. After the identification of the estimators $\underset{\sim}{\hat{\beta}}$ of $\underset{\sim}{\beta}$, they should be interpreted with caution since the latter differs from the typical linear regression model.

- $\hat{\beta}_0$: represents the likelihood of ”success” occurring when all independent variables are equal to 0.
- $\hat{\beta}_i$, $i = 1,...,p-1$: The independent variables X_i can be numerical or categorical. Depending on their type, but also on the sign of the coefficient $\hat{\beta}_i$, the corresponding interpretation is also given.

3.2.3 Linear vs Logistic Regression

In data science, linear and logistic regression are two of the most widely used models, and open-source programs like Python and R make the fit of such models rather cheap and simple.

Linear regression models are commonly employed in analyzing the relationship between a continuous dependent variable and one or more independent variables. This statistical approach aims to ascertain the connection between these variables. Simple linear regression is characterized by the presence of one or two independent variables alongside one dependent variable. Conversely, multiple linear regression arises from the inclusion of additional independent variables in the analysis. The objective of each

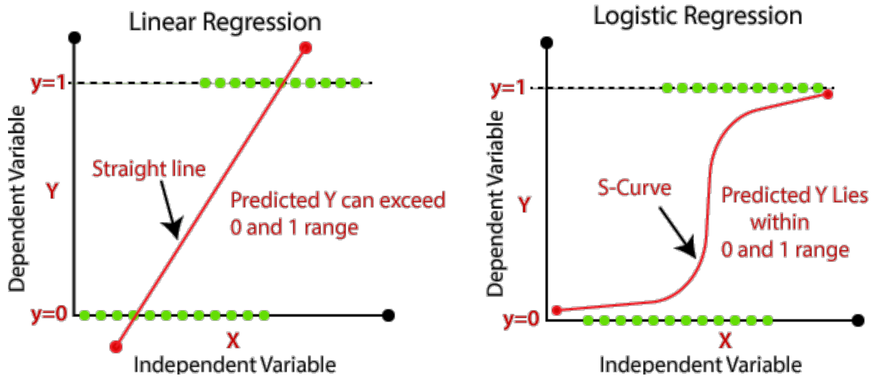


Figure 3: Linear vs Logistic Regression

sort of linear regression is to determine the optimal line of best fit (Figure 3, left image) for a given dataset, typically achieved by employing the least squares methodology.

Logistic regression, similar to linear regression, is utilized to evaluate the relationship between a dependent variable and one or more independent factors; however, it provides predictions for a categorical variable instead of a continuous one. Binary categorical variables, such as ”True or False,” ”Yes or No,” and ”1 or 0,” exemplify distinct cases within a categorical variable framework. The logit function transforms the sigmoidal curve into a linear form (Figure 3, right image). This contrast with linear regression, which yields a probability-based outcome.

Regression analysis utilizes both models to make predictions about future events; however, linear regression is commonly preferred due to its comparatively simpler comprehensibility. Linear regression does not require a sample size as big as that needed for logistic regression in order to accurately capture the values across all response categories. The statistical power of the model may be insufficient to effectively ascertain a significant effect unless a larger, more representative sample size is utilized.

3.2.4 Discriminant Analysis

The method of **Discriminant analysis**, a straightforward parametric statistical technique, has been designed to differentiate and distinguish between two distinct groups. The discriminant approach has garnered significant consensus among researchers as a widely established method for the classification of customers into either good or bad credit categories. This methodology has been extensively employed in diverse domains of credit scoring applications. In the realm of statistical analysis, a credit scoring model applying a discriminant approach serves as a fundamental method for segregating group variables into two or more distinct categories.

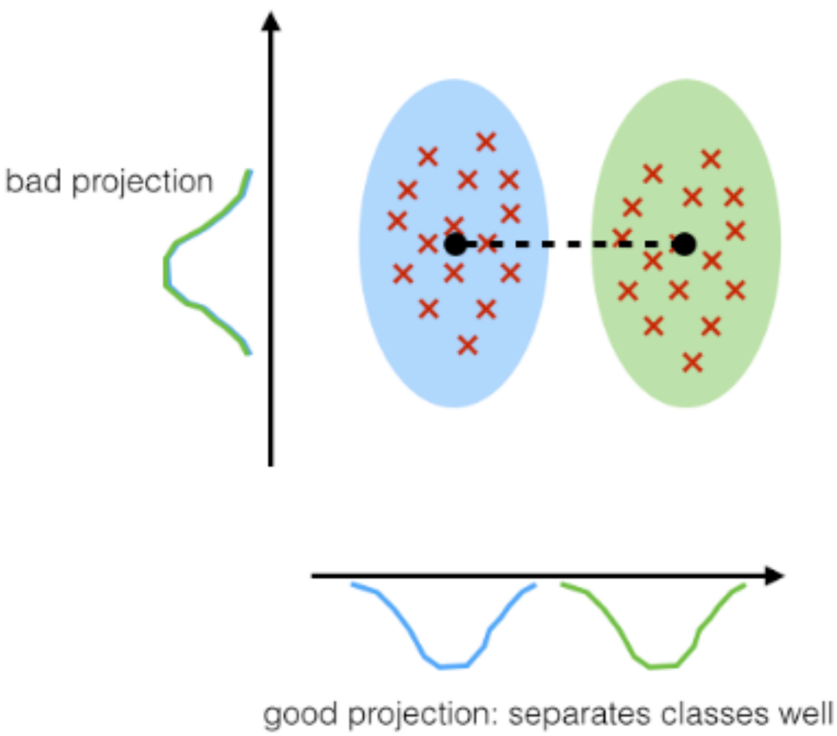


Figure 4: Discriminant analysis

Fisher [44] initially introduced discriminant analysis as a technique to facilitate discrimination and classification in academic research. Durand [40] conducted an early application of multiple discriminant analysis to credit scoring by examining car loan applications. An established technique frequently employed for corporate bankruptcy prediction involves the approach established by Altman [6], who conceived the initial empirical scoring model grounded upon a quintet of financial ratios extracted from eight distinct variables obtained from the financial statements of corporations. Moreover, he computed a Z-score which represents a weighted sum of the financial ratios using a linear combination methodology. Discriminant analysis has been deemed a legitimate technique utilized in the development of credit scoring models, as evidenced by studies conducted by Desai et al. [37] and Hand and Henley [55].

Nevertheless, a number of scholars have articulated objections against the utilization of discriminant analysis in the domain of credit assessment. Eisenbeis [42] identified several statistical challenges involved in the utilization of discriminant analysis, drawing upon his prior research conducted in 1977. Discriminant analysis requires careful consideration of a range of issues that may compromise its accuracy and reliability. These issues include the utilization of inappropriate linear functions rather than quadratic functions, inaccurate definition of groups, unsuitable prior probabilities, inadequate prediction of classification errors, among other potential problems. It is essential to address these issues to ensure the optimal accuracy and efficacy of the discriminant analysis methodology. Despite the aforementioned challenges, it is noteworthy that discriminant analysis remains ubiquitous in the realm of credit scoring, as indicated in prior research [53]

3.2.5 Probit Analysis

Probit analysis is a conventional statistical method that has been extensively employed in credit scoring applications over an extended period. Grablowsky and Talley [52] posited that the approach of probit analysis was initially utilized by Finney [43] in toxicology studies to examine the correlation between the level of poison administered and the probability of insect elimination. During the early 30s, a novel term known as ”probit” was coined, signifying probability unit. This terminology has been recorded by scholars such as Pindyck and Rubinfeld [100]. The Probit analysis methodology aims to determine the coefficients values that correspond to the probability of a dichotomous coefficient attaining a unit value. In the context of statistical analysis, the probit model involves converting a linear combination of independent variables to a cumulative probability value utilizing the normal distribution.

According to Grablowsky and Talley [52], the probit analysis assumes normal distributions of the ”threshold values”, while the discriminant analysis assumes multivariate normal distributions as well as equal variances. In addition, the likelihood ratio test can be utilized to test the significance of the estimates of coefficients under a probit function on an individual basis because of their ”uniqueness” factor. However, the situation differs when considering discriminant coefficients, as they are unable to undergo individual testing. This is in contrast to the possibilities presented by regression models, where both linear and logistic regression, as well as Poisson regression, permit such testing. It is worth noting that the probit function also allows for individual testing of coefficients, though this approach is considerably more complex than that employed by the aforementioned regression models. Multicollinearity has been identified as a factor that could result in erroneous coefficients signs in probit analysis. However, it has been observed that the probability values derived from the likelihood ratio tests remain unaffected. Finally, this issue is not deemed problematic in the context of discriminant analysis.

To sum up, probit analysis is an advanced type of regression analysis that is specifically utilized for the examination of binomial response variables. This type of variables only presents two possible responses, namely positive or negative outcomes. The aforementioned procedure facilitates the transformation of a response curve into a linear format, allowing for subsequent analysis via either least squares or maximum likelihood regression methods. Each detection probability (D_i) is computationally derived as follows:

$$D_i = \frac{n_{pos}}{n_{neg}},$$

where n_{pos} represents the count of replicates that are recorded as positive, while n_{neg} denotes the count of replicates that are recorded as negative. The current model is represented by an S-shaped or sigmoid curve, which correlates with the cumulative probability of a normal distribution. The plot illustrates the probability or hit rate against a standardized normal deviate, SD, is demonstrated in the graphical representation depicted in Figure 5.

3.2.6 Weight-of-Evidence Measure

The **weight-of-evidence measure** is a technique that utilized in credit scoring applications. Several studies have examined the efficacy of the weight-of-evidence measure in the respective field. However, the outcomes were

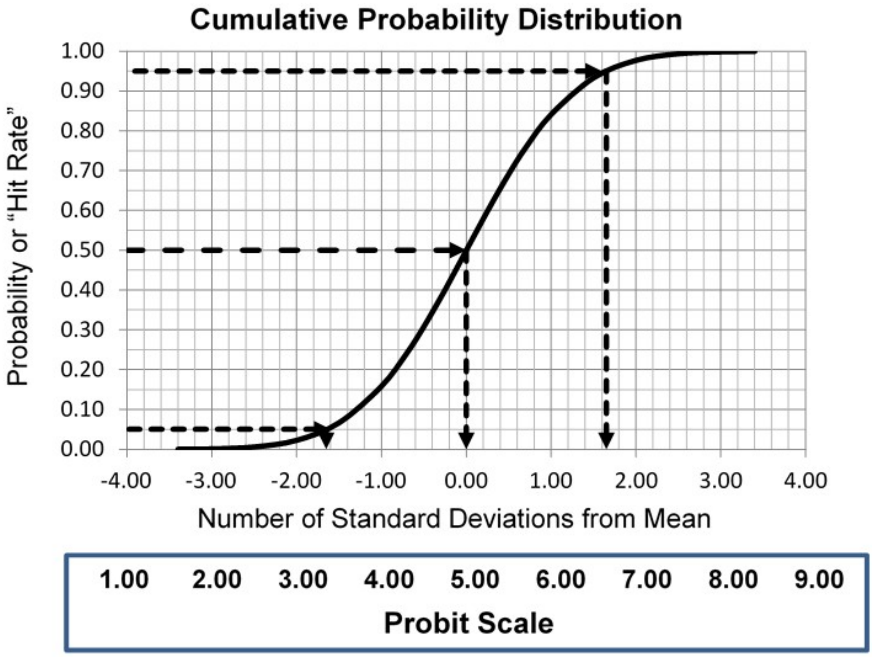


Figure 5: Probit Model

found to be similar to those obtained from other analytical techniques, as demonstrated in the works of Bailey [15], Banasik et al. [16], and Abdou [1]. The efficacy of applying probit analysis in statistical scoring has been subject to scrutiny in recent research. Indeed, prior studies conducted by Guillen and Artis [54], Greene [53], Banasik et al. [16], and Abdou [1] have investigated and compared the performance of probit analysis with other scoring models. The classification results obtained from probit analysis were found to be in close proximity to those of other techniques, as observed in Greene's [53] study. In contrast, probit analysis demonstrated superior predictive accuracy compared to techniques such as discriminant analysis, linear regression, and the Poisson model, as evidenced in Guillen and Artis's [54] research. Moreover, probit analysis has emerged as a proficient substitute for logistic regression.

The general mathematical expression of weight of evidence (WOE) is as follows:

$$WOE = log\left(\frac{\% of\ non - events}{\% of\ events}\right).$$

Especially, in the case of credit scoring, the quantification of the predictive efficacy of an independent variable in relation to the dependent variable is established through the weight of evidence. As it originates from the realm of credit scoring, it is commonly characterized as a metric for discerning the distinction between desirable and undesirable patrons. The term ”Bad customers” concerns those customers who have failed to meet the repayment obligations in the context of a loan agreement, while the term ”Good Customers” refers to those customers who have successfully fulfilled their financial obligations by repaying the loan. It is represented as:

$$WOE = log\left(\frac{Distribution of\ Goods}{Distribution of\ Bads}\right),$$

where:

- Distribution of Goods = % of Good Customers in a particular group,
- Distribution of Bads = % of Bad Customers in a particular group.

Consequently, a **Positive WOE** would imply that the Distribution of Goods is greater than Distribution of Bads and a **Negative WOE** would imply the exact opposite.

3.2.7 Decision Trees and Random Forest

Decision trees represent a widely utilized classification mechanism in the creation of credit scoring models, commonly referred to as recursive partitioning or classification and regression trees [55]. Breiman et al. [26] were among the pioneers who utilized a CART model, likely marking one of its earliest applications. Then, Rosenberg and Gleit [102] asserted that the initial decision tree-based model was instigated by Raiffa and Schlaifer [101] at the Harvard Business School. In addition, it was posited that a credit scoring model derived from decision trees was later developed by David Sparks in 1972 [109] at the University of Richmond. This nonparametric approach is known as a classification tree and serves to analyze categorical or dependent variables in accordance with continuous explanatory variables [9]. A classification tree is constructed through a dichotomous branching process that involves dividing data at each node according to a function that is determined by a singular input. The system evaluates all potential partitions to determine the optimal option, and the advantageous sub-tree is chosen on the basis of its comprehensive rate of error or minimal expenses incurred due to misidentification, as noted by Zekic-Susac et al. [125]. Also, Baesens et al. [11] elucidated further utilization of decision trees in the domain of credit scoring. Furthermore, Paleologo et al. [97] was conducted an analysis on credit requests submitted by corporate clients. This study sought to address the challenge of unbalanced data sets, and as part of the assessment, a subagging procedure was employed within a decision tree paradigm that incorporates extreme values to account for missing data.

Particularly, the decision tree is a data structure that follows a hierarchical approach and employs the divide-and-conquer technique. This nonparametric approach is proficient and applicable for classification and regression tasks. In modeling scenarios where the focus variable involves a categorical classification outcome, the decision tree algorithm utilized is commonly referred to as the classification tree. When making predictions for a particular instance, we usually employ a methodology whereby we allocate an observation within a designated area to the class that appears most frequently among training observations in that same area. This approach involves relying on the prevalence of classes in the training data to inform classification decisions for new observations. The completion of the aforementioned task may yield classification errors, specifically misclassification errors. The classification error rate is consequently calculated as the proportion of training observations within said region that are not attributable to the prevailing class. The quantification of misclassifications may be achieved through the application of the node impurity measure, denoted as Qm(T). This measure serves to evaluate the degree of homogeneity within the node's individual classes and is a quantitative indicator of the quality of a split. A split is considered to be pure when all branches resulting from the split exhibit homogeneity with respect to the classification of instances, whereby each instance selecting a specific branch of the split belongs to the same class.
As a conclusion, decision trees have the capability to perform classification and regression operations. The title inherently indicates that a flowchart with a tree-like structure is employed to depict the prognostications ensuing from a concatenation of attribute-based divisions. The process commences with a fundamental node and culminates with a resolution rendered by

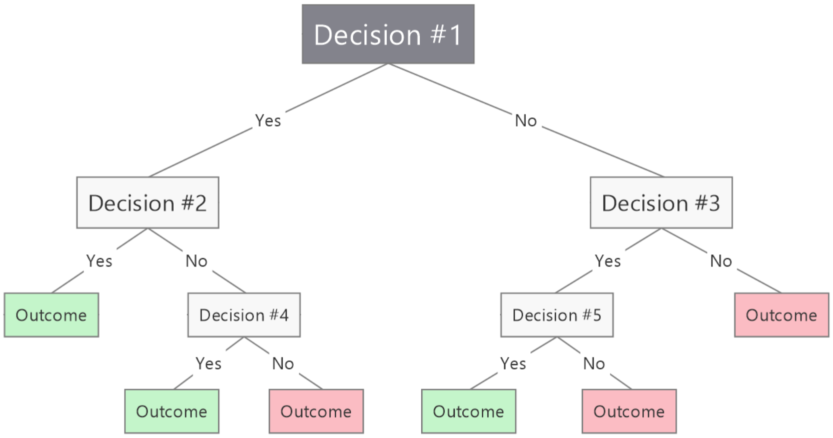


Figure 6: An example of a Decision Tree

terminal leaves.

Similar to decision trees, **random forests** represents a significant alteration of bagging (bootstrap aggregating) by constructing a diverse assortment of uncorrelated trees, which are subsequently combined via averaging. They provide a more streamlined and manageable training and tuning process. As a result, the implementation of random forests has led to notable outcomes in the field.
In the construction of a random forest, an ensemble of decision trees is generated through the creation of multiple decision trees on bootstrapped training data samples. Bootstrapping is an iterative resampling method which involves random sampling with replacement. For example, when evaluating potential splits in a tree structure, a sample of m predictors is randomly selected from the larger set of p predictors to serve as split candidates. Stated differently, the algorithm is precluded from taking into account all predictors, thereby precluding even a significant proportion of them during splitting. In the process of creating decision trees, a novel subset of individual predictors is regularly extracted at every split. The number of predictors within this set is often in the vicinity of the square root of the total number of available predictors, denoted by $m \approx \sqrt{p}$. The foregoing measure is instituted in order to preclude the division of numerous trees, particularly those commencing at the most robust predictor, resulting in trees that are correlated. In order to achieve a reduction in variance through bagging, it is imperative for random forests to exclusively take into account

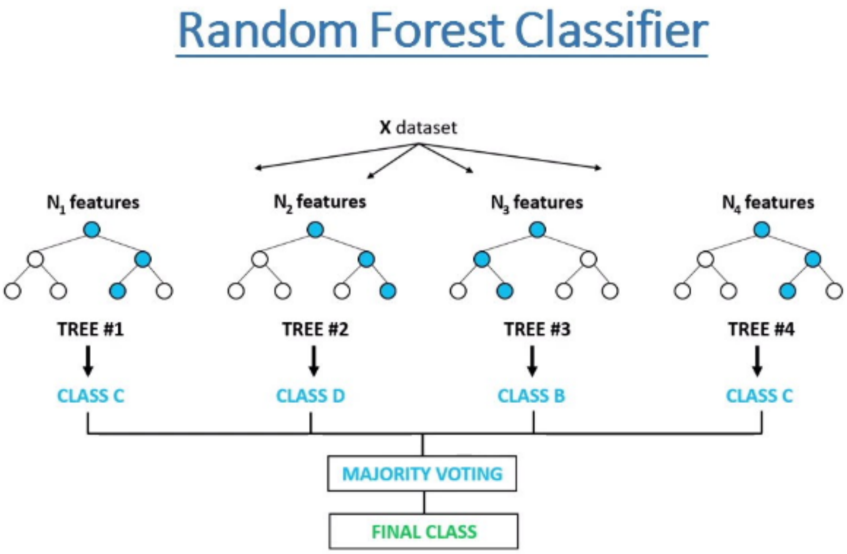


Figure 7: An example of a Random Forest

a subset of predictors.

3.2.8 k-Nearest-Neighbor

The **k-Nearest-Neighbor (kNN) method** represents a prevalent non-parametric technique utilized in resolving the classification problem. This approach was originally postulated by Fix and Hodges back in 1952 [45]. The application of the aforementioned concept was initially introduced in the context of credit scoring by Chatterjee and Barcun [31], and subsequently employed by Henley and Hand [59]. The approach of nearest neighbours involves the selection of a metric on the application data space, which serves as a means to quantify the distance or dissimilarity between two given applicants. Utilizing a representative sample of previous applicants, a novel applicant is subsequently categorized as either satisfactory or unsatisfactory depending on the distribution of satisfactory and unsatisfactory individuals within the k-nearest applicants from the representative sample. These nearest neighbors serve as a point of reference for the evaluation of the new applicant.

The crucial variables under consideration in the present context encompass the quantitative measure of d, the quantity of applicants k that comprise the nearest neighbor pool, and the corresponding proportion of this group required for classification as a good candidate. If a significant proportion of the applicant's neighbors display negative or undesirable behavior, the applicant may be categorized as belonging to the same group. The selection of an appropriate value for k can be determined via the process of cross-validation. Normalization of inputs is a crucial step when implementing the kNN algorithm, as the items being analyzed may have been measured in varying units. The set of k-nearest neighbors of x can be represented by S_x.

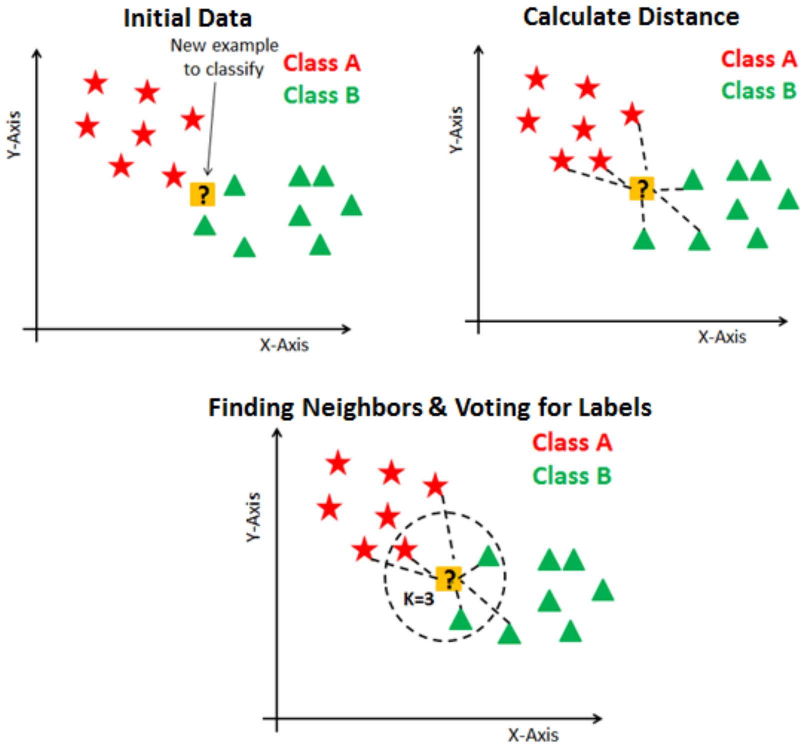


Figure 8: An example of the kNN algorithm

Mathematically, S_x is defined as $S_x \subseteq D$ s.t. $|S_x| = k$ and $\forall (x', y') \epsilon D \backslash S_x$,

$$dist(x, x') \geq max_{(x'', y'') \epsilon S_x} dist(x, x'').$$

As is represented in Figure 8, the kNN algorithm facilitates the identification of the nearest data points or clusters for a given query point. The identification of the most proximal clusters or points to a given query point necessitates the usage of a metric. In order to achieve this objective, the most common distance metrics employed include the following:

- Euclidean Distance

$$d(x,y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}. \tag{3}$$

- Manhattan Distance

$$d(x,y) = \sum_{i=1}^{N} |x_i - y_i|. \tag{4}$$

- Minkowski Distance

$$d(x,y) = \left(\sum_{i=1}^{N} (x_i - y_i)^p \right)^{\frac{1}{p}}. \tag{5}$$

Based on the aforementioned mathematical formula, it can be posited that if the variable p equals two, the resulting formula is equivalent to that of the Euclidean distance. Alternatively, if the variable p equals one, the resulting formula serves as the formula for the Manhattan distance.

3.2.9 Expert Systems

Expert systems have emerged as a novel technological solution for credit scoring applications in recent times. The resolution of intricate predicaments is reliant upon the competence, construal and cognitive approaches of expert human beings [102]. The existing body of research on expert systems within this discourse remains limited and regrettably lacks a significant level of elaboration. According to Hand and Henley [55], a notable advantage of expert systems is their capacity to offer explanations for outcomes, which, in turn, may be utilized as grounds for rejecting a credit applicant. Rosenberg and Gleit [102] provided a concise exposition of Nelson and Illingworth's [86] articulation of the fundamental components of an expert system, which are predicated on knowledge, encompassing both "facts and rules". The synthesis of this knowledge base necessitates the arrival at a conclusion, facilitated by an engine. Additionally, the system must offer an interface to afford users comprehension and explication of decisions and recommendations, while simultaneously updating said information.

Subsequently, various other applications employing expert systems have been disseminated in academic literature. Among the relevant literature is the study conducted by Ben-David and Frank [22], which entailed a comparative analysis between machine-learning models and a credit scoring expert system. Notably, the results of this study indicated that, although some of the machine-learning models demonstrated superior accuracy levels relative to the expert system model, a majority of the former did not. Another noteworthy study is that of Kumra et al [70], an inquiry was made into the application of an expert system approach towards commercial loan analysis, wherein it was established that said approach could bring forth a unique set of features within the underwriting process, distinguishing it from other methods employed earlier, as documented by Lovie [81] and Leonard [77].

As a result, an expert system typically encompasses two fundamental constituents:

- A knowledge repository that documents and retains specialized knowledge within a particular field of expertise, and

- An inference engine, composed of algorithms for the purpose of manipulating the knowledge represented within a knowledge base.

$$\text{Expert systems} = \text{Knowledge} + \text{Inference.}$$

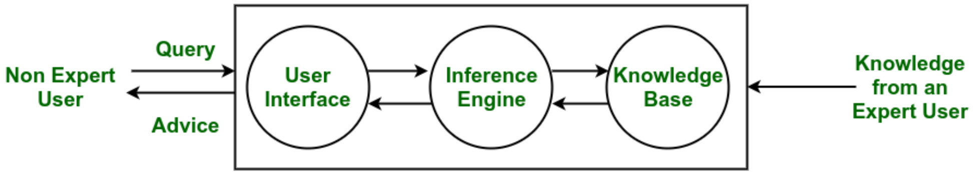


Figure 9: An example of an expert system

3.2.10 Neural Networks

Neural networks are a class of mathematical models that draw inspiration from the computational mechanisms of the human brain. Such models have been shown to be highly effective in addressing various types of complex problem-solving tasks. According to Gately [50], neural networks can be defined as a computer program aimed at solving artificial intelligence problems, which gains knowledge and expertise through a process of iterative learning and trial-and-error. The construction of neural networks necessitates a process of training, wherein the linear or nonlinear variables employed during the procedure serve to differentiate among the variables for the purpose of achieving a more refined and effective outcome in decision-making. In the domain of credit scoring, neural networks stand out from other statistical methodologies. An illustration was provided by Al Amari [5] to establish a clear distinction between regression models and neural network models. He depicted that the use of the ”inverse matrix” should be employed in the construction of an applicant score through regression models. In contrast, the utilization of the ”applicants” profiles is advocated in neural networks to appraise the relative scores of the individuals being assessed. Moreover, through employing neural networks, in case the results are deemed unsatisfactory, the estimated scores shall be subjected to modification by the networks until they reach an acceptable level, or until the optimal score of each candidate has been obtained.

The neural network equation can be defined as:

$$Z = W_0 + W_1X_1 + W_2X_2 + ... + W_nX_n, \tag{6}$$

where:

- Z represents the output of an Artificial Neural Network (ANN),
- X_i, $i = 1, ..., n$ is the independent variable,
- W_i, $i = 1, ..., n$ is the weight of X_i, and
- W_0 is the bias.

In the context of neural networks, a fundamental process involves the implementation of a set of three successive steps.

Firstly, the input variables and the aforementioned linear combination equation (6) are utilized to derive the output, or predicted Y values, denoted as Y_{pred}. Then, compute the loss or the error function. The error term refers to the discrepancy between the observed values and the predicted values. Finally, to optimize the performance of a model, it is imperative to minimize the loss function or the error term.

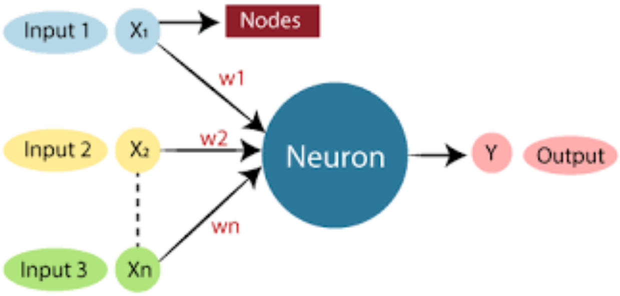


Figure 10: An example of an artificial neural network

In contemporary times, neural networks have emerged as a viable technological solution, exhibiting favorable results across diverse industries, including the financial sector in a broader sense, and more specifically within banking institutions. Gately [50] proposed various financial domains, including credit card fraud detection, bankruptcy and bank failure prediction, mortgage application, option pricing, among others, as prospective applications that could be effectively addressed using neural networks. Numerous issues are tackled by means of the utilization of feed-forward nets architecture, including pattern recognition. The predominant applications in this domain consist of multilayer feed-forward nets and probabilistic neural networks [84]. Several credit scoring models that utilize probabilistic neural networks have been scrutinized by researchers such as Masters [84] and Zekic-Susac et al. [125].

Numerous scoring models utilizing multilayer feed-forward networks have been employed by several researchers such as Trippi and Turban [115], Desai et al. [37], West [119], and Dimla and Lister [39] in a corresponding manner. The neural network models exhibit significantly higher accuracy rates as compared to discriminant analysis and logistic regression, as well as other techniques, in the aforementioned studies. It must be noted however, that differences in accuracy rates are often negligible.

Hybrid models, coupled with neural networks and sophisticated statistical methods, have been implemented in constructing scoring models, as evidenced by previous research by Stefanowski and Wilk [110], Lee et al. [72], Lee and Chen [73], Blochlinger and Leippold [25] and Trinkle and Baldwin [114]. Various studies have explored the comparison between conventional and contemporary statistical methodologies ([49], [72], [83], [125], [73] and [94]). The scope of comparisons has been broadened to encompass feed-forward networks and back-propagation networks, as both Arminger et al. [9] and Malhotra and Malhotra [83] have investigated. The findings of previous studies indicate that statistical association measures reveal neural network models to be superior in representing data compared to logistic regression and CARTs [125]. Conversely, despite displaying better classification capabilities, discriminant analysis falls short in terms of prediction accuracy. On the other hand, logistic regression exhibits a relatively higher degree of prediction ability [80]. In comparison to conventional methods such as discriminant analysis and logistic regression, neural network models have exhibited superior accuracy rates. Nevertheless, it is noteworthy to acknowledge that the disparity in outcomes between these methods was negligible. Several studies have been conducted on the topic under consideration, with notable contributions from Zekic-Susac et al. [125] and Crook et al. [36].

In the meantime, West [119] designed and implemented five distinct neural network architectures that were applied to credit scoring datasets obtained from Germany and Australia. According to West's analysis of credit scoring errors, it is recommended that scoring applications should consider both the mixture-of-experts (MOE) and radial basis function (RBF) neural networks, as they have demonstrated promising performance. Conversely, the multilayer perceptron (MLP) may not be the most accurate neural network model for scoring purposes. Although, logistic regression has been recognized as the most precise model among traditional statistical models.

3.2.11 Genetic Programming

The implementation of **genetic programming** in the area of credit scoring represents a technologically advanced approach. The origin of this method can be traced back to its initial manifestation as a subcategory of genetic algorithmic techniques, and can be characterized as an augmentation of the genetic algorithms protocol, as asserted by Goldberg [51] and Koza [68]. In the realm of computational intelligence, genetic algorithms are used to iteratively manipulate a given data-set through the application of specific genetic operations, guided by an established measure of fitness. In the context of genetic algorithms, the solution is represented by means of a "string," as described by Koza [68]. Genetic programming involves the generation of a set of competing programs through methods of mutation and crossover that emulate the principles of Darwinian evolutionary theory. These programs are subsequently assessed against one another in order to determine their efficacy. Commonly, genetic programming produces conflicting programs in the LISP or comparable programming language as a solution outcome [68]. The utilization of genetic programming has exhibited substantial growth in recent times [32], with a sizeable increase noted in the number of applications over the past few decades. These applications include among others bankruptcy prediction [88], scoring applications [63], classification problems [94], and financial returns [121].

Crook et al. [36], conducted a study to investigate the predictive accuracy of various classifiers utilizing credit scoring application data. Hence, based on the outcomes was obtained, it can be deduced that an optimum credit scoring methodology cannot be universally recommended for all data sets. This is primarily influenced by the underlying intricacies of the problem, the size and structure of the data, the usage of variables, the targeted market, and the selected cut-off point. In terms of statistical analysis, advanced techniques like neural networks and genetic programming exhibit superior performance compared to conventional statistical methods. Despite the prevalence and efficacy of modern analytical techniques, it is pertinent to acknowledge the continued relevance and applicability of conventional approaches such as linear discriminant analysis and logistic regression in certain studies.

In conclusion, numerous studies have conducted a comparative analysis between various techniques with the majority concluding that the utilization of advanced statistical techniques such as neural networks and genetic programming exhibits superior performance in comparison to traditional techniques that rely solely on the ACC rate criterion. The accuracy of classi-

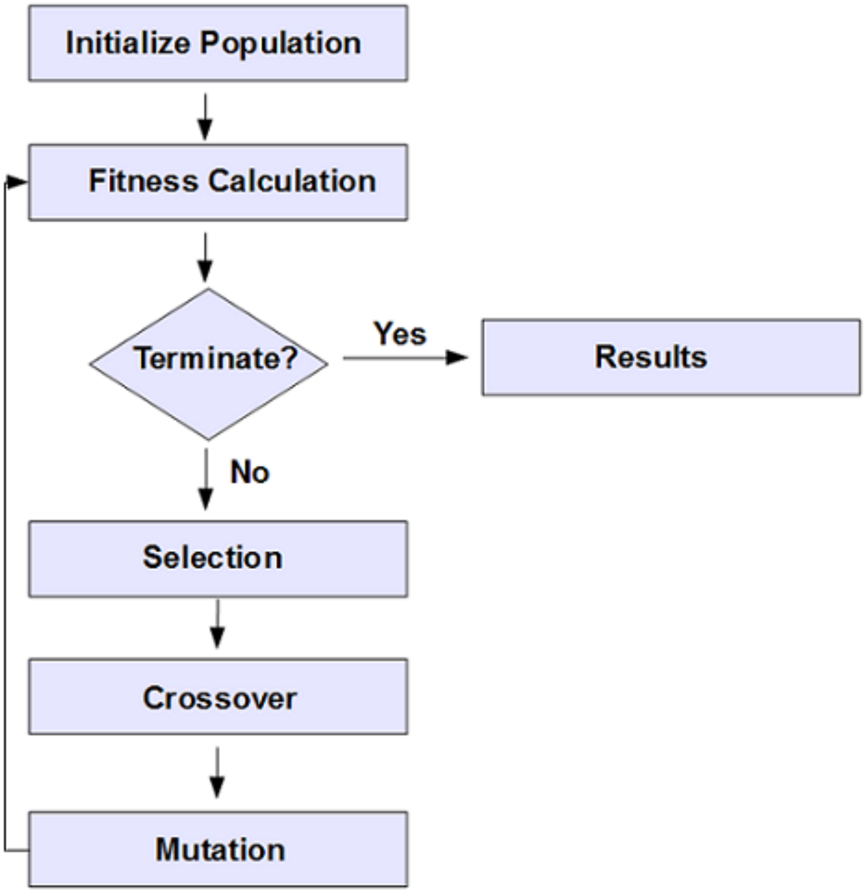


Figure 11: An example of a Genetic Algorithm

fication can occasionally be influenced by the initial group selected, which may be classified under the categories of ”bad” or a combination of ”good and bad”, as posited by various studies ([37] and [25]). Nonetheless, it is noteworthy that certain uncomplicated classification techniques such as linear discriminant analysis and logistic regression exhibit an impressive level of performance within this particular domain. In the vast majority of cases, their performance does not diverge significantly from that of other, more sophisticated techniques [12]. Finally, it is imperative to highlight that there exist diverse statistical techniques, such as support vector machines or smoothing nonparametric methods, time-varying models, mathematical programming, fuzzy rules, kernel learning method, Markov models, and linear programming [38].

3.3 Performance Evaluation Criteria of Credit Scoring

Additionaly, for evaluting the credit scoring performance is necessary to use suitable measures in order to conclude whether a model is effective or not. Various performance evaluation criteria are utilized in credit scoring applications in diverse disciplines. These criteria include the confusion matrix, accuracy (ACC) rate, estimated misclassification cost, mean-square error (MSE), root-mean-square error (RMSE), mean absolute error (MAE), the receiver operating characteristic (ROC) curve, GINI coefficient, and various other measures.

The utilization of the **confusion matrix** as a performance evaluation criterion, is prevalent in numerous fields including accounting and finance for the purposes of credit scoring, as well as marketing and health care in a broader sense. The **ACC rate** is a metric that gauges the ratio of accurately identified instances designated as either good or bad credit within a given dataset. It represents a substantial criterion with regards to assessing the classification proficiency of the suggested scoring models. The notion of accurate classification rates is derived from a tabular representation commonly referred to as a ”confusion matrix” [122], or alternatively a ”classification matrix” [1].

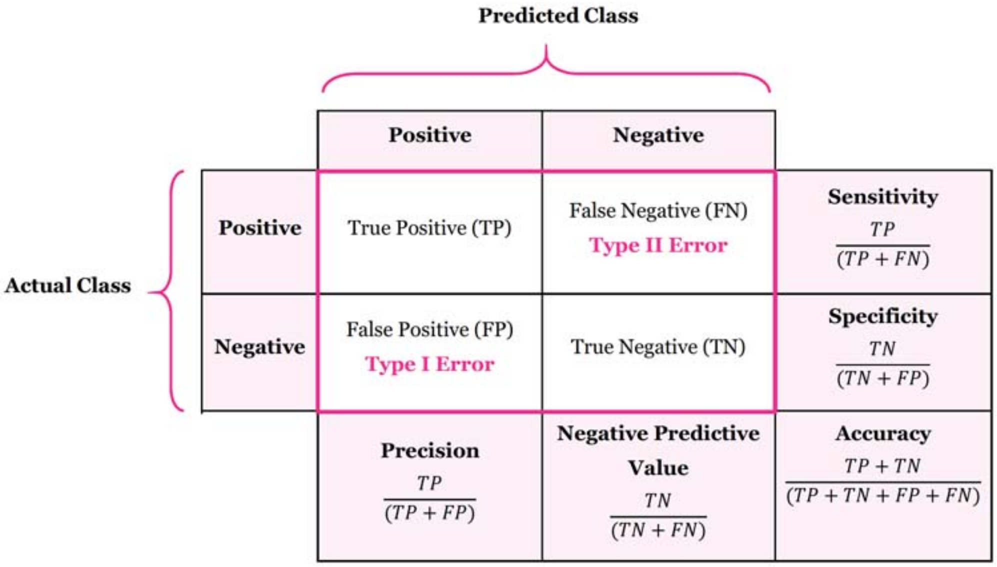
A confusion matrix displays the potential combinations of predicted and actual observations within a given data set. Also, the classification outcome which correctly identifies a valid positive example is referred to as a ”**True Positive**”. In this case, the model posits that the expected value corresponds with the empirically observed value, and the observed empirical evidence revealed a positive numerical outcome, while the machine learning algorithm effectively predicted a positive numerical outcome.

Furthermore, the term ”**True Negative**” refers to the accurate classification of an observation or test result as negative, when it is truly negative. This concept is integral in various fields of research, such as medicine, where the correct identification of negative findings plays a crucial role in the diagnosis and treatment of diseases. Indeed, in the context of predictive modeling, a true negative is recorded when the predicted value aligns with the actual value. The determined value was observed to have a negative magnitude, while the machine learning algorithm anticipated a negative output.

On the contrary, the phenomenon of falsely identifying a condition or attribute that is not actually present as positive is commonly referred to as a ”**False Positive**”. The machine learning algorithm generated an erroneous output. The observed outcome possessed a negative value while the machine learning model forecasted a positive value. So, a Type I error, commonly referred to as a false positive, is a significant concept in statistical hypothesis testing.

Finally, the occurrence of a negative response when a true positive result was expected, is referred to as a ”**False Negative**”. The artificial intelligence algorithm generated an erroneous inference. The obtained value exhibited positivity, whereas the model based on machine learning made a forecast that exhibited negativity. In brief, the Type II error, i.e., false negative, refers to the erroneous conclusion reached when a test incorrectly fails to reject a null hypothesis that is truly false.

After the confusion matrix has ascertained the quantities of True Positives (TP), True Negatives (TN), False Negatives (FN), and False Positives (FP), researchers are enabled to evaluate the classification accuracy, error rate, precision, and recall of the model.



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 12: Formulation of the confusion matrix

Additionally it is worth mentioning that, Yang et al. [122], compared the confusion matrix with two alternative measures, namely the Mahalanobis distance and the Kolmogorov-Smirnov statistics, in relation to an ROC curve. Moreover, this matrix has been previously evaluated in comparison to MSE and RMSE according to Fletcher and Goss [46] and Kumar et al. [69]. To sum up, it is common practice for the dominant credit scoring applications used in both accounting and finance, as well as other domains, to utilize the ACC rate as a performance evaluation metric [99].

The significance of the ACC rate as a key metric in credit scoring is widely acknowledged, particularly in the context of its utility for novel applications, as it underscores the precision of forecasting. However, the criterion based upon the ACC does not take into account varying costs incurred by a financial institution as a result of different categories of errors. Inadequate attention is paid to the issue of neglect, specifically with regard to the omission of important information or factors.

The notion of misclassification costs pertains to the incurrence of varying costs for misclassifying actual instances of goods predicted as bad, and the corresponding misclassification of actual instances of bad predicted as goods. In the realm of empirical evidence, it is a widely held belief that the expenses stemming from erroneous acceptance of the null hypothesis, also known as Type II errors, are significantly more substantial compared to those of incorrect rejection of the null hypothesis, which are classified as Type I errors [12].

The estimated criterion of misclassification cost serves as a means of gauging the comparative expenses incurred in the event of approving loan applications that exhibit poor repayment prospects as opposed to rejecting those applications that possess good repayment potential. The evaluation of scoring model's performance is determined by the **confusion matrix** criterion. This measure holds significant repercussions for banks, particularly concerning the absence of estimations in cases where actual bad predictions are made for observations assumed to be good. The assessment of the estimated misclassification cost criterion holds immense significance as it plays a pivotal role in the evaluation of the efficacy of credit scoring at large and aids in determining the minimum anticipated cost of misclassification for the proposed scoring models.

Several credit scoring applications have employed the **estimated misclassification cost** criterion in finance ([119], [73] and [1]) as well as in other disciplines [60]. According to Lee and Chen [73], the challenge of providing trustworthy and consistent estimates of misclassification costs poses a significant obstacle to the attainment of valid predictions. Hence, availability of a reliable prediction is not assured. According to Lee and Chen [73], there exists a widely held belief that the expenses related to Type I errors, which involve the misclassification of good credit as bad credit, and Type II errors, which involve the misclassification of bad credit as good credit, differ significantly. Furthermore, Lee and Chen [73] assert that the costs of misclassification associated with Type II errors are considerably greater than those associated with Type I errors.

$$\text{Misclassification Rate} = \frac{FP + FN}{TP + FP + TN + FN}.$$

The opposite of misclassification rate would be accuracy, which is calculated as:

$$\text{Accuracy} = 1\text{–Misclassification rate}.$$

According to West [119], Dr. Hofmann, who collated German credit data, reported a ratio of misclassification costs associated with Type II and Type I to be 5:1. This ratio has been subsequently utilized by Abdou [1]. The utilization of the relative cost ratio has been expanded in the context of sensitivity analysis to encompass elevated cost ratios, that is, those which exhibit a greater disparity between the costs under examination, for example numerical ratios 7:1 and 10:1, have been identified by Abdou [1].

Subsequently, the misclassification rate is a quantifiable measure indicating the proportion of observations that have been erroneously predicted by a given classification model.

The **ROC curve**, occasionally referred to as the ”Lorentz diagram”, is a graphical representation consisting of two dimensions. It depicts the sensitivity, which denotes the proportion of bad cases identified as bad and is plotted on the vertical axis, versus the proportion of good cases classified as bad, commonly referred to as ”one minus specificity” and plotted along the horizontal axis. This representation is established across all cut-off score values to create a comprehensive diagram. This proposition posits that sensitivity is equivalent to one minus the rate of type II error, while specificity is equal to one minus the rate of type I error. Notably, several studies have explored this relationship, including works by Baesens et al. [12], Crook et al. [36], and Yu et al. [124]. The ROC curve visualizes the comprehensive performance achieved by a diagnostic model in relation to all potential threshold values. Besides, the ROC curve portrays the operational characteristics of classifiers without considering the expenses incurred for misclassifications or disparities in class distributions. Consequently, it transcends these elements and effectively decouples classification performance from them, as noted by Thomas et al. [112] and Baesens et al. [12]. As a result, the ROC curve serves to discern optimal cut-off score values, which may effectively optimize the Kolmogorov-Smirnov statistic. Additionally, it is stated that the visualization of the Kolmogorov-Smirnov statistic is enhanced by the depiction of the ROC curve, according to the works of Hand and Jacka [113] and Blochlinger and Leippold [25].

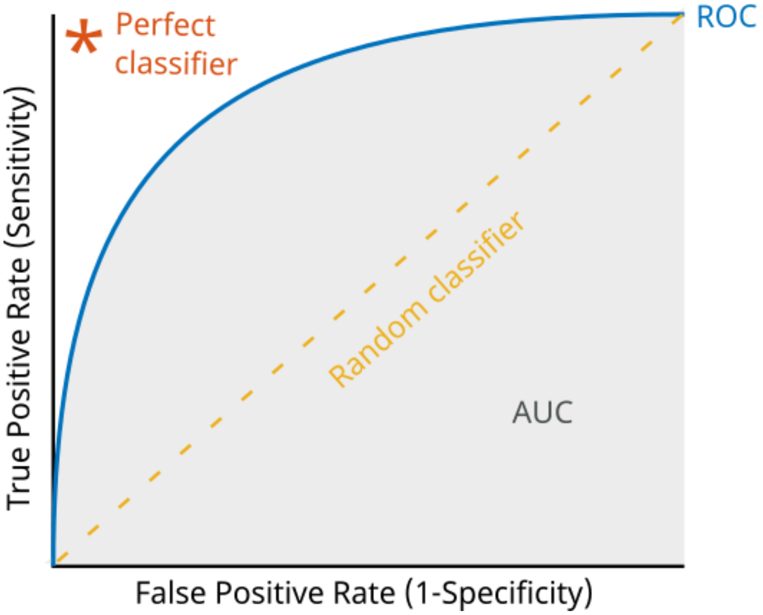


Figure 13: The ROC curve

According to Blochlinger and Leippold [25], the maximum distance separating the ROC curve from the diagonal is equal to a constant times the Kolmogorov-Smirnov statistic. However, this relationship only holds when the ROC curve is concave. In the event that the ROC curve lacks concavity, a general correspondence cannot be established. The ROC curve was initially employed as a method in the domains of psychology, healthcare, and manufacturing to assess the efficacy of signal recovery methodologies and diagnostic systems. Currently, the ROC curve is extensively utilized in the field of medical and health applications as evidenced by significant studies conducted by Shang et al. [105], Ottenbacher et al. [96], and Song et al. [108]. Also, the ROC curve has been observed to be utilized in various fields, including engineering applications, as reported by Yesilnacar and Topal [123]. Last but not least, the application of the ROC curve in finance and banking settings has been documented by scholars such as Baesens et al. [12], Blochlinger and Leippold [25], and Banasik and Crook [17].

One other common measure for assesing credit scoring is the **Gini coefficient**, commonly referred to as **Gini**, and represents a widely utilized metric within the financial sector for the purpose of assessing the efficacy of credit score models. The Gini coefficient serves as a measure that quantifies the discriminative ability of a model by assessing its efficacy in distinguishing between potential ”poor” loan recipients, who will likely default in the future, and ”strong” loan recipients, who will likely perform well in their loan repayment. This metric is frequently utilized in comparative analysis of model excellence and the assessment of their prognostic capabilities. As articulated by Thomas et al. [112], it offers a single numerical value that provides a comprehensive assessment of the scorecard’s efficacy across all cut-off points.

Despite the wide use of the Gini coefficient, there are plenty of researcers that lack a comprehensive understanding of its underlying mechanism and erroneously conflate it with a distinct metric that shares the same name. A common misconception among practitioners is to equate the Gini coefficient with the representation of the Lorenz curve, when in fact Corrado Gini’s original measure of inequality is being referred to. It is important to note that the Gini coefficient that is frequently employed by researceres is predominantly Somer’s D, which serves as a summary of the Cumulative Accuracy Profile (CAP) curve. The nomenclature ”Somer’s D” is attributed to the scholarly contributions of Robert H. Somers [3] in the field of statistical analysis. Specifically, this pertains to a quantitative assessment of the sequential correlation existing between two variables.

When it comes to credit score models, the measurement of the ordinal relationship between the prognostications of said models, as expressed in terms of the Probability of Default or the score assigned to a given borrower, and the factual outcome whether or not a default on payment occurs is pivotal. If the model is deemed effective, it is expected that low scores (representing high probability of default) are more strongly correlated with occurrences of defaults when compared to high scores (characterized by low probability of default).

The Somer's D metric is bounded within the interval of -1 to 1. The ordinal relationship of -1 represents a perfect negative correlation, while a perfect correlation is indicated by an ordinal relationship of 1. Empirically, a credit score model with Somer's D coefficient measure of 0.4 is considered to be of satisfactory quality.

Finally, note that in the context of binary classification, as with credit scoring where two distinct outcomes exist (default or no default), the **Gini index** serves as a useful measure:

$$G = 2p(1-p). \tag{7}$$

4 Optimization Techniques

4.1 Introduction

Optimization, more formally referred to as **mathematical programming**, is a comprehensive set of mathematical principles and techniques that are employed for resolving quantitative predicaments across diverse fields such as physics, biology, engineering, economics, and business. The topic has emerged as a result of recognizing that quantitative challenges in decidedly diverse fields share significant mathematical aspects. Due to this shared quality, a multitude of problems can be articulated and resolved through the employment of the consolidated range of principles and approaches that constitute the domain of optimization.

The term **mathematical programming**, often used synonymously with optimization, was first coined during the 40s when the term ”programming” did not yet have an association with computer programming. The field of mathematical programming encompasses the examination of the mathematical aspects underlying optimization problems, the creation of algorithmic approaches to address these problems, the study of mathematical properties inherent in these algorithms, and the practical utilization of these solutions with the aid of computer technology. The rapid progress of computer technology has greatly expanded the range and complexity of optimization problems that can be efficiently and successfully addressed. The progression of optimization techniques has occurred in tandem with advancements not only in computer science, but also in operations research, numerical analysis, game theory, mathematical economics, control theory, and combinatorics.

Furthermore, optimization is highly significant due to its huge range of applications, as well as the wealth of effective algorithms that are available for its implementation. From a mathematical point of view, optimization concerns itself with the optimization of a designated objective function, taking into account multiple decision variables that must satisfy distinct functional constraints. A common optimization model pertains to the distribution of sparse resources among feasible alternatives with the intention of optimizing an objective function, such as overall profitability.

The fundamental components of an optimization problem encompass the **decision variables**, **objective function**, and **constraints**.

The initial component encompasses a set of parameters, the numerical values of which can be changed to enhance the intended outcome. These variables are named **decision variables** of the optimization problem and correspond to the the variables within the objective function that the optimizer has the ability to modify. This relationship between the two elements is critical in the optimization process. Also, these variables may be referred to as design or manipulated variables. For example, various instances of decision-making involve determining the quantities of stock to purchase or vend, allocating differing quantities of resources across varied production activities, or charting out navigation routes for vehicles across traffic networks.

The second component entails a singular numerical measure, referred to as the **objective function**, which bears the objective of either being maximized or minimized. The primary aim of the objective function is to optimize the value it represents through various methodologies. The expression of the aforementioned concept is conventionally illustrated through the implementation of a functional representation, typically depicting a mathematical function denoted as $f(x)$. As an instance, the objective function could pertain to the maximization of investment profit or the minimization of energy consumption in a specific form, the projected yield on a stock portfolio, the operational expenses or gains of a corporation, or the arrival timing of a vehicle at a predetermined location. Such objectives are subject to evaluation and analysis to determine their feasibility and attainment.

The third component inherent to an optimization problem, pertains to a collection of **constraints** which serve as limitations on admissible variable values. As an example, in the context of a manufacturing process, it is imperative that the utilization of resources does not exceed the available amount and remains at a minimum of zero.

Optimization problems can exhibit diverse characteristics within the expansive framework in which they are situated. Unconstrained optimization problems denote problems lacking constraints, while the others are commonly denoted as constrained optimization problems. Especially, an optimization problem that imposes constraints on the decision variables to take on only integer values or a discrete set of values is identified as an integer or discrete optimization problem. In instances where the variables are unconstrained, the problem at hand constitutes a continuous optimization problem. It is plausible that certain predicaments may entail a confluence of discrete and continuous variables.

However, difficulties encountered in situations where the variables are of continuous nature, such as the circumstance of resource allocation, necessitate a distinct methodology as contrasted to challenges in which variables take on a discrete or combinatorial character, as in the scenario of choosing a vehicle route from a predetermined range of possibilities.

Another category is feasibility problems, which are typified by the absence of objective functions. On the other hand it is known that certain issues may encompass multiple objective functions and thus need to be tackled through the approach of simplifying them into either a singular objective optimization problem or a series of successive problems of the same nature.

4.2 Optimization Problems

The initial step involves presenting a broad characterization of an optimization problem. Suppose a function $f(x) : \mathbb{R}^n \to \mathbb{R}$ and a set $S \subset \mathbb{R}^n$ the issues of finding an $x^* \epsilon \mathbb{R}^n$ that satisfies:

$$\begin{gathered} min_x f(x) \\ s.t. \;\; x \epsilon S. \end{gathered}$$

The above is referred to as an optimization problem.

The function f is named as the **objective function**, and the set S is the **feasible region**, whereas it is referred to as **infeasible** if S is null. Also, the problem is called **unbounded**, if there exists a sequence $x^k \epsilon S$ such that $f(x^k) \to -\infty$ as $k \to +\infty$. In the case that the problem is not characterized by either infeasibility or unboundedness, it is possible to obtain a $x^* \epsilon S$ such that:

$$f(x^*) \leq f(x), \forall x \epsilon S. \tag{8}$$

This type of x is referred to as a **global minimizer** of the optimization problem. Note that if equation (8) becomes:

$$f(x^*) < f(x), \forall x \epsilon S, x \neq x^*,$$

then x^* is called a **strict global minimizer**.

Furthermore, in other occasions, it is necessary to find an $x^* \epsilon S$ that solves:

$$f(x^*) \leq f(x), \forall x \epsilon S \cap B_{x^*}(\varepsilon),$$

with $\varepsilon > 0$ and $B_{x^*}(\varepsilon) = \{x : \|x - x^*\| < \varepsilon\}$.

In such case x^* is called a **local minimizer** of the optimization problem. Similarly, a **strict local minimizer** can be defined in a manner that is analogous to the definition of a local minimizer.

Additionaly, the majority of instances involve the explicit depiction of the feasible set S through functional constraints, which may include both equalities and inequalities. One possible illustration of the variable S may be presented as:

$$S := \{x : g_i(x) = 0,\ i \in \mathcal{E}\, and\, g_i(x) \geq 0,\ i \in \mathcal{I}\},$$

where $\mathcal{E}$ and $\mathcal{I}$ are representative index sets that pertaining to both equality and inequality constraints.

Consequently, the optimization problem exhibits a generic structure as expressed in the following form:

$$\begin{aligned} min_x\ & f(x) \\ & g_i(x) = 0, i \in \mathcal{E} \\ & g_i(x) \geq 0, i \in \mathcal{I}. \end{aligned}$$

There are numerous factors that have an impact on the efficiency with which optimization problems can be solved. In the first instance, the number of decision variables, denoted as n, and the total number of constraints, represented by the sum of $\mathcal{E}$ and $\mathcal{I}$, have been found to be a reliable indication for assessing the complexity of solving an optimization problem. Also, there exist additional factors that are associated with the specific properties of the functions f and g_i, which are utilized to establish and characterize the problem at hand. However, difficulties in optimizing a linear objective function and adhering to linear constraints are relatively simple in comparison to complications that arise from non-linear objective functions and non-convex feasible sets. Similarly, when presented with a convex objective function and a convex set of constraints, the optimization process is comparatively less challenging. Due to specific features presented by particular problems, researchers have devised algorithms that satisfy these features, as opposed to commonly employed optimization algorithms with generalized purposes.

As such, it follows that optimization problems can be classified into multiple categories. In brief, **linear programming** is a significant category of optimization. The term ”linear” denotes a mathematical relationship between dependent and independent variables in which there are no instances of variables being raised to a higher power, such as squares. In the present category of problems, the objective is to minimize or maximize a linear function of real variables subject to linear equalities and inequalities being satisfied. Another essential category of optimization is that of **non-linear programming**. Nonlinear programming involves the consideration of real numbers as variables, with the objective function or some of the constraints represented by nonlinear functions. These functions may encompass various mathematical operations such as squares, square roots, trigonometric functions, or products of the variables. Furthermore, **stochastic programming**, **network optimization**, and **combinatorial optimization** represent significant categories of optimization problems. Stochastic programming addresses scenarios in which the objective function or constraints vary according to random variables, necessitating the determination of an optimum in a probabilistic or ”expected” sense. Network optimization, on the other hand, entails the optimization of a specific flow property within a network, such as the maximal quantity of substance that can be conveyed between two predefined locations in the network. Finally, combinatorial optimization confronts the challenge of identifying the optimal selection from among a finite, yet extensive set of feasible values.

In Figure 14, an outline of the steps that are followed for solving an optimization problem is presented.

4.3 Linear Programming

Linear optimization, also known as **linear programming**, is widely recognized as one of the simplest and most frequently encountered optimization problems. The issue at hand pertains to the optimization of a linear objective function under the presence of linear equality and inequality constraints. This pertains to the instance in question where the functions f and g_i are exclusively linear. However, if either the function f or any of the functions g_i are non-linear in nature, the problem at hand pertains to non-linear

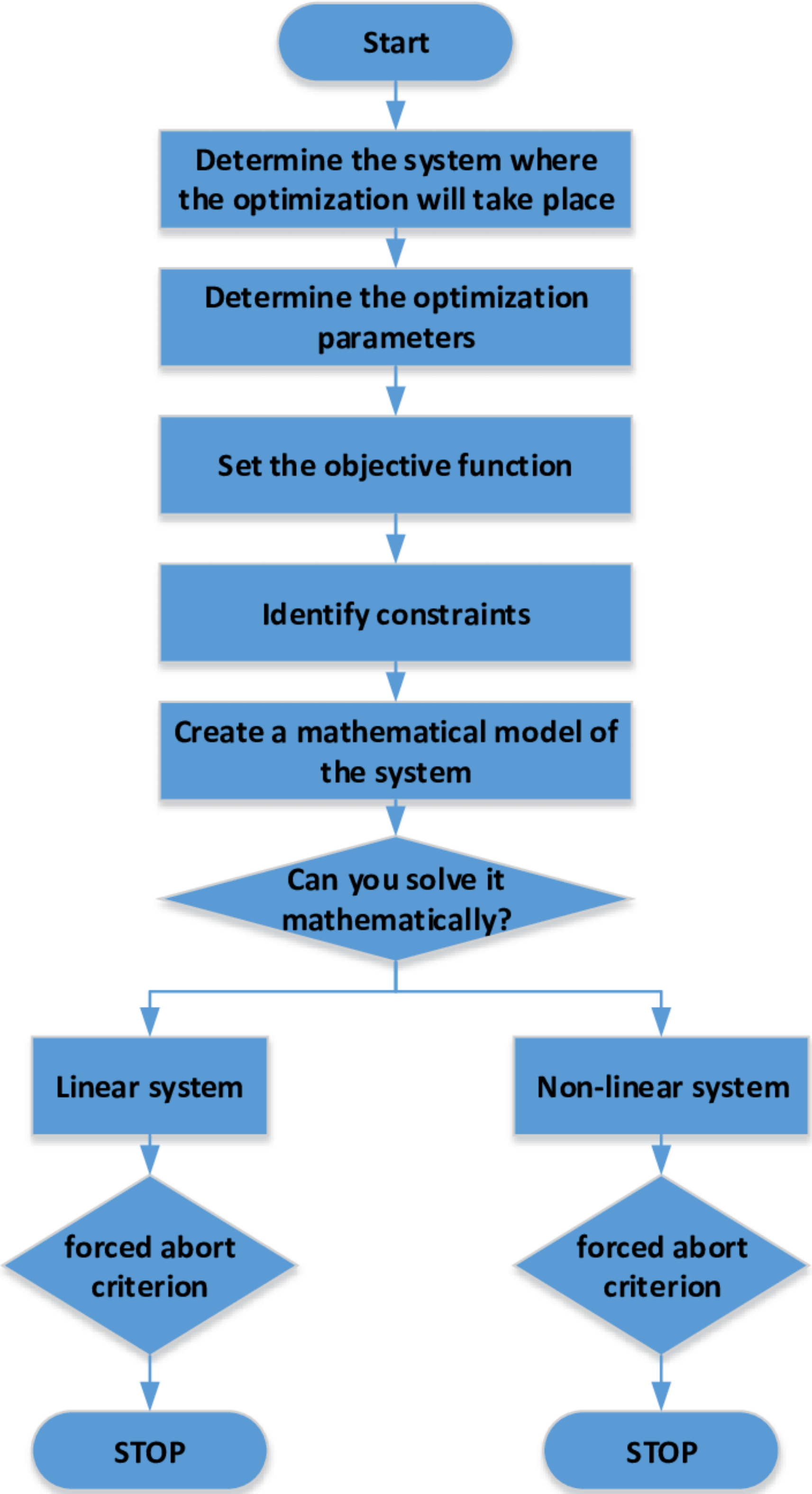

Start
Determine the system where the optimization will take place
Determine the optimization parameters
Set the objective function
Identify constraints
Create a mathematical model of the system
Can you solve it mathematically?
Linear system
Non-linear system
forced abort criterion
forced abort criterion
STOP
STOP

Figure 14: Steps of optimization process

programming.

Before 1947, linear programming was relatively obscure despite its current widespread application in resolving common decision-making challenges. Prior to this date, no substantial research was conducted despite the fact that Joseph Fourier, a notable French mathematician, appeared to have recognized the potential of the subject in early 1823. In 1939, an eminent mathematician from Russia, Leonid Vitalyevich Kantorovich, authored an elaborate dissertation titled ”Matematicheskie metody organizatsi i planirovaniya proizvodstva” (Mathematical Methods for Organization and Planning of Production), which is currently recognized as the seminal work towards the identification of well-defined mathematical structures underlying crucial classes of scheduling problems. Regrettably, Kantorovich’s recommendations were predominantly obscure both within the confines of the Soviet Union and beyond its borders for a duration of almost two decades. Moreover, significant advancements in the field of linear programming were observed in both the United States and Western Europe. After the conclusion of World War II, government officials within the United States began to recognize the necessity of utilizing scientific planning methodologies in order to effectively coordinate the collective energies and resources of the nation during a potential nuclear conflict. The inception of the computer rendered such an approach, viable.

Indeed, a comprehensive effort was initiated in the year 1947 within the boundaries of the United States Air Force. The proposition of the linear programming model was based on its relative simplicity in mathematical terms, while simultaneously offering a comprehensive and practical framework for representing interconnected activities that contend for limited resources. In the context of linear programming, the modeler adopts a systematic approach whereby the system under scrutiny is regarded as comprising multiple activities that entail flows of inputs, such as labor and raw materials, and outputs, including finished goods and services, which are proportionally linked to the level of the activity under consideration. It is posited that the levels of activity can be depicted through non-negative numerical values. The innovative aspect of this approach is its utilization of a linear objective function as a means of expressing the objective of the decision-making process. This involves the minimization or maximization of a specific outcome, such as the optimization of potential sorties for the air force or the maximization of profits within the industrial sector.

Up until 1947, the execution of practical planning was distinguished by a sequence of mandates on the procedures and priorities imposed by authoritative figures. The absence of general objectives is likely attributed to the arduous computations that would need to be conducted in order to effectively minimize an objective function while complying with constraints. In 1947, an algorithmic approach known as ”The Simplex Method” was introduced, demonstrating remarkable efficiency in resolving real-world mathematical problems. The phenomenon of linear programming attracted considerable attention in the 50s and subsequently gained traction among industrial enterprises, ultimately achieving widespread adoption by 1951. Presently, there exists almost no industry that does not employ some degree of mathematical programming, although discrepancies arise regarding the applications and extent thereof, even intra-industry.

The domain of linear programming has witnessed an extension of interest towards the discipline of economics. In 1937, the analysis of an expanding economy founded on a range of production techniques and established technological coefficients was conducted by the Hungarian-born mathematician, John von Neumann. In the context of the historical development of mathematics, the examination of linear inequality systems was a relatively unknown area of inquiry until the year 1936, garnering scant attention from scholars and researchers. Furthermore, in 1947, von Neumann postulated the equivalence between linear programming and matrix games, and introduced the pivotal notion of duality, while additionally offering several proposals intended for the numerical solution of both linear programming and game theoretical problems. Therefore, the assignment of a significant proportion of the fundamental contributions in the field of mathematics, can justifiably be attributed to von Neumann. Finally, it is worth mentioning that the year 1948 represents a significant turning point in the annals of mathematics, in which the concept of duality and its associated themes underwent a thorough and systematic investigation of a highly rigorous nature.

Mathematically, the linear programming is conventionally expressed in the standard form, which is presented as follows:

$$\begin{aligned} min_x\, c^T x \\ Ax = b \\ x \geq 0, \end{aligned} \tag{9}$$

where $A \epsilon \mathbb{R}^{m\times n}$, $b \epsilon \mathbb{R}^m$, $c \epsilon \mathbb{R}^n$ are given and $x \epsilon \mathbb{R}^n$ is the vector of variables, that must be setted. It is widely known that a k-vector is identified by a $k \times 1$ matrix. Also, the notation M^T is utilized to denote the transpose matrix of an $m \times n$ matrix M. Specifically, the transpose matrix is an $n \times m$ matrix that is characterized by entries $M_{ij}^T = M_{ji}$. Consequently, the aim of equation (9) is to minimize the linear function $\sum_{j=1}^{n} c_j x_j$.

Similar to the concept of optimization problem, the issue of linear programming is deemed *feasible* when its limitations are congruous. Furthermore, linear programming is classified as *unbounded* if there is an existence of a series of feasible vectors x^k such that $c^T x^k \to -\infty$. In instances where the linear program is feasible but not unbounded, it possesses an optimal solution. This optimal solution, denoted by x, is characterized by its ability to meet the constraints set forth while simultaneously minimizing the objective value in comparison to all other feasible vectors.

To sum up, *Linear Programming* is a widely researched topic in the field of *Operations Research* with the most effective and widely employed techniques for resolving linear programming problems being the interior-point and simplex methods.

4.3.1 The Simplex Method

The **Simplex method** is a widely used computational approach for solving linear programming problems manually. It involves the use of slack variables, tableaus, and pivot variables to facilitate the identification of the optimal solution of the optimization problem.

The simplex algorithm was originally developed by Dr. George Dantzig during the World War II. The linear programming models devised by him proved to be significant assistance in resolving transportation and scheduling difficulties faced by the Allied forces. In 1979, Leonid Khachian, a Soviet

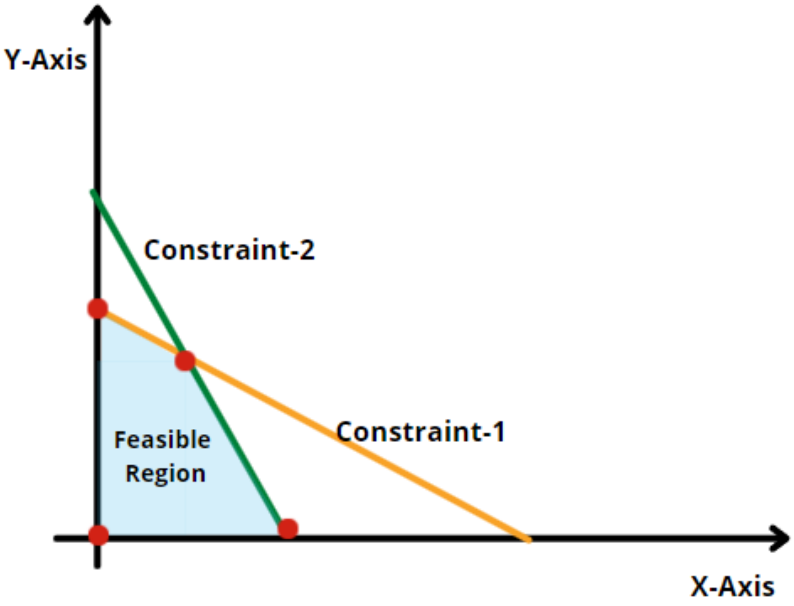


Figure 15: A presentation of linear programming

researcher, introduced the ellipsoid algorithm, posited to be a revolutionary method. However, subsequent analyses reveal that its efficacy is no superior to that of the simplex method. In 1984, Narendra Karmarkar, a research scientist at $AT\&T$ Bell Laboratories, formulated Karmarkar's algorithm, which has been verified to exhibit a four-fold increase in computational efficiency compared to the simplex method, for specific problem instances. However, it remains the case that the simplex method is the most effective solution for the majority of problems.

Simplex employs a methodology that exhibits a high degree of efficiency. The algorithm does not calculate the objective function value at each and every point within the domain, but rather begins its operation at a vertex of the feasible region wherein all primary variables assume a value of zero. It then proceeds to traverse the feasible region, successively moving from vertex to vertex, strengthening of the value of the objective function at each stage of the process. The aforementioned process will continue until the optimal resolution is achieved.

The simplex algorithm is a computational technique that employs iterative procedures based on mathematical calculations and logical reasoning to derive the optimal solution for a given linear programming problem. Also, it guarantees favorable consideration due to its exceptional efficiency and reliability. Furthermore, its utilization extends to mixed integer programming, particularly subsequent to relaxation of the pertinent constraints.

The simplex algorithm may be described as a problem solving approach that involves a series of fundamental stages:

1. **Define the problem.** It is necessary to formulate both the objective function and the inequality constraints.

2. **Transform the inequalities into mathematical equations.** The execution of this step entails the inclusion of an additional slack variable for every inequality present.

3. **Establish the starting simplex tableau.** The objective function should be expressed as the last row.

4. The entry with the most negative value in the bottom row represents the pivot column.

5. **Compute the quotients.** The identification of a row is based on the quotient that has the smallest weight. The point of intersection between the column identified in step 4 and the row marked in this step, is defined as the pivot element. The calculation of the quotients is achieved through the process of division wherein the far right column is divided by the identified within the 4th step column. Any quotient that equals zero, a negative number or includes a zero within the denominator, shall be disregarded.

6. Utilize **the method of pivoting** to transform all remaining values within this particular column to zero. This process is executed in a similar manner to our approach with the Gauss-Jordan methodology.

7. The termination criterion for the process entails **the absence of negative entries in the lowest row of the matrix**. In the event of their presence, it necessitates the repetition of the algorithmic sequence from step 4 and forward.

8. **Obtain the variables by utilizing the columns containing binary values of 1 and 0.** All other variables have a value of zero. The optimal value sought is situated in the lowermost and furthest right region.

4.3.2 Basic Solutions

In order to provide a comprehensive explanation, below sequential process is presented by utilizing a general model. Consider the following form of problem:

$$\begin{aligned} max\ \ & \mathbf{c}\,\mathbf{x} \\ \mathbf{A}\,\mathbf{x} & \leq \mathbf{b} \\ \mathbf{x} & \geq \mathbf{0}, \end{aligned}$$

where **A** denotes an $m{\times}n$ matrix, **b** represents an m-dimensional column vector and **c** represents an n-dimensional row vector. The variables of a given problem are denoted by an n-dimensional column vector **x**. Consequently, the vectors and matrices can be represented as follows:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, \mathbf{c} = \begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Subsequently, the incorporation of slack variables into the functional constraints generates the increased format of the problem. The vector of slack variables is denoted by $\mathbf{x_s}$.

$$\mathbf{x_s} = \begin{bmatrix} x_{n+1} \\ x_{n+2} \\ \vdots \\ x_{n+m} \end{bmatrix}.$$

Let the notation **I** refer to the identity matrix of dimension $m \times m$. The constraints in the augmented form can be expressed as follows:

$$\begin{bmatrix} \mathbf{A}, \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{x_s} \end{bmatrix} = \mathbf{b}, \begin{bmatrix} \mathbf{x} \\ \mathbf{x_s} \end{bmatrix} \geq \mathbf{0}. \tag{10}$$

However, there exist a plethora of possible resolutions for the specific equation system outlined in equation (10). The selection of $\mathbf{x} = \mathbf{0}$ and $\mathbf{x_s} = \mathbf{b}$, results in the fulfillment of the aforementioned equation, however, it does not necessarily ensure the satisfaction of all the corresponding inequalities. In a broader sense, partitions of the augmented matrix $[\mathbf{A}, \mathbf{I}]$ can be taken into consideration.

$$[\mathbf{A}, \mathbf{I}] \equiv [\mathbf{B}, \mathbf{N}].$$

The matrix denoted by $\mathbf{B}$ is a square matrix of dimension $m \times m$, whose columns are formed by a set of linearly independent vectors obtained from the concatenation of matrix $\mathbf{A}$ and the identity matrix $\mathbf{I}$. If the variable vector $\begin{bmatrix} \mathbf{x} \\ \mathbf{x_s} \end{bmatrix}$ is partitioned in the same way, then

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{x_s} \end{bmatrix} \equiv \begin{bmatrix} \mathbf{x_B} \\ \mathbf{x_N} \end{bmatrix}.$$

The equality constraints presented in equation (10) can be expressed in a restructured format as follows:

$$\begin{bmatrix} \mathbf{B}, \mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{x_B} \\ \mathbf{x_N} \end{bmatrix} = \mathbf{B}\mathbf{x_B} + \mathbf{N}\mathbf{x_N} = \mathbf{b}, \tag{11}$$

or, by applying the left-sided multiplication of $\mathbf{B}^{-1}$, the above equation ca be expressed as:

$$\mathbf{x_B} + \mathbf{B}^{-1}\mathbf{N}\mathbf{x_N} = \mathbf{B}^{-1}\mathbf{b}. \tag{12}$$

Through the above construction, we establish the equivalence of the subsequent three systems of equations, such that any solution to one system simultaneously satisfies the remaining two.

$$\begin{gathered} \begin{bmatrix} \mathbf{A}, \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{x_s} \end{bmatrix} = \mathbf{b} \\ \mathbf{B}\mathbf{x_B} + \mathbf{N}\mathbf{x_N} = \mathbf{b} \\ \mathbf{x_B} + \mathbf{B}^{-1}\mathbf{N}\mathbf{x_N} = \mathbf{B}^{-1}\mathbf{b}. \end{gathered}$$

The above linear systems can be characterized as alternative depictions of the first system through the utilization of the matrix $\mathbf{B}$. As previously noted, a clear resolution to the aforementioned system, and by extension the other two, is the variable values of $\mathbf{x_N} = \mathbf{0}$ and $\mathbf{x_B} = \mathbf{B}^{-1}\mathbf{b}$. In actuality, a solution for $\mathbf{x_N}$ can be acquired by setting its components to fixed values:

$$\mathbf{x_B} = \mathbf{B}^{-1}\mathbf{b} - \mathbf{B}^{-1}\mathbf{N}\mathbf{x_N}. \tag{13}$$

The independent variables, denoted by $\mathbf{x_N}$, can be deemed as selectable variables which may be freely chosen by the researcher. Subsequently, the dependent variables, $\mathbf{x_B}$, are unequivocally determined once these independent variables have been established. The description *"basic solution"* is assigned to a solution of the aforementioned systems that conforms to the following format:

$$\mathbf{x_N} = \mathbf{0}, \mathbf{x_B} = \mathbf{B}^{-1}\mathbf{b},$$

where **B** is a given basis matrix.

Additionally, if the restriction $\mathbf{x_B} = \mathbf{B}^{-1}\mathbf{b} \geq \mathbf{0}$ is satisfied, the resulting solution $\mathbf{x_B} = \mathbf{B}^{-1}\mathbf{b}$ and $\mathbf{x_N} = \mathbf{0}$ constitutes a feasible basic solution for the linear programming problem that described above. The variables $\mathbf{x_B}$ are referred to as the "basic variables", whereas the variables $\mathbf{x_N}$ are identified as the "non-basic variables". From a geometric perspective, it can be stated that fundamental feasible solutions are in direct correspondence with the extreme vertices of the set of admissible solutions $\{x : Ax \leq b, x \geq 0\}$. The extreme points of a given set are defined as those points that are incapable of being expressed as a convex combination of any two distinct points belonging to the same set.

The objective function, denoted as $\mathbf{Z} = \mathbf{cx}$, may be similarly expressed through utilization of the basis partition. The $\mathbf{c} = \left[\mathbf{c_B}, \mathbf{c_N}\right]$is employed to denote the division of the objective vector. The current sequence of comparable representations of the objective function equation is as follows:

$$\mathbf{Z} = \mathbf{cx} \Leftrightarrow \mathbf{Z} - \mathbf{cx} = \mathbf{0}$$
$$\mathbf{Z} - \left[\mathbf{c_B}, \mathbf{c_N}\right]\begin{bmatrix}\mathbf{x_B}\\ \mathbf{x_N}\end{bmatrix} = \mathbf{0}$$
$$\mathbf{Z} - \mathbf{c_B}\mathbf{x_B} - \mathbf{c_N}\mathbf{x_N} = \mathbf{0}.$$

By replacing $\mathbf{x_B} = \mathbf{B}^{-1}\mathbf{b} - \mathbf{B}^{-1}\mathbf{N}\mathbf{x_N}$ from equation (13):

$$\mathbf{Z} - \mathbf{c_B}(\mathbf{B}^{-1}\mathbf{b} - \mathbf{B}^{-1}\mathbf{N}\mathbf{x_N}) - \mathbf{c_N}\mathbf{x_N} = \mathbf{0}\mathbf{Z} - (\mathbf{c_N} - \mathbf{c_B}\mathbf{B}^{-1}\mathbf{N})\mathbf{x_N} = \mathbf{c_B}\mathbf{B}^{-1}\mathbf{b}.$$

Finally, it is wort to be note that the final equation lacks the fundamental variables. The aforementioned depiction enables us to ascertain the net impact on the objective function by altering a non-basic variable. The aforesaid characteristic serves as a crucial feature employed by the simplex algorithm. The vector that corresponds to the non-basic variables in the objective function, namely $\mathbf{c_N} - \mathbf{c_B}\mathbf{B}^{-1}\mathbf{N}$, is often referred to as the vector of reduced costs. This is due to the fact that its elements are the cost coefficients, c_N, that have been subjected to a reduction by the cross effects of the basic variables, represented by $\mathbf{c_B}\mathbf{B}^{-1}\mathbf{N}$.

4.4 Quadratic Programming

The *quadratic optimization* or **quadratic programming** problem has a more comprehensive scope as it entails the maximization or minimization of a quadratic function of variables for the objective function. The standard form of quadratic programming is commonly defined in the following manner:

$$\begin{aligned} min_x \ \frac{1}{2}x^T Qx + c^T x \\ Ax = b \\ x \geq 0, \end{aligned}$$

where $A \epsilon \mathbb{R}^{m\times n}, b \epsilon \mathbb{R}^m, c \epsilon \mathbb{R}^n, Q \epsilon \mathbb{R}^{n\times n}$ are considered known and $x \epsilon \mathbb{R}^n$.

In additional, from the relation

$$x^T Qx = \frac{1}{2}x^T(Q + Q^T)x,$$

it can be theorized, without compromising the generality of the analysis, that the variable Q possesses symmetry, with regard to its properties. This can be alternatively expressed as Q exhibiting a symmetrical characteristic $Q_{ij} = Q_{ji}$.

The aim of the quadratic progamming problem is to optimize a convex function of x, provided that Q is a positive semidefinite matrix. The aforementioned condition is tantamount to Q possessing exclusively non-negative eigenvalues. Once the aforementioned condition has been fulfilled, the quadratic programming dilemma satisfies the criteria of being a convex optimization issue, making it affordable to resolution within feasible time intervals via interior-point methodologies. The current theory pertains to an established concept utilized for assessing the intricacy of computational operations. For this reason polynomial time algorithms are deemed efficient owing to their capacity to derive optimal solutions in a time frame that is ascertained to be no more than a polynomial function of the input's magnitude.

4.5 Conic Programming

A further extension of the linear programming problem can be achieved through the substitution of non-negativity constraints $x \geq 0$ with general conic inclusion constraints. The problem under consideration is commonly referred to as a conic optimization problem. In order to achieve this objective, consideration is given to a closed convex cone denoted as C, which exists within a vector space X of finite dimensions. Furthermore, the conic optimization problem in question is expressed as follows:

$$\begin{aligned} min_x \; c^T x \\ Ax = b \\ x \epsilon C. \end{aligned} \tag{14}$$

If $X = \mathbb{R}^n$ and $C = \mathbb{R}^n_+$, the above equations are identified as the model of linear programming. However, it is possible to formulate a plethora of non-linear optimization problems in a more expansive manner using this approach. Moreover, certain algorithmic machineries that are highly efficient and resilient, originally developed for solving linear optimization problems, can be adapted to resolve these universal optimization problems. **Conic optimization problems** comprise a substantial subset of mathematical optimization, and are defined by the presence of a convex cone constraint on the variables in the optimization problem. Among these problems are two significant subclasses: (i) optimization over second-order cones, and (ii) optimization over semi-definite cones. The aforementioned instances refer to situations wherein C pertains to the second-order cone:

$$C_q := \{x = (x_1, x_2, ..., x_n) \epsilon \mathbb{R}^n : x_1^2 \geq x_2^2 + ... + x_n^2, x_1 \geq 0\},$$

and the cone of symmetric positive semidefinite matrices is:

$$C_s := \left\{ X = \begin{bmatrix} x_{11} & ... & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{n1} & ... & x_{nn} \end{bmatrix} \epsilon \mathbb{R}^{n \times n} : X = X^T, X \text{ is positive semi-definite} \right\}.$$

In the context of positive semi-definite matrices, the conventional inner products employed in the expressions $c^T x$ and Ax as outlined in equation (14) are substituted by a suitable inner product that pertains to n-dimensional square matrices.

4.6 Integer Programming

The class of **integer programs** encompasses optimization problems that necessitate one or more of the variables to exclusively correspond to integer values. The application of constraints on the variables can significantly increase the difficulty of solving problems. The attention will be centered towards integer linear programs, characterized by a linear objective function and constraints that are equally linear. A linear program that consists only of integer variables is referred to as a pure integer linear program and it is expressed as follow:

$$\begin{aligned} min_x \; c^T x \\ Ax \geq b \\ x \geq 0 \text{ and integrated,} \end{aligned}$$

where $A \epsilon \mathbb{R}^{m \times n}, b \epsilon \mathbb{R}^m, c \epsilon \mathbb{R}^n$ are given, and $x \epsilon \mathbb{N}^n$ is the variable vector to be decided.

A significant scenario arises in the instance wherein the x_j variables denote binary decision variables, specifically denoting that x be within the set of binary values, $\{0, 1\}^n$. Subsequently, the aforementioned issue is classified as a 0-1 linear programming problem.

Furthermore, the combination of continuous variables and integer constrained variables in a mathematical optimization problem is commonly referred to as a mixed integer linear program, and it is expressed as:

$$\begin{aligned} min_x\ c^T x \\ Ax \geq b \\ x \geq 0 \\ x_j \epsilon \mathbb{N},\ j = 1, \ldots, p. \end{aligned}$$

where the variables A, b, c represent predetermined data, while the integer value p, $1 \leq p < n$, is also an element of the input.

4.7 Dynamic Programming

Dynamic programming is a computational methodology that entails utilizing recurrence relations. The method in question was formulated by Richard Bellman during the initial half of the 50s [21]. The term ”dynamic programming” originated from the investigation of programming predicaments which necessitated a consideration of temporal alterations. Nonetheless, the method can be utilised irrespective of the temporal aspect of the issue at hand. The proposed approach involves partitioning the problem into discrete ”stages” as a means of conducting recursive optimization. Also, it is feasible to integrate stochastic components within the recursive procedure.

4.8 Optimization with Data Uncertainty

In all of the problem classes that have been discussed so far, with the exception of dynamic programming, it has been assumed implicitly that the data pertaining to the problem is known, including parameters such as Q, A, b, and c in quadratic programming. However, this phenomenon does not invariably occurs. Frequently, the parameters of a problem pertain to quantities that are yet to be actualized or that cannot be precisely determined during the period in which the problem must be established and resolved. Instances of this types, are frequently observed in models that pertain to financial quantities, including investment returns and risks. Two distinct methodologies will be examined herein, which pertain to optimizing with respect to data uncertainty. The utilization of stochastic programming is a methodological approach employed when the presence of data uncertainty can be attributed to random factors and can ultimately be delineated through the use of probability distribution techniques. Moreover, robust optimization is employed in situations where one seeks a solution that exhibits favorable performance outcomes across all conceivable iterations of uncertain data. The aforementioned approaches are not categorized as problem classes, such as linear programming and quadratic programming, but are rather regarded as modeling techniques that are utilized for the purpose of resolving data uncertainty.

4.9 Stochastic Programming

Stochastic programming is a form of optimization problem that involves a certain degree of probabilistic uncertainty in the underlying problem data. The optimization problem at hand may take the form of a linear, integer, or non-linear program. The stochastic linear programming example constitutes a significant and noteworthy case of research in the field of mathematical optimization.

The emergence of a stochastic program with recourse occurs, when certain decisions, namely recourse actions, are able to be made subsequent to the realization of certain or all random events. An illustration of a mathematical model in the field of operations research is a two-stage stochastic linear program with recourse, which can be expressed in the following manner:

$$\begin{aligned} max_x \quad & a^T x + \mathbb{E}[max_{y(\omega)}\, c(\omega)^T y(\omega)] \\ & Ax \;=\; b \\ & B(\omega)x + \; C(\omega)y(\omega) = \; d(\omega) \\ & x \geq 0, \qquad y(\omega) \geq 0, \end{aligned}$$

where the decisions made in the first stage are denoted by a vector x, while those made in the subsequent stage are denoted by a vector $y(\omega)$. It should be noted that the values of $y(\omega)$ are influenced by the occurrence of a stochastic event represented by the symbol "ω". Deterministic constraints on the first stage decisions x are determined by A and B, whereas stochastic linear constraints connecting the recourse decisions $y(\omega)$ to the first stage decisions are defined by $B(\omega)$, $C(\omega)$, and $d(\omega)$. The objective function comprises a fixed component $a^T x$ and the expected value of the secondary objective $c(\omega)^T y(\omega)$ across all possible outcomes of the stochastic event ω.

Moreover, after making the initial decision variables x and upon the occurrence of the random event ω, one may determine the optimal subsequent decision variables through the resolving of the consequent linear programming problem:

$$\begin{aligned} f(x,\omega) = max \quad & c(\omega)^T y(\omega) \\ c(\omega) y(\omega) &= d(\omega) - B(\omega) x \\ y(\omega) &\geq 0. \end{aligned}$$

Let us suppose that $f(x) = \mathbb{E}[f(x,\omega)]$ is the expected value of the optimal value of the problem. Thus, the two-stage stochastic linear program has the following form:

$$\begin{aligned} max \quad & a^T x + f(x) \\ Ax &= b \\ x &\geq 0. \end{aligned}$$

Therefore, in the case that the function $f(x)$ - which may be non-linear - is established, the problem may be converted to a non-linear programming dilemma. After, characterization of the finite distributions of the data $c(\omega)$, $B(\omega)$, $C(\omega)$, and $d(\omega)$, it is possible to prove that the function f incorporates properties of piecewise linearity and concavity. When the probability densities that describe the data are characterized by absolute continuity and possess finite second moments, it is proved that f is both concave and differentiable. In either scenario, the optimization problem at hand possesses convexity with constraints that are linear, and may be resolved by the use of specialized algorithms.

4.10 Robust Optimization

The concept of **robust optimization** concerns the process of formulating optimization problems having regard to the presence of uncertain data. The objective of this approach is to take a solution that can be deemed ”good” despite the variability in the unknown parameters across all potential scenarios. The present approach constitutes a deviation from the utilization of the randomness assumption in stochastic optimization for resolving uncertain parameters, as it assigns equitable significance to all feasible realizations. The description of uncertainty in parameters is expressed through the utilization of uncertainty sets that encompass the entirety or prevalent values that may possibly be actualized by such variables with uncertainty.

It is noteworthy that various definitions and interpretations of robustness exist, thereby leading to disparate models. A crucial concept in the field is the notion of constraint robustness, which is frequently referred to as model robustness. This pertains to solutions that persist to be viable for every possible value of the uncertain inputs. Additionally, this particular type of solution is deemed necessary in numerous engineering applications.

The following instance is extracted from Ben-Tal and Nemirovski [23], serving as an illustrative paradigm. Specifically, this centers on an complicated engineering process that involves multiple phases, such as a chemical distillation process, and the associated problem of optimizing the process. The under consideration optimization problem incorporates balance constraints, thereby ensuring that the amount of materials that enter a particular phase of the process does not overcome the amount utilized in that phase, along with what is reserved for subsequent phases. The quantities of end products pertaining to a specific phase may be subject to variation due to factors that exist outside the reach of control and thus, are deemed as uncertain. Irrespective of the values of uncontrolled factors, the constraints of balance must be observed. Consequently, it is imperative that the proposed solution exhibits constraint robustness against the fundamental uncertainties present in the underlying problem. In the study under discussion a mathematical model is proposed to identify constraint robust solutions. To elaborate, the approach involves formulating an optimization problem in the following format:

$$\begin{aligned} &min_x \quad f(x) \\ &\qquad\quad G(x,p) \epsilon K. \end{aligned}$$

In example above, the decision variables are denoted by x, while the objective function is represented by the symbol f. Additionally, the constraints of the problem are modeled by the structural elements G and K, which are assumed to be certain. Finally, the uncertain parameters of the problem are denoted by the variable p. It is recommended to contemplate an uncertainty set U, encompassing all conceivable values of the uncertain parameters p. A viable approach for obtaining a solution that is robust to constraints involves resolving the subsequent problem:

$$\begin{aligned} &min_x \;\; f(x) \\ &\quad G(x,p) \in K, \forall\, p \in \mathbb{U}. \end{aligned}$$

The concept of **objective robustness** is closely connected to situations in which the objective function involves uncertain parameters. The aforementioned phenomenon is commonly denoted as solution robustness within the academic literature. In order for solutions to be deemed robust, it is imperative that they maintain a high degree of optimality across all feasible realizations of uncertain parameters. After that, an optimization problem of the following form shall be contemplated:

$$\begin{aligned} &min_x \;\; f(x,p) \\ &\qquad\quad x \in S. \end{aligned}$$

In the aforementioned instance, S is deoicted as the set of feasible solutions and f signifies the objective function that is reliant on uncertain parameters p. Based on the above scenario, it can be assumed that the uncertainty set U encompasses the complete range of prospective values of indeterminate parameters p. Subsequently, a solution that is both objective and robust is attained through the process of solving:

$$min_{x \in S} \;\; max_{p \in U}\, f(x,p).$$

In summary, it should be noted that objective robustness constitutes a distinct subset of constraint robustness. The present formulation involves the introduction of a novel variable, designated as t, which is subjected to minimization within the context of *objective robustness*. Subsequently, a constraining factor is imposed on the function $f(x,p)$ such that its value remains less than or equal to t. This ultimately leads to an equivalent problem to objective robustness. The constraint-robust formulation of the resultant problem exhibits equivalence to *optimal robust optimization problem*.

Although, the notions of **constraint robustness** and **objective robustness** emerge in the context of risk avoidance decision making, the latter two may not be suitable for optimization problems that involve uncertain data.

5 A Real Case Study on Micro-Finance

Micro-finance, alternatively referred to as **micro-credit**, represents a form of banking assistance rendered to individuals or groups of limited economic means, who are typically precluded from obtaining access to conventional financial services. The aim of micro-finance is to provide the economically disadvantaged individuals with the means to attain self-sufficiency in the long run.

Participating institutions in micro-finance typically offer loan facilities, wherein micro-loans may vary in value from \$50 to \$50,000. In detail, micro-finance facilitates individuals in acquiring sensible small-scale business loans in a secure manner, that's upholding principles of ethical lending. While micro-financing operations are present worldwide, they are mostly concentrated in developing countries, such as Bangladesh, Cambodia, India, Afghanistan, the Democratic Republic of Congo, Indonesia and Ecuador.

Specifically, micro-finance solutions are extended to low-income or jobless individuals due to the fact that a large number of those ensnared in poverty, or with constricted financial means, are unable to engage in transactions with conventional financial institutions owing to insufficient earnings.

Similar to traditional financial institutions, micro-finance institutions are obliged to impose interest rates on their loans, and they establish distinct repayment schemes with payment deadlines occurring at fixed intervals. Certain lending institutions enforce in borrowers to allocate a piece of their earnings into a savings account, which serves as a safeguard in case that the individual is unable to repay the debt. Consequently, this leads to the individual who has received the loan successfully fulfills their obligation to repay the amount borrowed, and to accumulate additional savings.

However, due to the inability of numerous applicants to provide collateral, micro-lenders frequently unite borrowers as a means of creating a safeguard. Upon obtaining loans, individuals proceed to acquire a collective responsibility to repay their financial obligations. As the efficacy of the program is interdependent on the collective contributions of all partners, it engenders a form of peer pressure that may effectively raise borrower's obligation towards timely loan repayment.

Remarkably, despite being borrowers who typically fall under the category of impoverished individuals, the repayment figures on micro-credits surpass the average repayment rate offered by more conventional modes of funding.

It is important to mention, that the advantages of micro-finance extend beyond the immediate impact of providing individuals with a means to obtain capital. Successful business projects not only contribute to the creation of employment opportunities but also facilitate trade and generate positive economic development within the local community.

Fintech, an abbreviation for financial technology, refers to the utilization of advanced technologies with the goal of optimizing and simplifying the delivery and utilization of financial services. The key objective of fintech is to enhance the efficiency of financial operations for various entities such as companies, business owners, and consumers. This is achieved through the utilization of specialized software, algorithms, and digital platforms, which can be conveniently accessed via computer systems and smartphones.

In the context of mobile micro-loans, fintech facilitates the seamless application of loans via smartphones, thereby offering customers a convenient means of accessing financial services. Micro-loans commonly feature brief repayment periods, such as 7, 14, or 30 days. To ensure a seamless user experience and maintain responsible lending practices, customers are required to repay their previous debts before being able to reuse the serviceThe repayment process is facilitated with the assistance of the mobile operator. The dataset employed for analysis encompasses variables related to the customer's wallet, telecommunications company, and loan characteristics.

In brief, fintech enables users to obtain small-scale loans via a mobile application, thereby simplifying the provision of financial services and enhancing the efficiency of financial transactions for both loan recipients and providers.

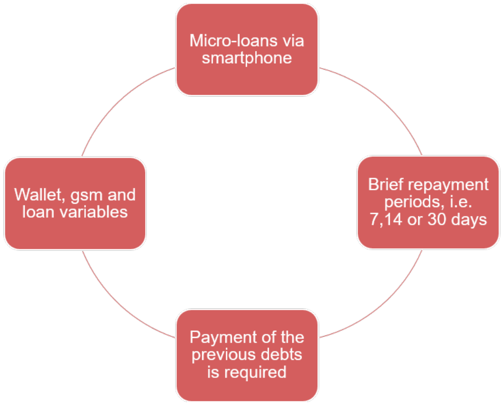

Micro-loans via smartphone
Brief repayment periods, i.e. 7,14 or 30 days
Payment of the previous debts is required
Wallet, gsm and loan variables

Figure 16: The fintech process for microloans

5.1 Case Study

In the present study, we focus on clients who have a prior record of micro-loan borrowing and those who do not. To that end, two datasets pertaining to real-world data were employed for the purpose of investigating the efficacy of logistic regression and determing the optimal results through the utilization of the simplex method. The initial dataset comprises individuals who have received loans as well as those who have not, whereas the subsequent dataset exclusively consists of customers who did not receive any loans. Furthermore, the initial dataset was subjected to two subsequent training phases, the first for all clients and followed a anologous procedure after the removal of the loan variables. Therefore, the primary objectives of this analysis are:

i The utilization of the logistic regression and random forest techniques;

ii The calculation of the probabilities for all clients of each dataset;

iii The construction of an optimization problem of the new limits of micro-loans; and

iv The unbiased distribution of the outcomed available amounts.

Please note that the computational statistics presented in this study were produced using a personal computer outfitted with an Intel(R) Core(TM) i5-7200U CPU, which operates at a base frequency of 2.50GHz and a maximum turbo frequency of 2.71GHz The computer system was equipped with a Random Access Memory (RAM) of 8.00 GB.

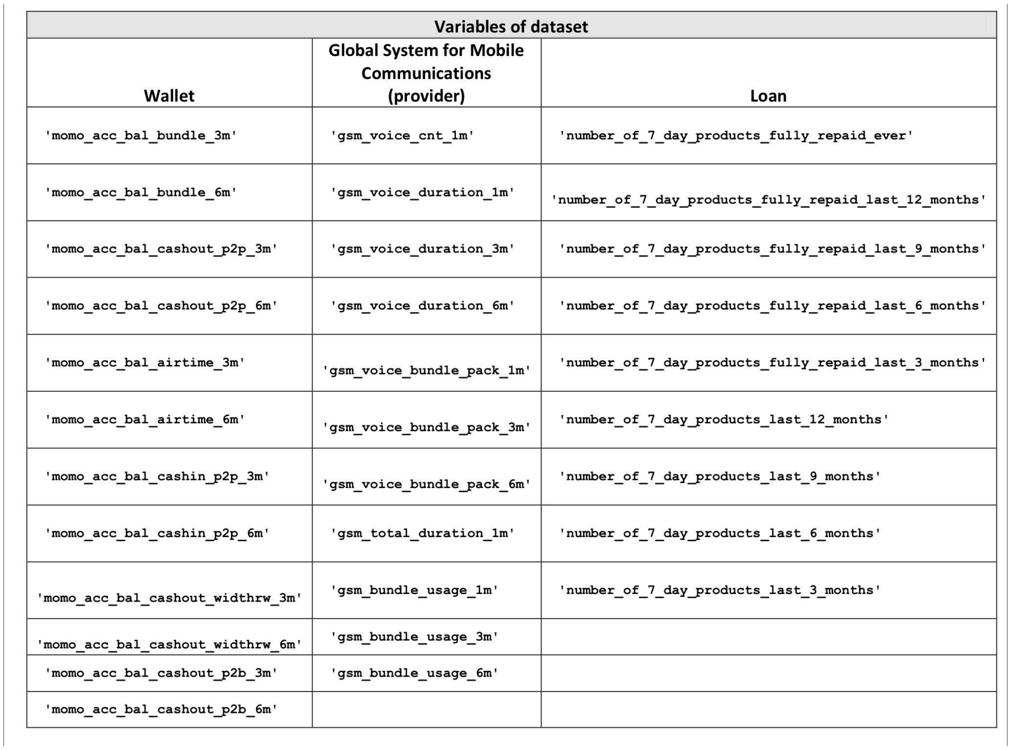
5.1.1 Data Pre-Processing

Data pre-processing is a crucial step in the training of machine learning models in order to enhance their performance and accuracy. The efficacy of the machine learning model depends upon the quality of the underlying data. The manner in which data are presented, has the potential to exert a significant impact on the capacity of machine learning models to learn from it. Models exhibit a tendency to achieve faster and more reliable convergence when numerical data is subjected to appropriate scaling techniques. To sum up, the selection and transformation methods employed in data analysis are critical factors that contribute to enhancing the predictive accuracy of models.

Therefore, the primary and pivotal step of this study involves the cleansing of the dataset. The first dataset comprises $50,000$ rows and 367 columns. Thus, there is a total of 367 variables, where one of these variables is identified as the dependent variable, while the remaining variables are characterized as independent. The discreteness of the dependent variable is attributed to its binary nature, representing only two possible outcome rates. Specifically, the dependent variable is identified as the characterization of customers as either 0:"good" or 1:"bad". The remaining variables are related to mobile wallet, gsm and loan variables, which are segmented based on repayment periods of 7, 14, and 30 days, indicative of the time required to completely repay the micro-loan.

The variables contained within the dataset, with the exception of the dependent variable appropriately labeled as the "default flag," are displayed in Table 1. The term "momo" pertains to a set of wallet variables characterized by a data type of Float, which is commonly utilized for the storage of numerical values containing decimal points or floating-point representations. Subsequently, "gsm" denotes the telecommunications company and simultaneously exhibits floating characteristics. The variables pertaining to loan characteristics, specifically referred to as the "number of 7 day repayments", which are represented by the data type Integer (int). This data type is utilized to store numerical values inclusive of whole numbers, whether positive or negative, in a format devoid of decimal points.

The **feature selection** process is initiated by the application of several



Variables of dataset		
Wallet	Global System for Mobile Communications (provider)	Loan
'momo_acc_bal_bundle_3m'	'gsm_voice_cnt_1m'	'number_of_7_day_products_fully_repaid_ever'
'momo_acc_bal_bundle_6m'	'gsm_voice_duration_1m'	'number_of_7_day_products_fully_repaid_last_12_months'
'momo_acc_bal_cashout_p2p_3m'	'gsm_voice_duration_3m'	'number_of_7_day_products_fully_repaid_last_9_months'
'momo_acc_bal_cashout_p2p_6m'	'gsm_voice_duration_6m'	'number_of_7_day_products_fully_repaid_last_6_months'
'momo_acc_bal_airtime_3m'	'gsm_voice_bundle_pack_1m'	'number_of_7_day_products_fully_repaid_last_3_months'
'momo_acc_bal_airtime_6m'	'gsm_voice_bundle_pack_3m'	'number_of_7_day_products_last_12_months'
'momo_acc_bal_cashin_p2p_3m'	'gsm_voice_bundle_pack_6m'	'number_of_7_day_products_last_9_months'
'momo_acc_bal_cashin_p2p_6m'	'gsm_total_duration_1m'	'number_of_7_day_products_last_6_months'
'momo_acc_bal_cashout_widthrw_3m'	'gsm_bundle_usage_1m'	'number_of_7_day_products_last_3_months'
'momo_acc_bal_cashout_widthrw_6m'	'gsm_bundle_usage_3m'	
'momo_acc_bal_cashout_p2b_3m'	'gsm_bundle_usage_6m'	
'momo_acc_bal_cashout_p2b_6m'		

Table 1: Variables of dataset

filters, wherein the independent variables are investigated for the presence of **duplicated**, **constant**, and **quasi-constant features**.

Following the implementation of the aforementioned filters, the resultant variables were excluded and the dataset now comprises 257 independent variables.

The subsequent stage entails an evaluation of **correlated values**. When two variables display a high correlation, it indicates a robust association between them. Conversely, a low correlation suggests a weakened relationship between the variables. Furthermore, it is important to consider instances of weak or zero correlation in which no correlation exists if one variable does not impact the other.

In the present study, Spearman's correlation coefficient was utilized to assess the degree of correlation between the variables at hand, given the non-linear nature of the associations under examination between two independent variables.

The last but not least step in the data preprocessing procedure involves reducing the dimensionality of variables. The process of **dimensionality reduction** serves as a methodology utilized to decrease the quantity of characteristics present within a given data set, whilst preserving a maximum amount of crucial information. In essence, the process involves the conversion of multidimensional data into a reduced dimensional space while maintaining its fundamental characteristics.

Generaly, in the domain of machine learning, datasets that are characterized by an excess of features or variables are denoted as high-dimensional. This phenomenon is characterized by the diminishing efficacy of a model when a large number of characteristics or attributes are incorporated into it. The cause of this is ascribed to the increase in model complexity, accompanied by a rise in the number of features, resulting in the progressively difficult task of achieving a feasible resolution. Additionally, there is a potential for overfitting to occur in datasets with a high number of dimensions, whereby the model becomes overly customized to the training data, thereby impeding its capacity to accurately generalize to novel data.

The employment of dimensionality reduction holds promising potential to mitigate these concerns through the reduction of model complexity and improvement its capacity to generalize. There are two predominant methodologies utilized for decreasing the dimensionality of a dataset, specifically, **feature selection** and **feature extraction**.

- The process of **feature selection** entails choosing a subset of the initial set of features that hold the utmost significance to the problem. The objective is to decrease the dimensionality of the dataset while preserving its critical characteristics. There exist diverse approaches for extracting features such as **filter methods**, **wrapper methods** and **embedded methods**.

- The process of **feature extraction** encompasses the creation of novel features through the combination or transformation of pre-existing features. Such a process is crucial in enhancing the effectiveness and efficiency of machine learning systems. The objective is to devise a collection of attributes that effectively represent the inherent characteristics of the source data within a reduced-dimensional framework.

> Numerous techniques are available for feature extraction, such as **principal component analysis (PCA)**, **linear discriminant analysis (LDA)**, and **t-distributed stochastic neighbor embedding (t-SNE)**.

In this work, we utilize the PCA method. In order to test the validity of applying the method to our case, Bartlett's test was conducted. The obtained p-value ($\cong 0.0$) was found to be less than the predetermined level of significance 5%, and thus, techniques related to dimension reduction can be utilized.

After conducting the Bartlett's sphericity test and determining that the dataset would benefit from dimension reduction, we proceeded to the implemention of the PCA method. The first stop was to calculate the eigenvectors and estimated communalities, followed by the removal of any variables that displayed communalities below 0.40. The resultant dataset consisted solely of the remaining variables. Subsequent to the initial iteration, the aforementioned process was reiterated until all extraneous variables were eliminated from the dataset.

Consequently, our ultimate dataset, obtained from the principal component analysis iterations, comprised 32 variables.

Upon completing the process of data cleansing and selecting only the noteworthy variables from the dataset,the next step was to train our model through the **logistic regression technique**.

5.1.2 Logistic Regression and Random Forest Model Fitting

The initial stage in the implementation of logistic regression necessitates the division of data into two distinct subgroups, namely **training** and **test set**, with proportion (here) 80% to 20%.

Afterwards, we employ the training set to instruct our model utilizing the logistic regression technique. Following the completion of a fitting process for our model on the training dataset, the accuracy of the model on the test dataset is subsequently calculated. This coefficient serves as an indicator of the degree of sufficiency with which the model at hand fits the observed data. The experimental results of the present study indicate that our model exhibits an **accuracy rate of 55.32%** and **AUC** is **0.61**. It is important to recall that the data utilized in this study are real and, consequently, the model's accuracy percentage is not particularly noteworthy.

The ROC curve, depicted by the dashed line in the Figure 17, serves as a reference point for a random classifier. A reliable classifier strives to attain a position as distant as possible from the dotted line, ideally in the top-left corner.

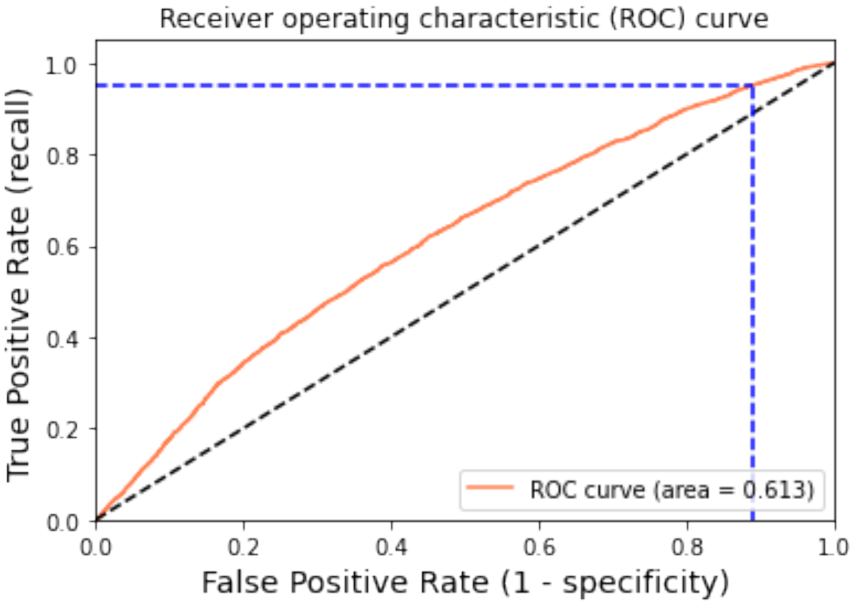


Figure 17: The ROC curve of the model using Logistic Regression model

It is noteworthy that, using a threshold of 0.388 guarantees a sensitivity of 0.950 and a specificity of 0.112, i.e. a false positive rate of 88.82%.

Following this, the probabilities of every customer within our dataset are calculated. The results obtained suggest the probability of a customer exhibiting unfavorable behavior. Consequently, to determine the probabilities associated with positive customer behavior, it is imperative to calculate the complement of the observed probabilities.

Figure 18 depicts the histogram of probability distribution of customers who possess behaviors indicative of good customers.

The subsequent step of our analysis entails the training of the model while

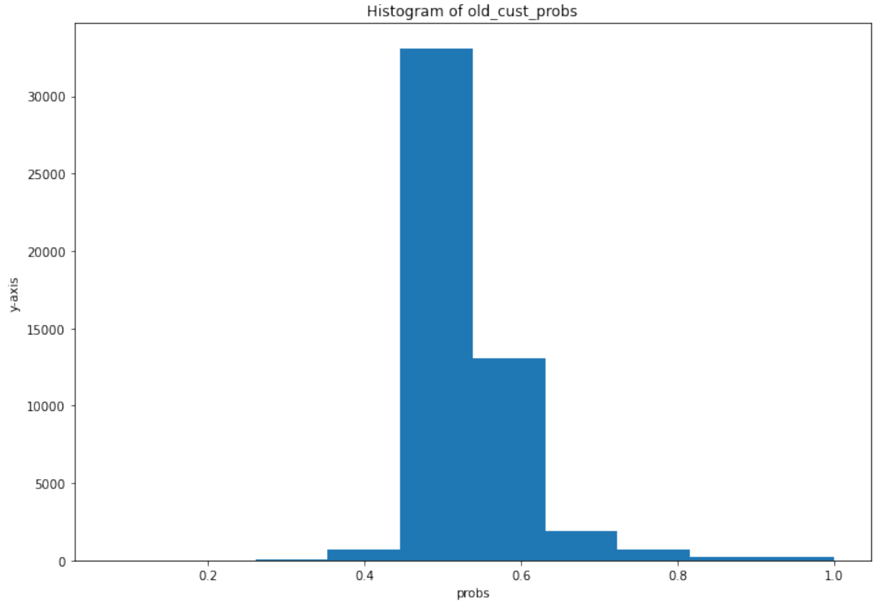


Figure 18: Histogram of probability of old customers using Logestic Regression model

excluding loan variables. In addition, we utilize the new set of data containing only new customers, specifically those lacking loan history, to compute their probabilities.

Following the extraction of the loan-related variables, the dataset encompasses a total of 24 variables, including the dependent variable. The previously employed methodology will be repeated in order to fit the model. Once again, the present dataset has been partitioned into distinct subsets, 80% for training and 20% testing.

The accuracy of the model has been estimated to be 55.24% a level of precision that for both models exhibits a insignificant difference. Furthermore, Figure 19 depicts the ROC curve, accompanied by the associated area under the curve (AUC) coefficient (61.32).

The utilization of a threshold value of 0.390 ensures a sensitivity rate of 0.950 and a specificity rate of 0.112, namely the rate of false positives has been determined to be 88.76% according to the results obtained.

After the completion of the training process for the novel model, which is

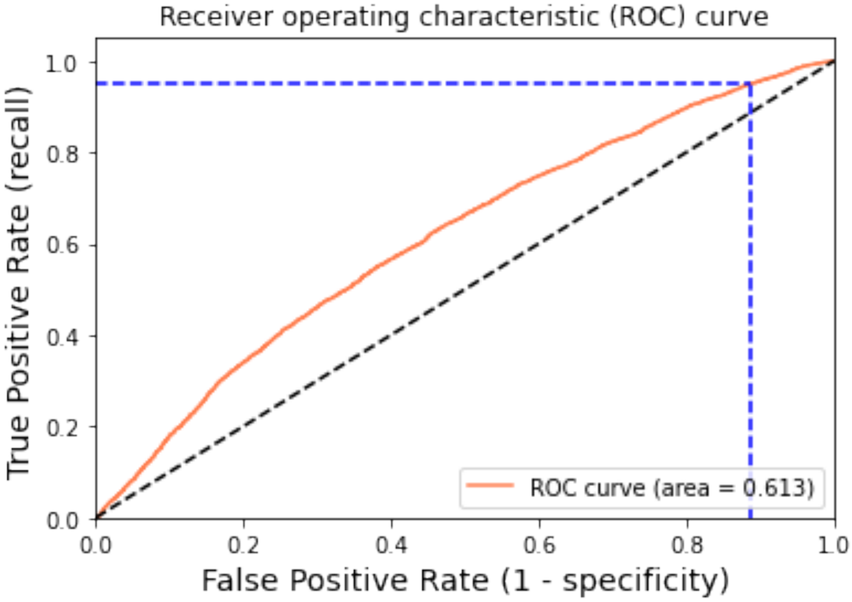


Figure 19: The ROC curve for dataset without loan variables using Logestic Regression model

devoid of the loan variables, the subsequent step entails the establishment of the probabilities of potential customers. As a result, the significant variables that appeared in the initial dataset are retained from the new dataset, and the probabilities are computed. The original probabilities pertain to the likelihood of a customer being considered as ”bad” and are consequently transformed by computing their complementary values. This approach is adopted to obtain a measure of the probability that a customer is ”good”.

Figure 20 illustrates the histogram of the probability distribution pertaining to newly acquired customers who display behaviors suggestive of being favorable customers.

The procedural steps employed within the logistic regression method are delineated in Figure 21.

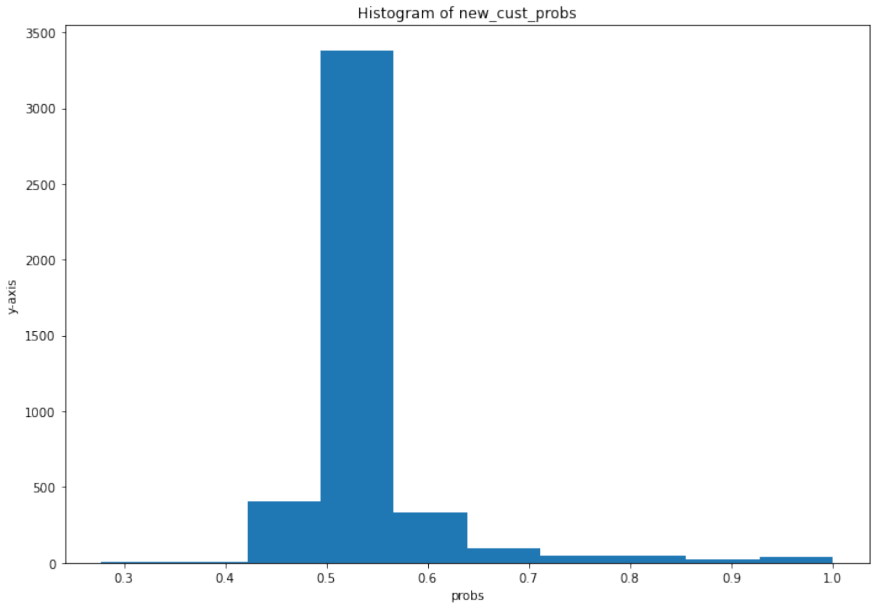


Figure 20: Histogram of probability of new customers using Logestic Regression model

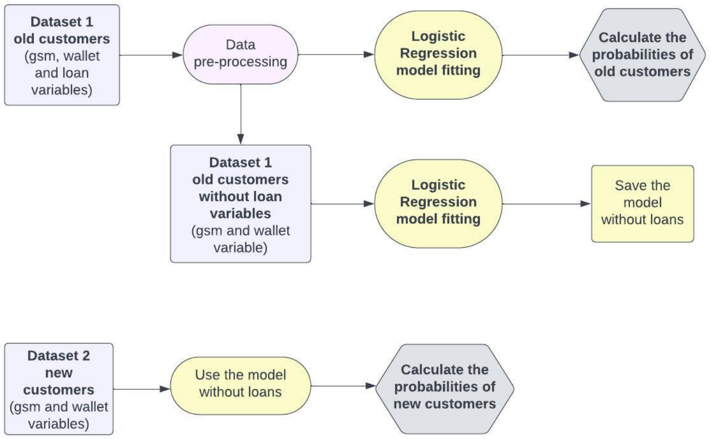

Dataset 1 old customers (gsm, wallet and loan variables)
Data pre-processing
Logistic Regression model fitting
Calculate the probabilities of old customers
Dataset 1 old customers without loan variables (gsm and wallet variable)
Logistic Regression model fitting
Save the model without loans
Dataset 2 new customers (gsm and wallet variables)
Use the model without loans
Calculate the probabilities of new customers

Figure 21: A diagrammatic representation of the first part of the application using Logistic Regression model.

For the purpose of evaluating and contrasting, we also employed the **Random Forest** statistical approach for training our models. In order to improve outcomes, the tuning technique is implemented on the Random Forest model. The optimization of Random Forest models through tuning facilitates the enhancement of their performance, regulation of complexity, achievement of an optimal trade-off between bias and variance, identification of significant features, and mitigation of specific challenges encountered in the dataset, such as class imbalance. The cross-validation method is employed to determine the optimal hyperparameters. The evaluated parameters encompassed the number of decision trees (n_estimators), the maximum depth of each tree (max_depth), and the maximum number of features considered in the tree's best split (max_features).

The initial dataset, incorporating loan variables, yielded an accuracy of $0.648(+/-0.008)$ and an AUC value of $0.708(+/-0.007)$. Furthermore, the ROC curve diagram is also furnished in Figure 22.

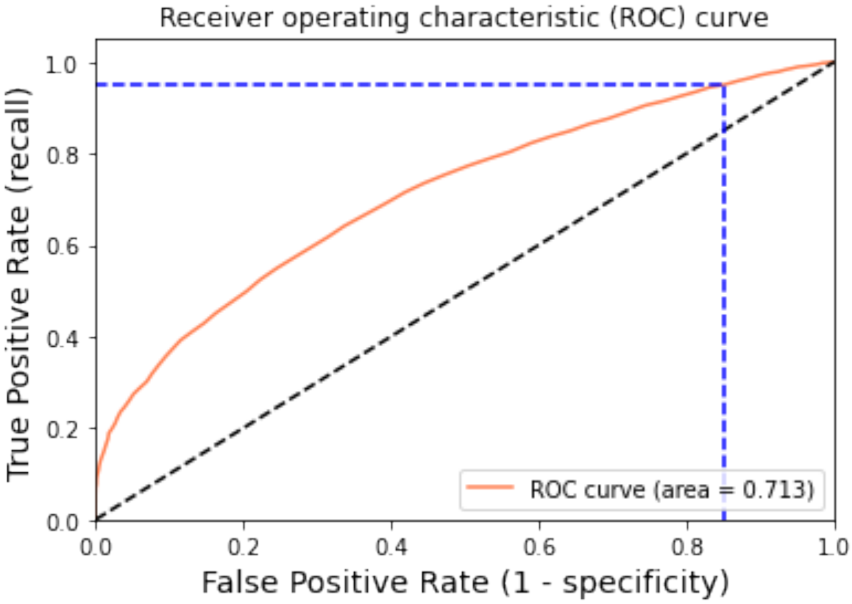


Figure 22: The ROC curve for dataset with loan variables using the Random Forest model

Note that, using a threshold of 0.290 guarantees a sensitivity of 0.950 and a specificity of 0.148, i.e. a false positive rate of 85.20%.

Subsequently, the probabilities of the previous clientele were computed and the resulting histogram of the Figure 23 was derived. It has been observed that the probability distribution in this instance exhibits notable differences when compared with its previous counterpart.

Additionaly, an identical approach was employed, as described earlier to

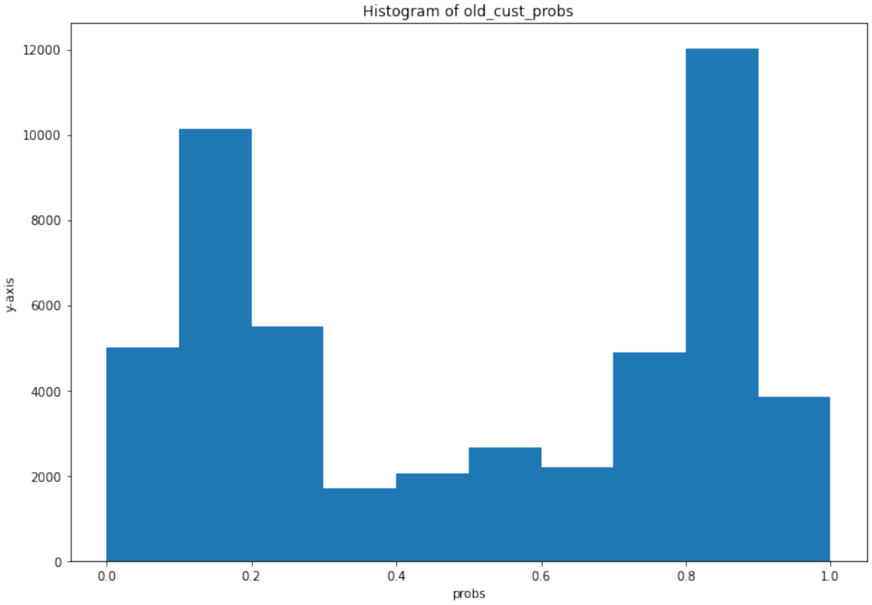


Figure 23: Histogram of probability of old customers using Random Forest model

train the model without the utilization of loan features, instead relying on Random Forest. In the present instance, the metric of accuracy attains a value of $0.612(+/-0.008)$, while the AUC metric exhibits equivalence to $0.656(+/-0.008)$. Furthermore, Figure 24 illustrates the plot of the ROC curve within this particular scenario.

In this case, using a threshold of 0.300 guarantees a sensitivity of 0.953 and a specificity of 0.116, i.e. a false positive rate of 88.38%.

Furthermore, it is worth noting that the probability distribution derived from the aforementioned methodology is depicted in the Figure 25.

The methodology employed in relation to the random forest technique is elucidated in Figure 26.

5.1.3 Optimization Problem

In order to determine the most suitable quantum of funds to extend as micro-finance loans to both existing and potential clients, we formulate an optimization problem.

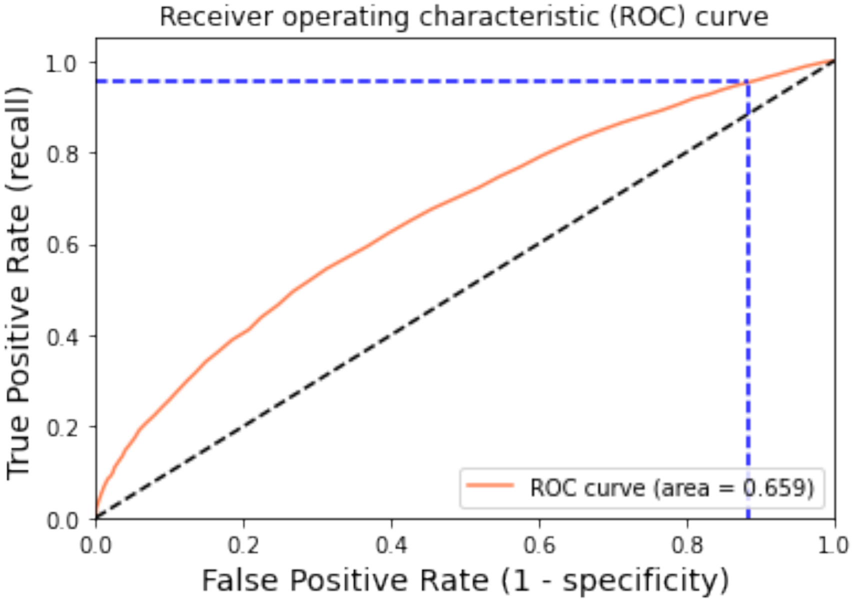


Figure 24: The ROC curve for dataset without loan variables using the Random Forest model

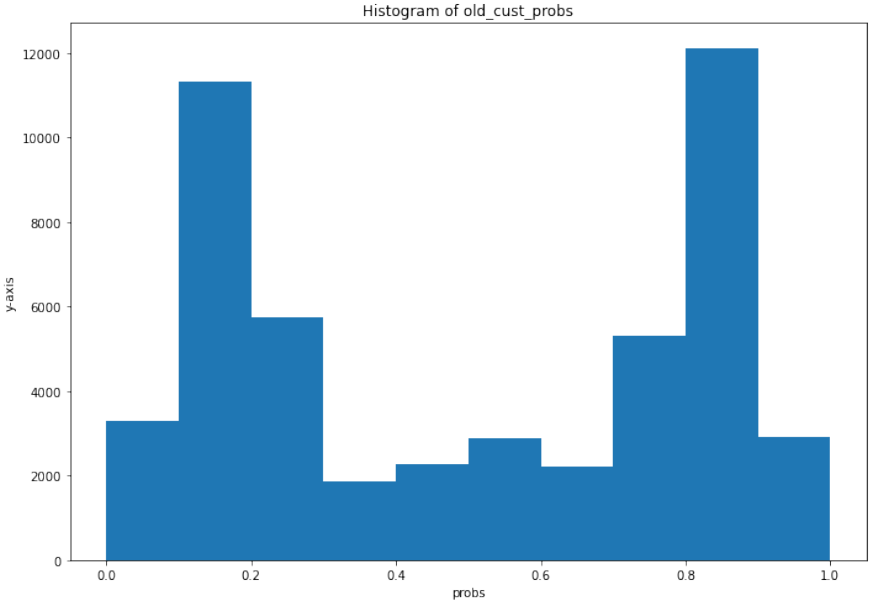


Figure 25: Histogram of probability of new customers using Random Forest model

It is important to note that we have access to the historical credit limits of our well-established clients, which aid in the determination of appropriate

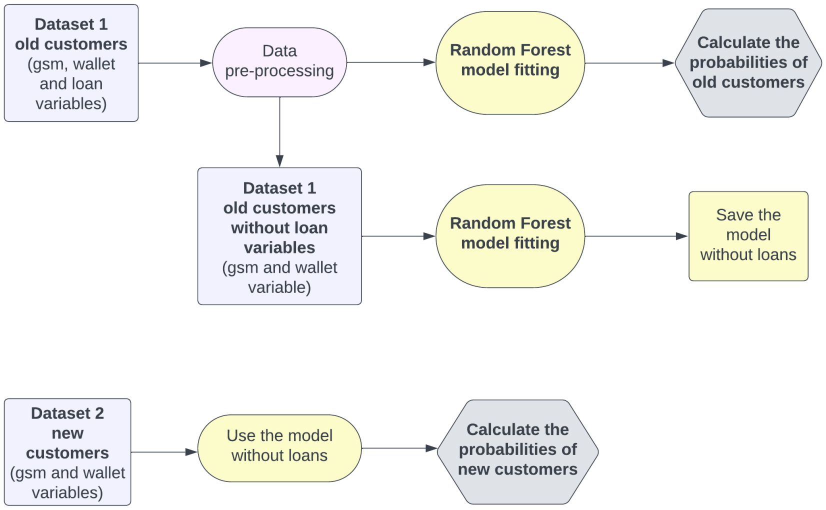


Figure 26: A diagrammatic representation of the first part of the application using Random Forest model.

levels for new credit limits to be conferred. In addition, it is imperative to acknowledge the specific timelines designated for loan repayment, inclusive of a grace period of 14 and 30 days. Hence, the issue at hand shall be partitioned on the grounds of the credit duration of the loan, thereby engendering the formation of two analogous optimization problems.

As concerns new customers, a grace period of 14 days shall be provided due to a lack of knowledge regarding their history and consequently, a lack of estimate with regards to their reliability.

Afterwards, we classify both new and old customers based on their probability distribution into three groups: ”good”, ”medium”, and ”bad” with the term ”good customer”, it is understood that there is a probability of repaying the entire borrowed amount, as well as the fee. For ”medium customers”, it is estimated that they will partially repay the amount, while with ”bad customers”, there is an expectation that they will not return the money borrowed.

To begin with, we investigate the case where the grace period is set to a duration of **14 days**. Herein are presented two discrete sets of data, one specifically associated with existing customers and the other with newly acquired customers. Therefore, we classify the old customers into three groups. The first group, denoted as XG1, consists of customers with a probability range of $[0, 0.50)$, which are considered to be the ”bad” ones. The second group, denoted as XG2, is composed of customers with a probability range of $[0.50, 0.70)$, which are regarded as ”medium”. Finally, the third group, identified as XG3, represents customers with a probability range of $[0.70, 1]$, which are considered as ”good”. Similarly, we allocate new customers into three groups denoted as YG1, YG2, and YG3 utilizing the same approach.

Consequently to this, various statistical measures pertaining to past clients within every group are calculated, including the count of individuals in each group and the number of recipients of loans during the most recent period. Moreover, the aggregate sum of borrowed funds acquired in the preceding interval, coupled with the entire magnitude of funds paid out to all customers, is also ascertained.

Following this, the objective coefficients of each team are computed and the constraints of the optimization problem are enforced with the aim of enhancing the satisfaction of existing customers.

The optimization problem of linear programming for the old customers comprises 3 decision variables and 8 constraints which are expressed as follows:

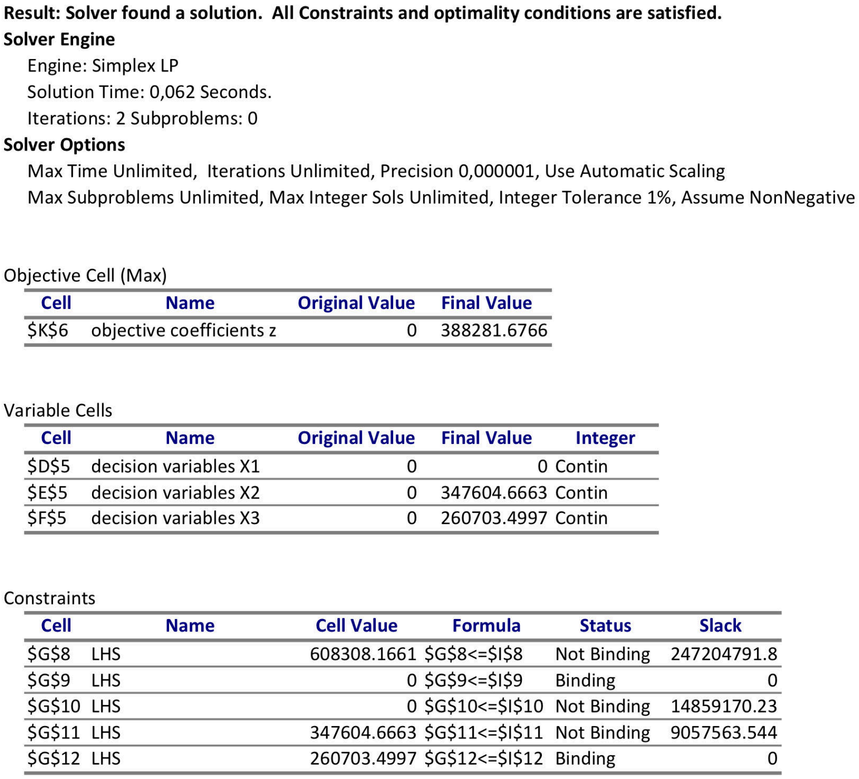
$$\begin{aligned}
\max\ z = \omega_1 X_1 + \omega_2 X_2 + \omega_3 X_3.\\
\text{s.t}\\
X_1 + X_2 + X_3 \leq 2t\\
X_1 + 0.75X_2 - X_3 \leq 0\\
X_1 \leq \sum_{i=1}^{n_1} \omega_i \lambda_i\\
X_2 \leq \sum_{i=1}^{n_2} \omega_i \lambda_i + 0.5sd\left(\sum_{i=1}^{n_2} \omega_i \lambda_i\right)\\
X_3 \leq \sum_{i=1}^{n_3} \omega_i \lambda_i + 2sd\left(\sum_{i=1}^{n_3} \omega_i \lambda_i\right)\\
X_1 \geq 0,\ X_2 \geq 0,\ X_3 \geq 0
\end{aligned}$$

where:

- ω_i, $i = 1,...,3$: objective coefficients.
- X_i, $i = XG1, XG2, XG3$: decision variables.
- z: the total amount of the objective function.

Furthermore, it should be noted that the variable "t" represents the aggregate sum of micro-loans paid out during the most recent time period in total for all 3 groups. Similarly, "λ_i", $i = 1,...,n_i$ denotes the quantity of funds received by a given individual during said time period for each group separately. Additionally, "ω_i", $i = 1,...,n_i$ pertains to the probability associated with each client, while "n_i" pertains to the population size of each group that availed themselves of micro-loans during the aforementioned time period.

The results arising from the implementation of the simplex method are illustrated in the subsequent Tables 2, 3 and 4. The present study establishes that the maximum target value that the objective function can achieve is equivalent to 388281.67. Of this amount, 347604.66 monetary units pertain to the mean customers, and the residual 260703.50 monetary units are



Result: Solver found a solution. All Constraints and optimality conditions are satisfied.

Solver Engine

Engine: Simplex LP

Solution Time: 0,062 Seconds.

Iterations: 2 Subproblems: 0

Solver Options

Max Time Unlimited, Iterations Unlimited, Precision 0,000001, Use Automatic Scaling

Max Subproblems Unlimited, Max Integer Sols Unlimited, Integer Tolerance 1%, Assume NonNegative

Objective Cell (Max)

Cell	Name	Original Value	Final Value
$K$6	objective coefficients z	0	388281.6766

Variable Cells

Cell	Name	Original Value	Final Value	Integer
$D$5	decision variables X1	0	0	Contin
$E$5	decision variables X2	0	347604.6663	Contin
$F$5	decision variables X3	0	260703.4997	Contin

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$G$8	LHS	608308.1661	$G$8<=$I$8	Not Binding	247204791.8
$G$9	LHS	0	$G$9<=$I$9	Binding	0
$G$10	LHS	0	$G$10<=$I$10	Not Binding	14859170.23
$G$11	LHS	347604.6663	$G$11<=$I$11	Not Binding	9057563.544
$G$12	LHS	260703.4997	$G$12<=$I$12	Binding	0

Table 2: Answers report for XG1, XG2 and XG3 in grace period 14.

attributed to the excellent customers. It is noticeable that micro-loans are not granted to individuals classified as bad clients.

Please note that the currency throughout the analysis concerns the Mozambican currency, which is commonly known as ”metical” (MTn). In terms of equivalence, 1 MTn holds a value of 0.014 euros.

Sensitivity analysis pertains to the study of the consequences that arise on the optimal solution due to variations in the parameters of a model. This process entails the pursuit of identifying the optimal solution and provides the opportunity for the exploration of scenarios and inquiries. The sensitivity analysis output yields significant insights into how modifications in variables and constraints influence the optimal solution for micro-loan management. Specifically, the outputs comprise crucial indicators, such us shadow prices, allowable increase/decrease, and reduced cost. The aforementioned indicators serve to quantify the rate of alteration in the value of the objective function as a result of modifications made to particular constraints. The provision of significant information pertaining to the sensitivity of the optimal solution to variations in constraints is a valuable resource which allows decision-makers to comprehend the trade-offs involved in the allocation of micro-loans.

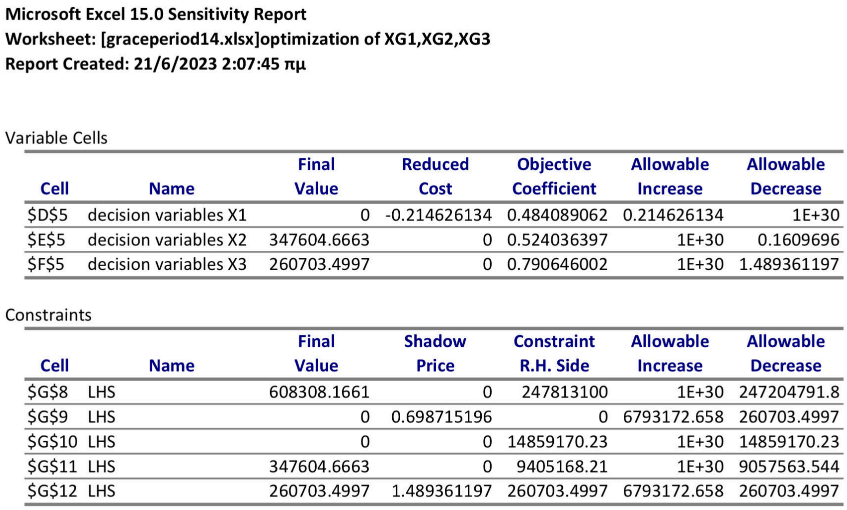
- **Shadow Prices** denote the magnitude of the transformation in the value of the objective function concerning the variation in a given constraint per unit. These entities denote the incremental benefit of loosening or imposing a restriction. The manifestation of a positive shadow price infers that the enhancement of the constraint value will yield a progression in the value of the objective function. Conversely, a negative shadow price suggests that diminishing the constraint value would result in a more favorable outcome.

- The **Allowable Increase and Decrease** parameters refer to the upper and lower bounds of change in the right-hand side of a constraint, such that the present optimal solution preserves its optimality. These values delimit the permissible boundaries within which the constraint may vary without impacting the optimal solution.

- The **Reduced Cost** quantifies the magnitude of adjustment required in the objective function coefficient of a non-basic variable in order for it to become a part of the solution at its present level. A reduction in cost to zero denotes optimality of the present solution, while a positive or negative reduction in cost suggests that modifying the corresponding objective function coefficient would lead to an improvement in the solution.

According to Table 3, it is evident that the objective coefficient pertaining to variable $X1$ can be augmented to a maximum value of 0.7, or conversely decreased to a minimum value of $-1E+30$, without eliciting any alterations to the feasible solution. Likewise, this assertion holds true for the remaining variables.

Additionally, it can be observed from Table 3 that the upper limit of the right-hand side for the initial constraint can be enhanced by the magnitude of $1E+30$, while it can be diminished to a minimum of 247204791.8, so the final result equals 608308.2, without effecting the feasible solution. Similarly, analogous interpretations are elicited for the remaining constraints.Similarly, analogous interpretations are elicited for the remaining



Microsoft Excel 15.0 Sensitivity Report
Worksheet: [graceperiod14.xlsx]optimization of XG1,XG2,XG3
Report Created: 21/6/2023 2:07:45 πμ

Variable Cells

		Final	Reduced	Objective	Allowable	Allowable
Cell	Name	Value	Cost	Coefficient	Increase	Decrease
$D$5	decision variables X1	0	-0.214626134	0.484089062	0.214626134	1E+30
$E$5	decision variables X2	347604.6663	0	0.524036397	1E+30	0.1609696
$F$5	decision variables X3	260703.4997	0	0.790646002	1E+30	1.489361197

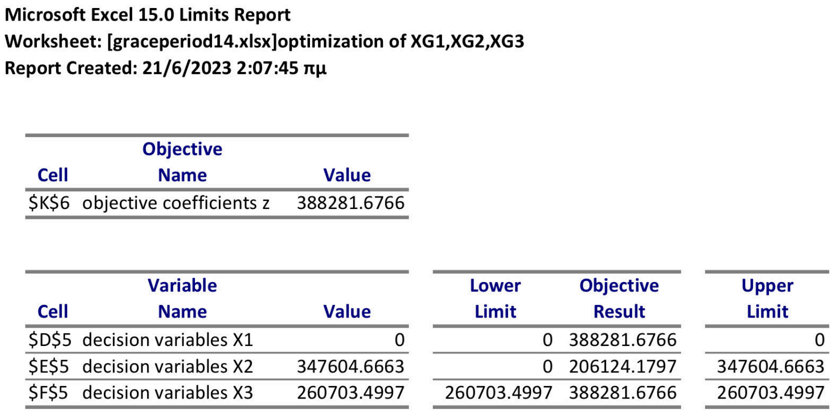
Constraints

		Final	Shadow	Constraint	Allowable	Allowable
Cell	Name	Value	Price	R.H. Side	Increase	Decrease
$G$8	LHS	608308.1661	0	247813100	1E+30	247204791.8
$G$9	LHS	0	0.698715196	0	6793172.658	260703.4997
$G$10	LHS	0	0	14859170.23	1E+30	14859170.23
$G$11	LHS	347604.6663	0	9405168.21	1E+30	9057563.544
$G$12	LHS	260703.4997	1.489361197	260703.4997	6793172.658	260703.4997

Table 3: Sensitivity analysis for XG1, XG2 and XG3 in grace period 14.

constraints.

The **Limits report** presents a distinct form of sensitivity analysis data with a specialized focus. This technique involves the iteration of the Solver model whereby every decision variable is systematically designated as the objective, both in the maximization and minimization instances, while keeping all other variables constant.



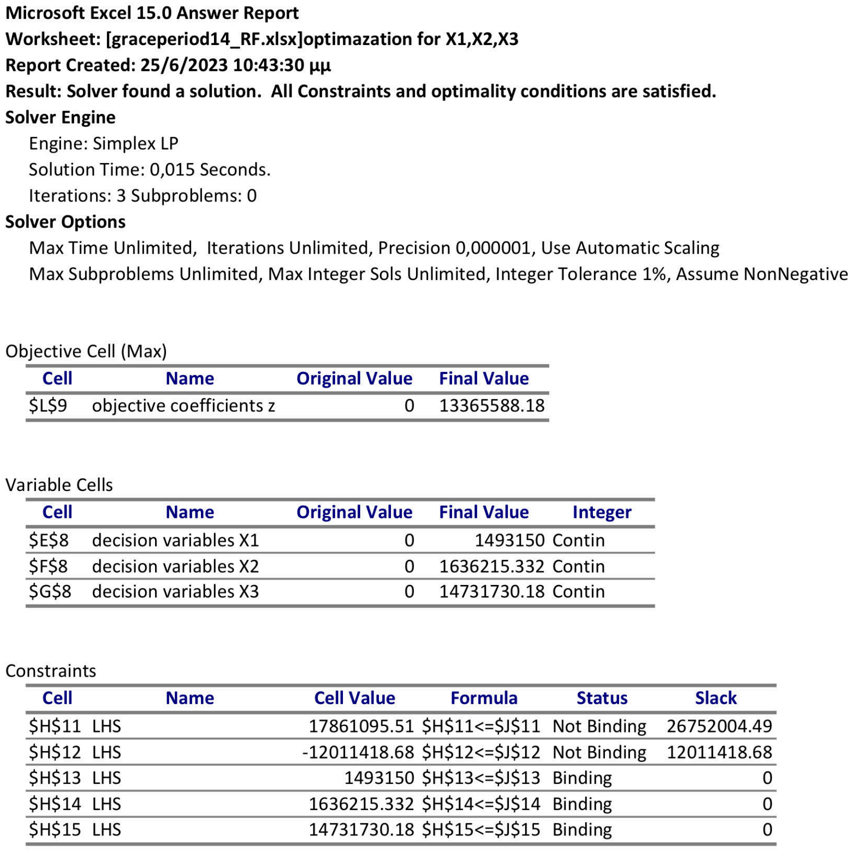
Microsoft Excel 15.0 Limits Report
Worksheet: [graceperiod14.xlsx]optimization of XG1,XG2,XG3
Report Created: 21/6/2023 2:07:45 πμ

	Objective	
Cell	Name	Value
$K$6	objective coefficients z	388281.6766

	Variable		Lower	Objective	Upper
Cell	Name	Value	Limit	Result	Limit
$D$5	decision variables X1	0	0	388281.6766	0
$E$5	decision variables X2	347604.6663	0	206124.1797	347604.6663
$F$5	decision variables X3	260703.4997	260703.4997	388281.6766	260703.4997

Table 4: Limits report for XG1, XG2 and XG3 in grace period 14.

The outcomes obtained for the clientele in accordance with the probabilities derived from the random forest model, whilst employing identical methods, are presented in the Tables 5, 6 and 7.



Microsoft Excel 15.0 Answer Report
Worksheet: [graceperiod14_RF.xlsx]optimazation for X1,X2,X3
Report Created: 25/6/2023 10:43:30 μμ
Result: Solver found a solution. All Constraints and optimality conditions are satisfied.
Solver Engine
Engine: Simplex LP
Solution Time: 0,015 Seconds.
Iterations: 3 Subproblems: 0
Solver Options
Max Time Unlimited, Iterations Unlimited, Precision 0,000001, Use Automatic Scaling
Max Subproblems Unlimited, Max Integer Sols Unlimited, Integer Tolerance 1%, Assume NonNegative

Objective Cell (Max)

Cell	Name	Original Value	Final Value
$L$9	objective coefficients z	0	13365588.18

Variable Cells

Cell	Name	Original Value	Final Value	Integer
$E$8	decision variables X1	0	1493150	Contin
$F$8	decision variables X2	0	1636215.332	Contin
$G$8	decision variables X3	0	14731730.18	Contin

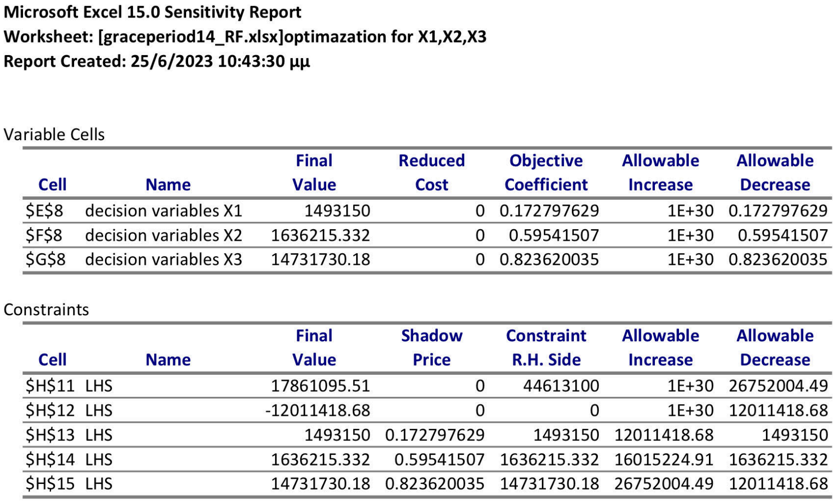
Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$H$11	LHS	17861095.51	$H$11<=$J$11	Not Binding	26752004.49
$H$12	LHS	-12011418.68	$H$12<=$J$12	Not Binding	12011418.68
$H$13	LHS	1493150	$H$13<=$J$13	Binding	0
$H$14	LHS	1636215.332	$H$14<=$J$14	Binding	0
$H$15	LHS	14731730.18	$H$15<=$J$15	Binding	0

Table 5: Answers report for XG1, XG2 and XG3 in grace period 14 using RF.

After determining the optimal amounts for each group of existing customers, it is now time to calculate the amounts that can be offered to new customers. Initially, we classify them into three corresponding groups, YG1, YG2, and YG3, based on their probabilities, as done in the old customers case.

The present investigation concerns an optimization problem that can be formally articulated as follows:



Microsoft Excel 15.0 Sensitivity Report
Worksheet: [graceperiod14_RF.xlsx]optimazation for X1,X2,X3
Report Created: 25/6/2023 10:43:30 μμ

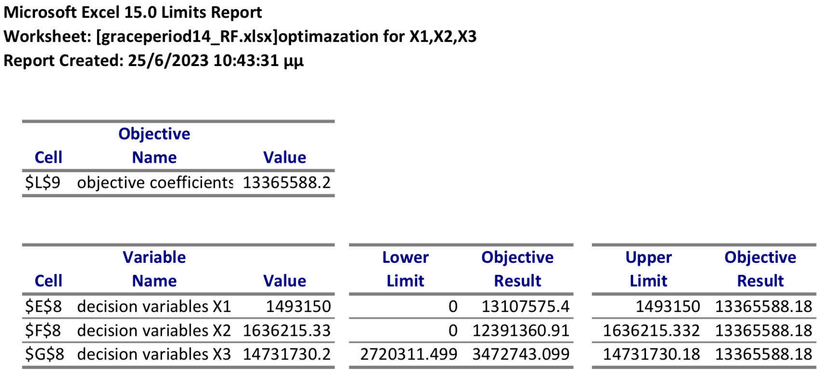
Variable Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
$E$8	decision variables X1	1493150	0	0.172797629	1E+30	0.172797629
$F$8	decision variables X2	1636215.332	0	0.59541507	1E+30	0.59541507
$G$8	decision variables X3	14731730.18	0	0.823620035	1E+30	0.823620035

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$H$11	LHS	17861095.51	0	44613100	1E+30	26752004.49
$H$12	LHS	-12011418.68	0	0	1E+30	12011418.68
$H$13	LHS	1493150	0.172797629	1493150	12011418.68	1493150
$H$14	LHS	1636215.332	0.59541507	1636215.332	16015224.91	1636215.332
$H$15	LHS	14731730.18	0.823620035	14731730.18	26752004.49	12011418.68

Table 6: Sensitivity analysis for XG1, XG2 and XG3 in grace period 14 using RF.



Microsoft Excel 15.0 Limits Report
Worksheet: [graceperiod14_RF.xlsx]optimazation for X1,X2,X3
Report Created: 25/6/2023 10:43:31 μμ

Cell	Objective Name	Value
$L$9	objective coefficients	13365588.2

Cell	Variable Name	Value	Lower Limit	Objective Result	Upper Limit	Objective Result
$E$8	decision variables X1	1493150	0	13107575.4	1493150	13365588.18
$F$8	decision variables X2	1636215.33	0	12391360.91	1636215.332	13365588.18
$G$8	decision variables X3	14731730.2	2720311.499	3472743.099	14731730.18	13365588.18

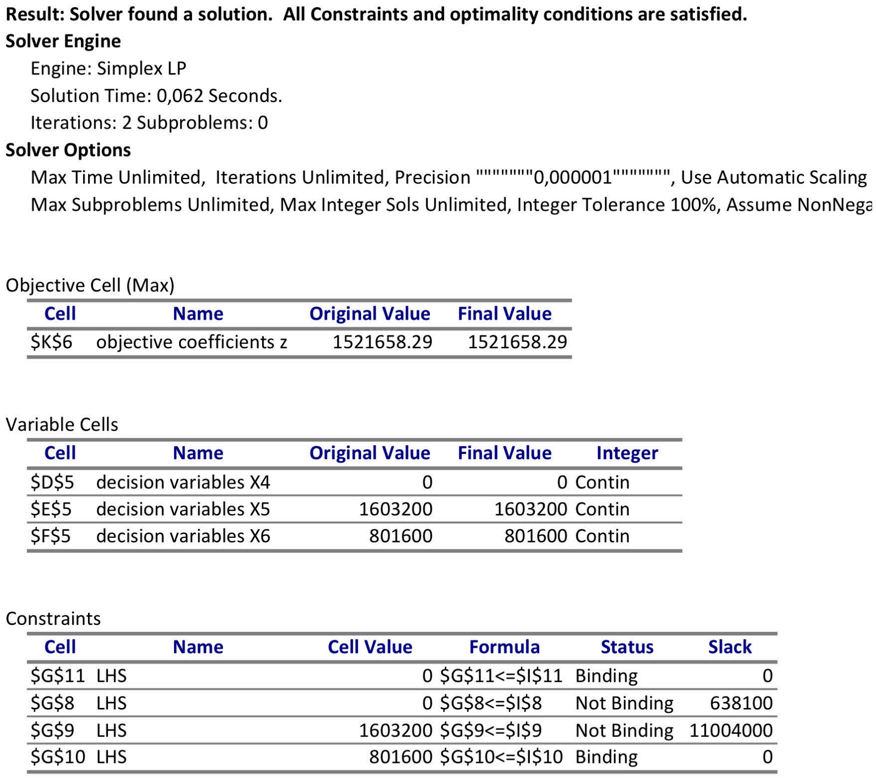
Table 7: Limits report for XG1, XG2 and XG3 in grace period 14 using RF.

$$
\begin{aligned}
\max\ z = \omega_4 X_4 + \omega_5 X_5 + \omega_6 X_6. \\
\text{s.t} \\
X_4 + 0.50 X_5 - X_6 \leq 0 \\
X_4 \leq 30\% \sum_{i=1}^{n_4} k_i \\
X_5 \leq \sum_{i=1}^{n_5} (k_i + 20\% k_i) \\
X_6 \leq \sum_{i=1}^{n_6} (k_i + 60\% k_i) \\
X_4 \geq 0,\ X_5 \geq 0,\ X_6 \geq 0
\end{aligned}
$$

where:

- ω_i, $i = 4, \ldots, 6$: objective coefficients.
- X_i, $i = YG1, YG2, YG3$: decision variables.
- z: the total amount of the objective function.

The term ”k”, represents the initial quantity allocated to newly acquired clientele. Based on our findings, it has been determined that the value of k corresponds to 3000 units of currency. Subsequently, the quantity of clientele within every cluster is ascertained by means of the notation ”n_i”. The simplex method was employed to attain the optimal value of the objective variable, as portrayed in Tables 8, 9 and 10.



Result: Solver found a solution. All Constraints and optimality conditions are satisfied.
Solver Engine
Engine: Simplex LP
Solution Time: 0,062 Seconds.
Iterations: 2 Subproblems: 0
Solver Options
Max Time Unlimited, Iterations Unlimited, Precision """"""""0,000001"""""""", Use Automatic Scaling
Max Subproblems Unlimited, Max Integer Sols Unlimited, Integer Tolerance 100%, Assume NonNega

Objective Cell (Max)

Cell	Name	Original Value	Final Value
$K$6	objective coefficients z	1521658.29	1521658.29

Variable Cells

Cell	Name	Original Value	Final Value	Integer
$D$5	decision variables X4	0	0	Contin
$E$5	decision variables X5	1603200	1603200	Contin
$F$5	decision variables X6	801600	801600	Contin

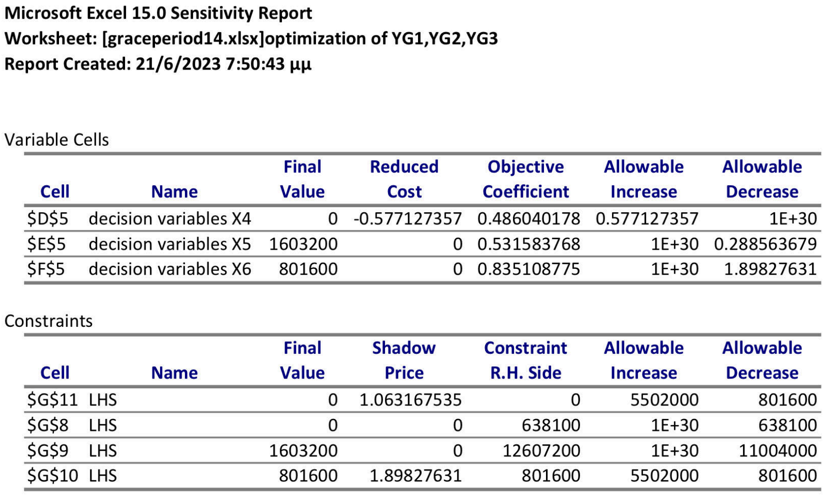
Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$G$11	LHS	0	$G$11<=$I$11	Binding	0
$G$8	LHS	0	$G$8<=$I$8	Not Binding	638100
$G$9	LHS	1603200	$G$9<=$I$9	Not Binding	11004000
$G$10	LHS	801600	$G$10<=$I$10	Binding	0

Table 8: Answers report for YG1, YG2 and YG3 in grace period 14.

The optimal value of the target variable refers to a monetary measurement of 1521658.29. This particular value represents the total amount that is allocated to all newly acquired customers. A sum of 1603200 monetary units is allocated to new customers of average status, whereas 801600 monetary units are allocated to customers of good standing. Last but not least, it has been determined that no micro-loans are extended to new customers with unfavorable credit histories.

In addition, Tables 9 and10 with the results of the sensitivity analysis and limits report, corresponding, are provided.



Microsoft Excel 15.0 Sensitivity Report
Worksheet: [graceperiod14.xlsx]optimization of YG1,YG2,YG3
Report Created: 21/6/2023 7:50:43 μμ

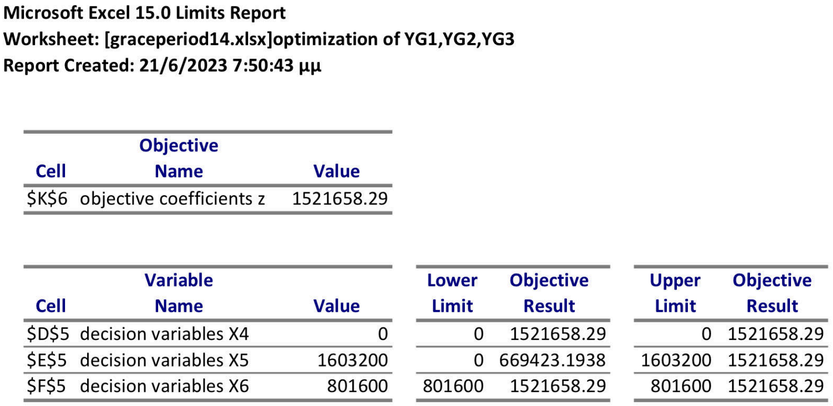
Variable Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
$D$5	decision variables X4	0	-0.577127357	0.486040178	0.577127357	1E+30
$E$5	decision variables X5	1603200	0	0.531583768	1E+30	0.288563679
$F$5	decision variables X6	801600	0	0.835108775	1E+30	1.89827631

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$G$11	LHS	0	1.063167535	0	5502000	801600
$G$8	LHS	0	0	638100	1E+30	638100
$G$9	LHS	1603200	0	12607200	1E+30	11004000
$G$10	LHS	801600	1.89827631	801600	5502000	801600

Table 9: Sensitivity analysis for YG1, YG2 and YG3 in grace period 14.



Microsoft Excel 15.0 Limits Report
Worksheet: [graceperiod14.xlsx]optimization of YG1,YG2,YG3
Report Created: 21/6/2023 7:50:43 μμ

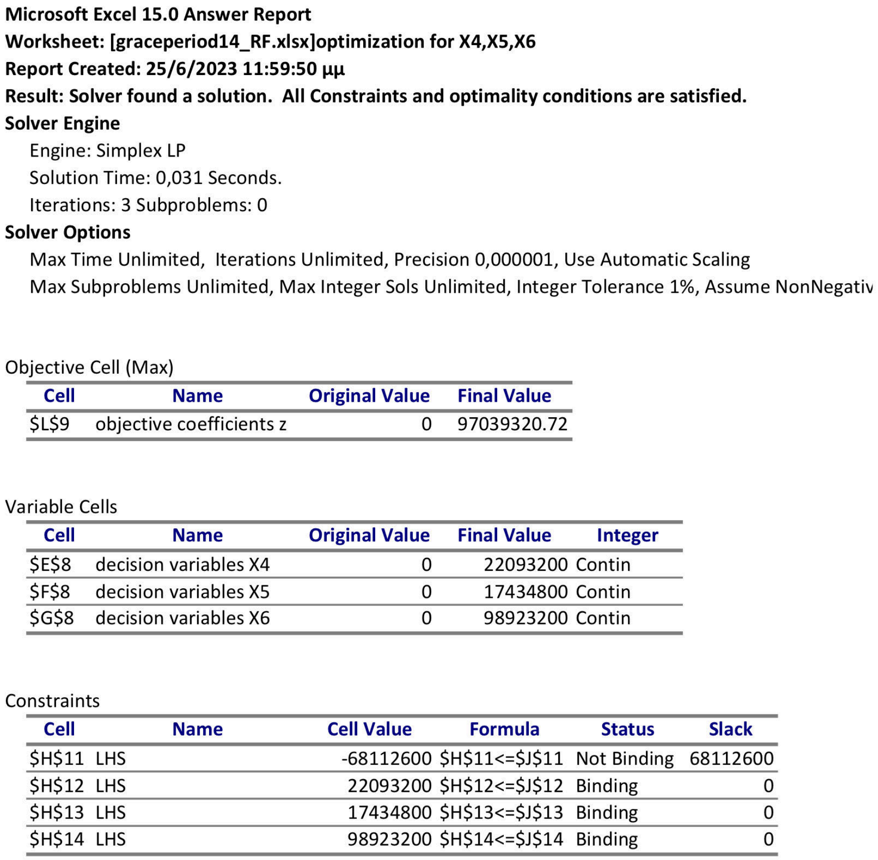
Cell	Objective Name	Value
$K$6	objective coefficients z	1521658.29

Cell	Variable Name	Value	Lower Limit	Objective Result	Upper Limit	Objective Result
$D$5	decision variables X4	0	0	1521658.29	0	1521658.29
$E$5	decision variables X5	1603200	0	669423.1938	1603200	1521658.29
$F$5	decision variables X6	801600	801600	1521658.29	801600	1521658.29

Table 10: Limits report for YG1, YG2 and YG3 in grace period 14.

The ultimate goal of the analysis involves the allocation of the resultant sums for each group to their respective clientele. The allocation of micro-loans to individual customers is determined based on the constraints of the group. The current study explores the differentiation of rewards for varying levels of customer loyalty. More precisely, customers with moderate loyalty are expected to receive a reward of $(\omega_i\lambda_i + 0.5sd\left(\sum_{i=1}^{n_i}\omega_i\lambda_i\right))$, whereas highly loyal customers are proposed to be rewarded with $(\omega_i\lambda_i + 2sd\left(\sum_{i=1}^{n_i}\omega_i\lambda_i\right))$. In contrast, for the allocation of new moderate customers, we assign a value of $k + 20\%k$, which equates to a sum of 3600 monetary units. Furthermore, for good new customers entails a value of $k + 60\%k$, with the monetary value amounts to 4800 monetary units.

The aforementioned methodology was applied analogously to new customers; their probabilities were derived from the Random Forest model and the resulting outcomes are displayed in the subsequent Tables 11, 12 and 13.



Microsoft Excel 15.0 Answer Report
Worksheet: [graceperiod14_RF.xlsx]optimization for X4,X5,X6
Report Created: 25/6/2023 11:59:50 μμ
Result: Solver found a solution. All Constraints and optimality conditions are satisfied.
Solver Engine
Engine: Simplex LP
Solution Time: 0,031 Seconds.
Iterations: 3 Subproblems: 0
Solver Options
Max Time Unlimited, Iterations Unlimited, Precision 0,000001, Use Automatic Scaling
Max Subproblems Unlimited, Max Integer Sols Unlimited, Integer Tolerance 1%, Assume NonNegativ

Objective Cell (Max)

Cell	Name	Original Value	Final Value
$L$9	objective coefficients z	0	97039320.72

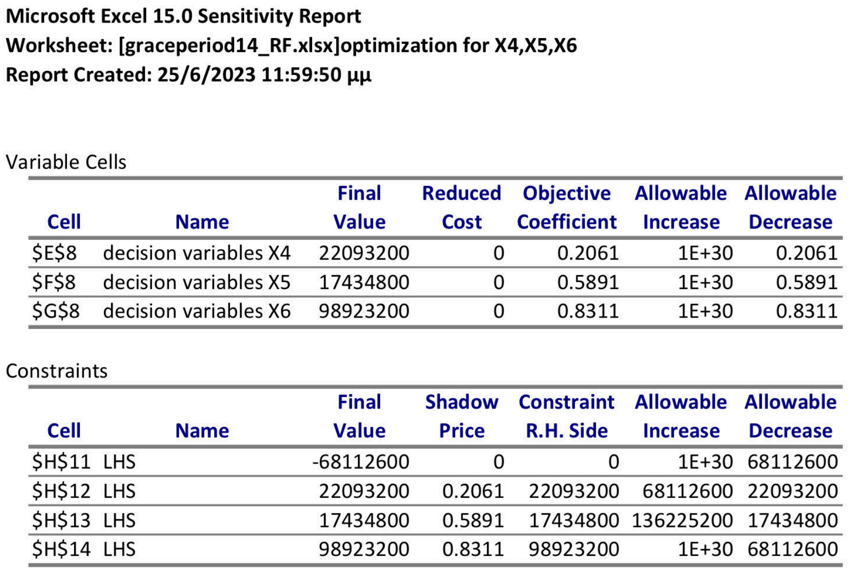
Variable Cells

Cell	Name	Original Value	Final Value	Integer
$E$8	decision variables X4	0	22093200	Contin
$F$8	decision variables X5	0	17434800	Contin
$G$8	decision variables X6	0	98923200	Contin

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$H$11	LHS	-68112600	$H$11<=$J$11	Not Binding	68112600
$H$12	LHS	22093200	$H$12<=$J$12	Binding	0
$H$13	LHS	17434800	$H$13<=$J$13	Binding	0
$H$14	LHS	98923200	$H$14<=$J$14	Binding	0

Table 11: Answers report for YG1, YG2 and YG3 in grace period 14 using RF.



Microsoft Excel 15.0 Sensitivity Report
Worksheet: [graceperiod14_RF.xlsx]optimization for X4,X5,X6
Report Created: 25/6/2023 11:59:50 μμ

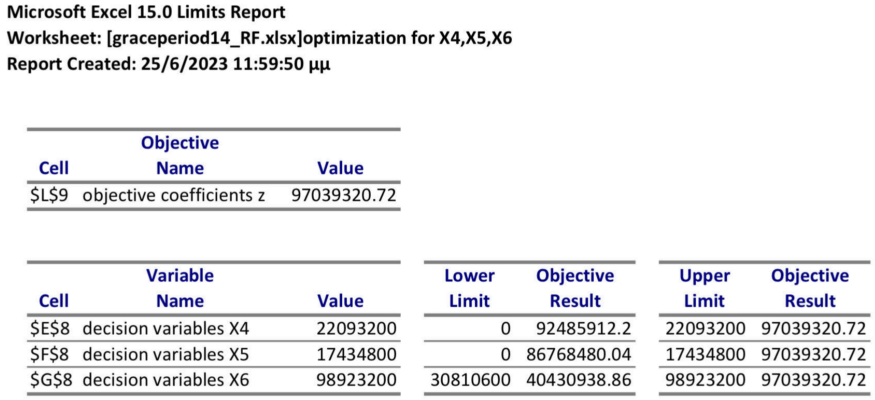
Variable Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
$E$8	decision variables X4	22093200	0	0.2061	1E+30	0.2061
$F$8	decision variables X5	17434800	0	0.5891	1E+30	0.5891
$G$8	decision variables X6	98923200	0	0.8311	1E+30	0.8311

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$H$11	LHS	-68112600	0	0	1E+30	68112600
$H$12	LHS	22093200	0.2061	22093200	68112600	22093200
$H$13	LHS	17434800	0.5891	17434800	136225200	17434800
$H$14	LHS	98923200	0.8311	98923200	1E+30	68112600

Table 12: Sensitivity analysis for YG1, YG2 and YG3 in grace period 14 using RF.



Microsoft Excel 15.0 Limits Report
Worksheet: [graceperiod14_RF.xlsx]optimization for X4,X5,X6
Report Created: 25/6/2023 11:59:50 μμ

Cell	Objective Name	Value
$L$9	objective coefficients z	97039320.72

Cell	Variable Name	Value	Lower Limit	Objective Result	Upper Limit	Objective Result
$E$8	decision variables X4	22093200	0	92485912.2	22093200	97039320.72
$F$8	decision variables X5	17434800	0	86768480.04	17434800	97039320.72
$G$8	decision variables X6	98923200	30810600	40430938.86	98923200	97039320.72

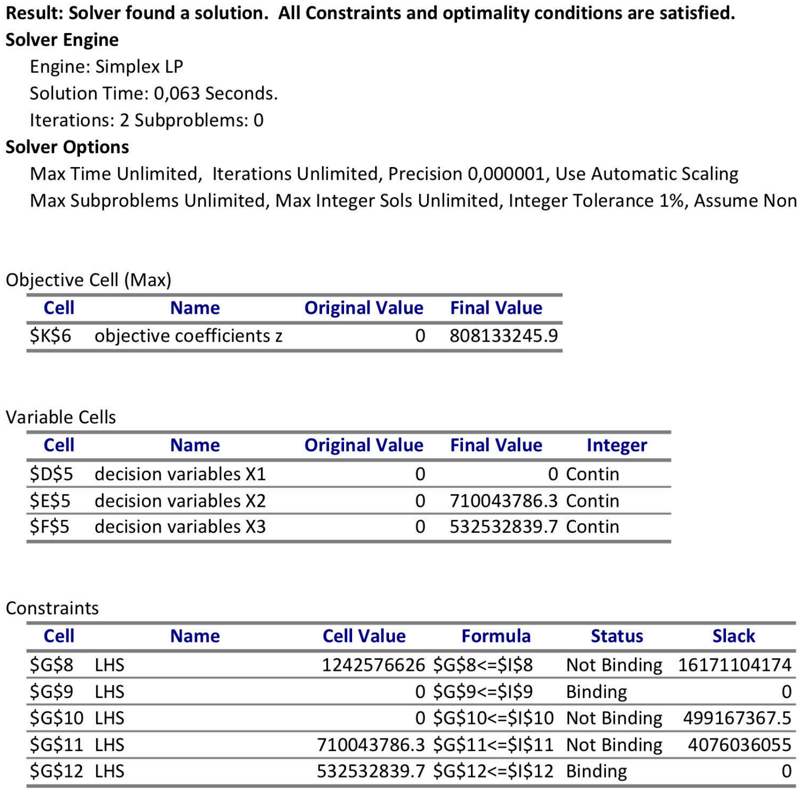
Table 13: Limits report for YG1, YG2 and YG3 in grace period 14 using RF.

Furthermore, the present study examine the scenario of a **grace period of 30 days**. It is pertinent to note that only former clientele with accessible historical information are considered to determine a reasonable loan amount that can be extended while ensuring their capacity to repay.

In this analysis, the identical methodology previously applied to old customers on the 14-day grace period is employed. To achieve this, the customers are segregated into their individual groups according to their probabilities. Subsequently, the necessary metrics are computed, followed by the

utilization of the simplex method.
The optimal value attained by the target allocation is 808133245.9 monetary units. It is imperative to acknowledge that, and in this scenario, the micro-loan is not extended to non-creditworthy borrowers. In this context, an allocation of 710043786.3 is made to the medium entities, while a sum of 532532839.7 is reserved for the higher-performing entities. This approach can be attributed to a deliberate effort to minimize the possibility of extending loans to individuals who exhibit a tendency towards default, thereby eliminating the risk of non-repayment and fee delinquency.

Subsequently, below are the detailed results that emerged after applying the simplex method (14), sensitivity analysis (15), and limits reporting (16).



Result: Solver found a solution. All Constraints and optimality conditions are satisfied.
Solver Engine
Engine: Simplex LP
Solution Time: 0,063 Seconds.
Iterations: 2 Subproblems: 0
Solver Options
Max Time Unlimited, Iterations Unlimited, Precision 0,000001, Use Automatic Scaling
Max Subproblems Unlimited, Max Integer Sols Unlimited, Integer Tolerance 1%, Assume NonNegative

Objective Cell (Max)

Cell	Name	Original Value	Final Value
$K$6	objective coefficients z	0	808133245.9

Variable Cells

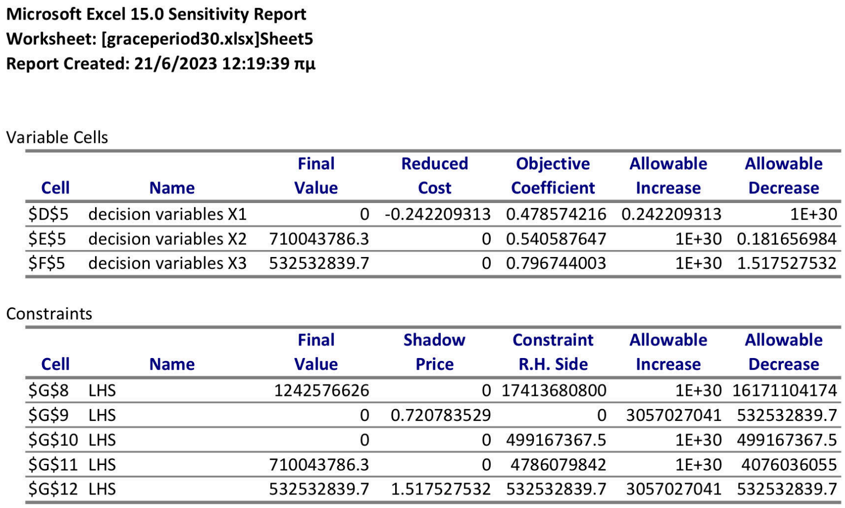
Cell	Name	Original Value	Final Value	Integer
$D$5	decision variables X1	0	0	Contin
$E$5	decision variables X2	0	710043786.3	Contin
$F$5	decision variables X3	0	532532839.7	Contin

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$G$8	LHS	1242576626	$G$8<=$I$8	Not Binding	16171104174
$G$9	LHS	0	$G$9<=$I$9	Binding	0
$G$10	LHS	0	$G$10<=$I$10	Not Binding	499167367.5
$G$11	LHS	710043786.3	$G$11<=$I$11	Not Binding	4076036055
$G$12	LHS	532532839.7	$G$12<=$I$12	Binding	0

Table 14: Answers report for XG1, XG2 and XG3 in grace period 30.

In a similar vein, in the instance where probabilities were computed using the Random Forest methodology, the solution to the optimization problem is presented below in Tables 17, 18 and 19.



Microsoft Excel 15.0 Sensitivity Report
Worksheet: [graceperiod30.xlsx]Sheet5
Report Created: 21/6/2023 12:19:39 πμ

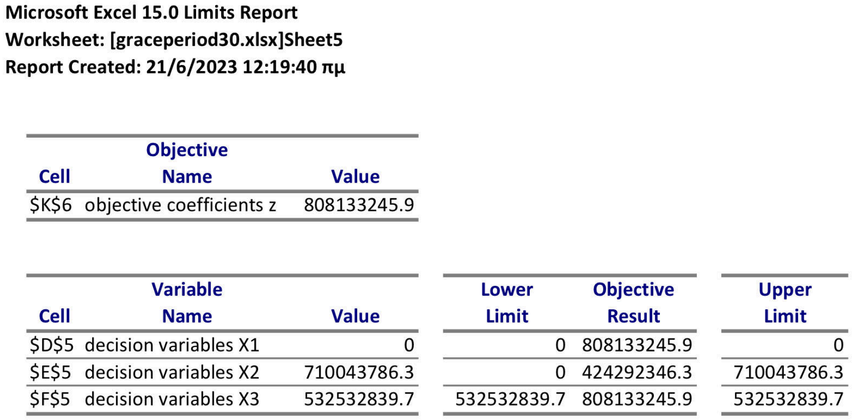
Variable Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
$D$5	decision variables X1	0	-0.242209313	0.478574216	0.242209313	1E+30
$E$5	decision variables X2	710043786.3	0	0.540587647	1E+30	0.181656984
$F$5	decision variables X3	532532839.7	0	0.796744003	1E+30	1.517527532

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$G$8	LHS	1242576626	0	17413680800	1E+30	16171104174
$G$9	LHS	0	0.720783529	0	3057027041	532532839.7
$G$10	LHS	0	0	499167367.5	1E+30	499167367.5
$G$11	LHS	710043786.3	0	4786079842	1E+30	4076036055
$G$12	LHS	532532839.7	1.517527532	532532839.7	3057027041	532532839.7

Table 15: Sensitivity analysis for XG1, XG2 and XG3 in grace period 30.



Microsoft Excel 15.0 Limits Report
Worksheet: [graceperiod30.xlsx]Sheet5
Report Created: 21/6/2023 12:19:40 πμ

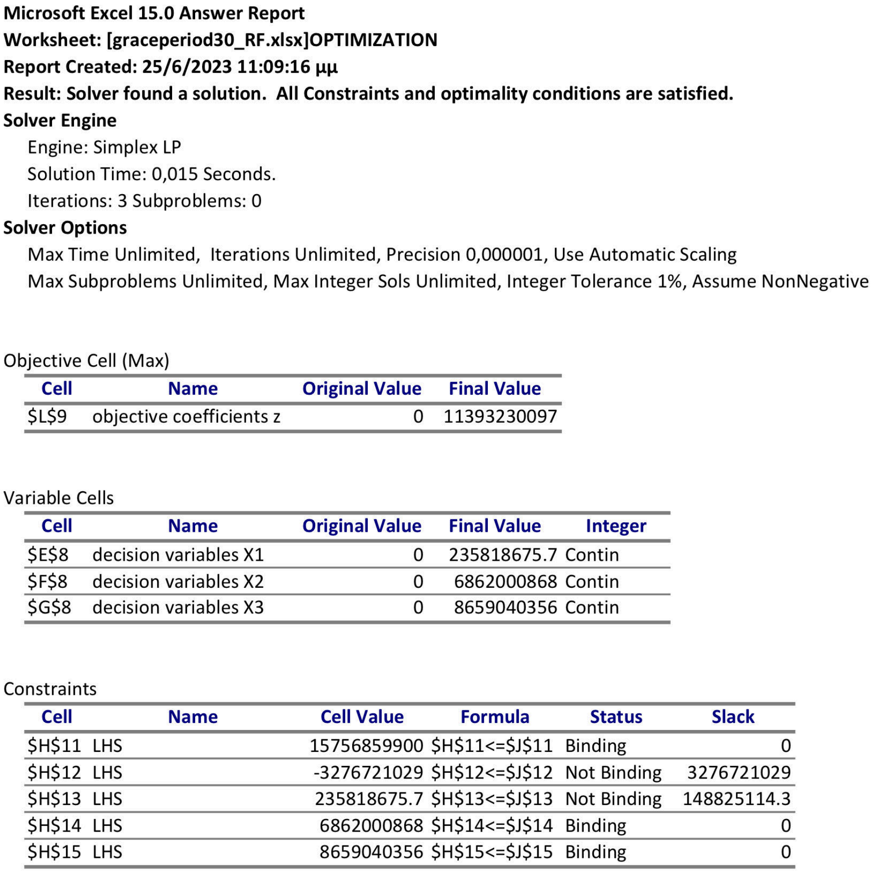
Cell	Objective Name	Value
$K$6	objective coefficients z	808133245.9

Cell	Variable Name	Value	Lower Limit	Objective Result	Upper Limit
$D$5	decision variables X1	0	0	808133245.9	0
$E$5	decision variables X2	710043786.3	0	424292346.3	710043786.3
$F$5	decision variables X3	532532839.7	532532839.7	808133245.9	532532839.7

Table 16: Limits report for XG1, XG2 and XG3 in grace period 30.

Furthermore, it is noteworthy to highlight that the Random Forest methodology produced superior accuracy and AUC compared to the results of the simplex approach. The outcomes of the latter technique are attributed to the observation that the achieved optimum value equated the maximum threshold, which implies that in a significant number of customers from each group will allocated to micro-loans.

A crucial final step in the analytical process involves the allocation of group amounts to respective clients. The methodology implemented is consistent with that which was utilized in the preceding period.



Microsoft Excel 15.0 Answer Report
Worksheet: [graceperiod30_RF.xlsx]OPTIMIZATION
Report Created: 25/6/2023 11:09:16 μμ
Result: Solver found a solution. All Constraints and optimality conditions are satisfied.
Solver Engine
Engine: Simplex LP
Solution Time: 0,015 Seconds.
Iterations: 3 Subproblems: 0
Solver Options
Max Time Unlimited, Iterations Unlimited, Precision 0,000001, Use Automatic Scaling
Max Subproblems Unlimited, Max Integer Sols Unlimited, Integer Tolerance 1%, Assume NonNegative

Objective Cell (Max)

Cell	Name	Original Value	Final Value
$L$9	objective coefficients z	0	11393230097

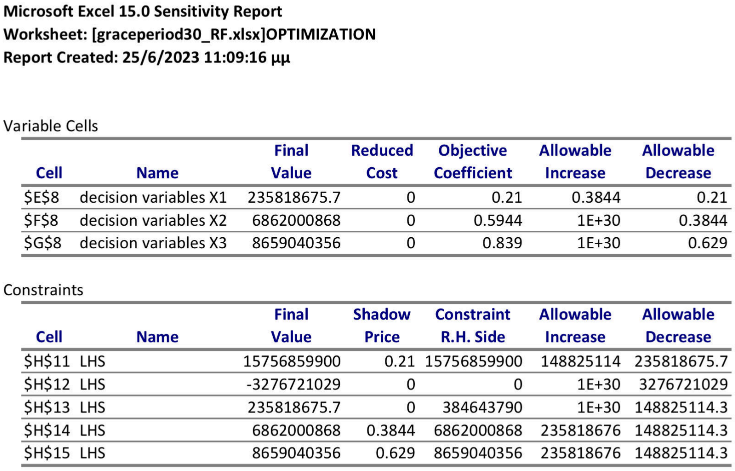
Variable Cells

Cell	Name	Original Value	Final Value	Integer
$E$8	decision variables X1	0	235818675.7	Contin
$F$8	decision variables X2	0	6862000868	Contin
$G$8	decision variables X3	0	8659040356	Contin

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$H$11	LHS	15756859900	$H$11<=$J$11	Binding	0
$H$12	LHS	-3276721029	$H$12<=$J$12	Not Binding	3276721029
$H$13	LHS	235818675.7	$H$13<=$J$13	Not Binding	148825114.3
$H$14	LHS	6862000868	$H$14<=$J$14	Binding	0
$H$15	LHS	8659040356	$H$15<=$J$15	Binding	0

Table 17: Answers report for XG1, XG2 and XG3 in grace period 30 using RF.



Microsoft Excel 15.0 Sensitivity Report
Worksheet: [graceperiod30_RF.xlsx]OPTIMIZATION
Report Created: 25/6/2023 11:09:16 μμ

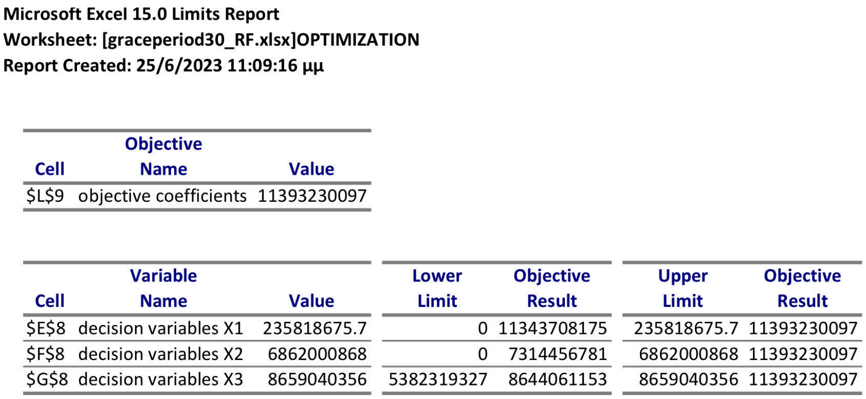
Variable Cells

		Final	Reduced	Objective	Allowable	Allowable
Cell	Name	Value	Cost	Coefficient	Increase	Decrease
$E$8	decision variables X1	235818675.7	0	0.21	0.3844	0.21
$F$8	decision variables X2	6862000868	0	0.5944	1E+30	0.3844
$G$8	decision variables X3	8659040356	0	0.839	1E+30	0.629

Constraints

		Final	Shadow	Constraint	Allowable	Allowable
Cell	Name	Value	Price	R.H. Side	Increase	Decrease
$H$11	LHS	15756859900	0.21	15756859900	148825114	235818675.7
$H$12	LHS	-3276721029	0	0	1E+30	3276721029
$H$13	LHS	235818675.7	0	384643790	1E+30	148825114.3
$H$14	LHS	6862000868	0.3844	6862000868	235818676	148825114.3
$H$15	LHS	8659040356	0.629	8659040356	235818676	148825114.3

Table 18: Sensitivity analysis for XG1, XG2 and XG3 in grace period 30 using RF.



Microsoft Excel 15.0 Limits Report
Worksheet: [graceperiod30_RF.xlsx]OPTIMIZATION
Report Created: 25/6/2023 11:09:16 μμ

Cell	Objective Name	Value
$L$9	objective coefficients	11393230097

Cell	Variable Name	Value	Lower Limit	Objective Result	Upper Limit	Objective Result
$E$8	decision variables X1	235818675.7	0	11343708175	235818675.7	11393230097
$F$8	decision variables X2	6862000868	0	7314456781	6862000868	11393230097
$G$8	decision variables X3	8659040356	5382319327	8644061153	8659040356	11393230097

Table 19: Limits report for XG1, XG2 and XG3 in grace period 30 using RF.

The following Table 20 provides a comprehensive overview of the results garnered from the investigation. It has been observed that a higher number of customers were granted micro-loans when the analysis was conducted based on the outcomes derived from Random Forest analysis.

Ultimately, the process diagram implemented in the analysis of our study is presented in Figure 27 and 28.

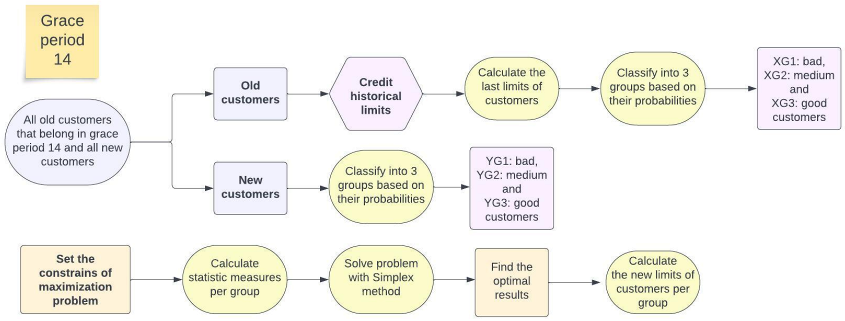
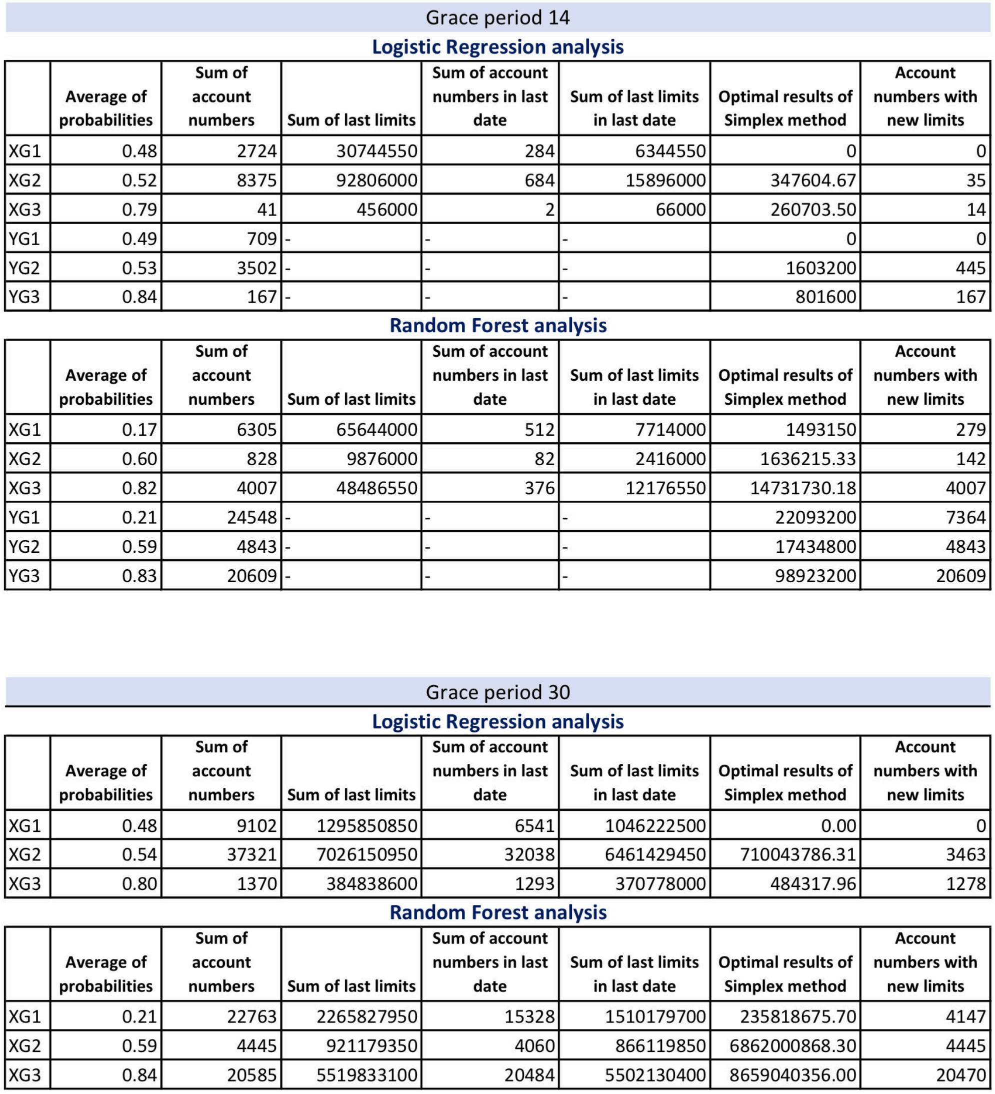

Grace period 14
All old customers that belong in grace period 14 and all new customers
Old customers
Credit historical limits
Calculate the last limits of customers
Classify into 3 groups based on their probabilities
XG1: bad, XG2: medium and XG3: good customers
New customers
Classify into 3 groups based on their probabilities
YG1: bad, YG2: medium and YG3: good customers
Set the constrains of maximization problem
Calculate statistic measures per group
Solve problem with Simplex method
Find the optimal results
Calculate the new limits of customers per group

Figure 27: A diagrammatic representation of procedure in grace period 14.



Grace period 14

Logistic Regression analysis

	Average of probabilities	Sum of account numbers	Sum of last limits	Sum of account numbers in last date	Sum of last limits in last date	Optimal results of Simplex method	Account numbers with new limits
XG1	0.48	2724	30744550	284	6344550	0	0
XG2	0.52	8375	92806000	684	15896000	347604.67	35
XG3	0.79	41	456000	2	66000	260703.50	14
YG1	0.49	709	-	-	-	0	0
YG2	0.53	3502	-	-	-	1603200	445
YG3	0.84	167	-	-	-	801600	167

Random Forest analysis

	Average of probabilities	Sum of account numbers	Sum of last limits	Sum of account numbers in last date	Sum of last limits in last date	Optimal results of Simplex method	Account numbers with new limits
XG1	0.17	6305	65644000	512	7714000	1493150	279
XG2	0.60	828	9876000	82	2416000	1636215.33	142
XG3	0.82	4007	48486550	376	12176550	14731730.18	4007
YG1	0.21	24548	-	-	-	22093200	7364
YG2	0.59	4843	-	-	-	17434800	4843
YG3	0.83	20609	-	-	-	98923200	20609

Grace period 30

Logistic Regression analysis

	Average of probabilities	Sum of account numbers	Sum of last limits	Sum of account numbers in last date	Sum of last limits in last date	Optimal results of Simplex method	Account numbers with new limits
XG1	0.48	9102	1295850850	6541	1046222500	0.00	0
XG2	0.54	37321	7026150950	32038	6461429450	710043786.31	3463
XG3	0.80	1370	384838600	1293	370778000	484317.96	1278

Random Forest analysis

	Average of probabilities	Sum of account numbers	Sum of last limits	Sum of account numbers in last date	Sum of last limits in last date	Optimal results of Simplex method	Account numbers with new limits
XG1	0.21	22763	2265827950	15328	1510179700	235818675.70	4147
XG2	0.59	4445	921179350	4060	866119850	6862000868.30	4445
XG3	0.84	20585	5519833100	20484	5502130400	8659040356.00	20470

Table 20: Total results of optimization problem.

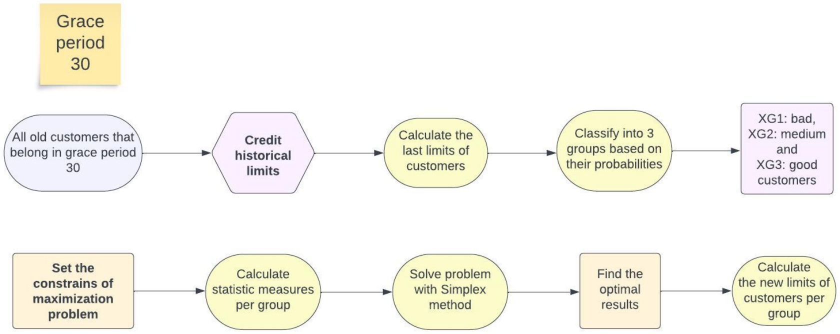

Grace period 30
All old customers that belong in grace period 30
Credit historical limits
Calculate the last limits of customers
Classify into 3 groups based on their probabilities
XG1: bad, XG2: medium and XG3: good customers
Set the constrains of maximization problem
Calculate statistic measures per group
Solve problem with Simplex method
Find the optimal results
Calculate the new limits of customers per group

Figure 28: A diagrammatic representation of procedure in grace period 30.

6 Conclusion & Possible Extensions

In summary, the present study endeavors to examine the credit risk associated with micro-loans and determine the optimal threshold outcomes for both newly acquired and existing clientele. The aim of the investigation constituted a targeted effort towards the optimization of precision and efficacy in credit risk evaluation within the domain of micro-finance.

In order to accomplish this aim, a multi-faceted methodology was utilized. This study delved into the theoretical underpinnings of credit risk and scrutinized the diverse credit rating techniques employed in practical settings, thus facilitating a holistic comprehension of the present credit risk evaluation terrain. The present study employs statistical methods, such as PCA, Logistic Regression analysis and Random Forest, to examine micro-loan data and establish predictive models for the purpose of creditworthiness evaluation.

Additionally, the simplex algorithm was employed to optimize the loan boundaries and constraints, ultimately leading to the attainment of optimal quantities of micro-loans to be disbursed and the necessary number of loans to be approved. A sensitivity analysis was conducted in order to assess the effects of fluctuating input parameters on the outcomes of optimization. Through this analysis, valuable information was obtained regarding the resilience and reliability of the proposed methodology.

The case study analysis conducted using real micro-loan data demonstrated the effectiveness of the proposed methodology in accurately assessing credit risk and establishing optimal loan boundaries for both new and existing clients. The results underscored the significance of integrating statistical methodologies and optimization algorithms into the analysis of credit risk in micro-finance, in order to enhance decision-making processes and mitigate potential risks.

The present study focused on prominent aspects of credit risk evaluation in micro-finance lending, however, potential areas for further inquiry persist. Potential domains for future research could encompass the examination of supplementary variables that influence credit risk, integration of alternative data sources for comprehensive analysis, and assessment of the enduring effects of micro-loans on the financial well-being and alleviation of poverty among borrowers.

In conclusion, the results obtained from this postgraduate thesis serve to further the discipline of credit risk evaluation in micro-finance, establishing a framework for more precise and effective loan allocation methodologies to bolster financial accessibility and promote enduring developmental objectives.

However, in the realm of credit risk analysis for micro-loans, prospective investigations can investigate multiple auspicious domains to augment the efficacy and influence of micro-finance programs. Dynamic risk assessment presents an opportunity to formulate credit risk models that can adjust and revise in accordance with the evolving conditions and circumstances of borrowers, their payment behaviors, and the market conditions. Dynamic models would enhance risk management strategies and ensure optimal lending practices by allowing for real-time adjustments to loan boundaries and terms.

Additionally, in order to ensure the ethical soundness of future research endeavors, ethical considerations must be given due attention. Due to the susceptibility of individuals who borrow micro-loans, it is imperative to examine the principles of equity, transparency, and impartiality in the decision-making procedures. The exploration of the ethical implications surrounding the utilization of alternative data, concerns pertaining to privacy, and the promotion of conscientious lending practices are critical components in guaranteeing the ethical foundations of credit risk evaluation in the domain of micro-finance.

Furthermore, a comprehensive analysis of the enduring effects of micro-finance on the social and economic landscape can provide valuable insights into its far-reaching implications. Conducting longitudinal studies to evaluate the impact on the financial well-being, livelihoods, and poverty alleviation of borrowers, as well as assessing the efficacy of credit risk models in facilitating the realization of sustainable development objectives, is expected to yield significant academic and practical insights. This research has the potential to provide valuable insights into the development and execution of micro-loan initiatives that can optimize their enduring beneficial effects.

Last but not least, the execution of comparative analysis across diverse micro-loan initiatives and geographical areas may generate valuable perceptions regarding the efficacy of diverse credit risk evaluation methodologies. Through comparative analysis of various models, methodologies, and risk management strategies, scholars and researchers can identify optimal practices and augment the comprehensiveness of credit risk knowledge within the context of micro-finance.

To sum up, by investigating potential research areas, advancement in the field of credit risk analysis pertaining specifically to micro-finance loans may be attained. This phenomenon is anticipated to result in a significant improvement in the accuracy, efficiency, and societal implications of micro-finance endeavours. The present academic undertakings possess the potential to yield significant advancements toward the development of financially viable and comprehensive structures that facilitate the empowerment of individuals and small enterprises situated in marginalized communities.

References

[1] Abdou, H. A. (2009). Genetic programming for credit scoring: The case of Egyptian public sector banks. Expert systems with applications, 36(9), 11402-11417.

[2] Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intelligent systems in accounting, finance and management, 18(2-3), 59-88.

[3] Adeodato, P., & Melo, S. (2022). A geometric proof of the equivalence between AUC, ROC and Gini index area metrics for binary classifier performance assessment. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1-6). IEEE.

[4] Agresti, A. (2012). Categorical data analysis (Vol. 792). John Wiley & Sons.

[5] Al Amari, A. (2002). The credit evaluation process and the role of credit scoring: A case study of Qatar (Doctoral dissertation, University College Dublin).

[6] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, 23(4), 589-609.

[7] Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). Journal of banking & finance, 18(3), 505-529.

[8] Anderson, R. A. (2022). Credit intelligence and modelling: Many paths through the forest of credit rating and scoring. Oxford University Press.

[9] Arminger, G., Enache, D., & Bonne, T. (1997). Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks. Computational Statistics, 12(2).

[10] Baesens, B., Roesch, D., & Scheule, H. (2016). Credit risk analytics: Measurement techniques, applications, and examples in SAS. John Wiley & Sons.

[11] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the operational research society, 54, 627-635.

[12] Baesens, B. (2003). Developing intelligent systems for credit scoring using machine learning techniques.

[13] Baestaens, D. E. (1999). Credit risk modeling strategies: the road to serfdom? Intelligent Systems in Accounting, Finance & Management, 8(4), 225-235.

[14] Baesens, B., Roesch, D., & Scheule, H. (2016). Credit risk analytics: Measurement techniques, applications, and examples in SAS. John Wiley & Sons.

[15] Bailey, M. (Ed.). (2004). Credit scoring: The principles and practicalities. White Box Publ.

[16] Banasik, J., Crook, J., & Thomas, L. (2003). Sample selection bias in credit scoring models. Journal of the Operational Research Society, 54(8), 822-832.

[17] Banasik, J., & Crook, J. (2007). Reject inference, augmentation, and sample selection. European Journal of Operational Research, 183(3), 1582-1594.

[18] Banasik, J., & Crook, J. (2010). Reject inference in survival analysis by augmentation. Journal of the Operational Research Society, 61(3), 473-485.

[19] Banasik, J., & Crook, J. (2010). Reject inference in survival analysis by augmentation. Journal of the Operational Research Society, 61(3), 473-485.

[20] Behrman, M., Linder, R., Assadi, A. H., Stacey, B. R., & Backonja, M. M. (2007). Classification of patients with pain based on neuropathic

pain symptoms: Comparison of an artificial neural network against an established scoring system. European Journal of Pain, 11(4), 370-376.

[21] Bellman, R. (1952). On the theory of dynamic programming. Proceedings of the national Academy of Sciences, 38(8), 716-719.

[22] Ben-David, A., & Frank, E. (2009). Accuracy of machine learning models versus “hand crafted” expert systems–a credit scoring case study. Expert Systems with Applications, 36(3), 5264-5271.

[23] Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). Robust optimization (Vol. 28). Princeton university press.

[24] Bieg, H., & Krämer, G. (2006). Banking Supervision in Europe: From Basel I to Basel II. Strategic Management—New Rules for Old Europe, 73-82.

[25] Blöchlinger, A., & Leippold, M. (2006). Economic benefit of powerful credit scoring. Journal of Banking & Finance, 30(3), 851-873.

[26] Breiman, L. (2017). Classification and regression trees. Routledge.

[27] Bureau, C. F. P. (2019). What is a debt-to-income ratio? Why is the 43% debt-to-income ratio important.

[28] Capon, N. (1982). Credit scoring systems: A critical analysis. Journal of Marketing, 46(2), 82-91.

[29] Carlone, G. (2020). Introduction to credit risk. CRC Press.

[30] Chandler, G. G., & Ewert, D. C. (1976). Discrimination on basis of sex and the Equal Credit Opportunity Act. Credit Research Centre, Purdue University, Indiana.

[31] Chatterjee, S., & Barcun, S. (1970). A nonparametric approach to credit screening. Journal of the American Statistical Association, 65(329), 150-154.

[32] Chen, M. C., & Huang, S. H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. Expert Systems with Applications, 24(4), 433-441.

[33] Chiang, W. Y. K., Zhang, D., & Zhou, L. (2006). Predicting and explaining patronage behavior toward web and traditional stores using neural networks: a comparative analysis with logistic regression. Decision Support Systems, 41(2), 514-531.

[34] Corneliu, B., & Maria, A. B. F. (2009). The management of operational risk in banks. Journal of Risk, 1(1), 63-72.

[35] Cornuejols, G., & Tütüncü, R. (2006). Optimization methods in finance (Vol. 5). Cambridge University Press.

[36] Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183(3), 1447-1465.

[37] Desai, V. S., Crook, J. N., & Overstreet Jr, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. European journal of operational research, 95(1), 24-37.

[38] Deschaine, L. M., & Francone, F. D. (2002). Comparison of Discipulus™ Linear Genetic Programming Soft-ware with Support Vector Machines, Classification Trees, Neural Networks and Human Experts. Register Machine Learning Technologies Inc.

[39] Dimla Sr, D. E., & Lister, P. M. (2000). On-line metal cutting tool condition monitoring: II: tool-state classification using multi-layer perceptron neural networks. International Journal of Machine Tools and Manufacture, 40(5), 769-781.

[40] Durand, D. (1941). Credit-rating formulae. In Risk Elements in Consumer Instalment Financing (pp. 83-91). NBER.

[41] Dvir, D., Ben-David, A., Sadeh, A., & Shenhar, A. J. (2006). Critical managerial factors affecting defense projects success: A comparison between neural network and regression analysis. Engineering Applications of Artificial Intelligence, 19(5), 535-543.

[42] Eisenbeis, R. A. (1978). Problems in applying discriminant analysis in credit scoring models. Journal of Banking & Finance, 2(3), 205-219.

[43] Finney, D. J. (1952). The estimation of the ED50 for a logistic response curve. Sankhyā: The Indian Journal of Statistics, 121-136.

[44] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of eugenics, 7(2), 179-188.

[45] Fix, E., & Hodges Jr, J. L. (1952). Discriminatory analysis-nonparametric discrimination: Small sample performance. California Univ Berkeley.

[46] Fletcher, D., & Goss, E. (1993). Forecasting with neural networks: an application using bankruptcy data. Information & Management, 24(3), 159-167.

[47] Frame, W. S., Gerardi, K., Sexton, D., & Tracy, J. (2020). Ability to Repay a Mortgage: Assessing the Relationship Between Default, Debt-to-Income.

[48] Freund, R. J., Wilson, W. J., & Sa, P. (2006). Regression analysis. Elsevier.

[49] Fritz, S., & Hosemann, D. (2000). Restructuring the credit process: Behaviour scoring for German corporates. Intelligent Systems in Accounting, Finance & Management, 9(1), 9-21.

[50] Gately, E. (1995). Neural networks for financial forecasting. John Wiley & Sons, Inc.

[51] Goldberg, D. E. (1989). Genetic Algorithmin in Search. Optimization and machine learning.

[52] Grablowsky, B. J., & Talley, W. K. (1981). Probit and discriminant functions for classifying credit applicants-a comparison. Journal of Economics and Business, 33(3), 254-261.

[53] Greene, W. (1998). Sample selection in credit-scoring models. Japan and the world economy, 10(3), 299-316.

[54] Guillen, M., & Artis, M. (1994). Count data models for a credit scoring system (No. 021).

[55] Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523-541.

[56] Hand, D. J., Sohn, S. Y., & Kim, Y. (2005). Optimal bipartite scorecards. Expert Systems with Applications, 29(3), 684-690.

[57] Hardgrave, B. C., Wilson, R. L., & Walstrom, K. A. (1994). Predicting graduate student success: A comparison of neural networks and traditional techniques. Computers & Operations Research, 21(3), 249-263.

[58] Henley, W. E. (1997). Construction of a k-nearest-neighbour credit-scoring system. IMA Journal of Management Mathematics, 8(4), 305-321.

[59] Henley, W., & Hand, D. J. (1996). AK-Nearest-Neighbour Classifier for Assessing Consumer Credit Risk. Journal of the Royal Statistical Society: Series D (The Statistician), 45(1), 77-95.

[60] Hill, T., & Remus, W. (1994). Neural network models for intelligent support of managerial decision making. Decision Support Systems, 11(5), 449-459.

[61] Hosmer, D., & Lemeshow, S. (1989). Applied Logistic Regression Wiley & Sons. New York.

[62] Hu, Y. C. (2008). Incorporating a non-additive decision making method into multi-layer neural networks and its application to financial distress analysis. Knowledge-Based Systems, 21(5), 383-390.

[63] Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. Expert systems with applications, 33(4), 847-856.

[64] Housing, H. U. D. (2005). Department of Housing and Urban Development. Government Printing Office.

[65] Interface, V. L. E. R. (2009). US Department of Veterans Affairs Veterans Benefits Administration VA Loan Electronic Reporting Interface VA Servicer Guide.

[66] Johnson, R. W. (1992). Legal, social and economic issues in implementing scoring in the US. Credit scoring and credit control, 19, 32.

[67] Jost, A. (1998). Vice President for Business Development HNC Software, Inc. Credit Risk Modeling: Design and Application, 129.

[68] Koza, J. R. (1994). Genetic programming II: automatic discovery of reusable programs. MIT press.

[69] Kumar, A., Rao, V. R., & Soni, H. (1995). An empirical comparison of neural network and logistic regression models. Marketing letters, 6, 251-263.

[70] Kumra, R., Stein, R. M., & Assersohn, I. (2006). Assessing a knowledge-based approach to commercial loan underwriting. Expert Systems with Applications, 30(3), 507-518.

[71] Landajo, M., de Andrés, J., & Lorca, P. (2007). Robust neural modeling for the cross-sectional analysis of accounting information. European Journal of Operational Research, 177(2), 1232-1252.

[72] Lee, T. S., Chiu, C. C., Lu, C. J., & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. Expert Systems with applications, 23(3), 245-254.

[73] Lee, T. S., & Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. Expert Systems with applications, 28(4), 743-752.

[74] Lee, T. H., & Sung-Chang, J. (1999). Forecasting creditworthiness: Logistic vs. artificial neural net. The Journal of Business Forecasting, 18(4), 28.

[75] Lenard, M. J., Alam, P., & Madey, G. R. (1995). The application of neural networks and a qualitative response model to the auditor's going concern uncertainty decision. Decision Sciences, 26(2), 209-227.

[76] Lensberg, T., Eilifsen, A., & McKee, T. E. (2006). Bankruptcy theory development and classification via genetic programming. European Journal of operational research, 169(2), 677-697.

[77] Leonard, K. J. (1993). Detecting credit card fraud using expert systems. Computers & industrial engineering, 25(1-4), 103-106.

[78] Levitin, A. J. (2012). The consumer financial protection bureau: An introduction. Rev. Banking & Fin. L., 32, 321.

[79] Lewis, E. M. (1992). An introduction to credit scoring. Fair, Isaac and Company.

[80] Liang, Q. (2003). Corporate financial distress diagnosis in China: empirical analysis using credit scoring models. Hitotsubashi journal of commerce and management, 13-28.

[81] Lovie, A. D. (1987). The bootstrapped model—Lessons for the acceptance of intellectual technology. Applied ergonomics, 18(3), 201-206.

[82] Maddala, G. S. (2001). Introduction to econometrics, John Wiley and Sons. West Sussex, England.

[83] Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. Omega, 31(2), 83-96.

[84] Masters, T. (1995). Advanced algorithms for neural networks: a C++ sourcebook. John Wiley & Sons, Inc.

[85] Mathew, A. (2017). Credit Scoring Using Logistic Regression.

[86] McCord, N. M., & Illingworth, W. T. (1990). A practical guide to neural nets. Reading: Addison-Wesley [Material Impreso en biblioteca].

[87] McCulloch, C. E., & Searle, S. R. (2004). Generalized, linear, and mixed models. John Wiley & Sons.

[88] McKee, T. E., & Lensberg, T. (2002). Genetic programming and rough sets: A hybrid approach to bankruptcy classification. European journal of operational research, 138(2), 436-451.

[89] Mishra, A. (2010). Hurlbert, Glenn H. 2010. Linear Optimization: The Simplex Workbook. Interfaces, 40(4), 331-333.

[90] Myers, J. H., & Forgy, E. W. (1963). The development of numerical credit evaluation systems. Journal of the American Statistical association, 58(303), 799-806.

[91] Nath, R., Rajagopalan, B., & Ryker, R. (1997). Determining the saliency of input variables in neural network classifiers. Computers & Operations Research, 24(8), 767-773.

[92] Nevin, J. R., & Churchill Jr, G. A. (1979). The equal credit opportunity act: An evaluation. Journal of Marketing, 43(2), 95-104.

[93] Nikolopoulos, K., Goodwin, P., Patelis, A., & Assimakopoulos, V. (2007). Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches. European journal of operational research, 180(1), 354-368.

[94] Ong C, Huang J, Tzeng G. 2005. Building credit scoring models using genetic programming. Expert Systems with Applications 29(1): 41–47.

[95] Orgler, Y. E. (1971). Evaluation of bank consumer loans with credit scoring models. Tel-Aviv University, Department of Envirnonmental Sciences.

[96] Ottenbacher, K. J., Linn, R. T., Smith, P. M., Illig, S. B., Mancuso, M., & Granger, C. V. (2004). Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture. Annals of epidemiology, 14(8), 551-559.

[97] Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. European journal of operational research, 201(2), 490-499.

[98] Palisade Corporation. (2005). Neural tools: neural networks add-in for Microsoft Excel. Version 1.0.

[99] Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. Expert systems with applications, 36(1), 2-17.

[100] Pindyck, R. S., & Rubinfeld, D. L. (1997). Econometric Models and Economic Forecasts. McGraw-Hill/Irwin.

[101] Raiffa, H., & Schlaifer, R. (1961). Applied statistical decision theory.

[102] Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: a survey. Operations research, 42(4), 589-613.

[103] Sarlija, N., Bensic, M., & Zekic-Susac, M. (2009). Comparison procedure of predicting the time to default in behavioural scoring. Expert Systems with Applications, 36(5), 8778-8788.

[104] Saunders, J. (1985). This is credit scoring. Credit Management, September, 23-26.

[105] Shang, J. S., Lin, Y. S. E., Goetz, A. M., Shang, J. S., Lin, Y. S., & Goetz, A. M. (2000). Diagnosis of MRSA with neural networks and logistic regression approach. Health Care Management Science, 3(4).

[106] Smith, A. E., & Mason, A. K. (1997). Cost estimation predictive modeling: Regression versus neural network. The Engineering Economist, 42(2), 137-161.

[107] Somers, M., & Whittaker, J. (2007). Quantile regression for modelling distributions of profit and loss. European Journal of Operational Research, 183(3), 1477-1487.

[108] Song, J. H., Venkatesh, S. S., Conant, E. A., Arger, P. H., & Sehgal, C. M. (2005). Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. Academic radiology, 12(4), 487-495.

[109] Sparks, D. L., & Tucker, W. T. (1971). A multivariate analysis of personality and product use. Journal of Marketing Research, 8(1), 67-70.

[110] Stefanowski, J., & Wilk, S. (2001). Evaluating business credit risk by means of approach-integrating decision rules and case-based learning. Intelligent Systems in Accounting, Finance & Management, 10(2), 97-114.

[111] Thomas, L., Crook, J., & Edelman, D. (2017). Credit scoring and its applications. Society for industrial and Applied Mathematics.

[112] Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). Credit Scoring and its Applications: SIAM monographs on mathematical modeling and computation. Philadelphia: University City Science Center, SIAM.

[113] Thomas, L. C., Hand, D. J., & Jacka, S. D. (1998). Statistics in finance.

[114] Trinkle, B. S., & Baldwin, A. A. (2007). Interpretable credit model development via artificial neural networks. Intelligent Systems in Accounting, Finance & Management: International Journal, 15(3-4), 123-147.

[115] Trippi, R. R., & Turban, E. (Eds.). (1992). Neural networks in finance and investing: Using artificial intelligence to improve real world performance. McGraw-Hill, Inc.

[116] Tsaih, R., Liu, Y. J., Liu, W., & Lien, Y. L. (2004). Credit scoring system for small business loans. Decision Support Systems, 38(1), 91-99.

[117] Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. Expert systems with applications, 34(4), 2639-2649.

[118] Warner, B., & Misra, M. (1996). Understanding neural networks as statistical tools. The american statistician, 50(4), 284-293.

[119] West, D. (2000). Neural network credit scoring models. Computers & operations research, 27(11-12), 1131-1152.

[120] Wonderlic, E. F. (1952). An analysis of factors in granting credit, Indiana Univ.

[121] Xia, Y., Liu, B., Wang, S., & Lai, K. K. (2000). A model for portfolio selection with order of expected returns. Computers & Operations Research, 27(5), 409-422.

[122] Yang, Z., Wang, Y., Bai, Y., & Zhang, X. (2004). Measuring scorecard performance. In Computational Science-ICCS 2004: 4th International Conference, Kraków, Poland, June 6-9, 2004, Proceedings, Part IV 4 (pp. 900-906). Springer Berlin Heidelberg.

[123] Yesilnacar, E., & Topal, T. A. M. E. R. (2005). Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). Engineering Geology, 79(3-4), 251-266.

[124] Yu, L., Wang, S., & Lai, K. K. (2009). An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring. European journal of operational research, 195(3), 942-959.

[125] Zekic-Susac, M., Sarlija, N., & Bensic, M. (2004). Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models. In 26th International Conference on Information Technology Interfaces, 2004. (pp. 265-270). IEEE.