UNIVERSITÀ
DI PAVIA

Dipartimento di Scienze Economiche e Aziendali

Corso di Laurea magistrale in Finance

Credit Scoring

Supervisor:

Prof. Paolo Giudici

Reviewer:

Victoria Orlova

Condidate

Nahid Fallah

517483

Academic Year 2024-2025

Contents

Abstract

Credit scoring and risk management are fundamental pillars of the financial sector, directly influencing lending decisions, investment strategies, and overall economic stability. In recent years, the adoption of advanced statistical models and machine learning techniques has significantly enhanced the ability to predict critical events such as credit defaults and downgrades, enabling a more dynamic and adaptable approach to evolving market conditions.

This thesis explores the evolution of credit assessment methodologies, with particular focus on the comparison between traditional approaches and advanced techniques. The analysis examines four key models: Logistic Regression, Random Forest, Ridge Regression, and Gradient Boosting Machines (GBM), applied to real-world datasets from modefinance and Cardo AI. The primary objective is to evaluate the predictive effectiveness of each model in identifying credit risk and forecasting rare events such as defaults and downgrades.

One of the major challenges addressed in this research was class imbalance, a common issue in financial datasets where critical events constitute only a small fraction of the total observations. To overcome this challenge, sampling techniques such as Synthetic Minority Oversampling Technique (SMOTE) and majority class undersampling were implemented. These approaches enhanced the models' ability to effectively recognize rare events, reducing bias and ensuring more reliable predictions.

The research findings demonstrate that integrating advanced machine learning techniques is essential to improving the accuracy and reliability of credit risk assessment. In particular, the random forest model outperformed others in balancing predictive accuracy and handling imbalanced classes. Model performance was evaluated using standard metrics such as Receiver Operating Characteristic (ROC) curves and variable importance analysis, providing a clear understanding of the key factors influencing financial risk.

This research highlights the transformative potential of machine learning techniques in credit risk management, offering a more detailed and nuanced understanding of financial risk. These tools provide valuable insights for financial

institutions, enabling better risk mitigation and increased resilience, as well as for lenders and investors, supporting more informed and balanced decision-making. The proposed approach not only improves predictive capabilities but also contributes to building a fairer, more stable, and sustainable financial system.

Abstract

La valutazione del credito e la gestione del rischio costituiscono pilastri fondamentali nel settore finanziario, influenzando direttamente le decisioni di prestito, le strategie di investimento e la stabilità economica generale. Negli ultimi anni, l'adozione di modelli statistici avanzati e di tecniche di machine learning ha apportato miglioramenti significativi nella capacità di prevedere eventi critici come il default e il downgrading creditizio, consentendo un approccio più dinamico e adattabile alle condizioni di mercato in evoluzione.

Questa tesi esplora l'evoluzione delle metodologie di valutazione del credito, ponendo particolare attenzione al confronto tra approcci tradizionali e tecniche avanzate. L'analisi si concentra su quattro modelli chiave: regressione logistica, random forest, regressione Ridge e gradient boosting machines (GBM), applicati a dataset reali provenienti da modefinance e Cardo AI. L'obiettivo principale è valutare l'efficacia predittiva di ciascun modello nell'identificazione del rischio di credito e nella previsione di eventi rari come default e declassamenti.

Uno dei principali ostacoli affrontati nella ricerca è stato lo sbilanciamento delle classi, un problema comune nei dataset finanziari in cui gli eventi critici rappresentano una frazione minima delle osservazioni totali. Per superare questa sfida, sono state implementate tecniche di campionamento, come il sovracampionamento sintetico della minoranza (SMOTE) e il sotto-campionamento della classe maggioritaria. Questi approcci hanno migliorato la capacità dei modelli di riconoscere efficacemente gli eventi rari, riducendo i bias e garantendo previsioni più affidabili.

I risultati della ricerca evidenziano come l'uso di tecniche avanzate di machine learning rappresenti un passo fondamentale per migliorare l'accuratezza e l'affidabilità della valutazione del rischio di credito. In particolare, il modello random forest ha dimostrato performance superiori nel bilanciare accuratezza predittiva e capacità di gestione delle classi sbilanciate. Le prestazioni dei modelli sono state valutate utilizzando metriche standard, come le curve delle caratteristiche operative del ricevitore (ROC) e l'analisi dell'importanza delle variabili, che hanno permesso di ottenere una visione chiara dei fattori determinanti il rischio finanziario.

La ricerca mette in luce il potenziale trasformativo delle tecniche di machine

learning nella gestione del rischio creditizio, offrendo una comprensione più dettagliata e sfumata del rischio finanziario. Questi strumenti forniscono informazioni di grande valore sia per le istituzioni finanziarie, che possono mitigare meglio il rischio e aumentare la loro resilienza, sia per i prestatori e gli investitori, che possono adottare decisioni più informate ed equilibrate. L'approccio proposto non solo migliora le capacità predittive, ma contribuisce anche a costruire un sistema finanziario più equo, stabile e sostenibile.

Chapter 1

Understanding Credit Scoring and Risk Assessment

This chapter explains how credit scores help determine the creditworthiness of individuals and businesses, and covers credit scoring and risk assessment in finance. It discusses how credit risk is measured, the impact of non-performing assets (NPAs), and the role of fintech, including machine learning (ML) and artificial intelligence (AI). The chapter also looks at the benefits of credit scoring, such as reducing bias and increasing efficiency, while highlighting issues such as age and geographical bias. It concludes by highlighting the importance of predicting credit downgrades for financial risk management.

1.1 Credit Scoring

Credit scoring is a pervasive and powerful practice that has captured the attention of almost every aspect of our financial lives. It assesses individuals' and companies' creditworthiness and has a significant impact of our financial activities. A credit score is a number that represents person's ability to repay debts, it's derived from statistical analysis of a credit report, which shows a person or company's borrowing and repayment history Trivedi [2020].

Corporate credit scoring is based on the financial results reported in financial statements. On the other hand, negative financial events such as mortgage default

or bankruptcy have a significant impact on the financial standing of individuals or companies, especially in an era where this information is easily accessible. Credit scoring is a critical element of a bank's credit management, analysing and categorising credit factors to guide lending decisions, assess customer risk, reduce default rates and ensure the success of its loan portfolio.

As Anderson [2007] explains, credit scoring combines the concept of credit coming from the Latin word credo , meaning I believe or I trust, with scoring a method of ranking and distinguishing cases based on specific attributes through numerical methods to ensure decisions are objective and consistent. Credit scoring employs statistical methods to translate relevant data into numerical indicators, which predict whether a potential borrower will default on a loan or not default Abdou and Pointon [2011]. This enables lenders to make informed decisions about who qualifies for credit, the amount that should be extended, and the best strategies to enhance borrower profitability Jemal et al. [2002].

1.2 Evaluation of Credit Risk

Assessing credit risk is vital for ensuring financial stability and promoting economic growth in banks, as noted by Crouhy et al. [2000]. Key factors, such as asset quality and effective resource allocation, are crucial in this process. Credit risk can stem from various sources, including borrower defaults, concentrated exposures, and sovereign risks, all of which require careful evaluation at each stage, from loan origination to repayment collection, especially in today's uncertain global economy.

Banks mitigate credit risk through a variety of strategies, including insurance, covenants, diversification, and risk based pricing. These measures have been shaped significantly by lessons learned from the 2007-2008 financial crisis and the subsequent Basel regulatory frameworks [1]. According to Crouhy et al. [2000] an effective credit risk assessment requires an analysis of repayment capacity, capital adequacy, loan terms, credit history and collateral. Leading credit rating agencies like standard & Poor's and Moody's play an integral role in this evaluation.

[1] *The Basel frameworks are international banking regulations developed by the Basel Committee to strengthen capital adequacy, supervision, and risk management in the banking sector.*

In India, the challenge of NPA[2] demands rigorous credit evaluations, particularly as public sector banks contend with a gross NPA ratio of 18.77 % as of March 2019. It is imperative to enhance credit scoring mechanisms in order to reduce NPAs. This will only be achieved by implementing realistic repayment schedules based on borrower's cash flows. Furthermore, the growth of digital transactions has increased the risk of fraud, making it essential to conduct comprehensive data analysis to improve credit risk evaluations Hu and Su [2022]. The landscape of credit risk assessment is transforming significantly, driven by advancement in fintech. Central to this evolution are:

1. Data Analytic

2. Machine Learning (ML)

3. Artificial Intelligence (AL)

which are becoming indispensable for identifying trends, forecasting potential risks and enhancing decision-making process. Fintech companies[3] are using these technologies to automate work-flows and deliver rapid, data-driven credit assessments with remarkable accuracy. The research literature clearly shows that various ML techniques such as (Neural Networks, Support Vector Machines(SVM) and Ensemble Learning) are highly effective at differentiating between creditworthy and non-creditworthy applicants. These methods are also highly effective at predicting NAP, enabling financial institutions to proactively manage risks and optimize credit portfolios Hu and Su [2022].

Furthermore, validating credit risk models is challenging due to the limited availability of historical data and extended forecasting periods. These challenges can and will be addressed through a processed cross-sectional simulation approach that will enhance model evaluation using robust statistical methodologies[4]. This

[2] *Non-Performing Assets (NPAs) refer to loans or advances for which the principal or interest payment has remained overdue for a period of 90 days or more, as per banking regulations.*

[3] *Fintech, short for financial technology, refers to companies that use advanced technology, such as artificial intelligence, blockchain, and data analytics, to enhance or disrupt traditional financial services.*

[4] *Robust statistical methodologies are designed to remain effective even when assumptions about the underlying data, such as normality or absence of outliers, are violated. They are widely used*

comprehensive examination clearly shows that banks are moving towards more sophisticated, data-driven strategies in credit risk evaluation.

The probability of default (PD) is the most important factor in credit risk assessment. It determines whether a company will fail to meet its financial commitments or cease to exist. This problem is typically addressed by looking at the firm's credit score. We can set a line to divide companies into one of two predictive classes:

1. Default

2. non-default

or in our case we have downgrade or improved Babaei et al. [2023]. To establish a binary target variable X, calculate the differences in rating between one time period (e.g., month or quarter) to another. If the change in from last period to the current period is more than zero, it is downgrade. Otherwise we have a zero value, which represent no change or improvement. The best credit scoring models are developed by combining different methods, including Logistic regression, Ridge regression, Random Forest, and Boosting. For seeing some improvements in model and reduce the risk of over fitting, we used some techniques such as over sampling and under sampling. These models are evaluated using the Receiver Operating Characteristic (ROC) curve, which provides a comprehensive measure of the model's ability to distinguish between different classes James et al. [2023].

1.3 Advantage of credit scoring

Credit scoring methods have the potential to offer distinct advantages in the landing decision-making process. This is primarily due to the fact that they rely on a smaller number of variables in comparison to judgmental approaches. This improved efficiency is a consequence of the fact that credit scoring models are designed to incorporate only those specific factors that have been statistically validated as predictors of repayment behavior.

---

in fields like finance, biology, and machine learning to enhance model resilience.

In contrast, judgmental methods rely on subjective assessments and lack the statistical backing necessary to narrow down the number of variables. As Thomas et al. [2017] notes, judgmental decisions lack statistical support, making it challenging to reduce the number of factors considered. Furthermore, there is no systematic way to evaluate which variables are significant in predicting repayment and which are not, especially when dealing with high-risk borrowers.

According to Abdou and Pointon [2011] credit scoring models are designed to reduce the impact of biases they may arise from evaluating of repayment histories and the related policies, among previously accepted applications. Unlike judgmental methods, which primarily focus only on the individual's behavior who have already been granted credit, credit scoring model are adjusted to predict how rejected applicants might have performed if they has been approved. This approach is the best way to understand repayment odds and avoid systemic bias, and it allows lenders to make the most accurate assessments of overall credit risk.

Moreover, it is simply not possible for humans to process the huge data sets that are used to train credit scoring models. These models use a wide range of explicitly defined and legally approved variables, so there is no chance of inadvertently applying prohibited criteria that could influence manual evaluations. Judgmental methods are not reliable, they depend on subjective factors and personal interpretations, which leads to inconsistent and potentially biased decisions.

The clear definition of the variables in credit scoring models enables a more direct link between these variables and repayment behavior. This is something that is often more difficult to achieve in judgmental methods, which frequently depend on subjective judgments rather than objective evidence. A significant advantage of credit scoring models, particularly those using machine learning [5] algorithms, is their capacity to process large value of data and analyze complex sets of customer characteristics at a scale that exceeds the human's capacity. These models incorporate objective factors in measuring credit risk, which offer for a more accurate and unbiased evaluation of the borrower's likelihood of loan repayment Abdou and Pointon [2011]. The application provides a strong framework for credit assessment, improving the accuracy, consistency and fairness of lending decisions, ultimately benefiting both lenders and borrowers by minimising the risk of default.

[5] *Machine learning refers to algorithms that enable computers to learn patterns from data and make predictions or decisions without explicit programming.*

Credit scoring systems optimize financial institution's cost by automating loan approval process. For example, in the past, evaluating a borrower's creditworthiness required experienced loan officers to manually assess applications, which was both time consuming and labor intensive. Credit scoring makes this process more efficient by allowing less experienced and lower paid staff, or even automated systems, to handle routine credit evaluations, which in turn reduce labor costs. Moreover, credit scoring provides faster and more reliable results, allowing loan officers to dedicate their time to more complex or unconventional applications. This significantly reduces the need for highly trained personnel, as the system can efficiently process standards cases Bumacov et al. [2017].

Credit scoring systems have the ability to evolve over time through a process known as algorithm refinement. As more information on borrower behavior becomes available, the scoring model will be updated to enhance its accuracy. For instance, if repayment patterns shift due to economic changes or societal trends, the credit scoring system will be retrained to incorporate these new insights Bumacov et al. [2017]. This flexibility is absolutely crucial in micro-finance[6], where loans to low income or informal sector borrowers often come with limited historical data. As more loans are granted and repaid, the system will learn from actual outcomes and improve its capacity to predict future loan performance. Modern credit scoring systems use artificial intelligence and machine learning to automatically adjust to new borrower profiles, allowing financial institutions to continuously refine their risk assessments as they gather more experience. This ability to adapt guaranteed the scoring model's continued effectiveness and risk mitigation, even as borrower characteristics, behaviors, or market conditions change Bumacov et al. [2017].

1.4 Disadvantage of credit scoring

Disparate impact is the term used to describe the unintended and often hidden negative effects that credit scoring systems can have on certain demographic groups. Age bias is a significant concern in this regard. Credit scoring models frequently fail to adequately account for the unique financial circumstances of older borrowers, who often have a long, consistent history of responsible borrowing and

[6] *Micro-finance provides small loans and financial services to low-income individuals or groups, often to support entrepreneurship and reduce poverty.*

timely repayment. However, the structure of these models is flawed. Factors like the length and type of recent credit activity can and do disproportionately lower their scores, regardless of their overall financial stability. This results in older adults being unjustly denied access to credit or being offered loans with higher interest rates that do not accurately reflect their low credit risk Avery et al. [2012].

Younger borrowers also face challenges within these system.Young individuals are unfairly penalized simply because they have not yet had the opportunity to build up extensive credit records, despite their responsible financial management.These age-related biased within credit scoring systems create a cycle of disadvantage, with older and younger borrowers alike being assessed by criteria that do not fully capture their true creditworthiness.

Credit scoring models can greatly disadvantage recent immigrants or foreign born individuals, creating significant obstacles to their financial inclusion. It is a fact that many of these immigrants often do not have an extensive credit history in their new country, this is because credit scoring systems typically depend on historical data to forecast future credit behavior. As a result, they are often misidentified or receive low credit scores due to a lack of available data. Recent immigrants often lack formal credit histories but prove themselves financially responsible and able to use credit effectively, these scoring models also make it impossible for them to get credit, which in turn make it much harder for them to get loans, mortgages, or even rental agreements, which further marginalizes them in the financial landscape Avery et al. [2012].

Traditional credit scoring systems rely heavily on historical financial data, such as credit card payments, loan histories and other formal financial products. However, this reliance often leaves out individuals with limited access to these products, such as low-income earners, recent immigrants and those with non-traditional financial behaviour. These individuals may have stable financial behaviour but lack the formal credit history that these systems prefer, and as a result are unfairly penalised by standard credit scoring models that do not take alternative financial behaviour into account. Recent challenges have been explored in using alternative data sources, such as rent payments, utility bills or even payment data from bank accounts, as a way to improve financial inclusion for these groups. However, concerns remain about the fairness and privacy of using such data; for example, while cash-flow data can provide insights into an applicant's financial reliability, its use could disproportionately affect people from disadvan-

taged backgrounds who may have unpredictable or fluctuating incomes, especially if the data is not properly contextualized Vissing-Jorgensen [2021].

While the use of machine learning (ML) in credit scoring systems is efficient and accurate, there are some downsides. One of the main risks associated with ML is that algorithms often rely on historical data that may reflect existing biases and inequalities in the financial system; for example, some ML models incorporate factors such as postcodes, which may inadvertently capture geographic and racial disparities. This means that the algorithm may unfairly penalise individuals from historically marginalized communities, even if their actual creditworthiness is strong. In addition, the complexity of these algorithms, which are designed to optimise predictive power, can lead to a lack of transparency about how decisions are made. This 'black box' effect makes it difficult for both regulators and consumers to understand why certain individuals are denied credit or charged higher interest rates, further entrenching financial exclusion Ostrowski [2021].

1.5 The importance of rating downgrade prediction

Forecasting credit rating downgrades by rating agencies such as Moody and S&P [7]plays a key role in managing financial stability, guiding investment decisions and maintaining strategic flexibility for both banks and corporates Kisgen [2019].Credit downgrades typically leads to higher borrowing costs and reduced access to wholesale and public debt market, creating funding constraints for financial institutions and corporates alike. For banks, these constraints often result in reduced lending capacity, particularly for institutions that rely on short-term funding and wholesale markets.

According to Adelino and Ferreira [2016] credit downgrading doesn't just impact funding and lending, however often triggers regulatory and contractual responses that increase the financial pressure on lenders. Many corporate and bond contracts include "rating-related" covenants that, when triggered by a downgrade, can force companies to increase collateral holdings or face default. By anticipat-

[7] *Moody's and S&P assign credit ratings to debt instruments, helping investors assess issuer creditworthiness. Both play a key role in global financial markets.*

ing downgrades, banks and supervisors can proactively address these challenges by increasing loan reserves, adjusting lending strategies, and maintaining overall financial stability amid market volatility.

Accurate downgrade forecasting is equally important for non bank companies; according to Kisgen [2019] research on the impact of credit ratings on corporate behavior suggests that companies anticipating a rating downgrade often restructure their capital strategies, reducing debt issuance in anticipation of higher borrowing costs and to hedge against reduced access to funding. By anticipating potential downgrades, companies can make informed strategic adjustments, such as:

1. Securing alternative funding

2. Adjusting leverage ratios

3. Re-balancing portfolios

to meet expected financial market and regulatory requirements. In addition, reliable downgrade forecast help companies reduce the negative impact on their valuation and reputation, maintain investor confidence and ensure financial resilience even in difficult economic conditions. The significant influence of credit rating agencies on financial decisions underscores the need for predictive models to forecast rating changes, helping institutions and corporates alike to navigate a volatile financial landscape with foresight and stability Adelino and Ferreira [2016].

1.6 Ratings

1.6.1 Rating Agencies

Credit rating agencies (CRAs) play a crucial role in the global financial system by evaluating and assigning ratings to various financial instruments such as bonds, loans and other forms of debt. These ratings provide an independent as-

sessment of the creditworthiness of borrowers, including a company, government or financial institution. As objective third-party evaluators, credit rating agencies provide valuable guidance on the relative risks of investing in different financial instruments, helping investors to make informed decisions Jory et al. [2016].

One of the key functions of credit rating agencies is to assess the ability and willingness of issuers to meet their financial obligations. Ratings are assigned based on comprehensive analyses of an issuer's financial health, industry position, economic factors and governance practices. These ratings, which reflect the likelihood of default on debt repayments, are expressed as letters or symbols (e.g. AAA, BBB or junk status). Agencies such as Moody's, Standard & Poor's (S&P) and Fitch are among the most prominent global players in the credit rating industry, and their ratings are widely used by investors, regulators and financial institutions to assess the risk associated with debt instruments. For example, a higher rating indicates lower risk and generally leads to lower borrowing costs for issuers, while lower ratings signal higher risk and higher interest rates for borrowers Chen et al. [2016].

CRA also provide insights into market stability and investor confidence, in addition to their core function of assessing creditworthiness. Their opinions are perceived as independent and impartial because credit rating agencies are not involved in the day-to-day activities of financial markets. This perceived objectivity is essential to maintain the credibility of their ratings and ensure that they are trusted by investors, analysts and regulators. The integrity of credit rating agencies is critical to the stability of financial markets, especially in times of economic uncertainty or financial crises Partnoy [2002].

Credit rating agencies contribute to economic growth and financial stability by facilitating informed investment decisions and ensuring market transparency. However, the accuracy and reliability of their ratings have been called into question in the aftermath of financial crises, with some agencies accused of underestimating the risks associated with certain financial products. However, in the global financial ecosystem, the role of credit rating agencies in assessing financial risks remains indispensable despite these criticisms.

1.6.2 Rating Methodologies

Credit rating methodologies are frameworks used by rating agencies to assess the credit risks associated with financial entities, corporations, or sovereign entities. These methodologies aim to form an informed opinion on an entity's creditworthiness, often expressed as credit ratings. To achieve this, most rating agencies use a combination of statistical models and analyst-driven assessments, with the choice depending on the complexity and requirements of the evaluation process Altman and Saunders [1997].

Rating agencies often incorporate financial measures such as profitability ratios, leverage ratios and liquidity measures into these models to predict the likelihood of default or downgrade. For example, assessing a bank's asset quality may involve analysing its loan portfolio, non-performing assets and financial statements Caridad et al. [2020]. Advanced quantitative techniques, such as machine learning and statistical regression, are increasingly being used to improve the predictive accuracy of these models. These techniques allow large datasets to be processed and identify subtle patterns that indicate credit risk Gupta et al. [2020]. However, a significant limitation of purely quantitative approaches is their inability to account for qualitative factors, such as changes in market dynamics or management quality, which are often critical in assessing credit risk.

In contrast, the analyst-driven approach relies heavily on expert judgment and qualitative assessment. Analysts assess a variety of data sources, including:

1. Published financial reports: Balance sheets, income statements, and cash flow analyses.

2. Market sentiment: Stock prices, credit default swap spreads, and bond yields.

3. Issuer interviews: Direct discussions with the management or governing bodies to understand strategic goals, risk management practices, and governance quality.

For example, when assessing a company's creditworthiness, analysts may evaluate

macroeconomic conditions, industry trends and competitive positioning in addition to financial performance. This qualitative approach is particularly important for entities such as municipalities, where non-financial factors - such as political stability, regulatory frameworks and demographic shifts - play a significant role in assessing credit risk Ratings [2022].

Many rating agencies favor a hybrid approach, combining quantitative models with analyst judgment to provide a holistic view of credit risk. The quantitative component provides consistency and objectivity, while the qualitative component provides the flexibility to incorporate real-time information and expert insight. For example, a statistical model may identify financial stress in a company, while analysts may contextualise the findings by examining the company's strategic responses, such as cost-cutting initiatives or new revenue streams Altman and Saunders [1997].

Emerging Trends in Rating Methodologies:

1. ESG Considerations[8]: Environmental, social, and governance (ESG) factors are now an integral part of credit rating processes, used to assess long-term sustainability and reputational risks. For example, Moody's and S&P Global have introduced methodologies that evaluate an entity's ESG performance as part of its credit rating.

2. AI and Big Data: The adoption of artificial intelligence (AI) and big data analytics is transforming credit risk evaluation. These technologies enable rating agencies to analyze non-traditional datasets, such as social media sentiment, real-time economic indicators, and geopolitical risks, to enhance prediction accuracy Gupta et al. [2020].

Credit rating methodologies vary widely depending on the type of entity being evaluated and the goals of the assessment. Statistical models offer precision and scalability, while analyst-driven approaches ensure depth and adaptability. The combination of these methodologies allows rating agencies to provide balanced and comprehensive credit opinions, addressing both financial and non-financial aspects

[8] *ESG (Environmental, Social, and Governance) considerations refer to the criteria used to evaluate a company's operations in terms of sustainability, ethical impact, and corporate governance.*

of credit risk. As the financial landscape evolves, the integration of ESG factors and advanced technologies will further refine these methodologies, enhancing their relevance and effectiveness in a dynamic environment.

Chapter 2

Methodology

The Methodology chapter discusses the role of Credit Rating Agencies (CRAs) in assessing creditworthiness and the methodologies they use, combining quantitative models and analyst judgment. The chapter further introduces statistical techniques such as:

1. Logistic Regression

2. Ridge Regression

3. Random Forest

4. Gradient Boosting Machines (GBM)

in addition, the chapter discuss about the class imbalance to address the challenge of predicting rare events such as defaults in financial forecasts.

2.1 Logistic regression

According to (lavalley2008) logistic regression is a widely used statistical method for modeling binary response data that is particularly effective for predicting outcomes that fall into two categories:

1. Success/Yes

2. Failure/No

The model is based on the Bernoulli distribution [1], which is a subset of the binomial distribution, making it ideal for binary or dichotomous outcomes. Logistic regression estimates the probability of an event occurring based on one or more explanatory variables, allowing researchers to understand the relationship between predictors and binary outcome.

The logistic regression model is expressed as follows:

$$\text{Logit}(p) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Where p is the probability of occurrence, $\frac{P}{1-P}$ represents the odds for an outcome to occur, and $\beta_0, \beta_1, \ldots, \beta_n$ represent coefficients assigned to the independent variables $X_1, X_2, \ldots, X_n$ Peng et al. [2002]. The logit unction, which combines odds ratios[2] with the natural logarithm, represents a linear combination of predictor variables. This approach restricts the predicted values to the interval [0,1], making it well suited for probability estimation by ensuring that predicted probabilities do not exceed these limits Peng et al. [2002].

Logistic regression is typically estimated using maximum likelihood estimation (MLE), a method that identifies the set of parameters that maximizes the likelihood[3] of the observed data under the model. MLE construct the log- likelihood function, which allows researches to sum over observations to find parameter

[1] *The Bernoulli distribution describes a random variable with two possible outcomes, typically labeled as 0 and 1, where the probability of success is p and the probability of failure is* $1 - p$.
[2] *They measure the odds of an event occurring in one group relative to another.*
[3] *It is a function that represents how likely it is to observe the data for different values of a statistical model's parameters.*

estimates that best fit the data Hilbe [2011]. The aim of using maximum likelihood estimation in logistic regression is to find the coefficients β_i that maximize the likelihood of observing the given sample of binary outcomes.

The key advantage of logistic regression is its ability to be interpreted, particularly in the form of odds ratios. Each coefficient β_i is the change in the log odds of the outcome for a unit increase in the predictor variable X_i assuming all other variables are held constant Hilbe [2011]. This power to interpret provides insight into the strength and direct of the association between predictors variables and the likelihood of the event, which can be invaluable in fields such as social science, health and marketing, where understanding the effects of predictors on outcomes is crucial. In addition, logistic regression has the advantage of being computationally efficient, especially when compared to more complex machine learning models. This efficiency makes logistic regression well suited to large data set and situations where rapid decision making is essential James [2013]. Unlike models that may require intensive computational resources, logistic regression can be implemented easily, making it accessible in both research and applied contexts where computational speed is a priority Menard [2010]. furthermore, logistic regression relies on fewer assumptions about the predictor variables then linear regression, which further contributes to its robustness, as it does not require the predictor variables to be normally distributed or homogeneous[4] Field [2022].

However, logistic regression has some limitations, for example, the model can be affected by multicollinearity, where high correlations between predictor variables make it hard to interpret the effect of each individual predictor, as their explanatory power may overlap. This problem can be mitigated by using adjustment techniques, such lasso[5] or ridge regression, which impose penalties on the coefficients to minimize over fitting and reduce multi-collinearity Tibshirani [1996]. In addition, the predictive power of logistic regression can sometimes be limited by its linear boundary, making it less effective for complex, non-linear relationships in the data. Nevertheless, logistic regression remains a fundamental tool in statistical analysis and is continually adapted to suit different analytic situations Hosmer Jr et al. [2013].

[4] *Homogeneous refers to elements that are uniform or identical in composition or structure.*
[5] *Lasso (Least Absolute Shrinkage and Selection Operator) performs variable selection by shrinking the coefficients of less important predictors to zero.*

2.2 Random forest

Random forests are a popular machine learning algorithm and ensemble method[6] introduced by Breiman [2001]. This approach combines multiple decision trees to improve the overall performance of the model, using majority voting for classification tasks or averaging predictions for regression tasksBreiman [2001]. Known for their versatility and effectiveness, random forests can handle both structured data and have been widely applied in various domains, including healthcare, finance and environmental science Yang et al. [2022]. Random forests combine numerous trees constructed from different random subsets of data and features to reduce over fitting, a common limitation of individuals decision trees. This approach, known as bagging (bootstrap), is critical for generating robust and accurate predictions across diverse data sets, especially those with high dimensionality Biau and Scornet [2016].

The random forest algorithm constructs multiple decision trees by randomly selecting subsets of data for training and randomly selecting feature at each node, a random subset of features is selected to determine the best split, minimizing over fitting by ensuring diversity across the trees and improving the model's ability to generalise. This randomness helps prevent over fitting by reducing the correlation among the trees, allowing the model to generalize better to unseen data Denisko and Hoffman [2018]. To improve predictive power and reliability, random forests use two main techniques:

1. Bootstrap Sampling: Each tree in the forest is trained on a randomly selected subset (with replacement) of the training data, leading to diversity in the training sets across trees Denisko and Hoffman [2018].

2. Feature Randomization: Rather than considering all features for each split, random forests limit the candidate features, thus increasing the independence among trees and enhancing model performance by focusing on the most relevant features Biau and Scornet [2016].

The final prediction is obtained by aggregating the predictions from each tree;

[6] *Ensemble methods combine multiple models to improve predictive accuracy*

in classification, a majority voting approach determines the final label, while in regression, the prediction is averaged across all trees Maturo and Porreca [2024].

Random forests have several advantages that make them one of the most widely used machine learning models:

1. High predictive accuracy: Random forests often outperform individual models by reducing variance, as errors from individual trees are averaged across the ensemble Biau and Scornet [2016].

2. Robustness to Noise and Over fitting: Due to their ensemble structure, random forests are less prone to over fitting, even when trained on high-dimensional data with a large number of features relative to the number of observations Denisko and Hoffman [2018].

3. Feature Importance: Random forests provide measures of feature importance, allowing researchers to identify and focus on the most important variables in their data, which is especially useful in complex fields like Genomics and Finance Scornet [2016].

4. Capability with Missing Data: Unlike other algorithms that require imputation, random forests can handle missing data by employing proximity weights in predictions, which allows the model to use similar observations in place of missing values Breiman [2001].

Despite their significant advantages, random forests also face challenges and limitations that can impact their performance and applicability, especially in real-world, large-scale settings:

1. Computational Complexity: Training a large number of trees can be computationally intensive, which may limit random forests' feasibility in applications with strict time constraints or on devices with limited processing power Maturo and Porreca [2024].

2. Lack of Interpretation: Compared to single decision trees, the ensemble nature of random forests makes them less interpretable, as the final output is a composite of many trees rather than a single, easy-to-follow decision path Biau and Scornet [2016].

3. Over fitting on Noisy Data sets: Although less susceptible than single trees, random forests can over fit when there is excessive noise in the data, which can lead to less stable predictions Scornet [2016]

2.3 Ridge Regression

Ridge regression, also known as Tikhonov regularisation, addresses problems of multi-collinearity and over fitting in linear regression models by introducing an L2 penalty term to the loss function. This modification ensures stability in coefficient estimates by shrinking their magnitude, making it particularly effective in data sets with highly correlated predictors or when the number of predictors exceeds the number of observations Hoerl and Kennard [1970]. The ridge regression objective function is defined as:

$$Loss = \sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.1}$$

The first term minimizes the residual sum of squares (RSS), while the second term imposes a penalty proportional to the square of the regression coefficients. The parameter λ controls the trade-off between RSS minimization and coefficient shrinkage. A higher λ increases the regularization, reducing the variance at the cost of increased bias, thereby improving model commonality Hastie [2009].

Ridge regression, a type of linear regression, differs from lasso regression in that it keeps all predictors in the model by shrinking their coefficients toward zero instead of eliminating some entirely. This feature makes ridge regression particularly effective in areas like genomics and financial modeling, where every variable may have some predictive value Martino et al. [2024]. In finance, it has been successfully used to predict credit downgrades and assess credit risk by accounting for correlated factors such as interest rates and inflation. By addressing multicollinearity, ridge regression enhances the stability and accuracy of predic-

tions, particularly in high-dimensional data sets James [2013]. The regularization parameter λ, which determines the degree of penalty on the coefficients, plays a key role in ridge regression. The optimal value of λ is typically chosen through cross-validation, which helps balance under fitting and over fitting. This process improves the model's ability to generalize to new data, making ridge regression a valuable tool in regression analysis Hastie [2009]. Unlike methods that promote sparsity, ridge regression's strategy of retaining all predictors is beneficial in situations where interactions between variables or their individual effects are small but collectively important Rajamani and Iyer [2023].

Despite its lack of feature selection, ridge regression remains a cornerstone for regression tasks in high-dimensional spaces, its effectiveness in controlling over fitting and mitigating the effects of multicollinearity ensures reliable and interpretable results, making it indispensable in various fields where predictive accuracy and robustness are critical. For example, ridge regression is widely applied in genomics, where all predictors may contribute to the outcome, and in financial modeling, where multicollinear economic indicators must be accounted for effectively Rajamani and Iyer [2023], Martino et al. [2024]. As a result, it continues to play a vital role in domains that require robust predictive modeling and consistent performance in high-dimensional data sets.

2.4 Gradient Boosting Machines

Gradient Boosting Machines (GBMs) are an ensemble learning method that combines multiple weak learners, usually decision trees, to form a powerful predictive model. This approach constructs trees in a sequential manner, with each new tree being trained to correct the errors (residuals) of the previous ones, allowing the model to improve progressively. By focusing on these residuals, GBMs are able to capture complex patterns in the data, making them highly effective for tasks like classification, regression, and ranking Friedman [2001]. However, a key challenge of GBMs is the potential for over fitting, especially if the model becomes overly complex or is trained on noisy data.

To mitigate over fitting and improve model generalisation, ridge regression is sometimes incorporated into the gradient boosting process. Ridge regression, a regularisation technique, introduces an L2 penalty to the model coefficients,

shrinking them and reducing the risk of over fitting. In the context of GBM, this regularisation is applied during each boosting iteration, allowing the model to control the size of its coefficients and thus the complexity of the learned model. By using regularisation, Ridge Gradient Boosting (Ridge GBM) helps to balance the bias-variance trade-off, improving stability and prediction accuracy Chen and Guestrin [2016], Friedman [2001]. This integration of regularisation into boosting has shown significant improvements, particularly in high-dimensional datasets or situations where the number of features is much larger than the number of observations.

According to Ke et al. [2017] ridge Gradient Boosting has demonstrated strong performance in various applications, including financial modeling, where high-dimensional and noisy datasets are common. It is particularly effective in tasks such as credit rating and downgrade predictions, where accurate, robust models are crucial. The combination of gradient boosting's ability to capture complex relationships with the regularization of ridge regression makes this approach especially useful in domains that require both high predictive power and interpretability. Ridge GBM's ability to generalize well while avoiding over fitting in high-dimensional spaces makes it an important tool in predictive analytics, especially when facing challenges like multicollinearity and noisy variables Bühlmann and Yu [2003].

2.5 Class Imbalance

Class imbalance is a critical issue in many machine learning applications, particularly in financial prediction tasks such as credit ratings, where downgrades and defaults are rare compared to stable or upgraded ratings. This imbalance can severely affect the performance of predictive models, as algorithms may bias their predictions toward the majority class (typically non-default or non-downgrade events), thereby failing to effectively detect the minority class. Misclassification of minority events is particularly problematic in domains like finance, where accurately predicting rare, high-impact events (such as defaults or downgrades) is crucial for effective risk management and informed decision-making He and Garcia [2009], Provost [2000].

To address class imbalance, several resampling techniques adjust the class

distribution in the training data. One prominent method is the Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic samples for the minority class by interpolating between existing data points. By increasing the minority class, SMOTE can improve model sensitivity to rare events Chawla et al. [2002]. However, SMOTE also carries risks such as introducing noise and over fitting by replicating patterns that may not generalise well to unseen data Blagus and Lusa [2013]. Variants of SMOTE, such as Borderline-SMOTE and ADASYN, have been developed to address these limitations by focusing on difficult-to-classify samples near decision boundaries Han et al. [2005], He et al. [2008]. Conversely, under sampling techniques that reduce the size of the majority class offer a different approach. While effective in balancing the dataset, under-sampling can result in the loss of valuable information from the majority class, potentially reducing the overall predictive power of the model Buda et al. [2018]. Hybrid techniques that combine over- and under-sampling aim to exploit the strengths of both methods while mitigating their weaknesses Yen and Lee [2009].

Another strategy for handling class imbalance involves modifying algorithms to account for imbalanced data. Weighted loss functions, for example, assign higher penalties to misclassification of the minority class, encouraging the model to prioritize these instances during training He and Garcia [2009]. Careful tuning of the weights is essential, as excessive emphasis on the minority class can lead to biased predictions and poor performance on the majority class. Cost-sensitive learning extends this idea by incorporating the misclassification costs of different classes directly into the learning process. This approach is particularly well-suited for financial domains, where the costs of false positives (predicting a downgrade when none occurs) and false negatives (failing to predict a downgrade) differ significantly Sun et al. [2007], Elkan [2001].

Ensemble methods, such as bagging and boosting, have also shown promise in dealing with class imbalance. These techniques combine the predictions of multiple models to improve overall accuracy. For example, Adaptive Boosting (AdaBoost)[7],can be adapted to unbalanced data by adjusting the weights of misclassified instances, ensuring that the model focuses on difficult cases Freund and Schapire [1997]. Similarly, Random Forests can be adapted to unbalanced datasets by modifying their sampling strategies or by using class-balanced de-

[7] *AdaBoost is an ensemble learning method that combines the predictions of multiple weak learners to create a strong predictive model. The goal is to improve the model's accuracy by focusing on the difficult-to-classify instances.*

cision trees Chen et al. [2004]. Recent advances in ensemble learning, such as Balanced Random Forests and EasyEnsemble, have demonstrated robust performance in unbalanced scenarios Chen et al. [2004].

Despite these strategies, dealing with class imbalance remains a complex challenge that requires careful consideration of both the data and the model. Over fitting and under fitting are common pitfalls, especially when using over sampling methods that may amplify noise or under sampling methods that discard informative data Fernández et al. [2018]. The choice of method often depends on the specific application and the relative costs of false positives and false negatives. For example, in financial forecasting, where the cost of a false negative (missing a default) can far outweigh a false positive, methods that priorities sensitivity to minority classes are often preferred Sun et al. [2007]. In addition, recent advances in deep learning have introduced novel approaches to dealing with class imbalance. Techniques such as focal loss, which dynamically down-weights the loss assigned to well-classified examples, have shown promise in improving model performance on unbalanced datasets Lin [2017]. In addition, transfer learning and pre-training on related tasks have been explored as ways to mitigate data scarcity and class imbalance in specialised domains such as financial risk assessment Weiss et al. [2016].

Chapter 3

Data Analysis

This chapter compares two companies: modefinance and Cardo AI. It begins with an analysis of modefinance, discussing their confusion matrix, AUC, sampling methods and feature importance. The same sections are then applied to the analysis of Cardo AI. This allows a direct comparison between the two companies' approaches.

3.1 Case 1 (modefinance)

modefinance is a European recognized credit rating agency (CRA and ECAI) known for its proprietary MORE methodology, making its rating legally valid. As Fintech Company, it's specialized in AI solutions, process automation, and data analysis for credit risk management. Modefinance develops custom models and cloud platforms, including APIs, to digitize credit risk, portfolio, and supply chain management.

modefinance offers financial companies and businesses a comprehensive solution for the daily management of exposure risk. Its services are based on rigorous evaluation process that ensure the highest standards of quality, transparency, and reliability. In this capacity, it has contributed to the European TranspArEEnS project (within Horison 2020), which has focused on the study and definition of guidelines for the evaluation of ESG ratings for small and medium-sized enterprises. To assist users in developing personalized management policies that

simultaneously guarantee maximum flexibility in configuration and the most sophisticated credit scoring capabilities, mode finance has developed Tigran, a risk management platform. Tigran integrates the functions and services of a Fintech rating agency into a modular web platform for counterpart risk assessment and investment and exposure portfolio management, designed for financial institutions, investment funds, Fintech companies, and holding companies.

3.1.1 Summary Statistics

Summary statistics are essential tools in data analysis, providing a concise representation of complex data sets. These metrics are invaluable in identifying patterns, trends and anomalies, allowing researchers and analysts to effectively compare data sets and gain deeper insights into their structure. By reducing large data sets to easily interpretable values, summary statistics facilitate decision-making in fields as diverse as finance, healthcare and social sciences McDermott et al. [2013]. Also it can be broadly can be divided into measures of central tendency and measures of variability:

1. Measures of Central Tendency:

    - These statistics identify the central point of a dataset.

    - The mean provides the average value, which is a fundamental descriptor of a data set's overall trend.

    - The median, the midpoint value, is especially useful in datasets with skewed distributions, as it is unaffected by outliers.

    - The mode, which identifies the most frequently occurring value, is valuable in categorical datasets or data with repeated patterns Stevens [2013], McDermott et al. [2013]

2. Measures of Variability:

- These provide insights into the spread or dispersion of data points.
- The range, defined as the difference between the maximum and minimum values, offers a simple yet effective overview of data variability.
- The variance measures the average squared deviation from the mean, quantifying the extent of data dispersion.
- The standard deviation, the square root of the variance, is a more intuitive measure, indicating how much data points typically deviate from the mean Stevens [2013]

In addition to numerical summaries, graphical tools such as histograms, box plots and pie charts play an important role in visually summarising data. Histograms show frequency distributions, showing how data points are distributed across ranges. Box plots provide a five-digit summary (minimum, first quartile, median, third quartile and maximum), highlighting the spread of the data and any outliers. Scatter plots and pie charts provide a deeper understanding of relationships and proportions within the data set McDermott et al. [2013]; these visual tools are indispensable for presenting data insights to non-technical audiences effectively.

The modefinance dataset, which contains 178,293 observations across 20 variables, which illustrates the value of summary statistics in financial data analysis, including:

- ISIN number
- current_rating
- current_rating_date
- dd
- cash_ratio_fuzzy
- cash_to_st_financial_debt_fuzzy

- current_ratio_fuzzy
- debt_to_capital_fuzzy
- ebit_interest_coverage_ratio_fuzzy
- financial_debt_to_ebitda_fuzzy
- interest_paid_weight_fuzzy
- leverage_ratio_fuzzy
- net_fixed_assets_turnover_fuzzy
- operating_revenue_total_assets_fuzzy
- roe_ratio_fuzzy
- roi_ratio_fuzzy
- shareholders_funds_fixed_assets_fuzzy
- cash_flow_ratio_fuzzy
- ebit_margin_fuzzy
- ebitda_margin_fuzzy

These variables represent financial ratios derived from publicly available company financial statements. To standardise and analyse these variables, modefinance uses a fuzzy transformation methodology; Fuzzy is an evaluation function that transforms ratio to the standard scale from 0 to 1 which (0 is being the best and 1 being the worse). This transformation mitigates the impact of extreme values and ensures that all variables are on a comparable scale, increasing the robustness of subsequent analysis.

Our primary objective is to examine how a company's credit rating evolves over time, with a particular focus on downgrades, which can have significant negative consequences for investors and the company's overall financial performance. Among the various financial indicators, a company's 'rating' is one of the most important and serves as a key measure of its creditworthiness.

This rating scale ranges from 1 to 22, with a rating of 1 signifying the highest level of creditworthiness, reflecting exceptional financial stability, robust performance, and a very low risk of default. On the other end, a rating of 22 indicates the lowest level of financial strength, pointing to a company's susceptibility to financial distress and a significantly higher risk of default. By utilizing this comprehensive rating system, we are able to conduct an in-depth assessment of a company's financial health over time, closely monitoring how their credit ratings evolve in

response to both internal factors (such as financial management, profitability, and growth strategies) and external factors (including economic conditions, market trends, and industry changes). This dynamic approach enables us to identify potential risks, detect emerging patterns, and anticipate trends that could affect a company's ability to meet its financial obligations, providing valuable insights for investors. Ultimately, this analysis helps investors make informed decisions by evaluating the potential risks and opportunities associated with investing in companies with varying levels of creditworthiness.

We have effectively compiled detailed profiles of the companies, including key information on the specific periods during which their ratings were downgraded shown in figure 3.1. In addition to identifying these critical times, the extent and direction of each rating change was carefully documented, recording both the size of the downgrading and the precise movement within the rating scale.

This comprehensive approach allows us not only to identify when these changes occurred, but also to quantify the degree of deterioration in a company's creditworthiness. By meticulously tracking these movements, we can analyse patterns and assess the broader impact of rating changes on financial stability and investor confidence in these companies. This detailed catalogue serves as a basis for the understanding of the following factors that drive rating changes and provides a rich dataset for further analysis of financial trends and risk factors.

The next phase of our analysis was to examine the frequency of rating downgrades across the companies in our data set. Our findings were striking: as can be noticed in figure 3.2, only 0.9% of companies experienced a rating downgrade, while around 99% either maintained their existing rating or saw an improvement. This suggests that the vast majority of companies have demonstrated a high degree of financial stability, or even growth, as reflected in their credit ratings.

These results are particularly significant as they are a strong indication of a favourable economic climate in which companies are effectively managing their financial health and overcoming external challenges. The extremely low downgrade rate reflects not only successful corporate management strategies, but also the generally positive market conditions that have supported their financial performance.

This high rate of rating stability or improvement signals that companies are

ISIN number	Downgrade_Periods	Ratings	Directions
AN8068571086	2016-04-01	5	Downgrade
AT0000641352	2021-11-01	10	Downgrade
AT0000720008	2013-10-01	9	Downgrade
AT0000741053	2012-08-01	8	Downgrade
AT0000743059	2016-03-01	8	Downgrade
AT0000746409	2013-08-01, 2016-04-01	8, 9	Downgrade, Downgrade
AT0000831706	2013-08-01	13	Downgrade
AU000000BHP4	2016-03-01	6	Downgrade
AU000000BLY8	2013-07-01, 2013-09-01, 2014-02-01, 2014-03-01, 2014-07-...	14, 15, 16, 17, 18, 18, 19, 20	Downgrade, Downgrade, Downgrade, Downgrade, Downgra...
AU000000CCL2	2014-04-01	8	Downgrade
AU000000EHL7	2015-02-01, 2015-08-01, 2016-01-01, 2016-09-01, 2017-03-...	15, 16, 17, 20, 21	Downgrade, Downgrade, Downgrade, Downgrade, Downgra...
AU000000FMG4	2015-04-01	12	Downgrade
AU000000IPL1	2016-11-01	10	Downgrade
AU000000MBN9	2013-10-01, 2013-12-01	20, 22	Downgrade, Downgrade
AU000000NCM7	2013-04-01, 2013-08-01	9, 10	Downgrade, Downgrade
AU000000NUF3	2014-07-01	13	Downgrade
AU000000ORG5	2013-02-01, 2015-08-01	9, 10	Downgrade, Downgrade
AU000000QAN2	2013-12-01, 2014-01-01	11, 12	Downgrade, Downgrade
AU000000QIN5	2017-05-01, 2017-06-01, 2017-07-01, 2017-08-01	17, 19, 20, 22	Downgrade, Downgrade, Downgrade, Downgrade
AU000000SBM8	2013-11-01, 2014-05-01	16, 17	Downgrade, Downgrade
AU000000STO6	2014-12-01, 2016-01-01	9, 10	Downgrade, Downgrade
AU000000WOW2	2015-08-01, 2016-03-01	8, 9	Downgrade, Downgrade
AU0000061897	2014-03-01, 2015-02-01	13, 14	Downgrade, Downgrade
BE0003593044	2013-03-01	10	Downgrade
BE0003822393	2016-09-01	8	Downgrade
BE0003826436	2012-08-01	14	Downgrade
BE0974276082	2019-05-01	13	Downgrade
BE0974293251	2016-05-01, 2018-12-01	7, 8	Downgrade, Downgrade
BE0974294267	2016-03-01, 2018-09-01, 2019-02-01, 2019-03-01	17, 17, 19, 22	Downgrade, Downgrade, Downgrade, Downgrade
BG1100005997	2015-04-01, 2015-07-01	15, 16	Downgrade, Downgrade

Figure 3.1: This table shows which companies had downgrade. Source: personal elaboration

increasingly able to optimise their operations, capitalise on market opportunities and effectively mitigate risks. This resilience and upward momentum in ratings underscores the strength of both individual companies' financial strategies and the broader economic environment in which they operate. On the other hand, these results also highlight a tendency among rating agencies to prioritize maintaining stability in their ratings. By assigning ratings that remain consistent or even improve over time, agencies may be signaling confidence in the broader market's resilience and reducing the likelihood of triggering panic or market disruption.

This tendency towards stability in ratings could suggest a more cautious approach by the agencies, focusing on minimizing volatility in the credit markets while still reflecting companies' generally positive financial trajectories. The minimal number of downgrades also suggests that companies are well-positioned to maintain or improve their creditworthiness, signaling confidence in their long-term financial stability and future growth potential.

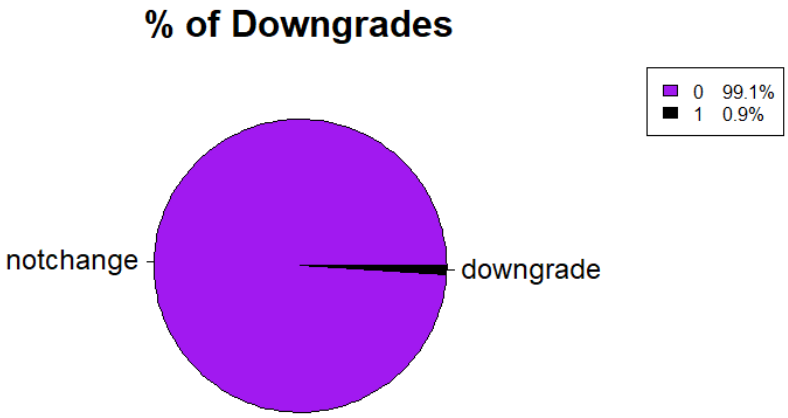

Figure 3.2: This pie chart shows the percentage of the total downgrade companies.
Source: personal elaboration

After the initial data pre-processing, which included addressing missing values and ensuring the dataset was clean, our final database contained 123,504 observations. With this refined dataset, we proceeded to analyse the relationships between multiple variables using a correlation matrix shown in figure 3.3. This matrix serves as a powerful tool to illustrate the statistical dependencies between variables, allowing us to assess the strength and direction of their associations Hadavand-Siri and Deutsch [2012].

The correlation coefficient, a key measure derived from the matrix, quantifies the linear relationship between two variables. It ranges from +1 to -1, where +1 indicates a perfect positive correlation (i.e., as one variable increases, the other also increases proportionally), and -1 indicates a perfect negative correlation (i.e., as one variable increases, the other decreases proportionally). Values closer to zero suggest little to no linear relationship between the variables, while values closer to +1 or -1 signal stronger relationships. It's important to note that the correlation coefficient cannot exceed these limits of ± 1, as this defines the maximum possible strength of a linear relationship Hadavand-Siri and Deutsch [2012].

The correlation matrix provides valuable insight into how different variables interact with each other, allowing us to identify potential redundancies, key influences, or patterns within the data. A high correlation between two variables, for instance, could suggest they are capturing similar underlying factors, while low

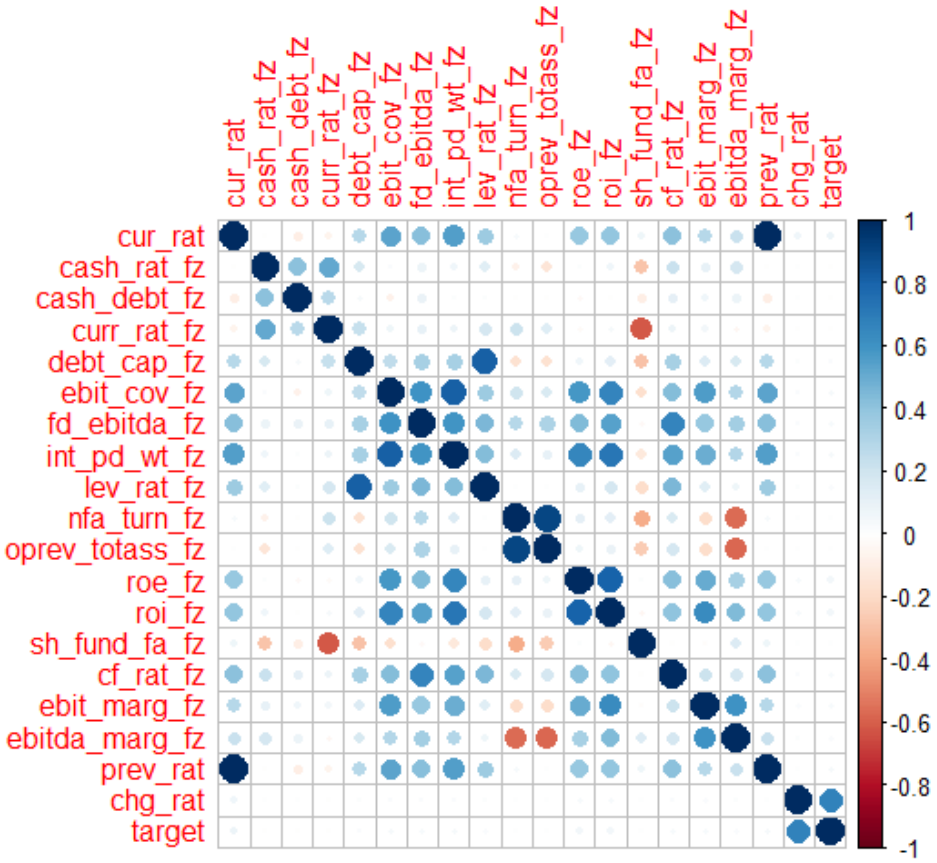

Figure 3.3: This correlation graph shows the correlation between variables. Source: personal elaboration

or near-zero correlation may indicate that the variables are independent of each other. This statistical tool, therefore, not only reveals the nature of dependencies but also helps guide further analytical steps, such as feature selection or modeling strategies James [2013], Cohen [2013].

Looking at the correlation matrix, we found that a significant proportion of the variables had high correlation coefficients, suggesting that the information supplied by them was redundant. This redundancy can lead to problems like multicollinearity, where strongly correlated variables confound the accuracy and interpretability of the model. To address this, we implemented a filtering process that eliminated variables with correlation coefficients greater than 70%. This criterion was chosen to ensure that each of the remaining variables provided unique, non-redundant information to the model; after this stringent filtering, 12 variables remained that met the selection criteria the graph can be seen in figure3.3, these 12 variables have been identified as the most appropriate to include in the next phase of analysis and modeling to build more accurate and reliable forecasts.

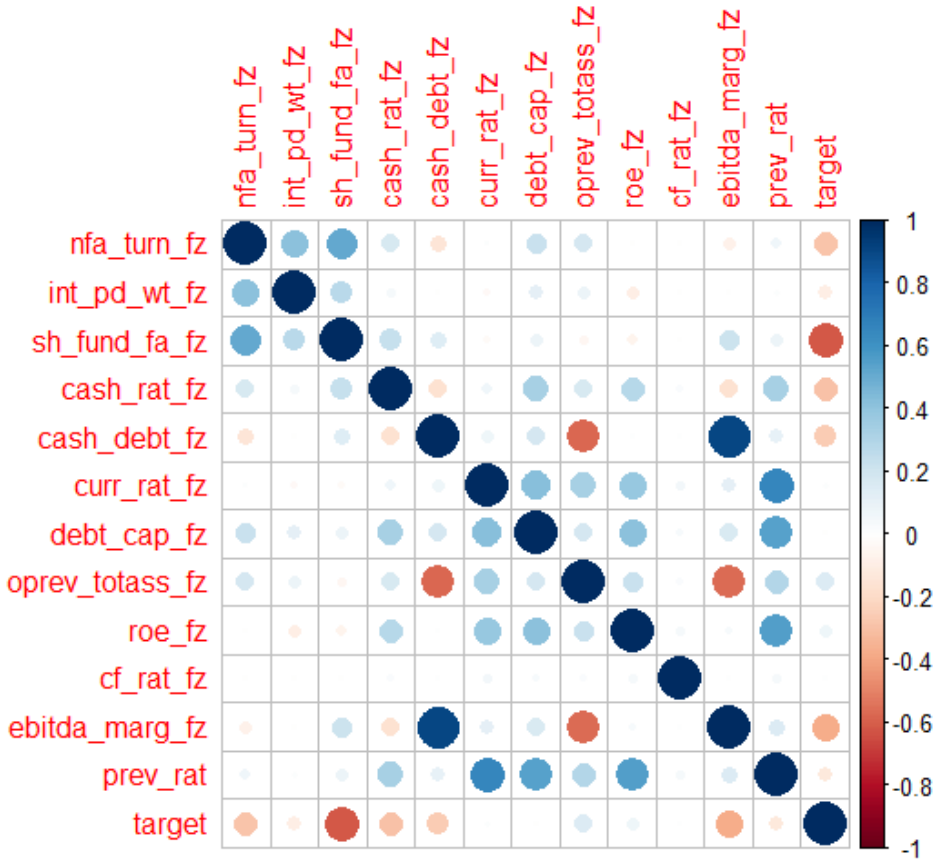

Figure 3.4: This correlation is after deleting the highly correlated variables. Source: personal elaboration

.

3.1.2 Train & Test

By using a spatially aware train-test split, we ensure that our models are trained and evaluated in a way that more accurately reflects real-world scenarios. This approach mitigates the risk of bias that can arise from ignoring spatial autocorrelation, a common problem with traditional random splits. By preserving the spatial structure of the data, the method ensures that both the training and test data sets are similarly complex to predict. This avoids the problem of splitting the data in a way that might result in overly easy test sets (due to spatial proximity) or overly difficult test sets (due to spatial separation). Such an approach is particularly important in fields where spatial relationships are intrinsic to the data, such as geostatistics, subsurface modeling and environmental science Babaei et al. [2023].

To begin the modeling process, we partitioned our refined database into a training set, which constituted 70% of the data (86,453 observations), and a testing set, which made up the remaining 30% (37,051 observations). This division enables us to develop robust models that are properly validated against unseen data, while also accounting for the spatial dependencies that are inherent to our dataset.

3.1.3 Model Selection

Model selection is a critical step in the machine learning process, where the goal is to identify the best performing model from a set of candidate models. Each model may use different algorithms, hyperparameters or techniques, but they all attempt to solve the same problem in a unique way. Since different models have different strengths and weaknesses depending on the dataset and the task at hand, it is crucial to select the one that best generalises to unseen data.

The primary goal is to select a model that not only performs well on the training set, but also makes accurate, reliable predictions on new, unseen data - a necessity for real-world applications Bengio [2009]. As noted by Zhang et al. [2023], model selection helps mitigate the risk of over fitting, where a model performs well on training data but fails to generalise to unseen data. By optimising performance and ensuring generalisation, model selection plays an important role in improving a model's ability to make accurate predictions.

This process often involves experimenting with different algorithms and hyperparameters, ultimately selecting the configuration that maximises key performance metrics such as accuracy, precision or recall. In addition, model selection helps to balance the bias-variance trade-off, ensuring that the chosen model is neither too simple (under fitting) nor too complex (over fitting), both of which can undermine the model's predictive power Hastie [2009].

The three main benefits of choosing a model are:

1. It improves prediction accuracy by ensuring that the best model is selected after comparing several alternatives. This is particularly important for complex tasks such as classification, regression, image recognition and natural

language processing Goodfellow [2016].

2. It provides flexibility by allowing different machine learning algorithms to be evaluated and tested for the best fit to the dataset.

3. Through the use of validation and testing procedures, model selection minimises over fitting and under fitting, ensuring that the model captures the true underlying patterns in the data without simply memorising them. As Raschka and Mirjalili [2019] highlights, confidence intervals can be used, for example by bootstrapping, to assess the uncertainty and variability of a model's performance on new, unseen data.

In this project, we began by applying several regression models to the data set. The first model applied was logistic regression, followed by ridge regression, gradient boosting machines (GBM), and finally, random forest. To address potential class imbalance in our binary classification task, we utilized resampling techniques across all models. These techniques adjust the data set by either reducing the number of instances in the majority class (under sampling) or increasing the number of instances in the minority class (over sampling). Under sampling helps mitigate the bias of having too many examples from the majority class, while over sampling generates new instances for the minority class, providing the model with a more balanced data set Chawla et al. [2002], Pranto et al. [2020].

Since our dataset involves binary classification, we utilized the (ROC) curve as a primary tool to assess the performance of our models. The ROC curve is a visual representation that plots the true positive rate (sensitivity) against the false positive rate (1-specificity) for all potential classification thresholds. It helps evaluate and compare model performance across different thresholds and is especially useful for assessing models at varying error rates Fawcett [2006].

The term "ROC" originates from communication theory and reflects its historical background James et al. [2023]. Along with the ROC curve, we calculated the Area Under the Curve (AUC), which measures the model's overall ability to distinguish between positive and negative classes. The AUC value ranges from 0 to 1, where a higher value indicates better performance. An AUC of 0.5 suggests the model has no discriminatory power (similar to random guessing), while an

AUC of 1.0 represents perfect classification Bradley [1997].

The results of our analysis showed that the regression models used (logistic regression, ridge regression and GBM) performed consistently, with error rates between 63% and 65%. However, Random Forest showed improved performance, especially after resampling, highlighting its ability to handle the complexity of our data set. This analysis highlights the importance of model selection and the role of techniques such as resampling in improving model robustness and generalisation.

3.1.4 Random Forest

Confusion Matrix Analysis

According to Weihs and Ickstadt [2018] the confusion matrix is a fundamental tool in machine learning and statistics for evaluating the performance of classification models. It provides a structured framework for comparing actual and predicted classifications, providing insight into a model's predictive strengths and weaknesses. The Confusion Matrix has four key components for binary classification problems:

1. True Positives (TP), where the model correctly predicts the positive class.

2. True Negatives (TN), where the model correctly predicts the negative class.

3. False Positives (FP), where the model incorrectly predicts the positive class.

4. False Negatives (FN), where the model fails to identify true positives.

By distinguishing between correct and incorrect predictions, this matrix provides a concise summary of the model's performance.

From the confusion matrix, several critical metrics can be derived to evaluate classification models comprehensively:

1. Accuracy, which measures the overall proportion of correct predictions, is the basic measure of performance.

2. Precision, also called Positive Predictive Value, quantifies the reliability of positive predictions by calculating the ratio of true positives to the total number of positive predictions.

3. Recall or Sensitivity reflects the model's ability to correctly identify true positive cases.

4. F1 Score, the harmonic mean of precision and recall, balances these two metrics, making it particularly useful for data sets with class imbalances Saito and Rehmsmeier [2015].

5. For multi-class classification, the confusion matrix is expanded to include additional classes, with rows representing actual classes and columns representing predicted classes, making it a scalable method for evaluating different scenarios Meedeniya [2023].

Following the training and fitting the Random Forest model, we assessed its performance using a Confusion Matrix, a validated approach to the evaluation of credit scoring models, as noted by Zeng [2020]. Significant insight into the classification performance of the model was provided by the Confusion Matrix results, for cases where the actual class was 0 (negative outcomes), the model demonstrated remarkable predictive accuracy, correctly identifying 34807 cases and classifying only 236 cases, achieving high reliability in predicting negative outcomes as shown in the figure3.5.

This strong performance for negative cases highlights the model's ability to deal effectively with majority class predictions. However, for cases where the true class was 1 (positive results), the model showed a significant performance gap. Only 84 true positives were correctly identified, while 1924 cases were misclassified as negative. This disparity suggests challenges in accurately identifying positive outcomes as shown in the figure3.5, which are often indicative of minority class cases.

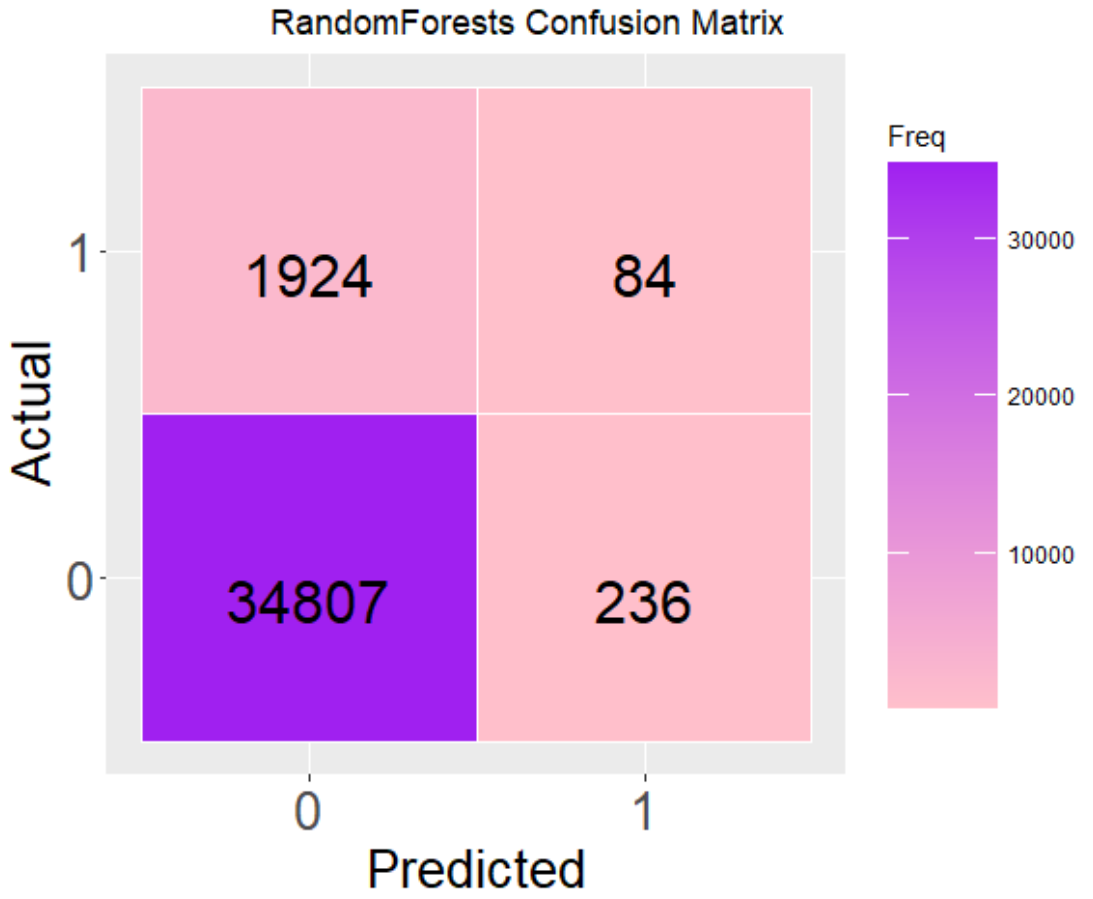

Figure 3.5: This table shows the confusion matrix in random forest. Source: personal elaboration

.

Several key metrics were derived from the confusion matrix to assess the performance of the model, including accuracy, precision, sensitivity and F1 score. While the model achieved good accuracy, this metric alone was insufficient due to class imbalance, as it primarily reflected the model's success in predicting the majority class. Precision, also known as positive predictive value, was notably low for the positive class, indicating a high prevalence of false positives; this result highlights the model's struggle to effectively distinguish between true positives and false positives. Similarly, recall (or sensitivity) was also low for positive cases, revealing the model's limited ability to identify true positives - a significant limitation in critical applications such as fraud detection or medical diagnosis López et al. [2013]. Finally, the F1 score, which balances precision and recall, reinforced the need to address class imbalance to improve the model's predictive performance for minority classes. Taken together, these metrics highlight the

importance of refining the model to ensure balanced and effective classification.

Receiver Operating Characteristic

Once the Random Forest model had been trained, we evaluated its performance by means of the (ROC) curve. This curve is a critical evaluation tool in machine learning, especially for binary classification problems, as it visually represents the trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate) at different classification thresholds. The ROC curve plots the True Positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis.

The ideal curve approaches the upper left corner, indicating high sensitivity (maximising true positives) and low false positive rates. This makes ROC analysis particularly valuable in data sets with class imbalance, where models may otherwise show misleadingly high accuracy by preferring the majority class James et al. [2023].

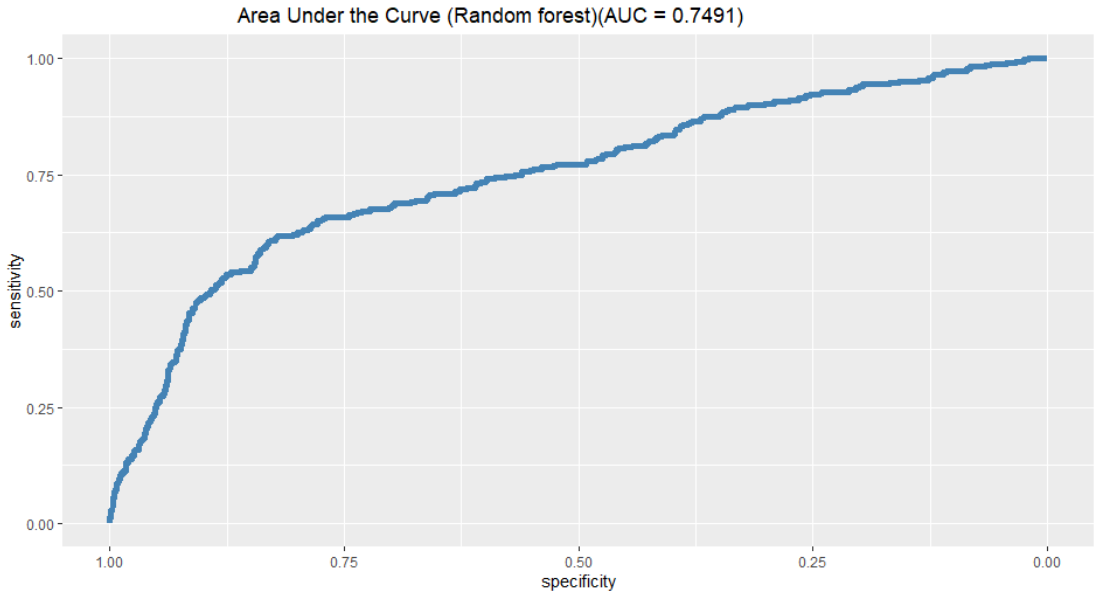

Figure 3.6: This graph shows the ROC curve for random forest. Source: personal elaboration

.

The (AUC) is a numerical metric derived from the (ROC) curve that summaries the classification performance of a model. AUC values range from 0.5,

indicating no discrimination between classes, to 1.0, indicating perfect discrimination; higher AUC values are indicative of better model performance as they reflect the model's ability to effectively separate positive and negative classes. In this context, a higher AUC suggests that the model is able to accurately classify instances from both classes, with minimal overlap between them James et al. [2023]. In our study, the initial Random Forest model achieved an AUC of approximately 75% shown in figure 3.6. While this value demonstrates a reasonable level of performance, it also shows that there is room for further improvement. Specifically, addressing class imbalance could improve the model's performance; when a dataset is imbalanced, the model may be biased towards predicting the majority class, which can limit its ability to accurately classify the minority class. Adjusting for this imbalance, either through resampling techniques or class weight adjustments, could improve the model's ability to discriminate between positive and negative outcomes, leading to a higher AUC and ultimately more reliable predictions.

Sampling

Recognizing that class imbalance was a significant challenge in our dataset, where positive cases (downgrades) represented only 1% of the total data, we implemented resampling strategies to rebalance the data and improve the model's ability to generalise effectively. When dealing with unbalanced datasets, there is a risk that the model will disproportionately favour the majority class, which can severely compromise its ability to accurately predict and classify instances from the minority class. As highlighted by He and Garcia [2009], this imbalance can lead to suboptimal performance, particularly in tasks such as fraud detection, medical diagnosis, or any other critical classification task where the minority class (in this case, downgrades) is of primary interest.

To address this issue, we used resampling techniques such as over sampling and under sampling, which are among the most commonly used methods in machine learning to deal with class imbalance. Over sampling artificially increases the number of instances in the minority class, while under sampling reduces the number of instances in the majority class. These strategies adjust the class distribution in the training data, ensuring that the model is exposed to a more balanced representation of both the majority and minority classes. In doing so,

these methods make the model more sensitive to patterns in the minority class and help prevent it from becoming biased towards the majority class. This results in a more robust model, capable of making more accurate predictions for both classes, thereby improving its overall performance in scenarios where both classes are of significant importance. Chawla et al. [2002], Batista et al. [2004].

Over sampling

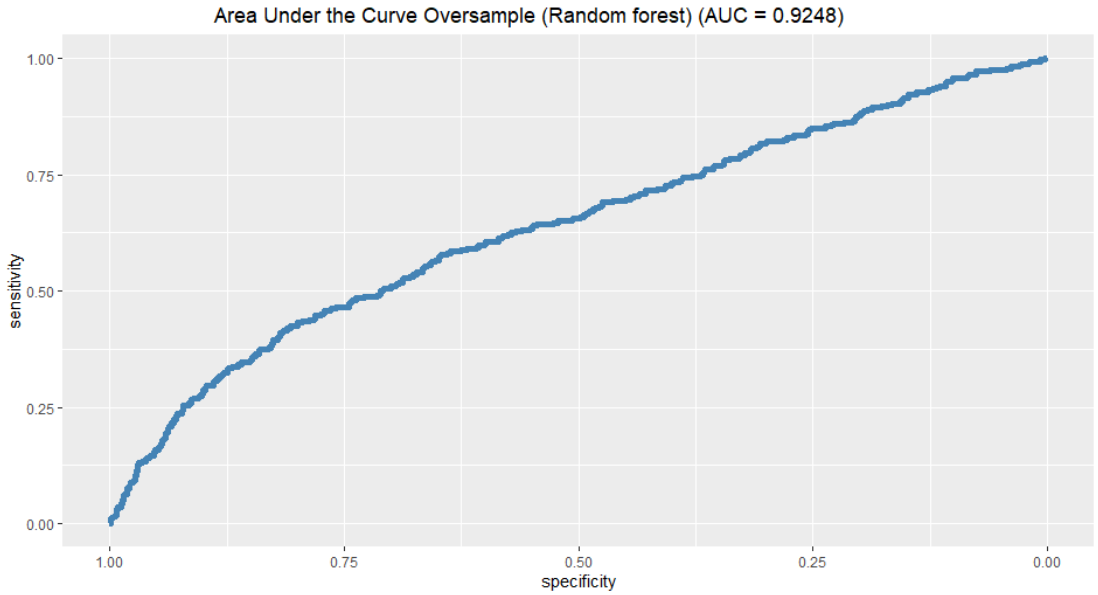

Figure 3.7: This graph shows the ROC curve for oversampling. Source: personal elaboration

.

For our dataset, SMOTE played a crucial role in addressing the class imbalance by increasing the proportion of downgrade events (positive cases) from just 1% to 50%. This adjustment was critical in creating a more balanced dataset, ensuring that the model was properly exposed to the patterns and characteristics associated with the minority class. Before applying SMOTE, the model had been trained on a highly imbalanced dataset where the vast majority of cases were from the majority class (non-downgrades). This imbalance resulted in a model that was biased towards predicting the majority class, thus limiting its ability to effectively classify the minority class (downgrades).

After applying the SMOTE technique, the dataset was resampled to produce an equal representation of both classes, allowing the Random Forest model to learn the underlying patterns of both the majority and minority classes. This

balanced representation allowed the model to better understand the features that differentiate the two classes, improving its predictive accuracy. The impact of SMOTE was evident in the evaluation of the model, where the AUC (area under the curve) increased significantly to 92.4% shown in figure 3.7. This was a significant improvement on the original model's performance, highlighting the effectiveness of SMOTE in improving the model's ability to classify the minority class with greater accuracy, whilst maintaining its generalisation ability.

The main benefit of using SMOTE was that it not only improved performance on the minority class (downgrading), but also prevented the model from overfitting on the majority class, which could have resulted in poorer generalisation to new, unseen data. The resampled dataset ensured that the Random Forest model had a well-rounded understanding of both classes, allowing it to make more reliable predictions across all types of outcomes. As a result, the application of SMOTE not only balanced the data, but also significantly improved the overall robustness and predictive power of the model, particularly in scenarios where class imbalance is a challenge.

Under sampling

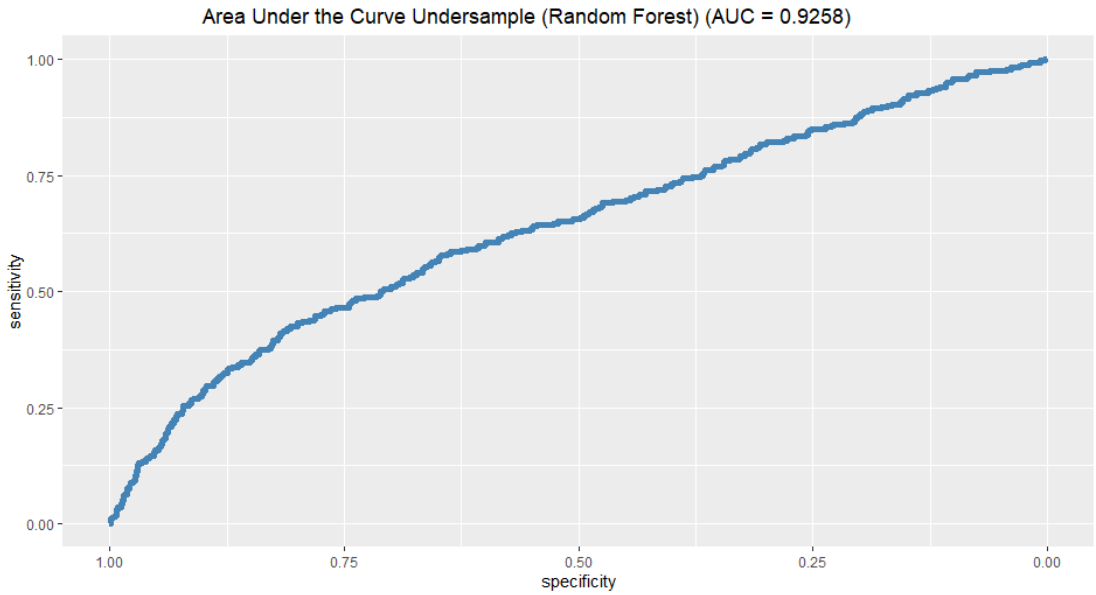

Figure 3.8: This graph shows the percentage of Roc curve in undersampling. Source: personal elaboration

.

We employed under sampling, a technique that reduces the number of major-

ity class instances to balance the size of the minority class. This approach removes superfluous data points from the majority class, allowing the model to focus on the most pertinent patterns. However, if used incorrectly, under sampling can result in the loss of valuable information, particularly in datasets with complex structures within the majority class Batista et al. [2004]. In order to conduct our analysis, we applied random under-sampling by selecting a subset of majority class samples that matched the size of the minority class. The under-sampled Random Forest model, despite the smaller dataset, achieved an AUC of 92.5% as it can be seen in figure 3.8, which was on the same level as the performance of the SMOTE-augmented model. These results demonstrate the effectiveness of under-sampling in addressing class imbalance and showcase the impressive performance of the Random Forest algorithm when applied to balanced datasets.

Model comparison

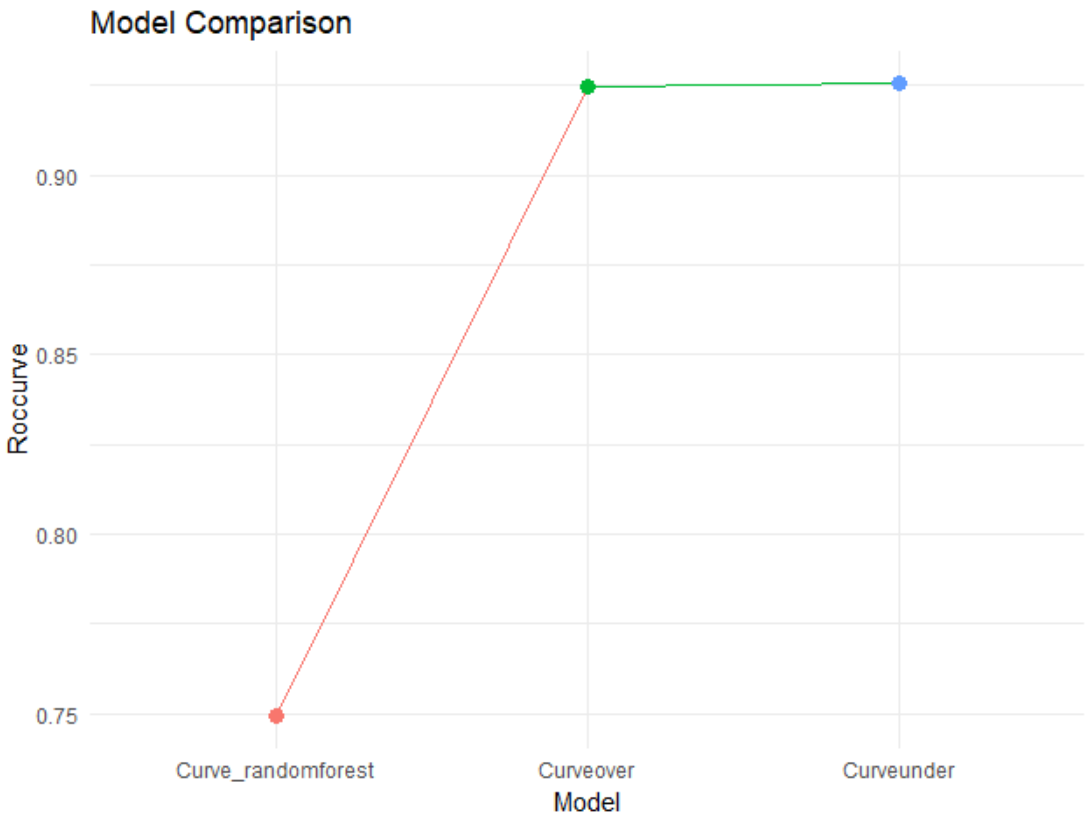

Figure 3.9: This graph shows the model comparison of random forest and its sampling. Source: personal elaboration

.

As shown in the comparison graph, it is clear how much the performance of the model improved after the resampling techniques were applied. Initially, the unbalanced model achieved an AUC of only 75%, highlighting the challenge of class imbalance in the dataset. However, both oversampling (SMOTE) and under sampling proved to be very effective in improving the performance of the model. The AUC values after applying these techniques were 92.4% (SMOTE) and 92.5% (under sampling) shown in figure 3.9, showing significant improvements in the model's ability to discriminate between positive and negative classes.

Both resampling strategies not only improved the AUC values, but also increased the generalisation ability of the model. By balancing the training dataset, the model became less reliant on majority class patterns and was better able to identify and correctly classify minority class instances. This overall improvement in model performance underscores the importance of addressing class imbalance and highlights the critical role of resampling techniques in ensuring robust and reliable predictions. These methods allow the model to better capture the underlying patterns in the data, resulting in more accurate and effective predictions on unseen data, as noted by Sun et al. [2009].

Feature Importance Analysis

The feature importance graph in a Random Forest model provides a valuable visualisation that highlights which variables are driving the model's predictive power. This graph helps to understand how each feature contributes to the model's decision making process, providing insight into which attributes have the greatest impact on the model's predictions Breiman [2001].

In our study, we used the mean reduction in the Gini index as a metric for determining feature importance; this approach quantifies how much each variable contributes to reducing node impurity, a key factor in constructing decision trees. The reduction in the Gini index is calculated by evaluating how much each feature helps to separate the data into pure (homogeneous) groups at each split in the tree building process. Features that produce the largest reduction in the Gini index are considered the most important in terms of predictive power.

How feature importance is calculated in Random Forests:

- Mean decrease in Gini index: In classification tasks, the Gini index measures the degree of impurity or misclassification at a node. The mean reduction in Gini quantifies how much each variable reduces impurity on average across all trees in the forest. Variables with a greater reduction in impurity are considered more important Breiman [2001].

- Mean decrease in accuracy: Another approach calculates the drop in model accuracy when the values of a given feature are randomly permuted. Features that cause larger accuracy drops are ranked higher in importance Strobl et al. [2007].

- Feature Importance Based on Node Impurity: Features are ranked based on the average reduction in impurity (e.g., Gini Index) they cause at the nodes of the trees in the forest Breiman [2001].

- Permutation Feature Importance for Regression Tasks: Similar to the classification case, but uses the model's prediction error (such as MSE) as a metric instead of accuracy James et al. [2023].

- TreeSHAP (SHapley Additive exPlanations): An advanced method for calculating feature importance that takes feature interactions into account, providing a more interpretable and accurate measure of how each feature contributes to the model's output Lundberg [2017].

.

As can be seen from the feature importance graph, the "previous rating" variable emerged as the most important feature shown in figure 3.10, as it encapsulates historical creditworthiness data and is directly correlated with the likelihood of a downgrade. This finding underscores the critical role of historical ratings in predicting future outcomes. Understanding the importance of features in random forests, especially in ensemble models, is critical to improving model interpretability and optimising performance. By identifying and prioritising the most important variables, we can improve model efficiency, reduce over fitting and potentially eliminate redundant features that contribute little to model ac-

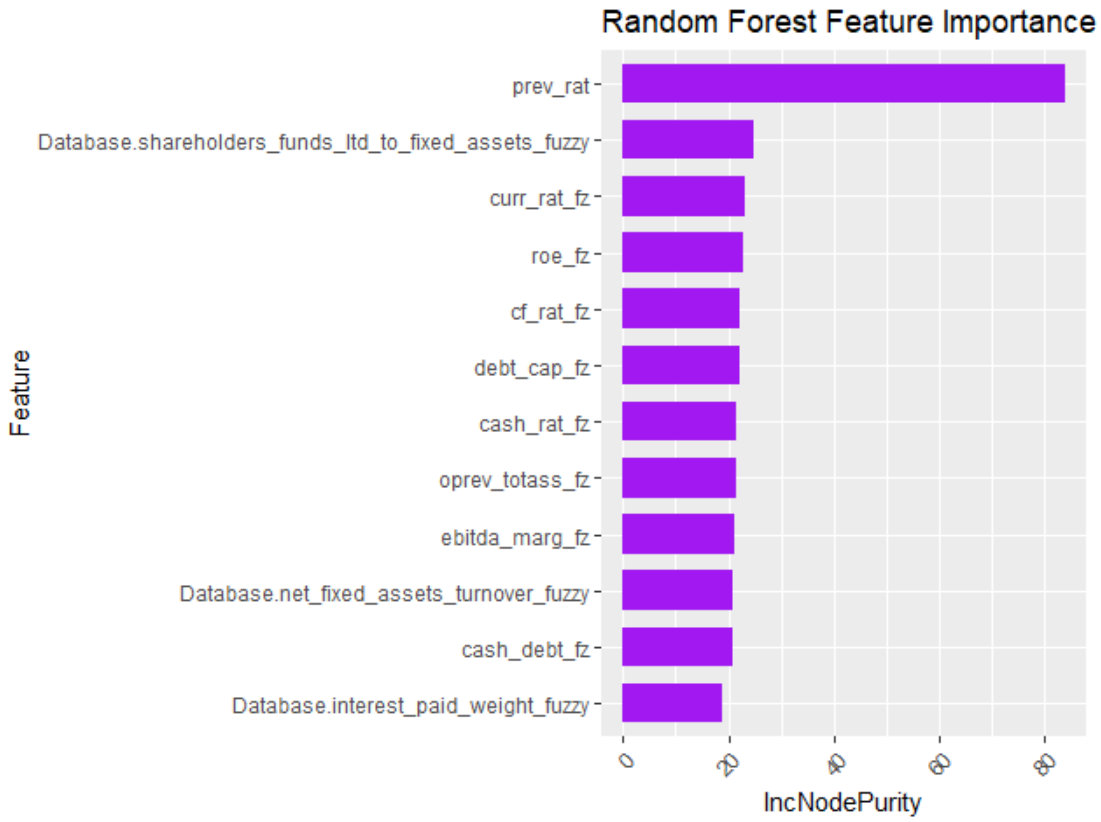

Figure 3.10: This table shows the feature importance of random forest. Source: personal elaboration

curacy, thereby streamlining the model and improving its generalisation to new data. This process is particularly beneficial when aiming to improve the overall performance and interpretability of machine learning models, facilitating better decision making in practical applications. By focusing on feature importance, we can further refine the model by prioritising high-impact variables, improving its predictive accuracy and interpretability. In addition, these insights guide future data collection efforts, ensuring that critical predictors such as historical credit scores are carefully captured and incorporated Breiman [2001].

3.2 Case 2 (Cardo AI)

Cardo AI is a fintech company based in Milan that specialises in advanced data management and analytics solutions for the private debt investment industry AI [n.d.]. Founded by Altin Kadareja, a risk management expert with experience at BlackRock and Prometeia, the company employs over 100 professionals, in-

cluding software engineers, data scientists and financial analysts.

Cardo AI provides a complete set of tools to assist financial companies in managing, analysing and optimising their private debt portfolios. The company's proprietary software, powered by artificial intelligence (AI) and machine learning (ML), provides advanced features such as delay prediction models, borrower behaviour analysis, and environmental, social, and governance (ESG) metrics for sustainable investment strategies. The platform enables clients to streamline and compare financial data across various debt instruments, enhancing both risk evaluation and operational effectiveness. Cardo AI is renowned for its commitment to research and partners with leading academic institutions, as well as participating in EU-funded programmes like Horizon Europe, in order to remain at the cutting edge of fintech innovation.

The company serves clients in Italy, Luxembourg and the UK, addressing the growing complexity of modern lending products such as 'buy now, pay later', income-based finance and salary-linked loans. Future plans include extending its technology suite to the private equity markets, further enabling seamless debt and equity management. Cardo AI's mission is to empower financial institutions with cutting-edge technology to make private debt investments more transparent, efficient and sustainable.

3.2.1 Summary statistics

Cardo AI's dataset, similar to the one provided by modefinance, serves as a valuable resource for loan default prediction tasks. The dataset records the monthly payment history of loans, tracking the behaviour of borrowers from month 1 to month M. However, data from month $M+1$ to month N is masked, adding complexity to the forecasting task. The objective is to determine whether a loan will default (objective = 1) or be repaid in full (objective = 0) by the loan's maturity (month N). This binary classification issue has been crucial in analysing financial risk and helps creditors to estimate the probability of default and to decide on lending.

The dataset contains comprehensive financial information on US personal loans, which tend to be paid back in monthly payments that cover both principal and interest. However, due to financial difficulties or other factors, borrowers

sometimes deviate from the planned payment schedule, either by making payments in excess of the required amount or by missing payments altogether. These variations in borrower behaviour add complexity to the data set and create challenges for accurate predictive modeling. The data is split into a training set containing 213,232 entries and a test set with 29,260 entries, spanning 20 variables. These variables encompass a range of features, including loan-specific attributes (e.g., loan amount, interest rate, term), borrower demographics (e.g., age, income, credit score), and behavioral indicators (e.g., payment-to-income ratio, number of missed payments). The 20 variables include:

- borrower_state
- home_ownership
- loan_id
- pbal_beg_period
- prncp_paid
- int_paid
- fee_paid
- due_amt
- received_amt
- pbal_end_period
- interest
- issue_date
- opencreditlines
- dti
- monthlyincome
- earliest_credit_date
- employment_years
- first fico
- last fico
- calculated_mob

To ensure model readiness, all variables were converted to numeric formats, enabling efficient processing by machine learning algorithms. Missing values (NA entries) were handled through careful preprocessing: for variables with missing

data, such as FICO scores [1], we merged the first and last fico to compute an average FICO score, thereby addressing the gaps. Additionally, some variables with a high percentage of missing values, as per company recommendations, were removed entirely from both the training and testing datasets to maintain data integrity and prevent biases during model training and evaluation. As a result of these adjustments, we retained 13 variables for further analysis and modeling.

To better understand the proportion of loan defaults within the dataset, we created a detailed visualisation showing the ratio of loans classified as defaulted to the total number of loans analysed. This graph clearly illustrated the significant imbalance between the two classes, showing that approximately 86% of loans were successfully repaid, while only 14% resulted in default as it can be seen in figure 3.11. Visualisations such as this are particularly important for interpreting class distributions, especially in financial datasets where imbalances between categories are common He and Garcia [2009]. The findings from this analysis reveal a strong trend towards repayment compliance among the majority of borrowers, indicating that most individuals have been able to meet their financial obligations and repay their loans in full by the end of the loan period.

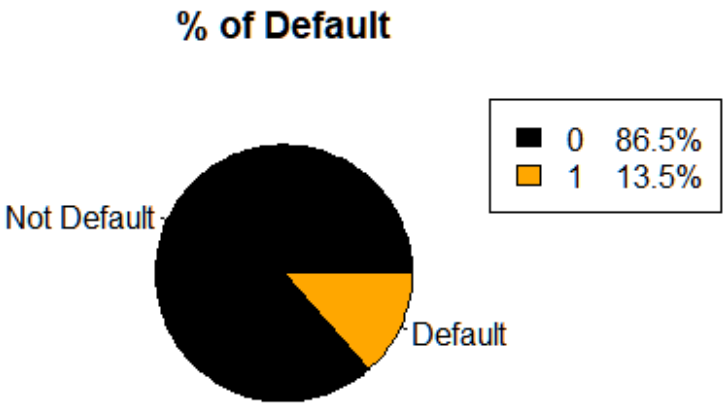

Figure 3.11: This pie chart shows the percentage of default in Cardo AI company.
Source: personal elaboration

[1] *FICO score is a credit score assigned to a borrower to assess their creditworthiness.*

.

The high repayment rate provides valuable insights into borrower behaviour and economic conditions during the period covered by the dataset, factors such as favourable economic conditions, well-structured loan agreements and robust borrower screening processes may have played a role in the observed trend. Moreover, the low default rate makes it clear that we must pay close attention to the small risks that can lead to bigger problems, such as late payment patterns, high loantoincome ratios or a decline in credit scores. Friedman [2001].

This analysis highlights the need for a balanced approach to credit management and risk assessment, while most loans are successfully repaid, it is crucial to understand the characteristics of defaulted loans to improve predictive models and guide lending strategies Hastie [2009], Ke et al. [2017]. This distribution provides a statistical overview, which is essential for financial institutions to make informed strategic decisions.

In order to understand the relationship between the target variable (loan default) and the other features in the dataset and their potential predictive power, an analysis of the correlations was conducted as shown in figure shown in figure 3.12. Also to gain a deeper understanding of how each variable relates to the target variable, we created an additional graph to visualize these correlations more clearly. This graph helps us see how each feature impacts the target, making it easier to spot which variables are most strongly connected to the outcome. By looking at this graph, we can get a better sense of which factors are key drivers and which ones might be less important. The graph, shown in Figure 3.13, gives us a clearer picture of the relationships in the data and helps us make more informed decisions as we move forward with the analysis.

Looking at the correlation heatmap, we can see that most of the variables have a weak correlation with each other, with values generally falling between 0 and 0.5. Only a few variables show a stronger correlation, either above 0.5 (indicating a strong positive relationship) or below 0 (suggesting a negative relationship). This indicates that, for the most part, the variables in the dataset are only loosely connected, with just a small number showing more meaningful connections.

However, in the second graph can be seen more closely to the variables and understand there is more positive correlation with `int_paid`, `dti` (debt-to-income

ratio) and `interest`. But, a few variables such as `avg_fico` (average FICO score), `calculated_mob` (calculated months on book) showed stronger negative correlations. For example, `avg_fico` showed a negative correlation with defaults, suggesting that higher credit scores reduce the likelihood of default as they typically reflect better financial stability. Conversely, `dti` had a positive correlation, suggesting that higher debt-to-income ratios increase the risk of default. These stronger correlations highlight the importance of these characteristics in understanding loan repayment behaviour James et al. [2023], Friedman [2001].

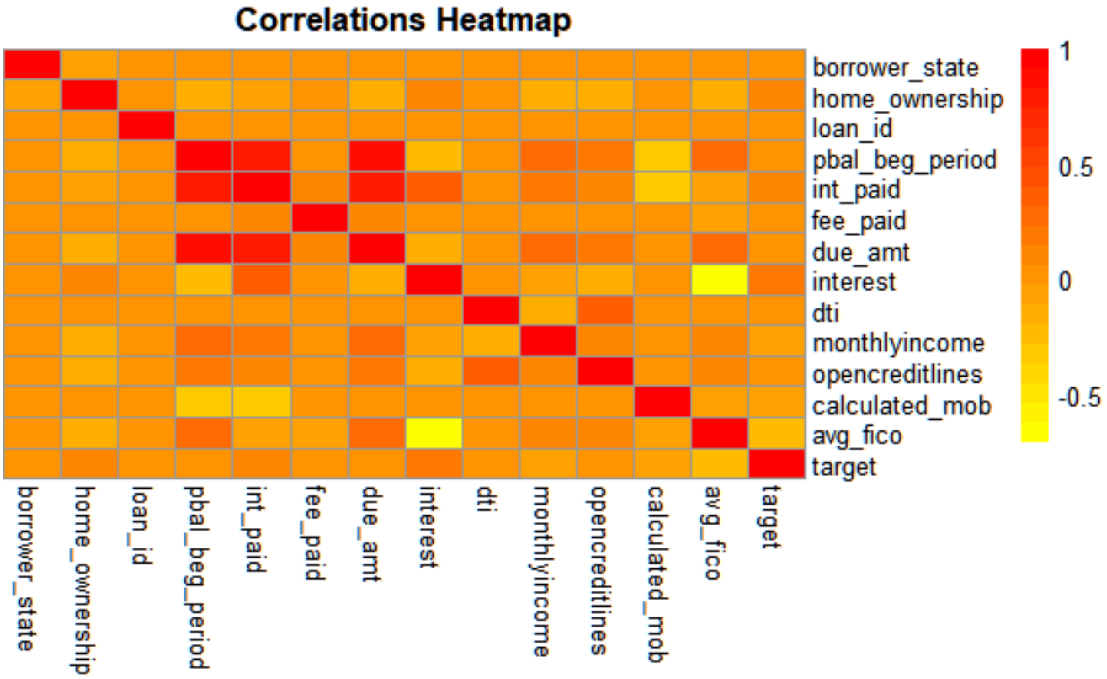

Figure 3.12: This chart shows the correlation of variables in Cardo AI company. Source: personal elaboration

.

.

The overall weakness of the correlations highlights the complexity of predicting loan default, as no single feature is a strong determinant of the outcome. This highlights the importance of using machine learning algorithms that can capture non-linear relationships and interactions between features, such as gradient boosting or random forests Hastie [2009]. In addition, weak correlations may reflect the influence of unmeasured factors, such as borrower behavior or macroeconomic conditions, which may also affect default rates Chawla et al. [2002].

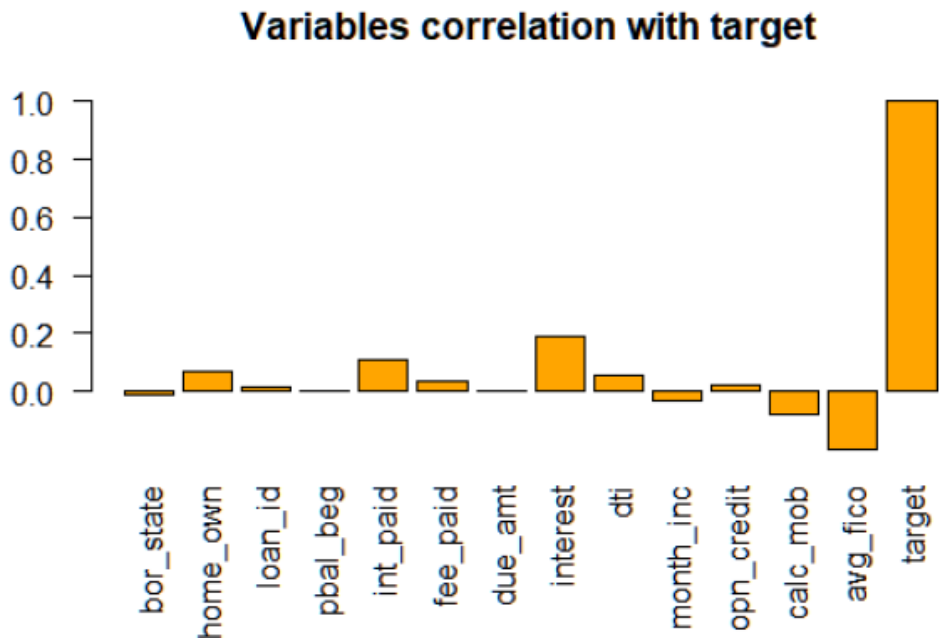

Figure 3.13: This graph shows the correlation between target variable and other response variables in Cardo AI company. Source: personal elaboration

For feature selection and development, it is essential to understand the correlations between features and loan defaults. Features with higher absolute correlation values can be prioritised during model development as they carry more predictive information. However, even weakly correlated features can have value when combined with others in multivariate models, especially in ensemble methods that aggregate information across many predictors Breiman [2001].

While individual features show limited correlation with the target variable, the interaction of multiple variables could provide significant predictive power. This analysis reinforces the need for robust modelling techniques that can handle high-dimensional data and extract meaningful patterns from weakly correlated variables Ke et al. [2017]

3.2.2 Model selection

In our analysis, we experimented with several models to find the most suitable one for the Cardo AI dataset. The models included Logistic regression, Ridge regression, Quadratic discriminant analysis (QDA), Random Forest and Gradient boosting machines (GBMs). These models were chosen to allow us to assess their

relative strengths and weaknesses on the dataset, as they represent a range of machine learning paradigms, from interpretable linear models to complex ensemble methods. Each model was evaluated using standard performance metrics such as accuracy, precision, recall, F1 score and area under the ROC curve (AUC-ROC). These metrics provided insight into how well the models balanced false positives and false negatives - critical in contexts such as credit default prediction, where the costs of misclassification are unequal Fawcett [2006].

While all of the models performed reasonably well, the Random Forest model emerged as the standout model, outperforming the other models on almost all of the metrics. Its ability to handle complex data structures, its robustness to missing values and its flexibility in capturing non-linear relationships made it particularly effective on this dataset. These strengths are consistent with existing research demonstrating the superior predictive power of Random Forest in financial modeling tasks such as credit scoring and default prediction [Louzada et al., 2016]. In particular, Random Forest achieved the highest AUC-ROC score, indicating excellent discriminative power between downgraded and non-downgraded classes. Its ensemble approach, which combines the predictions of multiple decision trees, also contributed to its robustness against over fitting.

Gradient Boosting Machines (GBMs), another ensemble method, also performed well, closely following Random Forest in terms of AUC-ROC and F1 scores. GBMs excel at capturing subtle patterns in the data by iteratively correcting the errors of previous trees. However, they required more careful hyperparameter tuning to prevent overfitting and ensure generalisability, which added computational complexity to the modelling process.

To address the class imbalance of the dataset, we incorporated resampling techniques - both oversampling the minority class and undersampling the majority class - in all models. These adjustments were critical in ensuring that the models could effectively identify downgrades without biasing predictions towards the dominant class. Random Forest showed exceptional resilience to class imbalance, further cementing its position as the preferred model for this dataset.

3.2.3 Random forest

The Cardo AI dataset presented unique challenges, with one of the most pressing being a significant class imbalance, only 14% of loans were classified as defaults. This imbalance posed a critical issue, as it could lead the model to favor the majority class (non-defaults) while struggling to identify rare but important cases like defaults. To address this, we implemented advanced sampling techniques to better balance the dataset. These methods helped the model recognize and classify rare default cases effectively without compromising overall accuracy. Achieving this balance was essential to develop a fair and reliable credit risk prediction system.

In addition to the class imbalance, the dataset included 20 different variables that spanned a wide range of information. These variables covered borrower demographics, such as income; loan characteristics, like interest rates and repayment schedules; and behavioral indicators, including repayment histories and missed payments. This diversity of data required careful preprocessing and thorough analysis to uncover the most important factors driving loan defaults. Identifying these key predictors not only improved the model's interpretability but also provided actionable insights for refining Cardo AI's credit risk management strategies.

This section delves into the performance of the Random Forest model applied to the dataset. We start by examining the Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC) metric, which measure the model's ability to distinguish between defaulted and non-defaulted loans. We then analyze the confusion matrix to get a detailed view of how the model performed, focusing on the balance between correctly and incorrectly classified cases. Following this, we discuss the sampling techniques used to address the class imbalance, exploring how they impacted model performance. Finally, we present a feature importance analysis, shedding light on the variables that played the most significant role in predicting loan defaults and providing deeper insights into the factors influencing credit risk.

Receiver Operating Characteristic

To get a clearer picture of the model's overall performance, we also used the Area Under the Curve (AUC) metric, which summarizes the ROC curve into a single value. In the case of the Cardo AI dataset, the model achieved an AUC of 80%, as shown in Figure 3.14, this means the model is pretty good at separating loans that are likely to default from those that are expected to be repaid fully.

An AUC of 80% is quite strong, especially when dealing with financial data where predicting defaults correctly is crucial. It shows that the model has a solid ability to correctly rank loans at risk of default higher than loans that are safe, even as we adjust the thresholds for decision-making. This kind of performance is essential in real-world applications, where the goal is to make accurate predictions, especially in cases where defaults are relatively rare compared to regular repayments. The high AUC reflects the model's reliability and its capability to handle the complexities of credit risk prediction effectively.

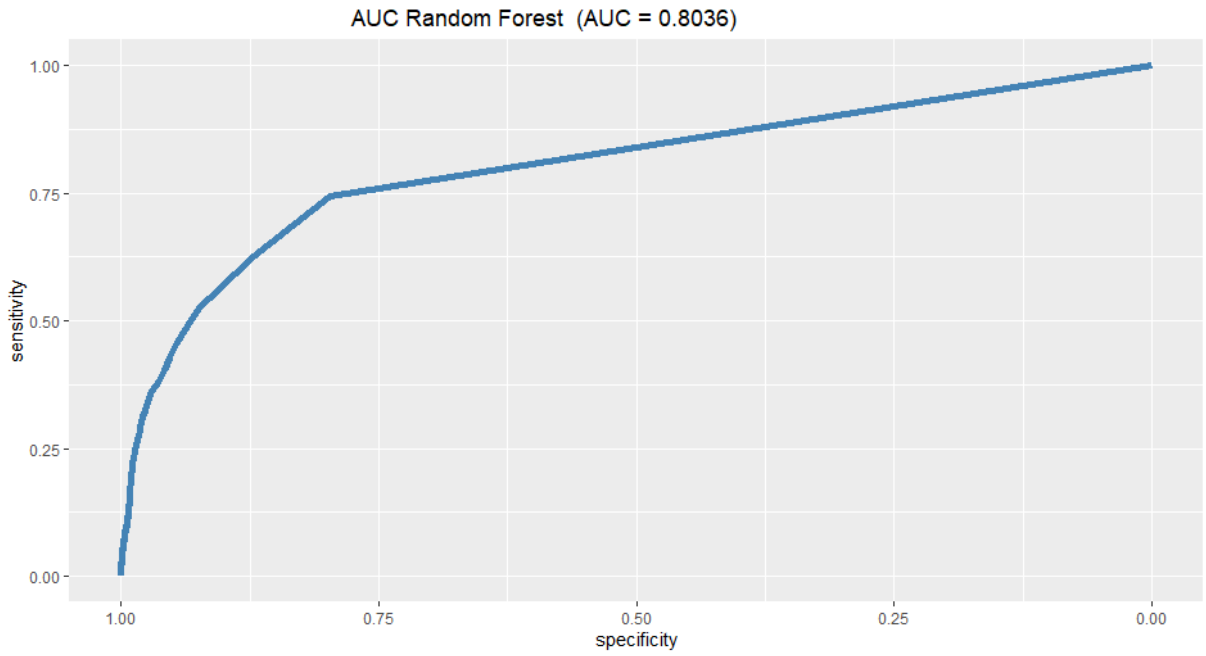

Figure 3.14: This graph shows the area under the curve random forest in Cardo AI company. Source: personal elaboration

.

Confusion Matrix Analysis

Following the AUC evaluation, the model's performance was analysed using the Confusion Matrix, which provides detailed insight into the prediction results for the Cardo AI dataset. The matrix classifies predictions into four categories shown in figure 3.15:

1. True Positives (TP): Correctly identified loan defaults (1698).

2. True Negatives (TN): Correctly identified non-default loans (877).

3. False Positives (FP): Non-default loans incorrectly classified as defaults(363).

4. False Negatives (FN): Defaults incorrectly classified as non-defaults (15569).

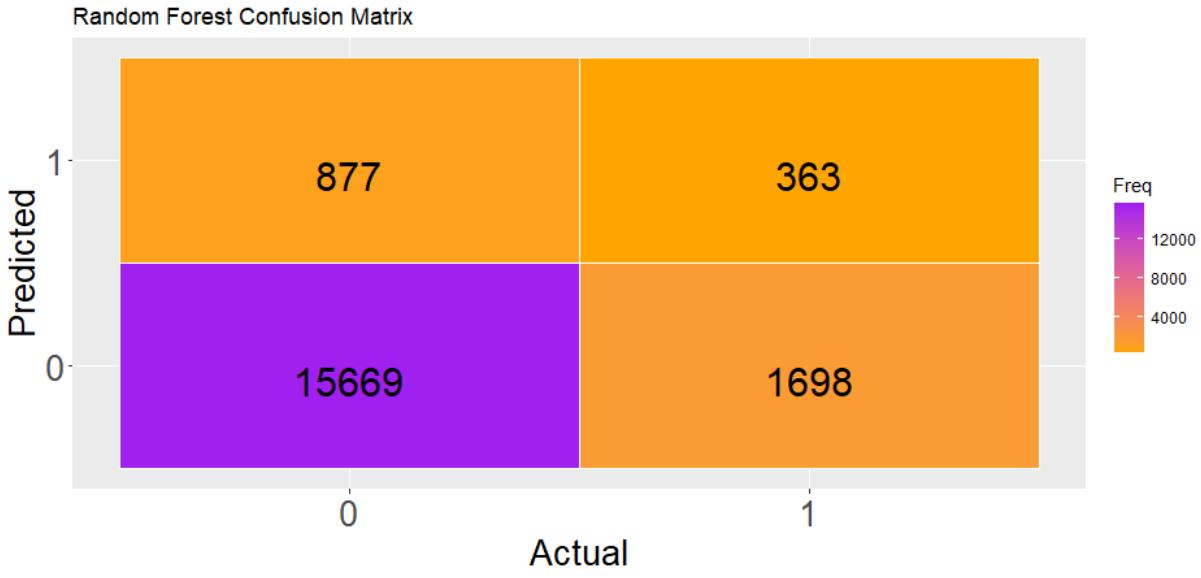

Figure 3.15: This table shows the confusion matrix in Cardo AI company. Source: personal elaboration

.

This breakdown shows high accuracy in identifying non-defaults, consistent with the class imbalance of the dataset, where 86% of loans were non-defaults. However, while the model showed high accuracy in predicting non-defaults, it

showed moderate difficulty in identifying defaults, as reflected in the relatively higher number of false negatives.

Sampling

Given that defaults made up only 14% of the Cardo AI dataset, we faced the challenge of class imbalance, which could cause the model to focus too much on predicting the majority class (non-default loans) and miss the minority class (defaults). To tackle this, we used sampling techniques to balance the dataset and make it easier for the model to detect loan defaults.

Like modefinance company we applied both over sampling and under sampling methods. In over sampling, we created synthetic examples of the minority class (defaults) using a technique called SMOTE, which helps to generate new, similar data points by blending existing ones. This gave the model more examples of defaults to learn from. On the other hand, under sampling involved reducing the number of non-default loans in the training set to ensure that the model was not overwhelmed by too many non-defaults. This forced the model to focus more on learning from the default cases.

We used these sampling techniques with the Random Forest model, which is great at working with complex data. By balancing the dataset, we made it easier for the model to focus on identifying loan defaults, helping it become more accurate at spotting these less common cases.

Over sampling To address the class imbalance, we applied the (SMOTE) to balance the dataset, we saw a noticeable improvement in the model's performance, with the AUC going up to 85% (as shown in Figure 3.16, compared to 80% without any sampling. This increase shows that the model got better at telling the difference between loans that were likely to default and those that were not. By giving more attention to the default cases, the model was able to make more accurate predictions for these rare events, while still being just as reliable in predicting non-default loans.

.

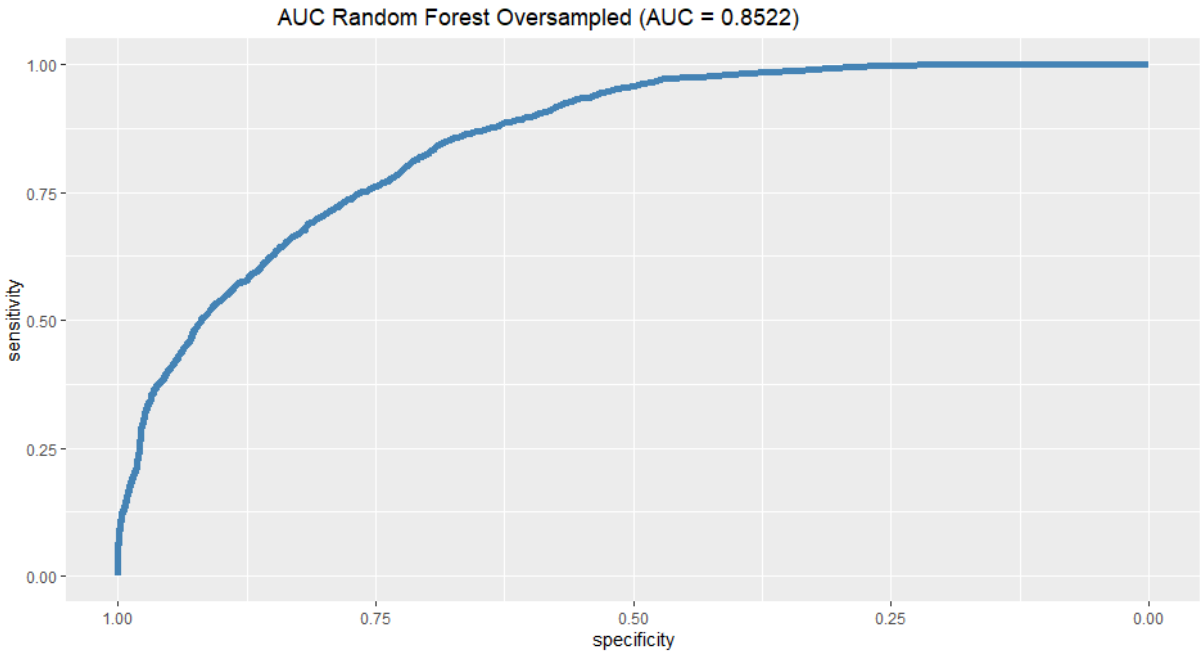

Figure 3.16: This graph shows the area under the curve random forest over sample in Cardo AI company. Source: personal elaboration

Under sampling

To compare, we also tried random under sampling, which involved reducing the number of non-default loans to match the number of defaults. While this helped balance the dataset, it did mean we lost some valuable information from the majority class. Even with this drawback, the model still performed quite well, achieving an AUC of 82.55% (as shown in Figure 3.17). This shows that even with fewer samples to work with, the model was still able to effectively differentiate between defaults and non-defaults. However, it's important to note that under sampling could have limited the model's learning potential, as it removed some of the non-default cases that might have helped improve overall prediction accuracy. Still, the model's strong performance highlights its ability to maintain accuracy despite fewer training samples.

.

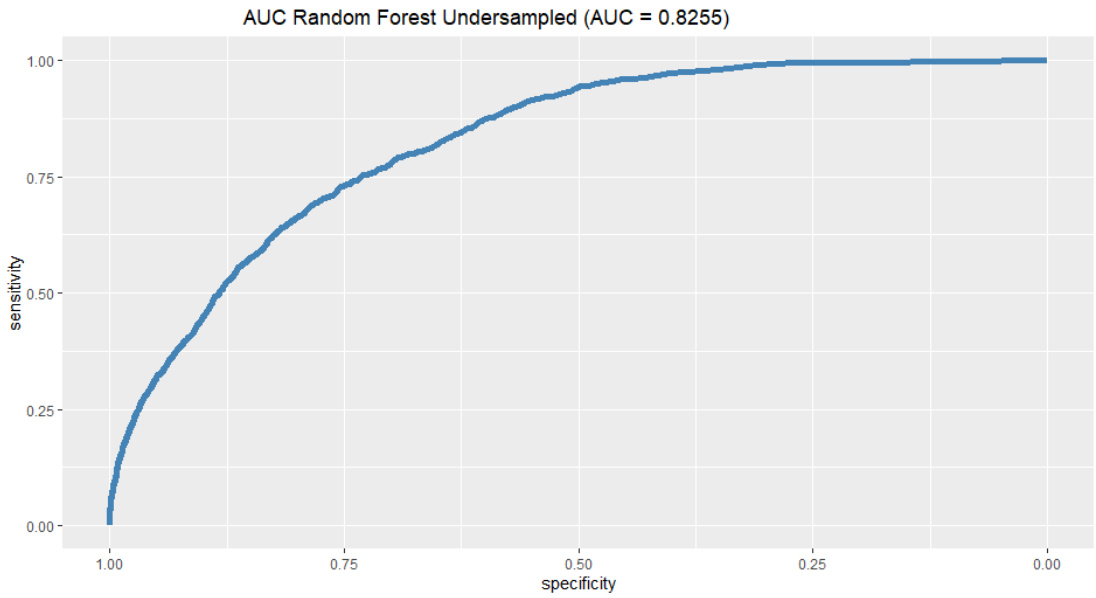

Figure 3.17: This graph shows the area under the curve random forest under sample in Cardo AI company. Source: personal elaboration

Feature Importance Analysis

After data sampling and feature selection, we analysed the importance of the features to identify the most influential variables in predicting credit risk. The results showed that Loan ID, DTI and monthly income were among the most significant predictors as it can be seen in figure 3.18. These variables provide important insights into a borrower's financial situation: loan ID uniquely identifies the loan and allows for specific tracking, DTI provides a clear measure of the borrower's ability to repay by comparing their monthly debt obligations to their income, and monthly income serves as a basic indicator of the borrower's financial capacity. The high importance of these characteristics suggests that they play a crucial role in assessing creditworthiness and risk in Cardo AI's modeling process.

.

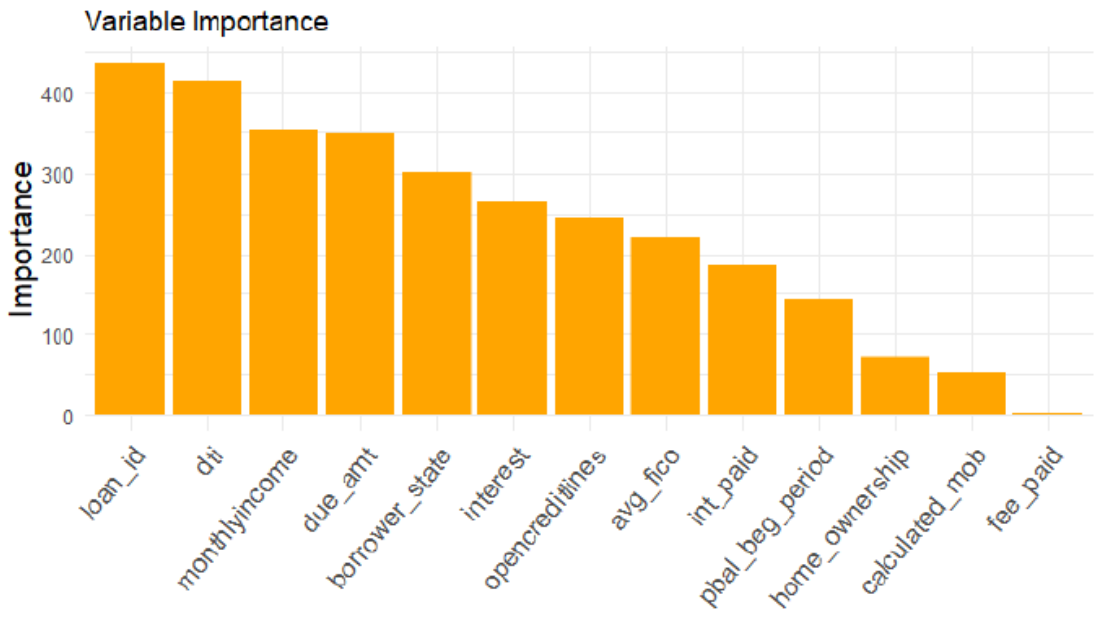

Figure 3.18: This graph shows the feature importance in Cardo AI company. Source: personal elaboration

3.3 Conclusion

This thesis examined the significance of credit risk in modern life, investigating its development, benefits, and drawbacks, as well as how it can be leveraged to produce better outcomes. By analyzing datasets from two different organizations, modefinance and Cardo AI, we developed and assessed various models to determine the most effective approach. The research provides a comparative analysis of conventional statistical techniques, such as logistic regression, Ridge regression, and quadratic discriminant analysis, alongside more advanced machine learning methods like random forests and (GBM). It also explores the evolution and practical use of credit scoring models.

A central focus of the study was addressing the ongoing issue of class imbalance, especially in predicting rare but significant financial events, such as credit rating downgrades. Through a systematic evaluation of model performance, this work offers meaningful insights into enhancing credit risk assessments and underscores the potential of advanced analytics in improving financial decision-making. The results highlight the transformational role of advanced machine learning model such as Random forest in modern credit risk evaluation:

1. Model superiority: Random forest emerged as the most effective model, outperforming others in terms of accuracy, robustness and adaptability. This advantage was particularly evident when dealing with class imbalance through techniques such as over and under sampling, which significantly improved the model's ability to identify rare events.

2. The limitations of traditional models: While three other models worked reliably on datasets with linear relationships, they struggled to capture the complexity of real-world financial data, where non-linear patterns and interactions dominate.

3. Importance of class imbalance reduction: Addressing class imbalance proved critical, with AUC values increasing from 75% in the unbalanced model to over 92% when resampling techniques were applied. These results highlight the need to balance data distributions to improve sensitivity and predictive reliability.

In addition to model performance, the study also highlights the critical role of feature selection and sampling strategies in improving the accuracy of the predictions. Variables such as previous ratings and fixed asset fuzzy consistently emerged as key predictors of credit downgrades, reinforcing their value in credit risk modeling. This research contributes to the financial sector by providing a comprehensive framework for assessing credit risk and demonstrates how advanced machine learning models can be integrated with traditional statistical methods to provide more accurate, data-driven insights. The methodology and findings of the study will enable financial institutions to improve their risk forecasting, mitigate financial shocks and make more informed credit decisions.

Future research could build on these findings by incorporating real-time financial data and factors such as environmental, social and governance (ESG) metrics. These additions would allow credit scoring models to adapt to evolving market dynamics and align with the growing focus on sustainable finance. In addition, exploring deep learning and hybrid modeling techniques could improve the granularity and interpretability of predictions, addressing the challenges of increasingly complex data sets.

Bibliography

H. A. Abdou and J. Pointon. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 18(2-–3):59—88, 2011. doi: 10.1002/isaf.325.

Manuel Adelino and Miguel A Ferreira. Bank ratings and lending supply: Evidence from sovereign downgrades. *The Review of Financial Studies*, 29(7): 1709–1746, 2016.

Cardo AI. About us, n.d. URL `https://cardoai.com`. Accessed: November 17, 2024.

Edward I Altman and Anthony Saunders. Credit risk measurement: Developments over the last 20 years. *Journal of banking & finance*, 21:1721–1742, 1997.

Raymond Anderson. *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford university press, 2007.

R. B. Avery, K. P. Brevoort, and G. Canner. Does credit scoring produce a disparate impact? *Real Estate Economics*, 40(s1), 2012. doi: 10.1111/j.1540-6229.2012.00348.x.

G. Babaei, P. Giudici, and E. Raffinetti. Explainable fintech lending. *Journal of Economics and Business*, 125–126:106126, 2023. doi: 10.1016/j.jeconbus.2023.106126.

Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

Y Bengio. Learning deep architectures for ai, 2009.

Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

Rok Blagus and Lara Lusa. Smote for high-dimensional class-imbalanced data. *BMC bioinformatics*, 14:1–16, 2013.

Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

Peter Bühlmann and Bin Yu. Boosting with the l 2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

V. Bumacov, A. Ashta, and P. Singh. Credit scoring: A historic recurrence in microfinance. *Strategic Change*, 26(6):543–554, 2017. doi: 10.1002/jsc.2165.

Lorena Caridad, Julia Núñez-Tabales, Petr Seda, and Orlando Arencibia. Do moody's and s&p firm's ratings differ? *Economics & Sociology*, 13:173–186, 2020.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Chao Chen, Andy Liaw, Leo Breiman, et al. Using random forest to learn imbalanced data. *University of California, Berkeley*, 110:24, 2004.

Sheng-Syan Chen, Hsien-Yi Chen, Chong-Chuo Chang, and Shu-Ling Yang. The relation between sovereign credit rating revisions and economic growth. *Journal of Banking & Finance*, 64:90–100, 2016.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Michel Crouhy, Dan Galai, and Robert Mark. A comparative analysis of current credit risk models. *Journal of Banking & Finance*, 24(1-2):59–117, 2000.

Danielle Denisko and Michael M Hoffman. Classification and interaction in random forests. *Proceedings of the National Academy of Sciences*, 115(8):1690–1692, 2018.

Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8): 861–874, 2006.

Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*, volume 10. Springer, 2018.

Andy Field. *An adventure in statistics: The reality enigma*. SAGE Publications Ltd, 2022.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of online learning and an application to boosting. *Journal of computer and system sciences*, 55:119–139, 1997.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Aaryan Gupta, Vinya Dengre, Hamza Abubakar Kheruwala, and Manan Shah. Comprehensive review of text-mining applications in finance. *Financial Innovation*, 6:1–25, 2020.

M. Hadavand-Siri and C. V. Deutsch. Some thoughts on understanding correlation matrices, 2012.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee, 2008.

Joseph M Hilbe. Logistic regression. *International encyclopedia of statistical science*, 1:15–32, 2011.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

Y. Hu and J. Su. Research on credit risk evaluation of commercial banks based on artificial neural network model. In *Procedia Computer Science*, volume 199, page 1168–1176, 2022. doi: 10.1016/j.procs.2022.01.148.

G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor. *An Introduction to Statistical Learning: With Applications in Python*. Springer International Publishing, 2023. doi: 10.1007/978-3-031-38747-0.

Gareth James. An introduction to statistical learning, 2013.

Ahmedin Jemal, Andrea Thomas, Taylor Murray, Michael Thun, et al. Cancer statistics, 2002. *Ca-A Cancer Journal for Clinicians*, 52(1):23–47, 2002.

Surendranath R Jory, Thanh N Ngo, and Daphne Wang. Credit ratings and the premiums paid in mergers and acquisitions. *Journal of Empirical Finance*, 39: 93–104, 2016.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Darren J. Kisgen. The impact of credit ratings on corporate behavior: Evidence from moody's adjustments. *Journal of Corporate Finance*, 58: 567–582, 2019. ISSN 0929-1199. doi: https://doi.org/10.1016/j.jcorpfin.2019.07.002. URL `https://www.sciencedirect.com/science/article/pii/S0929119918306837`.

T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, 250:113–141, 2013.

Francisco Louzada, Anderson Ara, and Guilherme B Fernandes. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2):117–134, 2016.

Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Roberta Martino, Annamaria Porreca, Viviana Ventre, and Fabrizio Maturo. Exploring intertemporal decision-making dynamics through functional data analysis: investigating variations in different discount function's dimensions. *Quality & Quantity*, pages 1–26, 2024.

Fabrizio Maturo and Annamaria Porreca. Demystifying functional random forests: Novel explainability tools for model transparency in high-dimensional spaces. *arXiv preprint arXiv:2408.12288*, 2024.

J. H. McDermott, M. Schemitsch, and E. P. Simoncelli. Summary statistics in auditory perception. *Nature Neuroscience*, 16(4):493–498, 2013. doi: 10.1038/nn.3347.

Dulani Meedeniya. *Deep learning: A beginners' guide*. CRC Press, 2023.

Scott W Menard. *Logistic regression: From introductory to advanced concepts and applications*. Sage, 2010.

Steffi Ostrowski. Judging the fed. *Yale LJ*, 131:726, 2021.

Frank Partnoy. The paradox of credit ratings. In *Ratings, rating agencies and the global financial system*, pages 65–84. Springer, 2002.

Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.

Badiuzzaman Pranto, Sk Maliha Mehnaz, Esha Bintee Mahid, Imran Mahmud Sadman, Ahsanur Rahman, and Sifat Momen. Evaluating machine learning methods for predicting diabetes among female patients in bangladesh. *Information*, 11(8):374, 2020.

Foster Provost. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, volume 68, pages 1–3. AAAI Press, 2000.

Santhosh Kumar Rajamani and Radha Srinivasan Iyer. Machine learning-based mobile applications using python and scikit-learn. In *Designing and developing innovative mobile applications*, pages 282–306. IGI Global, 2023.

Sebastian Raschka and Vahid Mirjalili. *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt publishing ltd, 2019.

Credit Ratings. S&p global ratings. *Institutions*, 98:96–104, 2022.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

Erwan Scornet. On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146:72–83, 2016.

James P Stevens. *Intermediate statistics: A modern approach*. Routledge, 2013.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8:1–21, 2007.

Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern recognition*, 40 (12):3358–3378, 2007.

Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009.

Lyn Thomas, Jonathan Crook, and David Edelman. *Credit scoring and its applications*. SIAM, 2017.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Shrawan Kumar Trivedi. A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society*, 63:101413, 2020.

Annette Vissing-Jorgensen. The treasury market in spring 2020 and the response of the federal reserve. *Journal of Monetary Economics*, 124:19–47, 2021.

C. Weihs and K. Ickstadt. Data science: The impact of statistics. *International Journal of Data Science and Analytics*, 6(3):189–194, 2018. doi: 10.1007/s41060-018-0102-5.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.

Mochen Yang, Edward McFowland III, Gordon Burtch, and Gediminas Adomavicius. Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. *INFORMS Journal on Data Science*, 1(2):138–155, 2022.

Show-Jane Yen and Yue-Shi Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36:5718–5727, 2009.

G. Zeng. On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics - Theory and Methods*, 49(9):2080–2093, 2020. doi: 10.1080/03610926.2019.1568485.

Jiawei Zhang, Yuhong Yang, and Jie Ding. Information criteria for model selection. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(5):e1607, 2023.