ISEG
Lisbon School
of Economics
& Management
Universidade de Lisboa

Master

Data Analytics for Business

Master´s Final Work

Dissertation

Loss Given Default Predictive Method Analysis Under the Perspective of Basel IV

Xixian Chen

March - 2024

MASTER
DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK
DISSERTATION

LOSS GIVEN DEFAULT PREDICTIVE METHOD ANALYSIS UNDER THE PERSPECTIVE OF BASEL IV

XIXIAN CHEN

SUPERVISION:
João Afonso Bastos

MARCH - 2024

To all my master's students
that started a dissertation
under my supervision, but
were abducted by extra-
terrestrials or crossed to
another dimension.

GLOSSARY

LGD – Loss Given Default

PD – Probability of Default

EAD – Exposure at Default

OLS – Ordinary Least Squares

RR – Recovery Rate

DRR – Discounted Recovery Rate

IRBA – Internal Ratings-Based Approaches

SMEs – Medium-sized Enterprises

ABSTRACT, KEYWORDS AND JEL CODES

In response to concerns over capital calculation variability among banks, the Basel Committee revised the Basel III framework in 2017, leading to substantial changes known as Basel IV. This paper explores Loss Given Default (LGD) within the context of Basel IV, focusing on its definition, impact, mathematical measurement, and Moody's LGD model. Additionally, it compares different machine learning models relevant to LGD. Utilizing Moody's Ultimate Recovery Database, which contains detailed recovery information for over 4,600 bonds and loans, this study aims to provide readers with a foundational understanding of LGD under Basel IV and conduct a comparative analysis of machine learning techniques for LGD estimation.

KEYWORDS: Loss given default; Basel IV; Production Function.

JEL CODES: C02; C10; C25; F65; F68; G21

TABLE OF CONTENTS

TABLE OF FIGURES

ACKNOWLEDGMENTS

First, I wish to thank Professor João Afonso Bastos for his encouragement and guidance.

I am also grateful to my colleagues at BNP PARIBAS for numerous discussions.

Finally, I am also thankful to my family for their patience and their support while I pursued this project.

1. INTRODUCTION

Credit risk is the oldest risk faced by modern banks, as well as the biggest risk that exists in almost all banks around the world. It has been making enormous economic cost to financial intuitions every year. For example, back to 2008 financial crisis, UBS was one of the banks that suffered significant losses during the subprime mortgage crisis due to investments related to U.S. subprime mortgages. The crisis led UBS to announce billions of dollars in losses and eventually required government assistance. More recently, Credit Suisse, Switzerland's second-largest bank, came to a head when it was announced that it was to be taken over by UBS. The downfall of this banking giant can be attributed to a sequence of events and scandals that are intricately linked to deficient and ineffective risk management practices within the realm of credit risk control.[1]

Banks are regulated at the national and regional levels, and since 1973, bank regulations have been coordinated globally by the Basel committee for the bank of international settlements. The organization is jointly owned by 63 central banks from countries that account for 95% of global GDP. Basel III is an international regulatory framework for banks, developed by the Basel Committee on Banking Supervision (BCBS) in response to the financial crisis of 2008. It includes several rules about capital requirement for banks to make sure the exposure of credit risk is under control. Basel III, also referred to as Basel III Endgame, Basel 3.1, or Basel IV, represents the finalization of post-crisis reforms in the banking sector. These reforms entail significant changes to international standards for bank capital requirements, which were agreed upon by the Basel Committee on Banking Supervision (BCBS) in 2017. The forthcoming changes are so comprehensive that they are widely regarded as constituting an entirely new regulatory framework. These reforms are expected to come into effect under transition rules starting from 2025.[2]

Basel III internal rating-based approach (IRB) for banks to calculate capital requirement which can effectively prevent banks corrupts from counterparties default. It consists of three key parameters, Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD). From bank's perspective, PD refers to the

[1] News: https://www.piranirisk.com/blog/credit-suisse-causes-of-the-recent-fall-of-the-swiss-bank
[2] BCBS: https://www.bis.org/basel_framework/

probability of default of a borrower in the future, usually over a one-year horizon; LGD is the credit loss occurred if a counterparty of the bank indeed defaults.

In 2017, the Basel Committee took a decisive step to revise the existing Basel III framework in response to concerns raised by academic studies indicating a growing unease regarding the variability in banks' calculations of their capital. This prompted the committee to institute substantial changes, primarily concentrating on global capital requirements. These modifications are so extensive that they are sometimes colloquially referred to as a new iteration, termed Basel IV. Notably, the European Union (EU) is set to enforce Basel IV before 2025, with prominent institutions such as BNP PARIBAS, a leading French bank, slated to adopt Basel IV from January 2024.

By doing reach on academic and literature review, with public data sources released by Moody’s, this paper goes through Loss Given Default (LGD) within the perspective of Basel IV, and have discussion on several key points – which are:

1. Definition and Impact of LGD in Basel IV: Exploring the concept of LGD and explaining alterations introduced by Basel IV.
2. Mathematical Definition and Measurement of LGD: A comprehensive examination of how LGD is mathematically defined and measured under Basel IV.
3. What is the approach in Moody’s model of LGD?
4. Comparison of Machine Learning Models: Investigating and comparing different machine learning models relevant to LGD.

The primary objective of this study is to investigate the impact of the new changes in Loss Given Default (LGD) associated with Basel IV, attempting to provide readers with a foundational understanding of the financial data provider model, along with a brief comparative analysis of three distinct machine learning models. The dataset employed in this investigation is derived from Moody's Ultimate Recovery Database, released by Moody’s. This comprehensive dataset encompasses detailed recovery information on nominal and discounted ultimate recoveries for over 4,600 bonds and loans, spanning more than 900 default events involving non-financial corporations in the United States. Through a rigorous analysis of this dataset, various machine learning techniques will be applied to discern the optimal model within the context of machine learning methodologies for estimating LGD.

Much of the prevailing literature in the field has primarily focused on Basel III in its examination of Loss Given Default (LGD), emphasizing various models and parameters within the confines of this regulatory framework. However, this research introduces an innovative perspective by diverging from the conventional emphasis on Basel III to thoroughly investigate LGD. This departure from the prevailing focus on Basel III represents a distinctive contribution, as the study aims to explore LGD considerations within the context of the impending Basel IV regulations.

In the forthcoming chapter 2, this study will conduct a comprehensive analysis of the literature pertaining to Loss Given Default and Basel III over the last two decades. Chapter 3 will provide an in-depth overview of the methods and procedures employed in this research, encompassing three distinct machine learning models and detailing the analytical approach applied to the dataset. The ensuing Chapter 4 will encapsulate the study's findings, presenting all observed outcomes and addressing the questions posited throughout the research. Finally, Chapter 5 conclusion will present a succinct synthesis of the study's responses to the research questions, highlighting the achievement of the stated aims, and acknowledging specific limitations encountered during the course of this investigation.

2. Literature Review

The year 2008 marked an unprecedented global financial crisis, profoundly impacting nearly every major country and resulting in significant losses for individuals and institutions alike. In response to this crisis, Basel III was introduced with the primary objective of imposing enhanced regulatory measures on financial institutions worldwide to mitigate the likelihood of a recurrence of such a severe economic downturn. It is crucial to recognize that the genesis of Basel III did not emerge from zero, but rather can be traced back to the evolutionary trajectory of Basel I and Basel II. Given the inherent deficiencies identified in the frameworks of Basel I and Basel II, particularly in addressing the global capital requirements for banks, the imperative for Basel III became apparent. Consequently, Basel III was conceived as a necessary evolution to address the prevailing issues within the global banking system stemming from its predecessors.

The Basel Committee on Banking Supervision (the BCBS or the Basel Committee) was established in 1974. King & Tarbert (2011) mentioned that throughout the 1980s and 1990s, many countries tried to deregulate bank and financial systems in their counter in order to allow banks to compete for larger market shares. However, this led to a rapid expansion of both domestic and foreign exposures by banks, as the regulatory landscape governing capital requirements was perceived to be inadequately stringent during this period. Consequently, these exposures were not adequately covered by corresponding capital bases. To address this challenge, the result was Basel I. Nonetheless, Basel I exhibited a significant flaw: its categorical risk weights were not only crudely calibrated but also permitted and, in fact, encouraged regulatory arbitrage.

Eventually Basel Committee decided to overhaul the framework, giving rise to the introduction of Basel II in 2004. This new framework of risk management introduced the risk in financial markets and business operations and brought the key concept of three pillars. The first pillar emphasized Minimum Capital Requirements, the second pillar focused on the Supervisory Review and Response to the First Pillar, and the third pillar underscored Market Discipline (or Market Constraints Mechanism), all with the overarching goal of promoting the stability of the financial system. Lall (2009) stated the failure of Basel II can be succinctly attributed to regulatory capture. A limited cohort of international banks succeeded in exerting influence over the Basel process, manipulating the rules of international capital regulation to maximize their profits at the expense of those entities lacking representation in the decision-making apparatus.

Regarding Basel III, also recognized as measures implemented in response to the 2007-8 financial crisis, the framework was unveiled in 2010. Basel III aimed to fortify regulations pertaining to capital adequacy, liquidity risk management, and systemic risk monitoring. The primary focus was on attaining higher and superior-quality capital, coupled with the imposition of more stringent requirements for the measurement and monitoring of liquidity risk. The implementation of these novel rules sought to augment proactive provisions for credit losses. On the other hand, Allen, CHAN, and Milne (2012) highlighted an ongoing debate over whether the new Basel III regulations concerning capital and liquidity would substantially escalate the cost of bank intermediation and curtail economic activity or if their impact on output growth would be more limited. This issue remains a subject of contentious discussion in the academic and financial sectors.

Loss Given Default (LGD) plays a critical role in Basel III when companies seek to quantify the extent of loss or determine the collateral required to offset expected losses. While the calculation of LGD is a complex process, contemporary financial solution providers offer LGD models that facilitate a direct computation of LGD. Moody's and S&P are among the prominent institutions that furnish widely used LGD models in the financial landscape. These models contribute significantly to enhancing the precision and efficiency of LGD calculations for businesses navigating the regulatory landscape defined by Basel III.

Moody’s has developed a sophisticated method knows as LossCalc. Gupton and Stein (2002) did research on this model 20 years ago. Their report concluded that LossCalc represents a multi-factor statistical model meticulously crafted using a database encompassing over 1900 defaulted instruments. This model provides estimated Loss Given Default (LGD) values for three distinct types of financial products: bonds, syndicated loans, and preferred stock. The analytical framework of LossCalc operates on four hierarchical levels, incorporating economic factors, industry factors, instrument-specific considerations, and capital structure. The authors of the study conducted a comparative analysis between LossCalc and alternative methods to discern the model's performance. Their findings revealed that LossCalc outperformed alternative methods in terms of LGD expectations, demonstrating superior accuracy, particularly in identifying instances of low recoveries when compared to historical average methods.

Expanding on the framework for Loss Given Default (LGD) analysis, Zheng & Huang (2014) indicated that the LossCalc LGD model includes factor transformation, modelling, and mapping. Factor transformation involves converting relevant influencing factors into model variables. From their study, it shows that Moody’s believes that using a composite index for forecasting macroeconomic variables will yield better predictive results of Loss Given Default (LGD) than using individual macroeconomic indicators alone. In terms of debt type and repayment order, using the average historical default loss rate is deemed to achieve better predictive results. As to the modelling phase, the LossCalc model primary applied regression methods. Regarding the last part of composite index forecasting method which is called as mapping, comparison between the output of LossCalc model and the statistical results of historical default rates will be used.

Regarding the last topic of this study, the application of machine learning techniques in risk management is addressed. Bastos and Matos (2022) discussed three specific machine learning techniques employed for predicting Loss Given Default (LGD). These techniques include fractional regression models, decision trees, and gradient boosting machine models. Bastos (2010) in his earlier paper mainly focused on regression tree models application for forecasting bank loan credit losses. Through extensive data analysis, Bastos concluded that regression tree models have a statistically significant predictive advantage over regression models in the dataset he used in terms of RMSE and MAE.

In addition to fractional regression models, decision trees, and gradient boosting machine models, other popular methods for predicting Loss Given Default (LGD) include random forest and Ordinary Least Squares (OLS). Töws (2016) raised a question whether normal methods, such as ordinary least squares (OLS) linear regression can be thought as an appropriate way for estimating LGD. To address this, a comparative analysis was undertaken between OLS and more complex methods, such as regression trees and multi-step models. The results indicated that the regression trees model exhibited a significantly better performance than OLS linear regression, particularly when dealing with large datasets.

Contrary to the perception that linear models are not considered the most suitable for predicting LGD, Yashkir (2013) focused on mainly on linear models. They compared several most popular LGD models including Tobit, LSM, Three-Tiered Tobit, Beta Regression, Inflated Beta Regression, Censored Gamma Regression. Besides, after comparison of these models, they concluded that the performance quality of the model depends mainly on the proper choice of model factors, but not on the fitting model. In summary, this means the performance quality of the model depends more on the choice of factors (variables) used to construct the model, rather than on the specific fitting model chosen. Even a simple linear model can perform well if the appropriate factors are selected.

3. Methodology

In this chapter, our focus will be primarily directed towards exploring the implications of the new Basel IV framework and its impact on LGD. We will begin by examining the latest changes introduced under Basel IV and how they influence LGD. Subsequently, we shall delve into the mathematical aspects concerning the definition and measurement of LGD. Finally, our attention will shift towards the mechanics behind Moody's model for LGD prediction, as well as the performance evaluation of machine learning models in this context. Through a comprehensive exploration of these topics, readers will develop a foundational understanding of LGD and the various methodologies employed in its prediction, catering to both novice and advanced learners alike.

3.1. New Changes and Impact of LGD in Basel IV

"Why are banks so important in our lives?" Some may ponder this question. Simply put, modern banks provide the financial means for ordinary people to make significant purchases such as homes and university tuition. They also finance companies, enabling business expansion and increased profitability. Furthermore, with the advancement of digital payment systems, banks have become increasingly vital to a country's financial infrastructure. Within the banking system, the Basel Accords, established by the Basel Committee on Banking Supervision (BCBS), serve as internationally agreed-upon rules. Initially supported by G10 Governors and central banks, the Basel Accords now involve 28 jurisdictions and 45 institutions.

3.2.1. Overview of Basel IV

Basel IV aims to strengthen risk management practices, impose higher capital requirements on banks, and address inconsistencies in measuring and reporting risk exposures, building upon the shortcomings identified in Basel III. The new accords strive for greater consistency, comparability, and transparency in risk measurement and capital adequacy assessment. Six key changes under Basel IV include enhanced standardized approaches, restricted use of Internal Ratings-Based (IRB) approaches, a leverage ratio buffer, a shift in operational risk calculation, a risk-sensitive floor, and minimum capital standards.

Compared to Basel III, some argue that moving away from IRB may not fully satisfy global financial institutions. This shift could be viewed as a reversal of the principles laid out in Basel II. Exiting IRB could potentially reintroduce ambiguity in bank risk management and capital allocation processes.

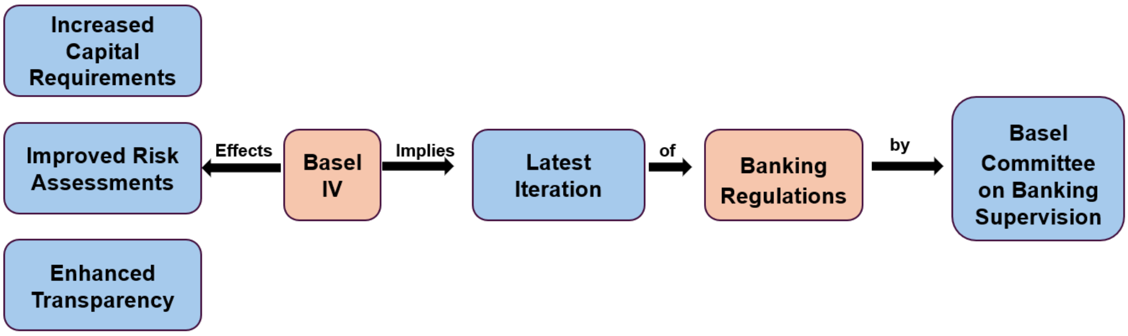
Increased Capital Requirements
Improved Risk Assessments
Enhanced Transparency
Effects
Basel IV
Implies
Latest Iteration
of
Banking Regulations
by
Basel Committee on Banking Supervision

Figure 1 - The effects and implication of Basel IV

3.3.2. LGD in Basel IV

In the context of LGD changes under Basel IV, Loss Given Default (LGD) input floors have been introduced. But first, what exactly are input floors? When banks employ internal models to estimate parameters such as the probability of default, they typically require enough observations for accurate estimation. If the number of input observations is too low, it indicates a significant risk of underestimation by the bank.

The 2008 financial crisis revealed that banks often lacked adequate data for analysing default probabilities, contributing to the crisis. Consequently, Basel III and Basel IV emphasize the importance of input floors to enhance the robustness and risk sensitivity of Internal Ratings-Based Approaches (IRBA) models used in Risk-Weighted Asset (RWA) calculations. As a result, input floors for LGD and Probability of Default (PD) were introduced.

Under the Basel IV framework, LGD input floors are set at values ranging from 25% to 50% for the unsecured portion of credit exposure, and from 0% to 15% for the secured portion. As a result, European Banking Authority (2019) indicated the increased significance of LGD input floors, particularly for positions subject to Internal Rating-Based Approaches Notably, these proposed LGD input floors are expected to have a substantial impact on Risk-Weighted Asset, especially in exposures to specialized banks, corporate Small and Medium-sized Enterprises (SMEs), and various retail categories. Chalpka, R., & Kopecsni, J. (2008) mentiond that this impact is particularly evident given the typically lower quality of data associated with LGD modelling compared to PD modelling.

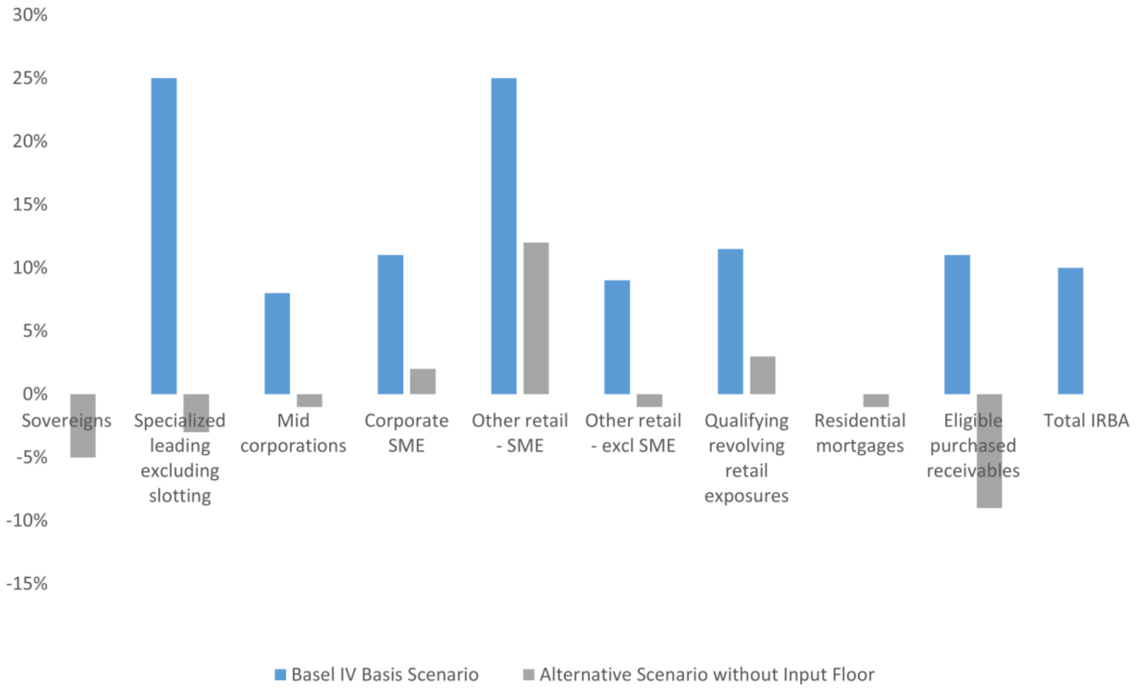

Figure 2 - Percentage change in IRBA SHE per exposure class without LGD floor.

3.2. Mathematical Definition and Measurement of LGD

Loss Given Default (LGD) represents the percentage of economic loss suffered by a bank in the event of a borrower's default. To fully grasp this concept, it's essential to define default, loss, and default risk exposure accurately. In the context of bank loans, default occurs when a borrower fails to meet contractual obligations, resulting in economic loss for the bank. Loss encompasses all relevant factors, including significant discount effects and direct and indirect costs incurred during the loan recovery process. Default risk exposure refers to the anticipated exposure to losses due to the potential default of a borrower. In summary, LGD can be defined as the ratio of the economic loss incurred by a creditor (bank) following a debtor's (borrower's) default on a specific transaction (obligation) to the risk exposure of that transaction.

3.2.1. Mathematical Definition

In mathematical terms, Loss Given Default (LGD) can be defined as the loss conditioned on the event of default, and its value is contingent upon the definition of default. In the following formula, we denote D as the event of default, L as the loss. LGD is represented as the random variable:

$$LGD = P(L|D = 1), \tag{3.1}$$

Where *D* = 1 in the event of default and *D* = 0 otherwise.

3.2.2. Measurement

There are several methods to measure LGD, and one prominent approach is Market LGD. This method involves measuring LGD by examining the market price of publicly traded bonds or loans that have defaulted. The market price inherently reflects investors' expectations regarding the recovery of bonds, including factors such as discounted principal, interest losses, and expenses associated with debt restructuring. The formula for Market LGD can be expressed as follows:

$$LGD = 1 - \frac{BP}{EAD}, \tag{3.2}$$

Where BP represents the bond price and EAD stands for exposure at default. Since they result from a market transaction, they are considered less susceptible to improper valuation.

The second method, known as workout LGD or the recovery discounting method, mentioned in Schuermann (2004)’s study, involves calculating LGD by discounting expected cash flows during the default settlement process, while accounting for various expenditures including fees. These cash flows are discounted to the point of default to ascertain the LGD value. The formula is as follows:

$$LGD = 1 - \frac{PV1 + PV2}{EAD}, \tag{3.3}$$

Where *PV1* refers to the "recovered principal and interest amount," and *PV2* signifies "the realized recovery amount from the liquidation of pledged assets after default."

The critical concept in this method is the discount rate, yet determining the appropriate rate can be challenging. Debt restructuring may involve the issuance of different assets, ranging from risky ones like equity or warrants to less risky options like notes, bonds, or cash. The correct valuation rate should ideally align with the risk level of the asset. Following default, the bank, now an investor in a defaulted asset, should value it accordingly, potentially employing the bank's hurdle rate. Unsuitable rates include the coupon rate (predetermined before default, often too low) and the risk-free (or Treasury) rate.

The last method is Implied Market LGD, also known as the spread estimation method in the study of Md (2023). This approach draws insights from the credit spreads of bonds that are still publicly traded in the market and have defaulted. It operates under the assumption that the market pricing of bonds is efficient and promptly reflects changes in the credit risk of the issuing company. The yield spread between corporate bonds and risk-free rates represents the risk spread of the bonds. This risk premium encapsulates both the Probability of Default (PD) and Loss Given Default (LGD), but its application is less common due to the requirement for substantial data support and the utilization of complex asset pricing models.

3.2.3. Single Influencing Factors of LGD

The technique for analysing the influencing factors of Loss Given Default (LGD) involves a statistical analysis method used to determine the various directions and degrees of influence within LGD affected by multiple factors. In simpler terms, the LGD amount is influenced by both idiosyncratic risk factors and systematic risk factors. This theory has evolved through three versions.

The first version was raised by Frye (2002) who believed that recovery rate (equal to 1-LGD) is affected by systematic factor and idiosyncratic risk factor, so the estimation of recovery rate is R_j.

$$R_j = \mu_j + \sigma p X + \sigma \sqrt{1 - p^2 Z_j}, \tag{3.4}$$

Where R_j is referred to the recovery rate, X represents the systematic risk factor, Z_j is the idiosyncratic risk factor. p represents the correlation coefficient between systematic risk factor and idiosyncratic risk factor $p = Corr(R_j, X)$.

In addition, this model assumed that same seniority class bonds have the same average default recovery rate and assumed that bond holders' idiosyncratic risk factors Z_j are independent and follows a normal distribution. Therefore, the recovery rate follows a normal distribution with mean μ_j, and the variance σ.

However, this model has a significant flaw which is that the value of μ_j is not limited within [0,1], it further caused the value range of R_j is between (-∞, +∞). Obiviouly this outcome didn't have any explainable economic meaning from Marc & Zöllner (2023).

Considering the fat tail in empirical research, Pikhtin (2003) assumed the default recovery rate follows Log-normal distribution, in this second version of theory.

$$R_j = \exp\left(\mu_j + \sigma p X + \sigma \sqrt{1 - p^2 Z_j}\right), \tag{3.5}$$

Although the assumption of a log-normal distribution is an improvement over the normal distribution assumption, it still does not address the issue of the range of values for recovery rates upon default. Afterwards, Schonbucher (2003) made a logit change on normal distribution, in the end third version of model was created out. The respective R_j can be represented as:

$$R_j = \frac{\exp(R_j')}{1 + \exp\ (R_j')}, \tag{3.6}$$

Where $R_j' = \mu_j + \sigma p X + \sigma\sqrt{1 - p^2 Z_j}$, and apparently this function satisfy the requirement of the value range of default recovery rate has to be fall within [0, 1], eventually it will make the LGD falls between the same value range.

In this section, we discussed a single influencing factor, which was simplified from the complex external world. While this simplification makes the model highly explainable, it fails to address the essence of the problem because describing the characteristics of only one external factor can be challenging. Therefore, further investigation into external factors is necessary. In the next section, we will delve into a multiple factors analysis model, using Moody's model as an example.

3.3. Moody's Model for Predicting LGD

To identify factors that best describe LGD and accurately estimate outcomes, we will explore one of the most renowned multiple factors models for predicting LGD, known as LossCalc. Initially developed by Gupton and Stein (2002), LossCalc was later adopted for practical use by Moody's. The accuracy of LossCalc estimation significantly surpasses traditional methods such as the Historical Moving Average Method. This improvement can be attributed to the extensive dataset maintained by Moody's, containing over 4000 records spanning loans, bonds, and preferred stock recovery data over a period of 20 years.

The LossCalc model incorporates nine explanatory factors to estimate LGD, encompassing aspects such as debt type and seniority, firm-specific capital structure, industry, and macroeconomic variables. LossCalc's forecasts for both immediate and one-year horizons are well-suited for various investor and risk management applications, offering enhanced insights into LGD prediction and risk assessment.

Debt Type and Seniority	
Historical average LGD by debt-type (loan, bond, and preferred stock) and seniority (secured senior unsecured, subordinate, etc.).	Historical Averages
Firm-Specific Capital Structure	

Seniority standing of debt in the firm's overall capital structure; this is the relative seniority of a claim. Note that this is different from the absolute seniority stated in Debt Type and Seniority above. The most senior obligation of a firm might be, for example, a subordinate note	Seniority Standing
Firm leverage (Total Assets / Total Liabilities)	Leverage
Industry	
Moving average of normalized industry recoveries. We have here controlled for seniority class.	Industry Experience
Banking industry indicator	Banking Indicator
Macro Economic	
One-year median RiskCalc default probability across time.	RiskCalc
Moody's Bankrupt Bond Index, an index of prices of bankrupt bonds	MBBI
Trailing 12-month speculative grade average default rate	Speculative-Grade Default Rate
Changes in index of Leading Economic Indicators	LEAD

Table 1: Explanatory factors in the LossCalc models.

3.3.1. Analytical Framework

When utilizing the LossCalc model to calculate LGD, the process can be divided into four distinct steps. Let's begin with Factor Transformation which refers to converting influencing factors (i.e., predictors) into variables that can be used in the model. This step aims to transform or process the raw data or influencing factors in a way that makes them suitable for model construction and analysis. For instance, Moody's suggests that using a composite index for macroeconomic variables yields better predictive performance compared to individual indicators. Similarly, when considering debt types and repayment order, employing the average historical default loss rate enhances predictive accuracy.

The defaulted debt prices used in the LossCalc model do not statistically follow a normal distribution. Therefore, to achieve better predictive accuracy, Moody's uses the Beta distribution instead of assuming a normal distribution. The Beta distribution, ranging from 0 to 1, is not constrained by symmetry assumptions and offers greater flexibility in describing data distribution. Specifically, it excels in characterizing data with higher probability distribution near the boundaries of 1 or 0, which is particularly useful for describing certain value ratios such as recovery rates.

In practice, due to substantial differences in the average recovery distributions among debt types, the LossCalc model initially groups variables based on different debts (e.g., loans, bonds, and preferred stocks). These variables are then transformed from Beta distribution to normal distribution. This transformation only requires observation of the mean (μ), standard deviation (σ), and bounding values of the recovery rates. The probability values of the transformed variables align with the probability values associated with the Beta distribution.

Once the transformation is complete and normal distribution and significance characteristics of explanatory variables are confirmed, the LossCalc model interprets the impact of independent variables and sub-models on the dependent variable through Linear Weighted Regression. The model can be represented as follows:

$$R' = a + bTYPE + cLEVG + dINDY + eMACRO + \varepsilon, \qquad (3.7)$$

Where R' is the recovery rate transformed form a Beta distribution to a Normal distribution, $TYPE$ represents Debt Type and Seniority, $LEVG$ is Capital Structure, $INDY$ is Industry, and $MACRO$ is referred to macroeconomic variables. a, b , c , d , e are all model parameters, ε is error.

In the next step, we focus on determining the parameters of the Beta distribution. It's important to note that R' is expressed through a normal distribution. When obtaining the mean value μ and variance value σ, an inverse action is necessary to transform it back into the original Beta distribution. We can derive the two parameters of the Beta distribution using the following formulas:

$$\mu = \frac{\alpha}{\alpha+\beta}, \tag{3.8}$$

$$\sigma = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2+1+\alpha+\beta}}, \tag{3.9}$$

Therefore, we can get the result of R (recovery rate) by addressing the following formula:

$$Beta(R,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(R)^{\alpha-1}(1-R)^{\beta-1}, \tag{3.10}$$

The final step involves calculating LGD. Once the value of R (recovery rate) is obtained, we can apply the relationship between LGD and RR, which states that LGD = 1 – R, to derive the value of LGD.

In summary: calculate R' using the linear regression model, assuming R' follows a normal distribution; compute the mean μ and variance σ^2 of R'; use these mean and variance values to solve for the parameters α and β of the Beta distribution; represent the default recovery rate R as following the Beta distribution with the solved parameters.

In evaluating the performance of the LossCalc model, two fundamental measures are commonly employed. The first measure assesses accuracy, gauging how effectively LossCalc predicts outcomes compared to actual data. The second measure evaluates efficiency, examining the width of confidence intervals around the model's predictions. It's important to note that these measures are interdependent and correlated to some extent. For instance, an increase in the confidence interval from 95% to 99% may positively impact accuracy.

3.3.2. Effective Validation

In the validation of the LossCalc model, Moody's employs the method of Walk Forward Validation. This involves using a period of data for model fitting and then testing the fitted model against subsequent data periods. This process is iterated continuously until testing reaches the current point in time. By doing so, the LossCalc model avoids the pitfall of fitting models to specific types of sample data and testing against homogeneous data, thus mitigating the risk of overfitting results. Moreover, Moody's selects data periods spanning multiple economic cycles for model validation to ensure robustness.

Moody's emphasizes that model validation is critical for establishing model credibility. Therefore, validation must be conducted rigorously and meticulously, with adjustments made for any unforeseen errors. For example, maintaining consistency in data sources is essential for comparative model testing. However, variations in predictive performance may arise for the same model across different sample tests. To minimize such differences and prevent misleading results, Moody's ensures the use of identical data sources and testing scopes when comparing the LossCalc model with other benchmark models, such as historical average recovery rate estimation methods.

3.4. Machine Learning Methods for Predicting LGD

In today's world, in addition to historical methods and external model approaches, Machine Learning methods are increasingly utilized across various industries worldwide. This trend is attributed to the flexibility of Machine Learning methods, which can adapt to different situations and be adjusted as needed to meet specific requirements. In this section, we will explore several popular Machine Learning methods for predicting LGD using datasets sourced from Moody's. Subsequently, we will conduct testing to identify the method with the best performance and compare their advantages and disadvantages. This section aims to provide readers with a fundamental understanding of the application of Machine Learning techniques in LGD prediction.

3.4.1. Overview of the Database

The dataset utilized in our research is known as Moody’s Ultimate Recovery Dataset (URD), comprising 2,784 bonds and 1,846 loans defaulted data spanning from 1987 to 2010 in the United States. The term "ultimate recovery rate" refers to the recovery values that creditors receive upon resolution of default. The coverage of default entities in the dataset includes US non-financial corporates with over $50 million in debt at the time of default. The dataset information is detailed in the following table, indicating that we have 27 columns or features. Some of these features are of data types "INT" or

"FLOAT," facilitating data analysis. However, some are of data type "OBJECT," necessitating conversion into dummy variables for analysis.

Column	Dtype	Column	Dtype
AD_ID	int64	totamount	float64
AC_ID	int64	debt	float64
name	object	instdebt	float64
industry	object	priabove	float64
date_default	datetime64[ns]	above	float64
type_default	object	pribelow	int64
instrument	object	cushion	float64
collateral	object	irindex	object
AIID	int64	spread	float64
datedefault	datetime64[ns]	effir	float64
ranking	int64	NRR	float64
origin	int64	FRR	float64
priamount	float64	DRR	float64
accamount	int64	RR	float64

Table 2: Data frame and data dype of URD dataset used in this paper.

In our research, we will segregate bonds and loans for analysis. Treating them similarly would implicitly assume that both financial products possess identical features and are impacted simultaneously by the same factors. Therefore, it is imperative to separate them during data analysis. Additionally, studies by Acharya, Bharath, and Srinivasan (2007) and Varma & Cantor (2005) have indicated that, generally, loans issued by banks have a higher probability of non-default compared to bonds. Furthermore, from a collateral perspective, secured lenders tend to recover more than unsecured creditors, aligning with common sense. Hence, it is essential to conduct separate analyses for bonds and loans to draw objective and meaningful conclusions.

3.4.2. Explanatory Data Analysis

In this section, we will conduct Exploratory Data Analysis (EDA) on our dataset before applying machine learning techniques. Jacobs & Karagozoglu (2018) indicated that EDA involves using numerical summaries and visualizations to explore the data and identify potential relationships between variables. The primary objectives of EDA include discovering anomalies in the data (such as outliers or unusual observations), identifying patterns, and proposing interesting questions or hypotheses.

By performing EDA, we can gain insights into the underlying structure of the data, understand its distribution, and uncover any potential issues or biases. This process allows us to make informed decisions about which machine learning techniques or statistical methods to apply and how to proceed with further analysis.

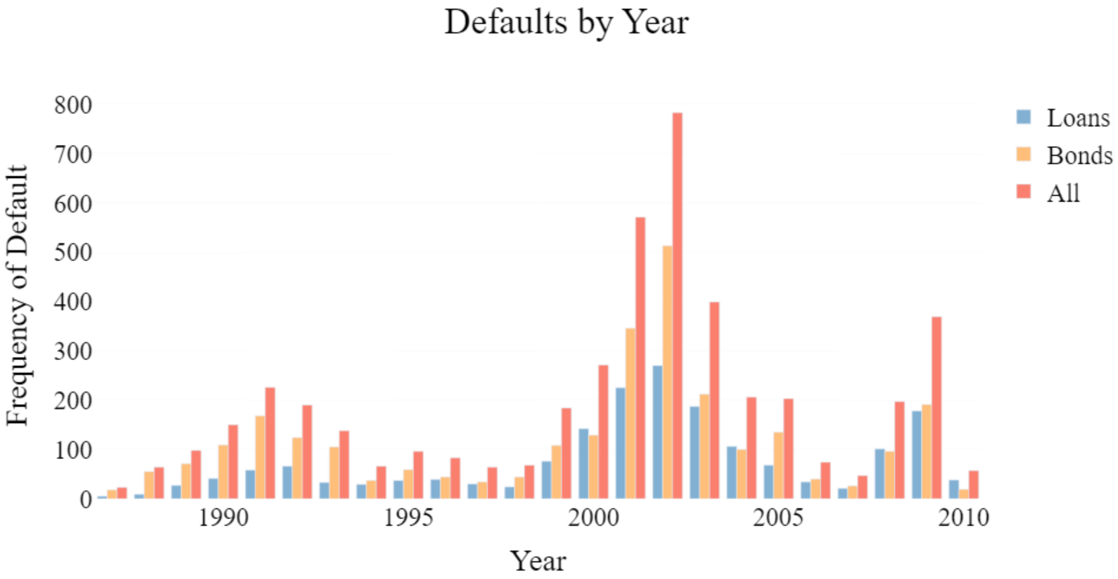

Figure 3 - Distribution of default for bonds and loans from 1989 to 2010.

From the histogram in Figure 3 above, depicting defaults data from 1987 to 2010, we observe distinct trends for loans and bonds. Notably, there are three prominent peaks within this period: 226 defaults in 1991, 783 defaults in 2002, and 370 defaults in 2009. The year 2002 stands out as the one with the highest number of default cases.

Upon closer examination and contextualizing with economic history, we can propose potential reasons for these default peaks. The brief recession experienced in the United States in 1991 resulted from a combination of factors, including war, financial crisis, manufacturing downturn, and debt issues. While the scale of this recession was relatively modest and its duration short, it nonetheless exerted a significant impact on the economy.

The year 2002 followed the bursting of the dot-com bubble, leading to global economic instability and numerous bankruptcies, particularly within the technology sector. Additionally, the September 11, 2001, terrorist attacks further exacerbated economic uncertainty, potentially contributing to the repercussions witnessed in 2002.

The peak of the global financial crisis occurred in 2008, with its effects intensifying in 2009. This crisis precipitated the collapse of financial markets, economic recession, and widespread corporate bankruptcies on a global scale. The ramifications of the financial crisis persisted for several years, with 2009 representing one of its peak periods.

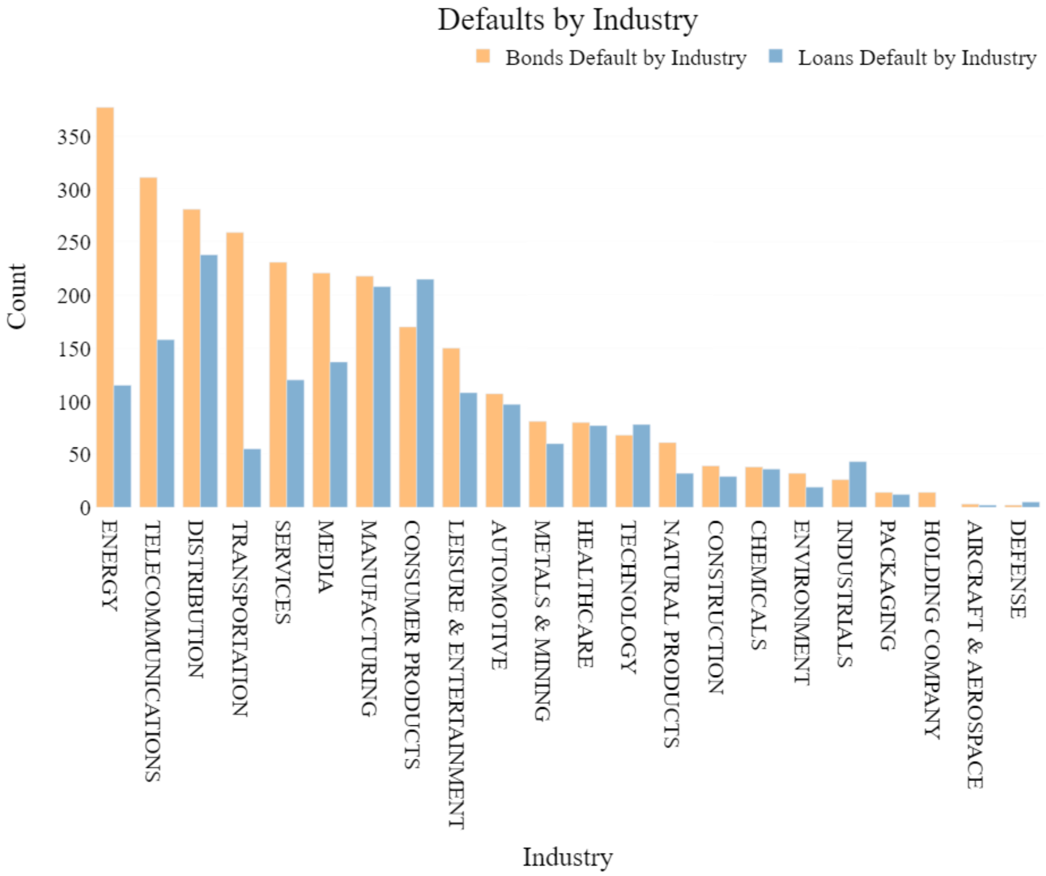

Figure 4 - The distribution of default for bonds and loans across various industries.

This plot provides distinct insights into the count numbers of defaults attributed to each industry in the total dataset. For loans, Distribution, Consumer Products, and Manufacturing accounted for the top three industries. Conversely, for bonds, the top three industries differed significantly from loans. The Energy sector occupied the first position, followed by Telecommunications, and Distribution.

This discrepancy in industry distribution between bonds and loans underscores the diverse default patterns across sectors. For instance, while Energy emerges as the leading industry for bond defaults, it does not hold the same position for loans. Such variations indicate differences in risk exposure and financial stability across different sectors. For investors, diversifying investments across various industries becomes crucial for effective risk management. Understanding the default distribution across different sectors empowers investors to make informed decisions regarding portfolio allocation, thereby mitigating risks associated with any single industry concentration.

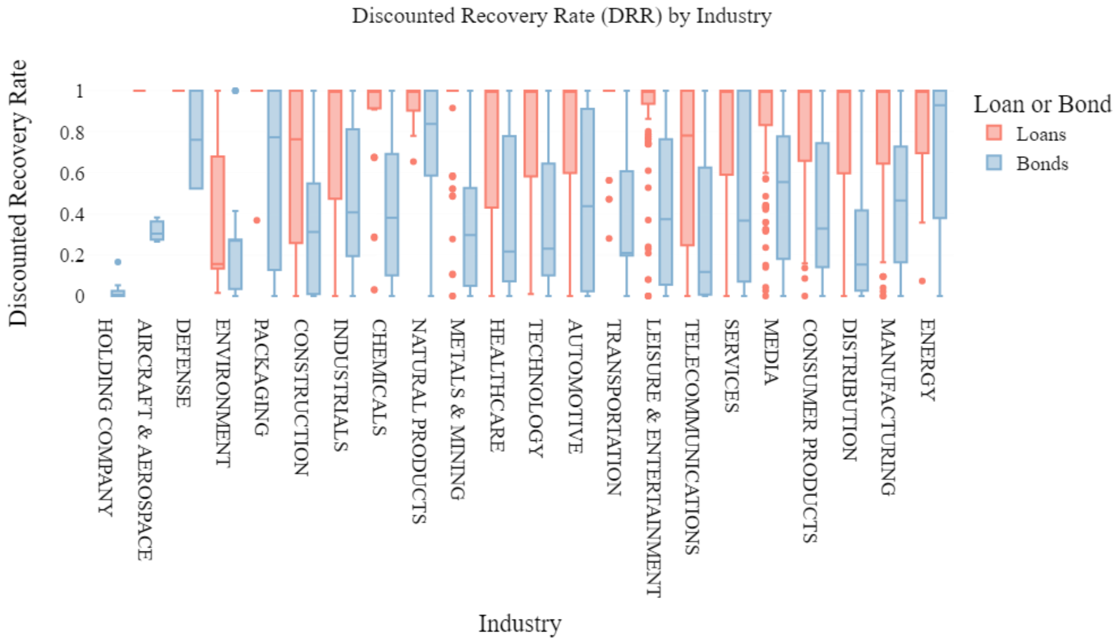

Figure 5 - Discounted recovery rate for bonds and loans across various industries.

Figure 5 presents the discounted recovery rate (DRR is the result of discounting the recovery amount of defaulted debt after default at a certain discount rate to its present value.) for bonds and loans across various industries using boxplots, which offer valuable insights into the variability across industries. The width of the boxes and the length of the whiskers depict the variability of DRR within each industry. Industries with longer whiskers or wider boxes tend to exhibit higher variability in DRR, indicating potential differences in recovery rates among entities within those industries. Additionally, any points lying outside the whiskers represent potential outliers in the data, signifying extreme values or unusual cases where the recovery rate significantly deviates from the norm within a particular industry.

Moreover, the line inside each box represents the median DRR for each industry. Comparing the positions of these lines across industries provides insights into the typical recovery rate within each sector. From this plot, we observe that for bonds, the services industry exhibits the greatest variability, while for loans, it is the telecommunications sector. In terms of median values, bonds in the natural products industry boast the highest median DRR at 83.9%.

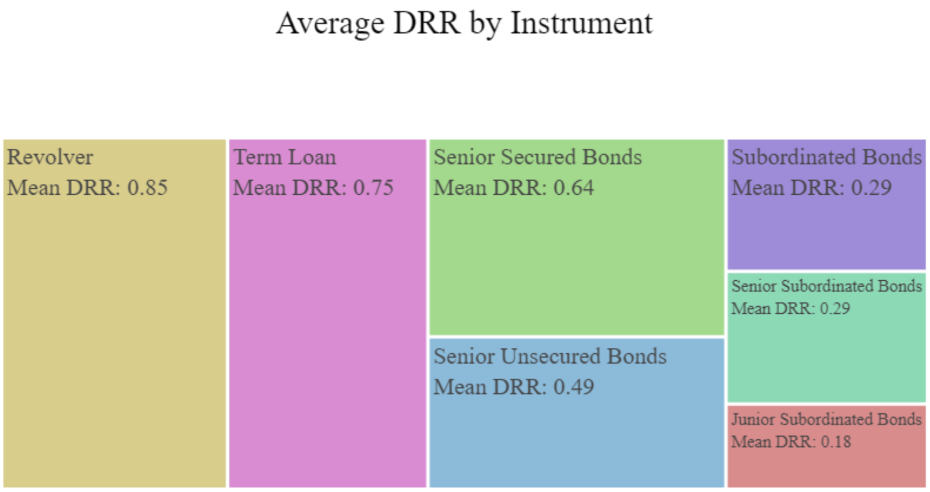

Figure 6 - Average discounted recovery rate by instrument.

Moving on to another influential variable affecting the recovery rate, namely the instrument type, we observe distinct categories for loans and bonds. For loans, we typically categorize them into two parts: Revolver and Term loan. A revolver allows a borrower to access funds up to a predetermined credit limit over a specified period, while a term loan provides a fixed amount of funds upfront, with repayment occurring over a specified period through regular instalments. In the case of bonds, they commonly exhibit five levels of instruments ranging from highest quality to lowest quality: senior secured bonds, senior unsecured bonds, subordinated bonds, senior subordinated bonds, and junior subordinated bonds. The tree map reveals that the DRR of each instrument aligns precisely with its level of quality. In other words, for bonds, senior secured bonds boast the highest DRR at 64% (indicating the lowest LGD), while junior subordinated bonds exhibit the lowest DRR at 18% (signifying the highest LGD).

Conversely, loans demonstrate relatively higher DRR values compared to bonds. Both term loans and revolvers exhibit DRR exceeding 75%. This suggests that loans generally exhibit lower LGD compared to bonds across various instrument types.

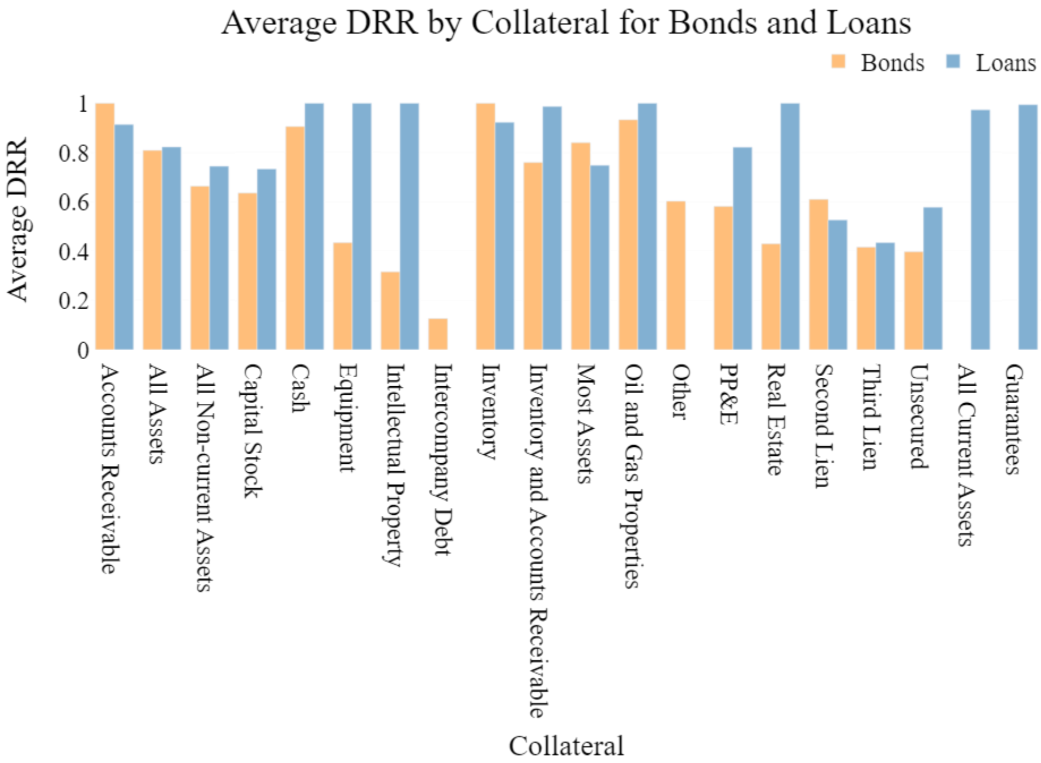

Figure 7 - Average discounted recovery rate by collateral for bonds and loans.

Let's delve into the analysis of the DRR by collateral, considering that banks often require collateral in the form of cash or equity, known as margin, to mitigate the risk of default. Examining Figure 8, we observe that collateral has a varied influence on DRR, depending on whether it pertains to bonds or loans. For bonds, collateral appears to exert a limited impact on DRR. However, specific types of collateral, such as second lien, third lien, and unsecured bonds, exhibit relatively low DRR, indicating higher LGD.

Conversely, for loans, collateral plays a more significant role in determining DRR. Accounts receivable, cash, and inventory secured loans emerge as the top three collateral types with the highest DRR, indicating lower LGD. Conversely, collateral types such as intellectual property and intercompany debt secured loans exhibit notably lower average DRR values. This analysis underscores the importance of collateral in determining DRR, particularly for loans, where certain collateral types significantly impact the recovery rate and, consequently, the LGD.

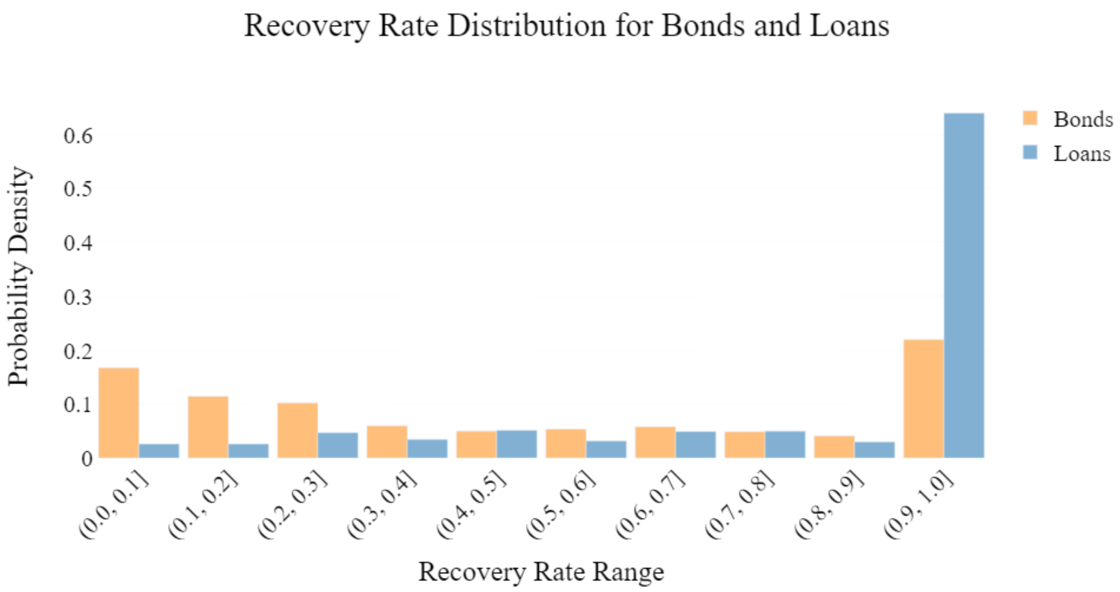

Figure 8 - Loan and bond recovery rate distribution.

Let's take a general look at the distribution of DRR in our dataset. It's evident that both bonds and loans exhibit strong skewness in their DRR distributions. For bonds, we observe a bimodal distribution with peaks at both ends. Approximately 17% of the distribution falls within the interval from 0 to 0.1, while 21% falls within the interval from 0.9 to 1. This suggests that bond default rates are notably high at both very low and very high values, indicating the presence of two distinct default scenarios or market behaviours. On the other hand, the DRR distribution for loans is unimodal, with the majority of defaults concentrated between 0.9 and 1, accounting for 63% of the distribution. This indicates that loan default rates are relatively concentrated, with most borrowers demonstrating high repayment ability. However, there are also a few cases of defaults in extreme situations.

Overall, significant differences exist in the default rate distribution between bonds and loans, reflecting different risk characteristics and market behaviours for the two asset classes. The bond market demonstrates a more diverse bimodal distribution, reflecting the diversity of default rates for different types of bonds. Conversely, the loan market exhibits a more concentrated unimodal distribution, possibly influenced by more consistent market factors.

3.4.3. Correlation Analysis

To assess the relationships between the explanatory variables and the LGD outcome, we will conduct a correlation analysis. We have identified six main factors as explanatory variables:

- Debt Cushion (cushion): This represents the portion of a company's total debt that ranks lower in priority for repayment in the event of default. It quantifies the amount of debt subordinate to other obligations and is often expressed as a percentage of the total debt.
- Ranking: This variable defines the rank of the debt in the capital structure, where lower numbers indicate higher priority for repayment.
- Spread: The spread refers to the difference between the loan interest rate and the interest rate index. It serves as an indicator of the loan's risk level and reflects market conditions, with higher spreads suggesting greater risk or market scepticism about the borrower's creditworthiness.
- Effective Interest Rate (effir): This is the actual annualized interest rate of a loan, accounting for the interest rate index, spread, and any additional fees or adjustments. It represents the interest rate paid by the borrower, considering all relevant factors.
- Principal Below (pribelow): This refers to the portion of debt that remains unpaid at default and is lower than the original principal amount.
- Principal Above (priabove): This denotes the portion of debt instruments that remain unpaid at default and exceeds the original principal amount.

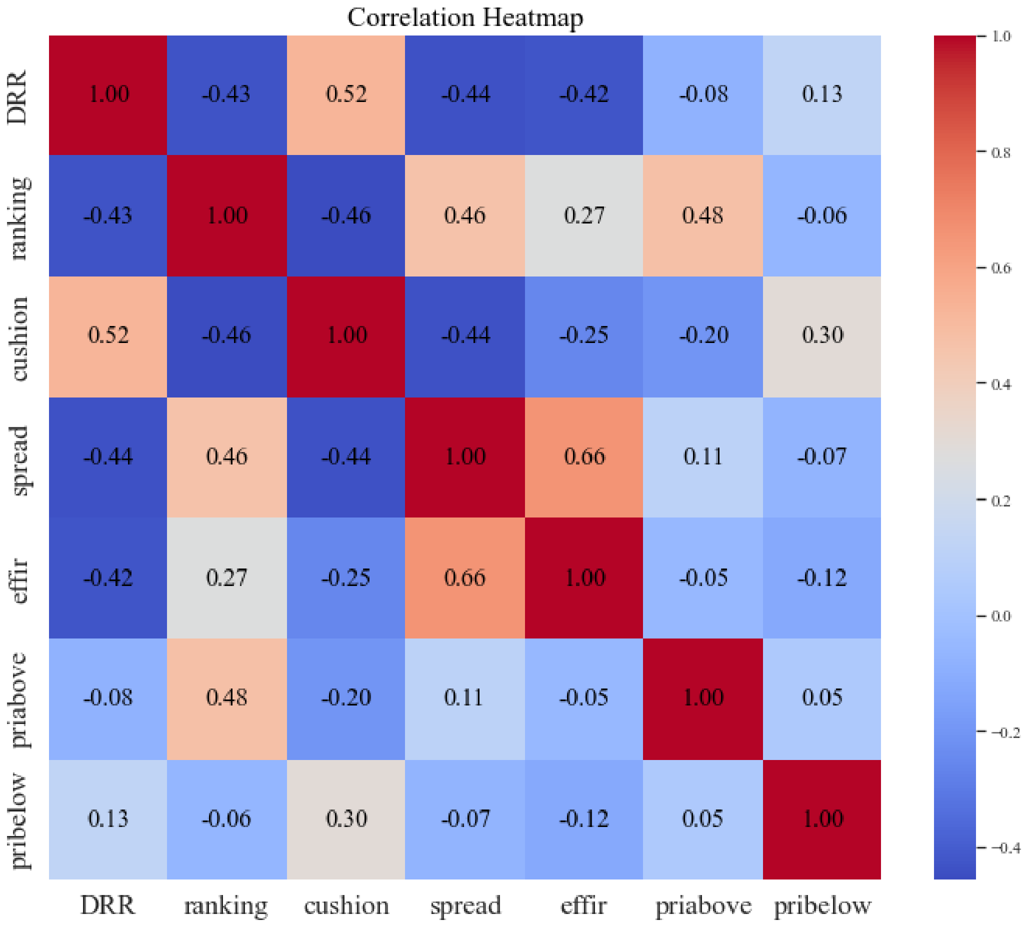

Figure 9 - Correlation heatmap between DRR, ranking, cushion, spread, effir, priabove, pribelow.

The correlation heatmap analysis has revealed several insights regarding the relationships between DRR and the selected variables:

1. Negative correlation with Ranking: The correlation coefficient of -0.43 indicates an inverse relationship between recovery rates and ranking. Here, "ranking" pertains to the hierarchical position of a debt within the corporate debt structure. A debt with a rank of 1 holds the highest priority for repayment, followed by subsequent ranks. Consequently, the observed negative correlation implies that higher recovery rates tend to align with lower-ranking debts within the debt structure of the firm.

2. Positive correlation with Debt Cushion: With a correlation coefficient of 0.52, there is a moderate positive correlation between DRR and Debt Cushion. This indicates that higher recovery rates may coincide with greater levels of debt cushion, suggesting that debt instruments with higher DRR may possess more substantial debt cushion.

3. Negative correlation with Interest Rate Spread: The correlation coefficient of -0.44 indicates that higher recovery rates may correspond to lower interest rate spreads. This suggests that debt instruments with higher recovery rates may have lower effective interest rates, reflecting lower perceived risk or market scepticism.

4. Negative correlation with Effective Interest Rate: With a correlation coefficient of -0.42, there appears to be a potential relationship between recovery rates and lower overall borrowing costs. This implies that debt instruments with higher recovery rates may have lower effective interest rates, contributing to lower borrowing costs for borrowers.

In summary, the correlation analysis has provided valuable insights into the associations between DRR and various variables. These associations offer valuable information for risk assessment and debt management strategies, highlighting the potential impact of DRR on debt instrument characteristics and borrowing costs. These factors will be instrumental in our subsequent machine learning analysis.

3.4.4. Linear Regression Model

In a linear regression model, we seek to uncover the relationship between a single dependent variable (Y) and multiple independent variables (X). In this section, we will employ multiple linear regression to conduct a straightforward prediction using a set of six numerical variables and some categorical variables, with the objective of exploring the predictive capability of these variables for Loss Given Default (LGD), represented as (1 - recovery rate). Multiple linear regression is a statistical technique designed to establish a linear relationship between independent variables and a dependent variable, thereby leveraging this relationship to predict the dependent variable. The model can be expressed as follows which are from the study of Svedberg, M., & Ljung, C. (2020):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon, \tag{3.11}$$

where y represents the dependent variable DRR (1-LGD), $x_1,\ x_2, \dots, x_p$ represent the independent variables, $\beta_0, \beta_1, \dots, \beta_p$ is the regression coefficients, ε represents the error term. By minimizing the sum of squared errors, optimal estimates of the regression coefficients can be obtained, thus establishing a linear relationship model between the independent and dependent variables.

3.4.5. KNN Regression Model

KNN (K-Nearest Neighbours) regression is a supervised learning technique employed for regression tasks. In KNN regression, when forecasting the target value of a new data point, the method considers the closest neighbours from the training data and calculates the average (or weighted average) of their target values to make the prediction. The key steps involved in KNN regression are as follows:

- Compute the distances between the new data point and all data points in the training set (typically using Euclidean distance or other distance metrics).

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + \cdots}, \tag{3.12}$$

- Find the K nearest training data points based on distance, known as the nearest neighbours.
- For regression problems, take the average (or weighted average) of the target values of these K nearest neighbours as the prediction for the new data point.

Feature scaling is also crucial in KNN regression. Since the KNN algorithm relies on distance metrics to determine nearest neighbours, the scale of features affects distance calculations. To ensure that each feature contributes roughly equally to distances, it's common practice to scale features to have similar magnitudes. Two commonly used methods for feature scaling are:

- Standardization: Transforming feature values to follow a standard normal distribution with a mean of 0 and a standard deviation of 1. This can be achieved by subtracting the mean and dividing by the standard deviation.

$$x_{new} = \frac{x - mean(x)}{std(x)}, \tag{3.13}$$

- Normalization: Scaling feature values to a fixed range, typically [0, 1] or [-1, 1]. This involves subtracting the minimum value and dividing by the range (the difference between the maximum and minimum values).

$$x_{new} = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{3.14}$$

The advantages of KNN regression are its simplicity and ease of understanding. It does not require a training process, but it does require storing all training data points. Additionally, during prediction, it needs to compute the distances between the new data point and all training data points, which can be computationally expensive for large datasets.

3.4.6. Decision Tree

Decision tree is a classic machine learning method widely used for addressing classification and regression problems due to its simplicity and interpretability. The methodology involves progressively dividing features during tree construction, where decision conditions are represented as internal nodes, and each leaf node represents an output category or value. Feature selection is crucial in this process, often employing metrics like information gain and Gini coefficient to evaluate feature importance and select the best splitting feature.

In the coding aspect, we initially segmented the database into Bonds and Loans as discussed in the previous chapter. Subsequently, we executed several essential preprocessing steps, including one-hot encoding for categorical variables and standardization for numerical variables to ensure data availability and accuracy. Following this, we integrated the preprocessing and decision tree model into a pipeline, training the model using training data and making predictions on test data.

Nath & Kumar Mohapatra (2017) discussed in their study that the advantages of decision tree models lie in their simplicity, intuitiveness, and ease of understanding and explanation, rendering them suitable for handling data with complex features and nonlinear relationships. However, decision trees also have some drawbacks, such as being prone to overfitting and sensitivity to data noise.

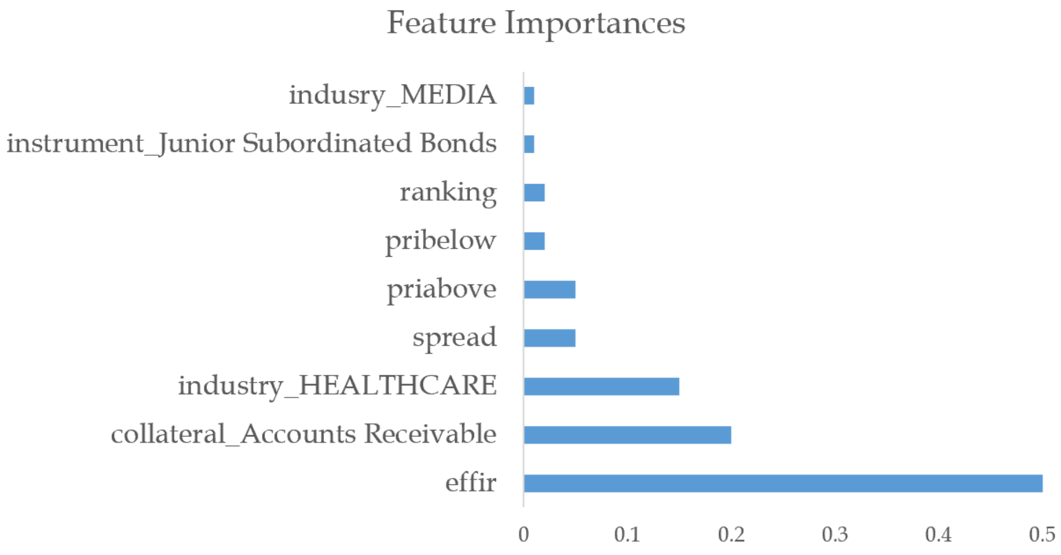

Figure 10 - Importance of regressors based on the sum of squares reduction in splits (Bonds).

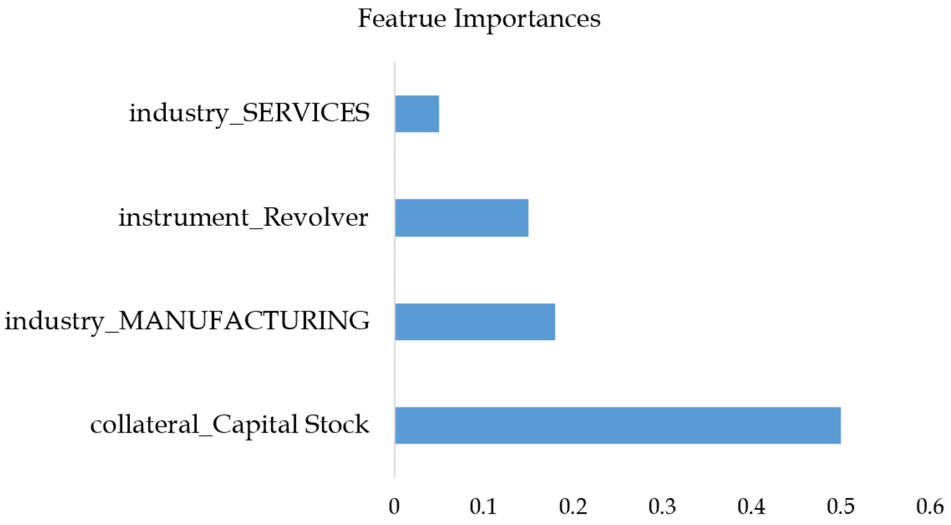

Figure 11 - Importance of regressors based on the sum of squares reduction in splits (Loans).

3.4.7. Random Forest

Before talking about random forest, let us have a look on Bagging (bootstrap aggregation) first. Bagging, also known as bootstrap aggregating, is an ensemble technique that operates on the original dataset by repeatedly selecting k new datasets with replacement for training classifiers. It utilizes a collection of trained classifiers to classify new samples, then aggregates the classification results of all classifiers using either majority voting or averaging their outputs. In detail, Bagging uses a set of trained classifiers to classify new samples. The process is as follows:

1. Create sub-datasets: Generate multiple sub-datasets from the original dataset using bootstrap sampling.

2. Train classifiers: Train an independent classifier on each sub-dataset.

3. Classify new samples: Input the new sample into all classifiers to obtain multiple predictions.

4. Aggregate results: Aggregate the predictions of all classifiers using majority voting (for classification) or averaging (for regression) to get the final prediction.

This method reduces bias and variance by combining the decisions of multiple classifiers, thereby improving the model's stability and accuracy. The highest-voted class or the average output is considered as the final label. Such algorithms are effective in reducing bias and can also lower variance.

Firstly, Random Forest employs CART decision trees as weak learners. In other words, Random Forest is essentially Bagging with CART decision trees as weak learners. Additionally, during the construction of each tree, only a random subset of features is considered, typically chosen as the square root of the total number of features, denoted as $\sqrt{m}$. In contrast, conventional CART trees utilize all features for modelling. Consequently, not only are the features randomized, but the randomness of features is also ensured.

The core idea of random forest is to construct a more powerful model by combining multiple decision trees. Each decision tree is built based on random sampling of the training data and random feature selection, ensuring that each tree has a certain level of diversity. During prediction, random forest integrates the predictions from all decision trees (For a given input sample, each decision tree predicts independently and obtains its own prediction result. These prediction results from all decision trees are collected to form a set.) and determines the final prediction result through voting (for classification problems) or averaging (for regression problems).

4. RESULTS

In this chapter, we will begin with a brief introduction to several error metrics commonly used for evaluating machine learning models. Subsequently, we will analyse the performance of the three models listed in the previous chapter. Finally, we will conduct a comparative analysis between each model.

4.4. Error Metrics

Regarding the predictive accuracy of models, the most common methods for assessment are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) from Reis & Quintino (2023).

Mean Squared Error (MSE) measures the average squared differences between the predicted values and the actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{4.1}$$

Where: n is the number of observations, y_i is the actual value of the target variable for observation *i*, $\hat{y}_i$ is the predicted value of the target variable for observation *i*.

Root Mean Squared Error (RMSE) is the square root of the MSE. It is in the same unit as the target variable and provides a more interpretable measure of error compared to MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{4.2}$$

Where n, y_i, $\hat{y}_i$, have the same meaning with above.

R-squared (R^2) measures the proportion of the variance in the target variable that is explained by the model. It ranges from 0 to 1, with higher values indicating better model fit. R-squared is a commonly used metric in regression analysis to evaluate the goodness of fit of a model. It represents the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{4.3}$$

Where $\bar{y}$ is the mean of the actual values of the target variable.

4.2. Model Performance

In this section, we will present three error metrics for each model, allowing for a comparative analysis to discern their respective advantages. All of them were being used k-fold cross-validation for performance measures.

4.2.1. Multiple Linear Regression Model

For this model, in addition to the six numerical variables discussed in the correlation analysis section, we introduced an additional categorical variable: Instrument. In the bond market, instruments denote the different priority levels or tiers of debt obligations issued by companies or governments. These tiers encompass senior secured bonds, subordinated bonds, and convertible bonds, each offering varying levels of risk and potential returns to investors based on their position in the capital structure. Since the instrument determines the order of repayment, it significantly impacts the DRR. As it is a categorical variable, we applied one-hot encoding to represent. Moreover, as previously discussed, loans and bonds possess different features, hence we will construct the model separately for each.

(RIGHT)

Metric	Value
MSE	0.094
RMSE	0.307
R^2	0.329

Metric	Value
MSE	0.062
RMSE	0.248
R^2	0.304

Table 3: Multiple linear regression for bonds (left) and loans (right).

For the bond market model, both MSE (Mean Squared Error) and RMSE (Root Mean Squared Error) are relatively low, indicating minimal error between predicted and actual values. Conversely, for the loan market model, MSE and RMSE are slightly elevated, accompanied by a marginally lower R² (Coefficient of Determination) compared to the bond market model. This implies a somewhat reduced explanatory power for the actual data in the loan market model.

4.2.2. KNN Regression Model

We implemented KNN regression with instrument as a categorical variable and six numerical variables. Initially, we encoded the categorical variable and standardized the numerical variables. We opted for 5 as the number of neighbours for the model.

Metric	Value
MSE	0.149
RMSE	0.386
R^2	0.439

Metric	Value
MSE	0.145
RMSE	0.381
R^2	0.341

Table 4: KNN regression performance of bonds (left) and loans (right)

The error metrics provide insights into the performance of the model: Mean Squared Error (MSE): is 0.086, indicating a relatively low level of prediction error on average. Root Mean Squared Error (RMSE): With an RMSE of 0.293. R-squared (R²): An R² of 0.389 suggests that the model can explain approximately 38.86% of the variance in the target variable, indicating moderate predictive performance.

Overall, the KNN regression model with the specified features and parameters performs reasonably well in predicting the target variable, as evidenced by the relatively low MSE and RMSE and the moderate R² value.

4.2.3. Decision Tree

The most important features for bonds are effective interest rate, collateral accounts receivable and industry healthcare. the most important features for loans are collateral capital stock, industry manufacturing and instrument revolver. We opted for 10 as the depth for bonds and 5 for loans.

Metric	Value
MSE	0.145

Metric	Value
MSE	0.091

RMSE	0.381	RMSE	0.302
R^2	0.458	R^2	0.336

Table 5: Decision tree regression performance of bonds (left) and loans (right)

From the performance score, we can see that decision tree model has a better performance for Loans which has all metrics lower than Bonds.

4.2.3. Random Forest

By utilizing the same set of variables as employed in the decision tree model previously discussed and specifying the number of estimators as 50 and the maximum depth as 20 for the random forest model, we obtained the subsequent outcomes.

Metric	Value	Metric	Value
MSE	0.051	MSE	0.031
RMSE	0.226	RMSE	0.177
R^2	0.635	R^2	0.643

Table 6: Random forest performance of bonds (left) and loans (right)

In general, in comparison with the results of the decision tree model, the random forest model exhibits reduced MSE and RMSE, indicative of enhanced predictive performance. Furthermore, it is noteworthy that the R-squared value of the random forest model surpasses that of the decision tree model, registering at 0.635. This suggests that the random forest model is capable of elucidating approximately 63.5% of the variance present in the dataset.

5. CONCLUSION

In conclusion, this study has provided a comprehensive overview of the concept of Loss Given Default (LGD) within the context of the Basel accords, with a focus on the changes introduced in Basel III and Basel IV. The discussion underscored the critical importance of LGD in the banking sector and examined the implications of LGD-related updates, particularly those outlined in Basel IV.

Furthermore, the research delved into the mathematical definition and measurement of LGD, tracing the evolution of LGD calculation formulas across different iterations of the Basel framework. This analysis illuminated the complexities involved in quantifying LGD and adapting measurement methodologies to meet regulatory requirements.

The study also explored Moody’s LossCalc model, a prominent tool for LGD prediction, elucidating its operational mechanisms and highlighting key factors influencing LGD estimates. By examining the model's architecture and underlying principles, valuable insights were gained into its predictive capabilities.

Additionally, empirical analysis was conducted using data from Moody’s Ultimate Recovery Database, employing various machine learning techniques such as linear regression, KNN regression, decision tree and random forest models. Through these analyses, the study evaluated the performance of each model using metrics such as R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

In addition to the analytical aspects discussed, this study employed exploratory data analysis techniques, including multiple data visualizations, to enhance understanding of the Moody’s Ultimate Recovery Database. These visualizations provided valuable insights into the distribution and relationships within the dataset, thereby facilitating a deeper comprehension of the underlying data patterns.

Overall, this research contributes to a deeper understanding of LGD within the Basel framework and provides practical insights into LGD prediction methodologies employed in real-world scenarios. By integrating theoretical discussions with empirical analyses, the study offers valuable insights for risk management practitioners, regulators, and researchers in the field of banking and finance.

References

Acharya, V., Bharath, S., & Srinivasan, A. (2007). *Does industry-wide distress affect defaulted firms?*

Allen, B., Chan, K., & Milne, A. (2012). Basel III: Is the cure worse than the disease? *International Review of Financial Analysis*, 159-166.

Bastos, J. (2010). Predicting bank loan recovery rates with neural networks. *Centre for Applied Mathematics and Economics (CEMAPRE)*.

Bastos, J. (2022). Explainable models of credit losses. *European Journal of Operational Research*, 386-394.

Chalpka, R., & Kopecsni, J. (2008). *Modelling bank loan LGD of corporate and SME segments: A case study*. Prague: Econstor.

European Banking Authority (2019). *Basel III Reforms: Impact Study and Key Recommendations.*

European Banking Authority. (n.d.). *EBA/GL/2017/16. "Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures, Article 131.* https://eba.europa.eu/.

Gupton, G. M., & Stein, R. M. (2002). *LossCalc: Model for Predicting Loss Given Default (LGD).*

Frey, J. (2002). *Depressing recoveries.* Risk.

Jacobs, M., & Karagozoglu, A. (2018). *Modeling Ultimate Loss Given Default on Corporate Debt.* USA: Hofstra University.

King, P., & Tarbert, H. (2011). *Banking & Financial Services Policy Report.*

Lall, R. (2009). *Why Basel II failed and why any Basel III is doomed.* Oxford.

PIkhtin, M. (2003). *Unexpected recovery risk.* Risk.

Marc, G., & Zöllner, M. (2023). *Tuning White Box model with Black Box models: Transparency in credit risk modeling.* University of Braunschweig – Institute of Technology.

Md., A. (2023). *Determinants of Bank Credit Risk: Empirical Evidence from Scheduled Commercial Banks in Bangladesh.* University of Dhaka.

Nath Pandey, T., & Kumar Mohapatra, S. (2017). *Credit Risk Analysis using Machine Learning Classifiers.* India: International Conference on Energy, Communication, Data Analytics and Soft Computing.

Reis, B., & Quintino, A. (2023). *Evaluating Classical and Artificial Intelligence Methods for Credit Risk Analysis.* Lisbon: Journal of Economic Analysis.

Schonbucher. (2003). *Credit Derivatives pricing Models: Model, Pricing and Implementation.*

Schuermann, T. (2004). *What Do We Know About Loss Given Default.* New York.

Svedberg, M., & Ljung, C. (2020). *Estimation of Loss Given Default Distribution for Non-Performing Loans Using Zero-and One Inflated Beta Regression Type Models.* Stockholm.

Töws, E. (2016). *Advanced Methods for Loss Given Default Estimation.* PhD thesis, Universität zu Köln.

Varma, P., & Cantor, R. (2005). *Determinants of recovery rates on defaulted bonds and loans for North American corporate issuers: 1983-2003.*

Yashkir, O., & Yashkir, Y. (2013). Loss Given Default Modelling: Comparative Analysis. *Journal of Risk Model Validation*, Vol.7, No.1.

Zheng, Y., & Huang, D. (2014). Research on default loss rate measurement methods by S&P and Moody's. Shanghai: Credit Quarter.