Study on Loan Risk Control by Markov chain and the variational method

Kun Wang

Thesis submitted for the fulfillment of the requirements for the degree of Master Research

Victoria University, Australia
Institute of Sustainable Industries and Liveable Cities
July 2023

Abstract

It is widely acknowledged that loan companies face various risks, and there are several models available to help them analyze customer behavior and control these risks. My research focuses on building up models to predict customs behavior as well as control the risk. In particular, we answer the following questions: which kind of factors do we need to use to set up a model? What is the Markov chain model, and how to use it? How to use the variational inference method to estimate the transition matrix?

We give an introduction to the research problem, highlighting its background, the significance of my research, and its objectives and outcomes in Chapter 1. The aim of our model is to overcome some common challenges encountered in the machine learning area, such as the concept drift problem and the data imbalance problem. One of the key contributions of our model is the provision of interval estimations for the coefficients within the transition matrices. This feature offers a distinct advantage over conventional models as it allows for greater flexibility in terms of coefficient selection.

In Chapter 2, we conduct a literature review from various perspectives on the research problem. We show that there are many factors that could affect customs behavior, such as income, job situation, health shocks, divorce, accidents, the market value of house prices, current loan-to-value, and so on. Previous researchers used different models to analyze the risks for the loan company. We will particularly use the Markov chain model, logistic regression model, and random forest tree model in our following work. We will apply the variational inference method in the Markov chain transition matrix, which gives us more flexibility to the model.

Chapter 3 focuses on developing an intelligent, machine learning-based Markov chain model to investigate loan risk and strategies for credit risk control. We reviewed the Markov chain model and the variational inference method in this chapter. We combined these two methods together and set up a new model. Our model involves the utilization of a Markov transition matrix to model state transitions of loan accounts. We optimize collection actions for each state and age of every consumer type to maximize the lender's expected value. Additionally, we tackle the challenge of imbalanced data by employing the variational inference method and logistic regression model. This approach bridges the gap commonly found in traditional machine learning processes when dealing with imbalanced datasets, thus improving prediction accuracy and reliability. The results of this chapter have been submitted to a reputable journal. Our model offers a novel approach to credit risk management. We anticipate that our study will significantly impact credit risk management practices and lay the foundation for future advancements.

Chapter 4 explores an effective approach to scheduling collection actions on consumer term-loan accounts. To achieve this, we employ a Markov decision model that facilitates efficient decision-making over time. The utilization of a Markov chain model in managing consumer loans rests upon the understanding that loan accounts naturally transition through various delinquency states over time. For example, an account in good standing remains so with timely payments but transitions to a delinquent state if payment is not received by the due date. Transition probabilities between states can be estimated using historical data, such as through maximum likelihood estimation. Our collection model gives an estimation of the profit for the loan company so can help to minimize their risk by adopting different actions at different levels. Our optimization approach takes into account the time value of money, balancing interest revenue and borrowing costs. It also ensures time consistency throughout the optimization process. We address competing risks that may arise between different account states and consider penalties that may be incurred due to late payments.

Chapter 5 focuses on studying the Random Forest model and applying it to a specific data set. We reviewed the random forest model and the logistic regression model, and compared two of them. We create a scoring model aimed at predicting loan default for new applicants. We are provided with a dataset containing information about title loan customers, including their performance status (default) and other factors. Initially, we analyze the relationship between the factors and select appropriate ones. We then apply the random forest model to the data, achieving an AUC over 80% by carefully choosing the hyper-parameter combination.

Finally, in Chapter 6 we give a conclusion about our research work and mentioned some future research directions in this area.

DECLARATION BY AUTHOR

I, Kun Wang, hereby declare that my Master of Research thesis entitled “Study on loan risk control by Markov Chain and the Variational Method” is no more than 50,000 words in length including quotes and exclusive of tables, figures, appendices, bibliography, references and footnotes. The work presented in this dissertation is the work of my own except where otherwise acknowledged. This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is my own work.

I have conducted my research in alignment with the Australian Code for the Responsible Conduct of Research and Victoria University Higher Degree by Research Policy and Procedures.

Signature

X [redacted]

Kun Wang

Date: November 16, 2023

DEDICATION

TO

MY PARENTS, MY HUSBAND, AND MY KIDS

I dedicate this thesis to my family. Thank you all very much for your relentless love and compassion throughout your life to support me.

Acknowledgement

I would like to express my deepest gratitude and appreciation to all those who have supported me throughout my journey in completing this master's thesis.

First and foremost, I am immensely grateful to my supervisor, Professor Hua Wang, Professor Yanchun Zhang, and Professor Sudha Subramani, for their guidance, invaluable insights, and continuous encouragement. Their expertise in the field has been instrumental in shaping the direction of this research and refining the quality of my work.

I would also like to extend my heartfelt thanks to the faculty members of Victoria University, whose knowledge and teachings have contributed immensely to my academic growth. Their dedication to excellence in education has been inspiring and has played a crucial role in shaping my intellectual development.

I am indebted to my friends and family for their unwavering support, encouragement, and understanding throughout this challenging endeavor. Their belief in my abilities and their constant motivation has been the driving force behind my success.

Furthermore, I would like to thank the research participants who generously contributed their time and insights to this study. Their willingness to share their experiences has been invaluable in shedding light on the subject matter and enriching the findings of this thesis.

Finally, I am grateful for all the sources and references that have contributed to the body of knowledge surrounding my research. Their extensive work has laid the foundation for this study and has helped me gain a deeper understanding of the subject.

Completing this master's thesis would not have been possible without the support and contributions of all these individuals and organizations. I am truly thankful for their involvement and encouragement throughout this journey.

DETAILS OF INCLUDED PAPERS: THESIS WITH PUBLICATION

Chapter No.	Publication Title	Publication Status
Chapter 3	Study on credit risk control BY variational inference	Accepted by Springer Lecture Notes in Computer Science (see [1])
Chapter 4	A collection model using Markov chain and Variational Inference methods	Submitted
Chapter 5	Comparison of Random forest model and logistic regression model	Submitted

Contents

1 INTRODUCTION

1.1 Background

A loan is a financial arrangement where one party, typically a financial institution such as a bank, lends a specific amount of money to another party, usually an individual or a business. The borrower is expected to repay the loan amount over a predetermined period of time, along with any interest or fees agreed upon.

Loans can be used for various purposes, such as purchasing a house, financing a vehicle, funding education, starting or expanding a business, or covering unexpected expenses. The terms and conditions of a loan, including the interest rate, repayment period, and collateral requirements, can vary depending on the type of loan and the lender's policies.

It's important to carefully consider your financial situation and ability to repay a loan before borrowing. Be sure to review and understand all the terms and conditions, including interest rates, repayment schedules, and any associated fees, to make an informed decision.

The compatibility of the loan market refers to how well the loan products and services offered by financial institutions align with the needs and preferences of borrowers. A highly compatible loan market would provide a wide range of options and flexible terms that meet the diverse requirements of borrowers.

In recent years, advancements in technology and changes in the financial industry have significantly increased the compatibility of the loan market. Here are a few factors contributing to its compatibility:

1. Diverse Loan Products: Lenders now offer a wide variety of loan products tailored to different purposes, such as personal loans, mortgages, auto loans, student loans, and small business loans. This enables borrowers to choose the loan type that best suits their needs.

2. Flexible Terms: Loan providers are offering more flexible terms, including different repayment periods, interest rate options, and repayment structures. This allows borrowers to find loan terms that align with their financial capabilities and long-term goals.

3. Online Access and Digital Services: Many lenders now provide online platforms and digital services, making it easier for borrowers to apply for loans, track their repayment progress, and access account information [2], [3]. This convenience and accessibility enhance compatibility by meeting the expectations of tech-savvy borrowers.

4. Personalization and Customization: Financial institutions are increasingly focused on personalized loan offerings. They assess borrowers' financial situations, creditworthi-

ness, and preferences to provide customized loan solutions, making the loan market more compatible with individual needs.

5. Transparent Information: The loan market has become more transparent, with lenders providing clear and detailed information about interest rates, fees, repayment terms, and conditions. This allows borrowers to make informed decisions and compare different loan options effectively.

It's important to note that the level of compatibility may still vary across different regions and lenders. It's advisable for borrowers to research and compare loan options, understand the terms and conditions, and assess their own financial circumstances before making borrowing decisions.

There are many motivations for loan companies to expand their business market. First, expanding into new markets allows loan companies to tap into a larger customer base, potentially increasing their loan portfolio and generating more revenue. By reaching more borrowers, they can generate more interest income and earn additional profits. Second,the loan industry is highly competitive, and expanding into new markets helps loan companies gain a larger market share. By establishing a presence in new regions or segments, they can strengthen their competitive position and potentially outperform their rivals. Third, expanding into new markets enables loan companies to diversify their risk exposure. By operating in multiple markets, they can reduce their reliance on a single market's performance. This diversification strategy helps them mitigate risks associated with economic downturns or changes in specific industries. Fourth, loan companies may identify underserved or untapped markets with a high demand for loans. Expanding into these markets allows them to address the needs of borrowers who may not have easy access to credit, thereby fulfilling a demand gap and establishing a strong presence in a relatively unexplored area. Fifth, expanding into new markets can provide loan companies with access to different customer segments. For example, they may target specific demographics, such as small business owners, students, or individuals with limited credit history. By catering to these segments' unique needs, loan companies can broaden their customer base and attract new borrowers. Sixth, with expansion, loan companies may achieve economies of scale. By spreading their fixed costs over a larger customer base, they can potentially reduce operational expenses per loan and enhance their overall efficiency. This can contribute to improved profitability and competitiveness. In summary, it's important for loan companies to conduct thorough market research and analysis before expanding into new markets. Understanding the local regulations, cultural nuances, and customer preferences is crucial for a successful expansion strategy.

There are various models used in controlling the risk of loans. The specific models used can vary depending on the institution or organization, but some commonly used models include: a) Credit Scoring Models: These models assess the creditworthiness of borrowers based on their credit history, payment behavior, income, and other relevant factors. b) Default Probability Models: These models estimate the likelihood of a borrower defaulting on their loan based on historical data and various risk factors. c) Loss Given Default Models: These models predict the potential loss a lender may incur in the event of a borrower default. They take into account factors such as collateral value, recovery rates, and liquidation costs. d) Stress Testing Models: These models simulate the impact of adverse economic scenarios on loan portfolios, helping to assess potential risks and evaluate the resilience of the portfolio under different stress conditions. e) Risk Rating Systems: These models assign risk ratings to loans based on a combination of quantitative factors (such as credit scores) and qualitative factors (such as industry risk, management quality, and market conditions). f) Value at Risk (VaR) Models: These models estimate the potential loss in the loan portfolio over a specified time horizon, typically using statistical techniques to calculate the maximum expected loss at a given confidence level. g) Machine Learning and Artificial Intelligence Models: These advanced models use algorithms to analyze large volumes of data and identify patterns and trends that may contribute to loan risk. They can help in credit scoring, fraud detection, and portfolio risk management. It's important to note that these models are generally used in combination, taking into account different aspects of loan risk, to ensure a comprehensive assessment and control of lending risks (see [4], [5], [6], and [7]).

The main objective of this project is to establish several models that can effectively analyze and understand the behavior of loan customers. Given that obtaining a loan is a significant and routine aspect of our everyday lives, it is crucial to investigate and comprehend the various factors that influence borrowers' decisions and actions.

Almost everyone encounters situations where they require financial assistance, such as purchasing a house or apartment, buying a car, or even acquiring a cash loan for specific expenses. These instances emphasize the importance of understanding loan customers' behavior to manage potential risks effectively.

The research conducted within this project encompasses both secured loans and unsecured loans. Secured loans involve collateral, which serves as a guarantee for repayment, examples of which are mortgage loans and auto loans. On the other hand, unsecured loans do not require collateral. The study aims to analyze both types of loans comprehensively, considering the distinctive characteristics and risk factors associated with each type.

By establishing accurate models to investigate loan customers' behavior, lenders, financial institutions, and policymakers can make informed decisions, develop appropriate risk management strategies, and ensure responsible lending practices for the benefit of both borrowers and lenders.

1.2 Research Motivation and Significance

Housing wealth is typically the largest component of wealth for households and mortgages are their main source of credit. During the recent global financial crises, unhealthy debt triggered a wave of foreclosures that impacted household consumption, balance sheets as well as the transmission of monetary policy to the real economy. When a crisis happens, consumer welfare has been jeopardized, and the ability of consumers to meet their loan obligations has been severely affected. As such, many economies experience rising impaired loans, which are choking the financial sector and the overall performance of their economies. Thus, the risk needs to be controlled and it is worth studying customers behavior during different periods.

My research is to build several models focusing on the following issues:

1. Predict a customer's behavior when he/she is applying for a loan to make a decision on if going to approve the loan or not.

2. Predict a customer's delinquency possibilities when he/she has already started a loan. This could help the loan company to control its cash flow.

3. Using the Markov chain model to fulfill basic ideas and applying the Bayesian method, and logistic regression method to build the model.

4. Using the Random Forest model to study a loan example.

5. Use the proposed model to help the loan company to expand its business market.

In general, it is a very complicated problem to predict an individual's behavior since different individual has a different habit and their behavior could depend on all kinds of factors for different individual. Many research work have done in various fields such as in Antisocial Behavior [8], [9], health related problems and cyber security issues [10]–[13]. Ignoring some very rare factors (e.g. accident, serious illness, etc.), for the majority there will be some kind of regular connections between their payment behavior and their current status (which may change at any time), such as unemployed, financial crisis, getting divorced, bankrupt, getting through Covid-19 pandemic, experiencing a hurricane, getting over an earthquake, etc. Therefore, there are many questions related to this problem we need to investigate.

In this project, the following topic will be focused on and discussed:

- Which information is related? What factors are important to us? How can we specify them?
- How a factor will impact a customer's behavior? How to quantitate a factor?
- Which model is appropriate? How bias is a factor?

Our models are very powerful and could be used in various aspects including but not limited to the following examples:

1. Market Research and Analysis: Conduct research on target markets, customer demographics, and industry trends. Gather data on consumer preferences, competitor analysis, and market demand to assist the marketing team in developing effective strategies and campaigns.

2. Customer Segmentation: Help the marketing team identify and segment the target audience based on various criteria, such as age, income level, location, or specific loan needs. This allows them to tailor marketing messages and initiatives to specific customer segments and improve the overall targeting and effectiveness of their campaigns.

3. Content Creation: Assist in creating compelling and informative content for marketing materials, such as website pages, blog posts, social media content, and email campaigns. This can involve writing engaging copy, producing relevant visuals, and ensuring that the content aligns with the company's branding and messaging.

4. Digital Marketing Support: Provide assistance in digital marketing activities, including search engine optimization (SEO), pay-per-click (PPC) advertising, social media management, and email marketing. Help optimize online presence, increase website traffic, and improve lead generation efforts through effective digital marketing strategies.

5. Analytics and Performance Tracking: Help set up and analyze marketing analytics tools to track the performance of marketing campaigns, website traffic, lead generation, and conversion rates. Provide insights and reports on key performance indicators (KPIs) to help the marketing team assess the effectiveness of their initiatives and make data-driven decisions.

6. Customer Feedback and Surveys: Assist in gathering and analyzing customer feedback through surveys or other feedback channels. This information can help the marketing team understand customer needs, preferences, and pain points, allowing them to refine their messaging and tailor marketing efforts accordingly.

7. Collaborate on Campaigns: Collaborate with the marketing team to brainstorm ideas, develop marketing campaigns, and execute promotional activities. Offer input on creative concepts, messaging, and strategies based on your understanding of the target audience and market dynamics.

1.3 Significance of the Research

The lending environment has been far from stable within the subprime home equity industry. The covid-19 pandemic brings the credit risk control area new challenges for both banks and loan companies. It even counts as a life-or-death choice for some companies. This research is focusing on setting up a new model which could produce forecasts of the outstanding balance in each delinquency state. This prediction is useful for predicting the cash flow situation for banks or companies. This project will be a pioneer study to use machine learning methods in the credit risk control field.

In this project, we have made significant advancements by developing a novel model that combines the traditional Monte Carlo Markov chain method with the variational inference method. By integrating these two powerful approaches, we have enhanced the accuracy and efficiency of our analysis.

One of the key contributions of our model is the provision of interval estimations for the coefficients within the transition matrices. This feature offers a distinct advantage over conventional models as it allows for greater flexibility in terms of coefficient selection. This flexibility, in turn, helps address common challenges encountered in machine learning, such as the concept drift problem and the data imbalance problem.

The concept drift problem refers to the phenomenon where the statistical properties of the target variable change over time, making it difficult for traditional models to adapt. However, by integrating the Monte Carlo Markov chain method, our model excels at capturing and adapting to these changes in real-time. This enables us to effectively monitor and predict loan customers' behavior, even in dynamic and evolving scenarios.

Moreover, the data imbalance problem, which arises when the number of instances belonging to one class significantly outweighs the other, can also be mitigated through the incorporation of the variational inference method in our model. By employing this technique, we optimize the model's ability to handle imbalanced datasets, ensuring reliable predictions and minimizing biases.

Overall, the inclusion of these advanced techniques in our model significantly improves the precision and robustness of loan customer behavior analysis. We are confident that our innovative approach will have far-reaching applications in the field of risk management, allowing financial institutions to make informed decisions and effectively control loan-related risks.

The novel framework proposed in this research is expected to help the industry develop a new method for evaluating a particular loan or pricing a new loan since it gives a prediction of customer payment information. The company could decide whether or not to approve the loan or approve the loan with what level of interest. For example, if the customer is with less credit, then the company can reduce its own risk by increasing the interest rate for the customer, and vice versa. The model could also help clients to analyze their ability for future payment and help them to make choices. In this way, it may help the industry significantly improve decision-making efficiency, reduce the human working intensity, and save the cost of system management.

This research may be an effective response to the public's concern about the loan risk control: (i) avoiding refinancing of FRMs mortgages for households; (ii) avoiding intermediary organizational constraints and lack of effective competition, especially within the refinancing market resulting in limited debt relief or refinancing; (iii) averting agency conflicts in servicing of largely securitized mortgages; and (iv) reducing moral hazard concerns in that by offering debt relief to distressed borrowers, many solvent borrowers could stop making payments to enjoy similar benefits.

The significance of my research on how to expand the loan company's market lies in its potential to provide valuable insights and inform strategic decision-making.

Our research can help identify untapped market segments, emerging trends, and new opportunities for the loan company to expand its customer base. It allows the company to understand market dynamics and assess the potential demand for its loan products in different regions or target demographics.

By conducting thorough research, you can gain insights into customer needs, preferences, and pain points. This understanding is crucial for tailoring loan products, services, and marketing initiatives to effectively meet customer expectations. It helps the loan company address specific customer needs and differentiate itself from competitors.

Research can help the loan company anticipate potential challenges or risks associated with expanding into new markets. It allows the company to assess regulatory requirements, competitive landscapes, economic conditions, and cultural factors that may impact its operations. By being well-informed, the loan company can develop strategies to mitigate risks and overcome obstacles.

Our research provides data and insights that can support data-driven decision-making within the loan company. It enables the company's management and marketing teams to make informed choices based on market trends, customer behavior, and industry analysis. This helps reduce guesswork and enhances the chances of success in expanding the loan company's market.

Through our research, the loan company can gain a competitive edge by better understanding its target market, competitors, and customer preferences. It allows the company to develop unique value propositions, differentiate itself from competitors, and position its loan products or services more effectively.

Overall, the significance of your research lies in its potential to inform the loan company's expansion strategy, improve decision-making, and increase the company's chances of success in new markets. By providing valuable insights, you can help the loan company navigate the complexities of expansion and achieve its growth objectives.

1.4 Outcomes of the Research

The project is expected to fill several research gaps in the field of machine learning methods used in credit risk control. A new approach will be applied to dig out more information from the current. This study will be groundbreaking in addressing the issue of data mining in the area of credit risk analysis. Unlike existing studies that have achieved good results in overall prediction accuracy, this project will discuss the performance of these models on minority classes. Its contribution to knowledge is mainly reflected in many aspects.

New models based on the Markov chain idea will be built. An innovative machine framework adopting the Bayesian estimation method will be designed to adjust the parameters in the Markov chain model. We will apply the Variational Bayes strategy in the logistic model for approximate Bayesian inference. This research delves into the development of an intelligent and machine learning-based Markov chain model that is specifically designed to examine loan risk and establish effective strategies for controlling credit risk. The model involves the careful modeling of state transitions for loan accounts using a Markov transition matrix, along with optimizing collection actions at each state and age for different consumer types. The objective is to maximize the expected value for the lender.

To address the challenges posed by imbalanced data, especially in the minority class, we have incorporated innovative algorithms and devised a consecutive incremental batch learning framework within the model. This framework aims to enhance the performance of the model and improve its accuracy in predicting credit risk.

Moreover, in order to bridge the gap in imbalanced data during the traditional machine learning process, we have employed the variational inference method and logistic regression model. These techniques aid in addressing the imbalance and further refine the model's predictions. By combining these advancements, we anticipate that the results of this study will contribute significantly to the field of credit risk. Our proposed machine learning-based prediction method not only offers enhanced accuracy but also provides a more effective approach towards managing credit risk.

Our research focuses on the efficient scheduling of collection actions for consumer term-loan accounts using a Markov decision model. Each consumer loan account, at any given age, can fall into different account states such as current, delinquent, early payoff, default, or bankrupt. To analyze the behavior and progression of loan accounts, we employ a Markov transition matrix to model state transitions. This matrix allows us to determine the probabilities of accounts moving between various states over time. By leveraging this framework, we develop an optimization method that aims to maximize the lender's expected value by determining the most suitable collection action at each state and age for each consumer type.

Our optimization approach takes into account multiple factors such as default risk, operational cost, time value of money, tradeoffs between interest revenue and borrowing cost, time consistency in optimization, competing risks among different account states, and penalties for late payment. These considerations enable us to create a more comprehensive and effective collection policy.

Through extensive analysis, we compare the performance of our dynamic collection policy with a static collection policy. The results clearly demonstrate that our method provides significantly higher value, particularly for accounts with high interest rates and medium to high loan amounts. Moreover, the effectiveness of our approach is particularly pronounced when stronger collection effects are observed. Additionally, our research sheds light on how the implementation of an optimal collection policy is influenced by factors such as interest rates, loan amounts, and collection effects. By examining the interactions between these variables and collection actions, we gain insights into the dynamics of loan accounts and can make informed decisions to maximize the lender's returns.

In Chapter 5 of our study, we extensively explore the Random Forest model and its application in the context of our research. We recognize the significance of this model as it offers unique advantages in terms of predicting and analyzing loan risks. To demonstrate its effectiveness, we apply the Random Forest model to a carefully curated dataset and conduct a thorough comparative analysis of the results obtained from this model with those generated by the conventional logistic regression model, which is commonly used in similar studies.

By utilizing the Random Forest model, we aim to harness its capabilities in handling complex relationships and capturing non-linearities within the dataset. This exploration allows us to evaluate the performance of the Random Forest model in terms of accuracy, precision, recall, and other relevant evaluation metrics. By comparing these outcomes with those obtained through the logistic regression model, we can draw meaningful conclusions about the suitability and efficacy of the Random Forest model in addressing loan risk control and credit risk management.

Furthermore, during our investigation, we analyze the predictive capabilities of both models, considering important factors such as the feature importance and variable interaction effects. This detailed analysis provides valuable insights into how each model handles the intricacies of loan risk and contributes to refining our understanding of their practical implications and limitations.

Through this empirical analysis, we aim to establish a sound foundation for adopting the Random Forest model as an effective alternative to the conventional logistic regression model in the context of loan risk assessment and credit risk control.

Overall, our study contributes to the development of more efficient and effective strategies for scheduling collection actions in consumer term-loan accounts. By considering a wide range of factors, we provide a comprehensive framework that can be employed by lenders to optimize their collections and improve their overall financial performance.

2 Literature Review

This section reviews literature related to the topic on looking for factors affecting mortgage payment. The history and recent advances as well as gaps between existing research and real-world applications will be presented.

2.1 Factors Study

Many studies have looked at factors affecting mortgage payment with most of them focusing on default or foreclosure especially in the USA real estate market and a sizable number on the UK market. Early literature shows the key drivers of default were home values and interest rates ([14], [15]). Riddiough (1991) [16] provided early insights on the modeling of "trigger events" such as job loss, health shocks, divorce, and other accidents. Similarly, Kau et al. ([17]); Schwartz and Torous (1993) reported loan vintage and housing index returns volatility as the key drivers of observed default behavior. Deng et al. (2000) [18] argued that negative events such as job losses and divorces were significant predictors of mortgage default. Using data on mortgages originated between 2003 and 2007, Mayer et al. (2009) [19] found unemployment and house prices as the key predictors of delinquency in the USA market (see [20]).

In response to the mortgage default and foreclosure crises which began in 2007, an increased number of researchers analyzed and documented numerous factors as the determinants of the observed default and foreclosure behavior ([21]). One of the key hypotheses regarding the causes of mortgage delinquency is that homeowners will not continue serving a mortgage if they enter into negative equity, for instance, if the value of the property drops below the mortgage value (Kau et al. 1992 [17]; Kelly and O'Malley 2016). Chan et al. (2014) [22] found that loan and individual characteristics such as borrowers' credit history, current loan-to-value, race, ethnicity, and income are key drivers of foreclosure. Guiso et al. (2009) [20] found severe negative equity, gender, future employment expectations, race, and morality as key determinants of ruthless or strategic default. Long-term unemployment as well as falling home prices which led to negative equity were also found to be key drivers of observed mortgage defaults, foreclosures, and housing vacancies (Jones et al. 2016 [21]; Tian et al. 2016 [23]).

Foote et al. also assessed this concept of negative equity on mortgage default decisions and found that some mortgagors who were in negative equity did not default and argued that this could partly be explained by price expectations. Consistent with that, Foote et al. also indicated that mortgagors who were in negative equity and defaulted could have done so not only because of negative equity but due to a “double-trigger” effect (negative equity combined with some adverse event such as loss of employment, health issues, death of the spouse, divorce, etc.). In agreement with the double trigger hypothesis, numerous other studies also documented that mortgages could be in negative equity and still not default (see [24]).

2.2 Loan marketing area

Loan market strategy refers to the approach taken by financial institutions and lenders to effectively manage and navigate the loan market. It involves the formulation and implementation of plans and tactics to achieve specific objectives, such as increasing loan portfolio growth, mitigating risk, and maximizing profitability. One key aspect of loan market strategy is the target market selection. Lenders analyze various factors like demographics, creditworthiness, and market trends to identify the most promising customer segments to focus their lending efforts on. This helps optimize resource allocation and tailor loan products to the specific needs and preferences of the target market.

Another crucial element of loan market strategy is product development and pricing. Lenders evaluate market demand and competition to design loan products that cater to the evolving needs of borrowers. They also analyze cost structures, interest rate trends, and risk assessments to set competitive pricing for their loan offerings. Balancing affordability for borrowers and profitability for lenders is a critical consideration in this process.

Risk management is a fundamental component of loan market strategy. Lenders employ various risk assessment models, such as credit scoring, underwriting standards, and collateral evaluation, to evaluate the creditworthiness of borrowers and manage the risk associated with loans. Effective risk management practices help reduce loan delinquencies, defaults, and overall portfolio risk. Marketing and customer acquisition strategies also play a pivotal role in loan market strategy. Lenders employ targeted marketing campaigns, advertising, and customer relationship management techniques to attract potential borrowers and establish long-term relationships. Emphasizing customer satisfaction and retention is crucial for building a loyal customer base and gaining a competitive edge in the loan market.

Furthermore, technology and digital transformation have become increasingly important in loan market strategy. Financial institutions leverage advanced technologies, such as artificial intelligence, machine learning, and automation, to streamline loan origination, underwriting processes, and enhance customer experience. This digitalization enables faster loan approval, improved operational efficiency, and better data-driven decision-making.

Digital marketing and online presence play a crucial role in the loan industry, specifically in the promotion and processing of loan applications. With the widespread use of the internet and digital technologies, borrowers are increasingly turning to online channels to find and apply for loans. As a result, loan companies need to have a strong digital marketing strategy and a prominent online presence to attract and cater to these tech-savvy borrowers.

Digital marketing for loan applications involves various strategies and techniques that aim to increase visibility, generate leads, and drive conversions. These strategies may include search engine optimization (SEO) to improve search engine rankings and ensure loan company websites appear in relevant search results. Additionally, pay-per-click (PPC) advertising campaigns can be employed, using targeted keywords and demographics to reach potential borrowers.

An engaging and user-friendly website is essential for loan companies to provide detailed information about their loan products, application processes, interest rates, terms, and conditions. A well-designed website can inspire trust and encourage visitors to submit loan applications.

To enhance the online presence, loan companies can also leverage social media platforms to engage with potential borrowers. They can create informative and engaging content, including blog posts, videos, and infographics, to educate and inform borrowers about loan options, financial management, and relevant industry trends. Utilizing social media advertising and retargeting can also help reach a wider audience and drive traffic to loan application pages.

Furthermore, the use of online loan application forms simplifies the loan application process, making it more convenient and accessible for borrowers. Companies can employ secure and user-friendly online application systems that enable borrowers to input their information, upload necessary documents, and track the progress of their application.

An integral part of a successful digital marketing and online presence strategy for loan applications is maintaining a strong online reputation. Positive customer reviews and testimonials can significantly influence the decision-making process of potential borrowers. Loan companies should actively monitor and respond to customer feedback, addressing concerns, and providing excellent customer service.

Last but not least, regulatory compliance is an essential consideration in loan market strategy. Lenders must stay up to date with relevant legal and regulatory frameworks to ensure compliance with consumer protection laws, anti-money laundering regulations, and fair lending practices. Adhering to these regulations helps foster trust, maintain a positive reputation, and avoid legal repercussions. Overall, loan market strategy encompasses a range of interconnected activities aimed at achieving sustainable growth, managing risk, meeting customer needs, and staying competitive in the dynamic financial market. It requires a comprehensive understanding of market dynamics, borrower behavior, and industry trends, along with continuous monitoring and adaptation to changing circumstances.

2.3 Research on Random forest model

The random forest model is a very useful and important model. There are many famous research works about the random forest model. For example, “Random Forests” by Leo Breiman (see [25]), this seminal paper, published in 2001, introduced the random forest algorithm and demonstrated its effectiveness in building ensemble models for classification and regression tasks; “Elementary Statistical Learning” (see [26]) by Trevor Hastie, Robert Tibshirani, and Jerome Friedman: This influential book provides a comprehensive overview of various machine learning algorithms, including the random forest, and serves as a fundamental reference in the field; “Extremely Randomized Trees” (see [27]) by Pierre Geurts, Damien Ernst, and Louis Wehenkel: Presented in 2006, this paper introduces the concept of extremely randomized trees, a variation of the random forest model that adds extra randomness to the tree-building process; “Random Forests for Large Data Sets” (see [28]) by Andy Liaw and Matthew Wiener: Published in 2002, this paper focuses on the application of random forest models to large datasets and provides practical insights on efficient algorithm implementations. These research works have significantly contributed to the development and popularization of the random forest model in the machine learning community.

2.4 Previous research about Markov Chain model

Markov chain is a very famous model. There is a long history of research works on MCMC. For example, “Foundations of Markov Chain Monte Carlo Methods” by Persi Diaconis and David Freedman: This influential paper, published in 1999, provides a comprehensive overview of the theoretical foundations and applications of Markov chain Monte Carlo (MCMC) methods; “Markov Chains and Mixing Times” (see [29]) by David Levin, Yuval Peres, and Elizabeth L. Wilmer: This book, published in 2009, explores the mathematical theory of Markov chains, including their mixing times and applications in various fields like computer science, statistics, and physics; “Introduction to Stochastic Processes with R” (see [30]) by Robert P. Dobrow: This book, published in 2016, offers an introductory approach to understanding stochastic processes, including Markov chains, with a focus on practical applications using the R programming language; “Markov Decision Processes: Discrete Stochastic Dynamic Programming” (see [31]) by Martin L. Puterman: This influential book, first published in 1994 and later updated in 2005, provides a comprehensive introduction to Markov decision processes and their applications in decision analysis and operations research; “Markov Chains: (see [32]) Gibbs Fields, Monte Carlo Simulation, and Queues” by Pierre Brémaud: This book, published in 1999, covers the theory and applications of Markov chains, including the study of Gibbs fields, Monte Carlo simulation methods, and queueing systems; “An Introduction to Markov Chain Monte Carlo Methods and Their Actuarial Applications” (see [33]) by Shaun Zhang and Gareth W. Peters: This research paper, published in 2009, explores the use of Markov chain Monte Carlo (MCMC) methods in actuarial science and insurance, providing insights into pricing models, risk assessment, and portfolio optimization.

Please note that these are just a few examples of famous research works in the field of Markov chain modeling, and there are numerous other seminal works and contributions by researchers in this area.

2.5 Variational inference

Variational inference is a powerful method used in probabilistic modeling to approximate complex posterior distributions. It has gained significant popularity in recent years due to its ability to address computational challenges posed by high-dimensional and intractable Bayesian models (see [34]).

The main idea behind variational inference is to approximate the true posterior distribution with a simpler and more tractable distribution from a predefined family of distributions. This approximation is achieved by minimizing the divergence between the true posterior and the approximation, typically measured using the Kullback-Leibler (KL) divergence. The family of distributions used for approximation is often chosen to be a parametric family, such as Gaussian or exponential family distributions (see [35], [36], [37], [35] and [38]).

To perform variational inference, one formulates an optimization problem where the goal is to find the parameters of the approximating distribution that minimize the KL divergence. This optimization problem is typically solved using iterative methods such as gradient descent or coordinate ascent. The result is an optimized approximation that provides an efficient representation of the posterior distribution.

Variational inference offers several advantages. Firstly, it allows for efficient posterior inference even for complex models with a large number of latent variables. Secondly, it provides a principled framework for model selection through the use of model comparison criteria such as the evidence lower bound (ELBO). Additionally, variational inference lends itself well to computational scalability, making it suitable for analyzing massive datasets.

Variational inference has found diverse applications across various domains, including machine learning, computer vision, natural language processing, and computational biology. Researchers have extended the basic variational inference framework to handle specific challenges, such as incorporating hierarchical structures, handling missing data, and addressing posterior multimodality.

Notable research works in the field of variational inference include "Variational Inference: A Review for Statisticians" by David M. Blei et al. (see [39]), and Christopher K. I. Williams, and "Auto-Encoding Variational Bayes" (see [40]) by Diederik P. Kingma and Max Welling.

Overall, variational inference is a versatile and widely adopted technique for approximating complex posterior distributions, enabling efficient and scalable probabilistic modeling across various applications.

My research works demonstrate the potential of combining Markov chain models with variational inference techniques to tackle a range of modeling and inference problems in various fields, such as finance, biology, and natural language processing.

2.6 Customer relationship management

Loan company customer relationship management (CRM) is a strategic approach that focuses on building and maintaining strong relationships with customers in the loan industry. It involves utilizing various tools, technologies, and methodologies to manage interactions and track customer information throughout the loan lifecycle. The primary goal of CRM in a loan company is to enhance customer satisfaction, improve customer retention, and drive profitability.

CRM enables loan companies to effectively manage customer interactions by centralizing customer data and providing insights into customer behavior, preferences, and needs. By leveraging CRM systems, loan companies can gather and analyze data from various touchpoints, including loan applications, customer inquiries, loan repayments, and customer feedback.

With CRM, loan companies can personalize their communication with customers, ensuring that offers, recommendations, and services are tailored to individual needs. This personalization enhances customer experience and fosters loyalty. CRM systems also facilitate efficient customer onboarding by streamlining application processes and automating approvals, reducing response times, and improving overall efficiency.

CRM tools further enable loan companies to proactively identify and address customer concerns, allowing for timely resolution of issues and personalized assistance. By tracking customer interactions and transaction history, loan companies can also monitor customer satisfaction levels and identify opportunities for cross-selling or upselling additional loan products or services.

Additionally, CRM helps loan companies streamline marketing campaigns, enabling targeted outreach and effective lead generation. By analyzing customer data, loan companies can identify potential prospects and segment their customer base to deliver relevant marketing messages, increasing the chances of success.

Furthermore, CRM provides valuable analytics and reporting capabilities, allowing loan companies to extract insights from customer data. These insights help in making informed decisions related to loan product development, pricing strategies, risk assessment, and overall business strategy.

In summary, loan company customer relationship management plays a crucial role in nurturing and maintaining strong relationships with customers, enhancing customer satisfaction, and driving business growth and profitability. It enables loan companies to leverage customer data, personalize interactions, streamline processes, and make data-driven decisions to effectively serve their customers in the competitive loan market.

2.7 Marketing trends and Performance Measurement

Loan market trends refer to the patterns and shifts in the loan industry that impact borrowing and lending activities. These trends are influenced by factors such as economic conditions, regulatory changes, customer preferences, and technological advancements. It is important for financial institutions and lenders to stay informed about these trends in order to adapt their strategies and remain competitive in the market.

Market performance measurements are crucial indicators that help assess the overall health and success of the loan market. These measurements provide insights into the efficiency, profitability, and risk levels of loan portfolios and lending activities. Common market performance measurements in the loan industry include (see [41]): a). Loan origination volume: This measurement tracks the total value of new loans originated within a specific period. It helps gauge the level of lending activity and market demand. b). Default rates: Default rates indicate the proportion of loans that borrowers fail to repay.

Monitoring default rates provides an understanding of credit quality and potential risks in the loan market. c). Net interest margin: This performance measurement reflects the profitability of lending activities. It represents the difference between the interest earned from loans and the interest paid on deposits or borrowings. d). Loan portfolio growth: This metric measures the rate of expansion in a lender's loan portfolio. It helps assess the institution's ability to attract borrowers and capture market share. e). Loan delinquency rates: Delinquency rates indicate the percentage of loans that are past due or in default. High delinquency rates suggest potential credit quality issues and collection challenges. f). Market share: Market share measurement tracks a lender's portion of the overall loan market. It provides insights into the institution's competitiveness and ability to attract and retain borrowers.

Loan market analysis is a comprehensive assessment of the loan market, involving the examination and interpretation of various factors that influence lending activities. It encompasses a wide range of data and information, including market trends, economic indicators, industry regulations, borrower profiles, and competitive landscape. The primary objective of loan market analysis is to gain insights into the current state of the market, identify potential risks and opportunities, and make informed decisions to optimize lending strategies.

The analysis begins with gathering relevant data, including loan volumes, interest rates, default rates, and borrower characteristics. This data is then analyzed using statistical techniques and financial models to identify patterns, correlations, and potential relationships. Market trends are identified by tracking changes in loan demand, loan types, and market conditions over time.

Additionally, loan market analysis involves examining external factors such as economic indicators (e.g., GDP growth, inflation rates) and regulatory changes that may impact lending practices. By understanding these influencing factors, lenders can adjust their strategies and risk management practices accordingly.

Furthermore, competitive analysis plays a crucial role in loan market analysis. It involves studying the strategies, products, and services offered by competing lenders to identify unique selling points, market positioning, and potential areas for differentiation.

Market performance measurements are also an integral part of loan market analysis. These measurements may include profitability metrics, efficiency ratios, and credit quality indicators. Profitability measurements assess the profitability of loans by analyzing interest income, net interest margin, and loan loss provisions. Efficiency ratios evaluate the effectiveness of loan origination and servicing processes, while credit quality indicators gauge the level of risk associated with borrowers and loan portfolios.

Conducting regular loan market analysis allows lenders to identify emerging trends, understand borrower behavior, assess creditworthiness, and make informed decisions regarding loan product development, pricing, and risk management. It provides valuable insights that help lenders stay competitive, manage risks effectively, and maximize profitability in the loan market.

2.8 Compliance and ethical markerting

Loan company compliance is a crucial aspect of ensuring ethical business practices and maintaining the trust and confidence of customers. It involves adhering to legal and regulatory requirements set forth by governing bodies and industry associations to safeguard the rights and interests of borrowers.

Compliance in the loan industry encompasses various aspects, including responsible lending practices, transparency in loan terms and conditions, fair collection practices, and protection against discriminatory practices. Loan companies need to establish robust internal controls, policies, and procedures to ensure compliance with these regulations. This may involve implementing systems to verify borrower information, assessing creditworthiness, and providing clear and accurate loan disclosures.

In addition to legal compliance, ethical marketing is essential for loan companies to build and maintain a positive reputation. Ethical marketing ensures that loan products and services are marketed in a fair, honest, and transparent manner. This includes providing accurate and complete information about interest rates, fees, repayment terms, and any potential risks associated with the loan.

Ethical marketing practices also involve treating customers with respect and ensuring that marketing messages do not exploit or mislead borrowers. Loan companies should avoid deceptive or predatory practices that may lead to financial harm for borrowers. Instead, they should focus on providing educational resources and guidance to help borrowers make informed decisions about their loans.

By prioritizing compliance and ethical marketing, loan companies can demonstrate their commitment to responsible lending and customer-centric practices. This not only protects borrowers but also helps build trust and loyalty among customers, leading to long-term success in the industry.

2.9 Methodology Review

Consumer credit risk assessment employs various methods to evaluate the creditworthiness of borrowers. One of the pioneering models in this field is the Z score discriminant analysis model introduced by Altman in 1968. However, logistic regression has now become the industry standard for credit risk assessment (Crook et al., 2007 [42]; Noh et al., 2005 [43]; Lessmann et al., 2015 [44]).

Bajari et al. (2008) ([45], [46]) developed a scoring model specifically for the US subprime market using a bivariate probit model. This model considers borrowers defaulting either due to a mortgage to equity ratio surpassing a certain threshold or due to inadequate income and limited access to other credit options, potentially leading to an improvement in their overall financial condition.

To enhance default prediction accuracy, researchers have explored various statistical and machine learning approaches, as well as alternative sets of predictor features [47]–[50]. Discriminant analysis, support vector machines [51], [52], artificial neural networks [53], [54], decision trees, genetic programming [55], [56] and standard models [57], [58] using external credit ratings have all been successfully applied (Arminger et al., 1997 [59]; Hand and Henley, 1997 [60]; Kruppa et al., 2013 [61]; Lessmann et al., 2015 [44]; Butaru et al., 2016 [62]; Baesens et al., 2003 ([63]); Abellán and Castellano, 2017 ([64]).

While there have been numerous advancements in credit risk modeling, fewer studies have focused on capturing default as a dynamic process, despite evidence suggesting that dynamic models can offer better results. Models that account for the dynamism in default, such as Markov chain-based approaches, have been shown to be conceptually more appropriate (Du Jardin and Séverin, 2011 [65]; Volkov et al., 2017 [66]). Grimshaw and Alexander (2011) ([15]) specifically modeled the transition matrix of loan movement between delinquent states as a Markov chain but did not forecast transition probabilities using loan-level covariates, instead utilizing a term structure of credit risk spreads (see also [67]).

In summary, the field of consumer credit risk assessment has seen advancements in various methodologies, including logistic regression, bivariate probit models, and machine learning approaches. While many models have focused on static credit risk, there is still potential for further exploration of dynamic models to better understand and predict default probabilities.

In addition to traditional approaches, advanced methods such as survival models have emerged as superior techniques for consumer credit risk assessment. These models have shown their effectiveness by incorporating time-varying covariates, including macroeconomic conditions that impact loan payment performance over time (Castro, 2013 [68]). Moreover, survival models have the capability to forecast event occurrences such as default, recovery, prepayment, and foreclosure at the next moment in time, considering that these events have not yet taken place (Bellotti and Crook, 2013 [69]; Chamboko and

Bravo, 2016 see [70], [71], [72]).

Survival models have been predominantly employed for modeling default risk (Bellotti and Crook, 2013 [69]; Noh et al., 2005 [43]; Sarlija et al., 2009 [73]; Tong et al., 2012 [74]; Chamboko and Bravo, 2019, see [75], [5]). Furthermore, these models have been utilized to model foreclosure in mortgage loans (Gerardi et al., 2007 [76]) as well as the transition from delinquency to a current loan status (Ha and Krishnan, 2012 [77]; Chamboko and Bravo, 2016 [78], 2019b [79]; Ha, 2010 [80]; Miguel, [81]).

To handle the dual risks of early payment and default on loan contracts, the competing risks survival framework has been employed in various studies (Deng et al., 1996 [82]; Stepanova and Thomas, 2002 [83]). It allows for a comprehensive understanding of the different potential outcomes and enables more accurate risk assessment.

Overall, the utilization of these advanced methods in loan market analysis enhances the ability to predict and manage credit risk more effectively, considering the dynamic nature of borrower behavior and external economic factors.

The option-based model of default has also been widely used in the USA (Kau et al. 1992 [17]; Deng et al. 2000 [18]) and UK markets (Ncube and Satchell 1994 [84]) to characterise mortgage default as ruthless or mainly influence by the relation of price of houses and value of mortgages. These models define default as an American option with the strike price equal to the value of the mortgage and then assume that a borrower will default as soon as the property value falls to or below the mortgage value (Kelly and O'Malley 2016 [85]), particularly when the lender has no recourse. A major limitation of these models is that defaults are usually defined the same way as foreclosures, thus ignoring additional important options for borrowers, for instance, cure or prepayment (see also [86]).

When conducting a methodology review for loan market expansion, it involves examining different research methodologies and approaches used to analyze and implement strategies for expanding a loan company's market. There are also many research on the following areas: a) Case Studies: Look for case studies that examine successful loan market expansion initiatives. Case studies provide detailed insights into real-world scenarios, outlining the strategies, challenges, and outcomes of loan companies that have successfully expanded their markets. Analyze the methodologies used in these studies to gain practical insights and identify best practices. b) Surveys and Interviews: Explore research that utilizes surveys or interviews to gather data from loan companies, borrowers, or industry experts. Assess the methodologies employed to design survey questionnaires, select participants, and analyze the collected data. This approach can provide valuable qualitative or quantitative insights into market expansion strategies, customer preferences, and industry trends. c) Data-driven Approaches ([87]): Examine research studies that employ market analysis techniques, such as market segmentation, trend analysis, and data-driven decision-making. Assess the methodologies used to collect and analyze market data, identify target markets, and develop data-backed strategies for loan market expansion. d) Comparative Analysis: Look for studies that conduct comparative analyses of loan companies that have expanded into different markets. These studies may compare factors such as market characteristics, regulatory environments, customer preferences, and competitive landscapes. Analyze the methodologies used to compare and evaluate the effectiveness of different market expansion approaches (see [88]). e) Simulation and Modeling: Explore research that utilizes simulation or modeling techniques to assess the potential outcomes of loan market expansion strategies. Such methodologies involve creating hypothetical scenarios or mathematical models to predict the impact of different factors on market expansion. Evaluate the methodologies used to develop and validate these models. f) Mixed Methods Approaches: Consider research that combines multiple methodologies, such as a combination of qualitative and quantitative approaches. These studies may employ surveys, interviews, data analysis, and other techniques to gain comprehensive insights into loan market expansion strategies. Assess the strengths and limitations of these mixed methods approaches. g) Action Research and Experiential Approaches: Look for studies that employ action research or experiential methodologies to explore loan market expansion. These methodologies involve actively engaging with loan companies or participating in market expansion initiatives to gather firsthand insights and experiences. Assess the methodologies used to document and analyze the outcomes of these experiential approaches.

2.10 Research gaps

As evidenced by the existing literature, prior studies predominantly focused on examining mortgage delinquency, default, or foreclosure as the primary outcomes of interest. However, these studies often lacked in-depth exploration or the inclusion of additional information over time. In addressing these limitations, our research project aims to employ machine learning methods to investigate a broader range of factors and provide more accurate foreclosure predictions.

By modeling the occurrence of multiple loan events simultaneously and considering their recurrences, our approach offers a more comprehensive and nuanced perspective compared to traditional approaches. An essential aspect of our research is identifying the relevant factors to incorporate into our modeling process. To achieve this, we extensively analyze published data and perform thorough correlation analyses to establish the relationships between various factors.

Another crucial aspect of our modeling process is determining appropriate coefficients. In Chapter 3, we present a novel approach using the variational inference method to obtain a line segment estimation for these coefficients. This technique enhances our ability to understand and interpret the impact of different variables on loan outcomes. In Chapter 5, we further evaluate the performance of various models, allowing for a comprehensive comparison.

Additionally, in Chapter 4, we delve into the development of loan collection decision models, building upon the insights gained from our previous findings. By integrating these models into practical decision-making processes, our research contributes to the development of effective loan collection strategies.

Overall, our research project addresses key questions related to factor selection, coefficient estimation, and practical implementation, paving the way for a more comprehensive understanding of loan dynamics and providing valuable insights for loan companies and industry practitioners.

3 STUDY ON CREDIT RISK CONTROL BY VARIATIONAL INFERENCE

Abstract

This paper presents the development of an intelligent, machine learning-based Markov chain model to investigate loan risk and strategies for controlling credit risk. The model involves modeling state transitions of loan accounts using a Markov transition matrix, and optimizing collection actions at each state and age for each consumer type to maximize the expected value for the lender. To enhance the performance on the minority class, we have designed some new algorithms and developed a consecutive incremental batch learning framework within the model. In addition, we use the variational inference method and logistic regression model to address the imbalance data gap during the traditional machine learning process ([89]). We expect that the results of this study will lead to a more accurate and effective machine learning-based prediction method and make a significant contribution to the credit risk field.

3.1 Background

Loan risk control is an essential aspect of lending and banking operations. Lenders face the risk of loss from borrower default or non-payment, which can result in significant financial losses. As such, managing loan risk is critical to the success and stability of lending institutions.

Historically, loan risk control has been a key concern for banks and other lending institutions. In the early days of banking, lenders primarily relied on personal relationships and trust to manage loan risk. However, as lending institutions grew and loan portfolios became more complex, lenders began to develop more sophisticated methods for managing loan risk.

The development of credit scoring in the mid-20th century revolutionized loan risk control by enabling lenders to evaluate the creditworthiness of borrowers more objectively. Credit scoring models used statistical algorithms to analyze a borrower's credit history and other factors to generate a numerical score that indicated the likelihood of the borrower repaying the loan. This approach made loan risk management more efficient and enabled lenders to make better-informed lending decisions.

Over time, lenders developed additional strategies for managing loan risk, including loan collateral, loan covenants, diversification, and risk-based pricing. These strategies enabled lenders to further reduce their credit risk exposure and manage their loan portfolios more effectively.

In recent years, advances in technology have further transformed loan risk control. Lenders can now use data analytics and machine learning algorithms to analyze vast amounts of data and identify patterns that can indicate credit risk. These technologies have made loan risk control even more efficient and effective. Today, lenders rely on a combination of historical approaches and cutting-edge technologies to manage loan risk and ensure the stability and success of their lending operations.

In this paper, we aim to investigate loan risk and control credit risk using a machine learning approach based on the Markov chain model. Specifically, we will utilize both Variational Bayes and logistic regression methods to develop a more robust and precise model.

To begin, let's provide an overview of the Markov chain model. This model involves estimating the transition matrix, which can be achieved through machine learning methods such as logistic regression, Bayesian estimation, or EB estimation. These techniques can help us develop a more accurate understanding of loan risk and enable us to implement effective risk management strategies.

3.2 Markov Chain model

Let $\{X_n\}$ denote a Markov chain where X_n is the delinquency state of a loan in month n. Let $\pi(n)$ denote the unconditional probability distribution of a loan in month n, and is a vector whose entries correspond to the different Markov chain delinquency states. If the delinquency state for month n is known, then $\pi(n)$ is a row vector with a one indicating this month's delinquency state for loan i and zeros elsewhere. The transition matrix moving from month n to month $n+1$ of the Markov chain is denoted by $P(n, n+1)$, a matrix containing the probabilities of movement between delinquency states ([90]). If the transition matrix is known, a forecast of the delinquency state probability distribution for next month can be formed given the previous month's delinquency state probability distribution. That is, for loan i, the delinquency state probability distribution of month $n+1$ is computed from $\pi_i(n+1) = \pi_i(n)P_i(n, n+1)$ if $\pi_i(n)$ and $P_i(n, n+1)$ are known. Associated with loan i is an outstanding balance at month n denoted by $w_i(n)$. The 'one month ahead' forecast outstanding balance by delinquency state of loan i is the vector

$$w_i(n+1) = w_i(n) \cdot \pi_i(n+1).$$

Markov Chain related research work can be found in many fields [91]–[93].

3.3 Machine learning method

Machine learning (ML) is a branch of artificial intelligence (AI), containing multiple data analytical algorithms, such as linear regression, logistic regression, neural networks, support vector machine and decision tree [94]–[96]. ML algorithms can identify patterns and build mathematical decision-making models by automatically learning from data [97]–[100].

A typical machine learning algorithm can be formulated as the following equation

$$\hat{y}^{(t)} = f_{\Theta}(x) \tag{1}$$

where $x \in R^n$ is a n-dimensional input feature vector extracted form raw data; $f_{\Theta}(\cdot)$ is a mathematical mapping function from input to output, which is decided by the specific machine learning algorithm; Θ is the parameters of function $f_{\Theta}(\cdot)$, which could be learnt from data; $\hat{y}$ is the predictive output of the machine learning model [96], [101], [102].

To describe the learning (training) process of a supervised machine learning model, denote a labelled dataset as $D = \{X, Y\}$, where $X = [x_1, x_2, \cdots, x_m]$ is the input matrix containing m input vectors and $x \in \mathbb{R}^{m \times n}$; $Y = [y_1, y_2, \cdots, y_m]$ is the corresponding output vectors, and $Y \in \mathbb{R}^{m \times 1}$; $y_i(i = 1, 2, \cdots, m)$ is also called the ground truth label of input x_i $(i = 1, 2, \cdots, m)$. If y_i is a real number, the machine learning model is a regression model while if y_i is a discrete value, the machine learning model is a classification model. Specifically, if $y \in \{0, 1\}$, the machine learning model is a binary classification model [103]–[105].

Taking the classification model (classifier) for example, to train a classifier $f_{\Theta}(\cdot)$ is to find out the optimized value of Θ. To learn the parameters from a given a labelled dataset as $D = \{X, Y\}$, a cost function used to measure the differences between the output of the model and the ground truth needs to be defined. Equation (2-2) is the widely used cross-entropy cost function.

$$L_{(X,Y)}(\Theta) = -\frac{1}{m}\sum_{j=1}^{m}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{2}$$

Then, the classifier training problem turns into a minimization problem, $\min L_{(X,Y)}(\Theta)$. Algorithms, like gradient descent, can be used to solve this optimal problem. ML algorithms have achieved great progress in some traditional areas like information retrieval, machine translation, automatic speech recognition and computer vision. Machine learning algorithms learn latent data distribution and patterns from labelled data.

3.4 Research gaps of Machine learning based access control

Coefficients limitation problem The logistic regression model is commonly employed to address various problems, and it provides a solid foundation for modeling. However, the traditional approach of estimating model coefficients may have limitations in terms of accuracy. To address this concern, we propose the use of the inference variational method in our paper. This method bridges the gap by offering a more robust alternative. By employing this approach, we can now assign a segment or range to each coefficient, allowing for greater flexibility and providing valuable insights in practical applications.

By incorporating the inference variational method, our model gains the ability to capture the inherent uncertainty in the coefficient estimates. This not only enhances the accuracy of our predictions but also provides a more comprehensive understanding of the underlying data.

In summary, our paper introduces the inference variational method as a means to improve logistic regression models. By providing segments for each coefficient, we enhance the model's flexibility and offer valuable insights, resulting in more accurate and nuanced analyses.

There are another two challenges when applying ML algorithms to estimate the transition matrix in the Markov chain models. The first one is concept drift problem, and another is data imbalance problem.

Concept drift problem Concept drift refers to the phenomenon where the statistical distributions of the data that a machine learning (ML) algorithm attempts to describe and predict change over time in an arbitrary manner. It is often observed in data streams, which are sequences of data organized chronologically. Let's denote a data stream as $S = \{d_1, d_2, \cdots, d_t, d_{t+1}, \cdots\}$, where $d_i = \{x_i, y_i\}$ is a labelled sample observed at time step i. If the relationship between $S_{(0,t)} = \{d_1, d_2, \cdots, d_t\}$ follows a certain function $f_{\Theta(0,t)}(\cdot)$, and $f_{\Theta(0,t)}(\cdot) \neq f_{\Theta(t+1,t+a)}(\cdot)$, where a is an arbitrary positive number, then the concept drift occurs at time step $t + 1$. [106].

If concept drift exists, traditional batch learning method for ML algorithms will cause performance to decrease as time passes. To deal with concept drift, there are two ways. One is called lazy strategy, which means, the ML model's parameter Θ will not be updated until a concept drift was detected. Obviously, the performance of lazy strategies subjects to the accuracy of drift detection. To avoid the inaction on concept drift caused detection failure, researchers proposed active strategies to cope with concept drift, which means ML models keep updating themselves with concept drift adaption algorithms once new labelled data is available. Active strategies are essentially online learning methods, compared with batching learning methods.

To address the concept drift problem in the context of a Markov chain model, a common approach is to divide a portfolio of loans into segments. The objective of segmenting the loans is to group together those that share similarities and are expected to exhibit the same transition matrix. This segmentation can be based on various factors, such as characteristics of the financial product or insights gained from data mining (see also [107], [108], [109], [110]).

For instance, consider the differentiation between fixed-rate loans and adjustable rate mortgage (ARM) loans. These two types of loans differ in several ways, including interest rate structures, payment schedules, and potential risks. Given these differences, it is reasonable to assume that fixed-rate loans and ARM loans will have distinct transition matrices when modeling their loan performance over time.

By segmenting the loans based on such characteristics, the Markov chain model can account for the unique dynamics within each segment. This segmentation enables the model to capture and adapt to the varying transition patterns exhibited by different types of loans, thereby improving its ability to address concept drift effectively.

In practice, the determination of loan segments and their associated transition matrices can be informed by domain expertise, empirical analysis, historical data, or a combination of these factors. The goal is to create segments that provide meaningful insights into the distinct behavior and transitions within the loan portfolio, ultimately enhancing the accuracy and predictive power of the Markov chain model in the presence of concept drift.

Data imbalance problem In addition to concept drift, another significant challenge that can lead to a severe decrease in performance for machine learning (ML) algorithms is data imbalance. Data imbalance occurs when the number of samples belonging to each class label in a labeled dataset is not balanced or equivalent. This imbalance or skewness in the dataset can have a detrimental impact on the performance of predictive ML models.

When dealing with a classification problem, having an imbalanced dataset can introduce biases and distortions in the learning process. ML algorithms tend to prioritize the majority class, resulting in exaggerated performance on the majority class while the performance on the minority class is often unsatisfactory ([111], [112]).

To address data imbalance, it is crucial to employ appropriate ML algorithms and ensure the availability of a balanced training dataset. There are various techniques and strategies to mitigate data imbalance, such as resampling methods (e.g., oversampling the minority class or undersampling the majority class), synthetic data generation, or employing specialized algorithms designed for imbalanced data [113]–[115].

When it comes to studying loan states or predicting loan behavior, the available data often suggests that loans are more likely to remain current as they age. This characteristic, known as non-stationarity, can be effectively modeled by incorporating covariates such as the "number of months since last delinquency" or "number of months since origination." Additionally, other covariates that capture differences between loans over the repayment period, such as credit quality, repayment history, and loan age, can contribute to a more comprehensive understanding of loan dynamics and improve the accuracy of predictive models in the loan domain.

By considering and incorporating relevant covariates, ML models can better capture the underlying patterns and dynamics in loan repayment behavior, thereby enabling more accurate predictions and informed decision-making.

3.5 Variational Bayes for loan-level classification modeling

3.5.1 Refining Markov Model Estimation with Loan-Level Models

As briefly discussed in the introductory chapter, it is reasonable to incorporate covariates that may influence specific transition probabilities for a loan throughout its repayment period. For instance, loan-level models can be employed to estimate crucial transaction probabilities, such as transitioning from 'Current' to 'DPD30' or from 'DPD30' to 'DPD60'.

Binary classification models, including logistic and probit regression, are among the most prevalent loan-level models in credit risk and loss forecasting. Diggle *et al.* [116] outline a general methodology for modeling longitudinal categorical data as a Markov chain, where the transition matrices $P(n, n+1)$ feature multinomial logistic models for each row, using n as a covariate. Smith and Lawrence [117] model all transition probabilities, however, as previously mentioned, this approach is not essential for modeling loan repayment, as only a few key transition probabilities exhibit non-constant probabilities. Moreover, modeling all transition probabilities may be unwarranted in cases where there is insufficient data to construct a predictive model, such as when loans with a 'DPD120+' status in the current month transition to 'Current' the following month. In this paper, we outline the construction of loan-level models, incorporating essential explanatory variables within a Bayesian inference framework (see [118]), to estimate transaction probabilities and integrate them into the transaction matrix.

3.5.2 Addressing Model Output Uncertainty with Variational Bayes Classification Models

Bayesian logistic or probit regression offers the advantage of providing a posterior distribution, as opposed to a single point estimate or a confidence interval found in the classical or frequentist approach (see [119], [120]). By integrating prior beliefs, we can quantify uncertainty surrounding point estimates of transaction probabilities. This enables us to gain insights into the potential ranges of outstanding balances for each delinquency state on a monthly basis.

Variational Bayes is a widely used approach for approximate Bayesian inference. However, straightforward methods are typically limited to specific model classes, particularly those with conditionally conjugate structures within an exponential family (see [118], [121], [39]). Instead of using Gaussian priors, which are a common and popular choice in variational Bayes approximation, Li *et al.* [34] integrate the intrinsic priors [122] introduced by Berger and Pericchi into Variational Bayes to construct probit regression models.

However, models with logit components seem to be a significant exception to this class, primarily due to the lack of conjugacy between the logistic likelihood and the Gaussian priors for the coefficients in the linear predictor (see [123]). To enable approximate inference within this widely used class of models, Jaakkola and Jordan [124] proposed a straightforward variational approach that leverages a family of tangent quadratic lower bounds for the logistic log-likelihood, thereby reestablishing conjugacy between these approximate bounds and the Gaussian priors. Durante and Rigon [125] offer insights into the effectiveness of this strategy by presenting a formal connection between the aforementioned bound and the Pólya-gamma data augmentation [126] for logistic regression (see also [127]).

3.5.3 Background on Variational Bayes Methods

Variational methods have their origins in the 18^{th} century with the work of Euler, Lagrange, and others on the *calculus of variations* ([128], [129]). [1] *Variational inference* is a body of deterministic techniques for making approximate inference for parameters in complex statistical models. Variational approximations are a much faster alternative to

[1] The derivation in this section is standard in the literature on variational approximation and will at times follow the arguments in Bishop (2006) and Jordan *et al.* (1999).

Markov Chain Monte Carlo (MCMC), especially for large models, and are a richer class of methods than the Laplace approximation, see [130].

Variational inference is a probabilistic modeling technique used to estimate the posterior distribution of hidden variables and model parameters. It is commonly employed in Bayesian inference and probabilistic programming to estimate the posterior distribution of complex models, especially when dealing with high-dimensional data.

The core idea behind variational inference is to reformulate the problem of posterior distribution estimation as an optimization problem. It assumes a simplified posterior distribution, often a parameterized distribution like a Gaussian distribution, to approximate the true posterior distribution. The goal is to adjust the distribution's parameters to make this approximate distribution as close as possible to the true posterior distribution. This process typically involves maximizing a function known as the variational lower bound (ELBO), which can be mathematically expressed as:

$$\text{ELBO} = \mathbf{E}[\log p(\text{data, latent})] - \mathbf{E}[\log q(\text{latent})]$$

Here: $\mathbf{E}$ represents the expectation operation, "data" stands for the observed data. "latent" represents the hidden variables. p represents the true posterior distribution. q represents the approximate posterior distribution. By maximizing the ELBO, variational inference aims to make the approximate posterior distribution q as close as possible to the true posterior distribution p. This maximization process often involves optimization algorithms like gradient descent.

In summary, variational inference is a method for estimating posterior distributions in complex probabilistic models. It simplifies the problem by finding an approximate distribution and then uses optimization to approximate the true posterior distribution. This approach is widely used in machine learning and statistical modeling.

3.5.4 Obtaining a tractable approximation for the posterior

Suppose we have a Bayesian model and a prior distribution for the parameters. The model may also have latent variables, here we shall denote the set of all latent variables and parameters by $\boldsymbol{\theta}$. And we denote the set of all observed variables by $\mathbf{X}$. Given a set of n independent, identically distributed data, for which $\mathbf{X} = \{\mathrm{x}_1, ..., \mathrm{x}_n\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n\}$, our probabilistic model (e.g. probit regression model) specifies the joint distribution $p(\mathbf{X}, \boldsymbol{\theta})$, and our goal is to find an approximation for the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ as well as for the marginal likelihood $p(\mathbf{X})$. For any probability distribution $q(\boldsymbol{\theta})$, we have the following decomposition of the log marginal likelihood (see also [131])

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \text{KL}(q||p)$$

where we have defined

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \ln \Big\{ \frac{p(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \Big\} d\boldsymbol{\theta} \tag{3}$$

$$\text{KL}(q||p) = -\int q(\boldsymbol{\theta}) \ln \Big\{ \frac{p(\boldsymbol{\theta}|\mathbf{X})}{q(\boldsymbol{\theta})} \Big\} d\boldsymbol{\theta} \tag{4}$$

We refer to 3 as the lower bound of the log marginal likelihood with respect to the density q, and 4 is by definition the Kullback-Leibler divergence of the posterior $q(\boldsymbol{\theta}|\mathbf{X})$ from the density q. Based on this decomposition, we can maximize the lower bound $\mathcal{L}(q)$ by optimization with respect to the distribution $q(\boldsymbol{\theta})$, which is equivalent to minimizing the KL divergence. And the lower bound is attained when the KL divergence is zero, which happens when $q(\boldsymbol{\theta})$ equals the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$. It would be hard to find such a density since the true posterior distribution is intractable.

3.5.5 Factorized distributions

The essence of the variational inference approach is approximation to the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ by $q(\boldsymbol{\theta})$ for which the q dependent lower bound $\mathcal{L}(q)$ is more tractable than the original model evidence. And tractability is achieved by restricting q to a more manageable class of distributions, and then maximizing $\mathcal{L}(q)$ over that class.

Suppose we partition elements of $\boldsymbol{\theta}$ into disjoint groups $\{\boldsymbol{\theta}_i\}$ where $i = 1, ..., M$. We then assume that the q density factorizes with respect to this partition, i.e.,

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i). \tag{5}$$

The product form is the only assumption we made about the distribution. Restriction (5) is also known as *mean field* approximation and has its root in Physics. [132]

For all distributions $q(\boldsymbol{\theta})$ having the form (5), we need to find the distribution for which the lower bound $\mathcal{L}(q)$ is largest. Restriction of q to a subclass of product densities like (5) gives rise to explicit solutions for each product component in terms of the others. This fact, in turn, leads to an iterative scheme for obtaining the solutions. To achieve this, we first substitute (5) into (3) and then separate out the dependence on one of the factors $q_j(\boldsymbol{\theta}_j)$. Denoting $q_j(\boldsymbol{\theta}_j)$ by q_j to keep the notation clear, we obtain

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_{i=1}^{M} q_i \Big\{ \ln p(\mathbf{X}, \boldsymbol{\theta}) - \sum_{i=1}^{M} \ln q_i \Big\} d\boldsymbol{\theta} \\ &= \int q_j \Big\{ \int \ln p(\mathbf{X}, \boldsymbol{\theta}) \prod_{i \neq j} q_i d\boldsymbol{\theta}_i \Big\} d\boldsymbol{\theta}_j - \int q_j \ln q_j d\boldsymbol{\theta}_j + \text{constant} \\ &= \int q_j \ln \tilde{p}(\mathbf{X}, \boldsymbol{\theta}_j) d\boldsymbol{\theta}_j - \int q_j \ln q_j d\boldsymbol{\theta}_j + \text{constant}\end{aligned} \tag{6}$$

where $\tilde{p}(\mathbf{X}, \boldsymbol{\theta}_j)$ is given by

$$\ln \tilde{p}(\mathbf{X}, \boldsymbol{\theta}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \boldsymbol{\theta})] + \text{constant}. \tag{7}$$

The notation $\mathbb{E}_{i \neq j}[\cdot]$ denotes an expectation with respect to the q distributions over all variables $\mathbf{z}_i$ for $i \neq j$, so that

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \boldsymbol{\theta})] = \int \ln p(\mathbf{X}, \boldsymbol{\theta}) \prod_{i \neq j} q_i d\boldsymbol{\theta}_i. \tag{8}$$

Now suppose we keep the $\{q_{i \neq j}\}$ fixed and maximize $\mathcal{L}(q)$ in (6) with respect to all possible forms for the density $q_j(\boldsymbol{\theta}_j)$. By recognizing that (6) is the negative KL divergence between $\tilde{p}(\mathbf{X}, \boldsymbol{\theta}_j)$ and $q_j(\boldsymbol{\theta}_j)$, we notice that maximizing (6) is equivalent to minimize the KL divergence, and the minimum occurs when $q_j(\boldsymbol{\theta}_j) = \tilde{p}(\mathbf{X}, \boldsymbol{\theta}_j)$. The optimal $q_j^*(\boldsymbol{\theta}_j)$ is then

$$\ln q_j^*(\boldsymbol{\theta}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \boldsymbol{\theta})] + \text{constant}. \tag{9}$$

The above solution says that the log of the optimal q_j is obtained simply by considering the log of the joint distribution of all parameter, latent and observable variables and then taking the expectation with respect to all the other factors q_i for $i \neq j$. Normalizing the exponential of (9), we have

$$q_j^*(\boldsymbol{\theta}_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \boldsymbol{\theta})])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \boldsymbol{\theta})]) d\boldsymbol{\theta}_j}. \tag{10}$$

The set of equations in (9) for $j = 1, ..., M$ are not an explicit solution because the expression on the right-hand side of (9) for the optimal q_j^* depends on expectations taken with respect to the other factors q_i for $i \neq j$. We will need to first initialize all of the factors

$q_i(\boldsymbol{\theta}_i)$ and then cycle through the factors one by one and replace each in turn with an updated estimate given by the right-hand side of (9) evaluated using the current estimates for all of the other factors. Convexity properties can be used to show that convergence to at least local optima is guaranteed. [133] The iterative procedure is described in Algorithm 1.

Algorithm 1 Iterative procedure for obtaining the optimal densities under factorized density restriction 5. The updates are based on the solutions given by 9.

```
Initialize $q_2^*(\boldsymbol{\theta}_2), ..., q_M^*(\boldsymbol{\theta}_M)$. Cycle through
    $q_1^*(\boldsymbol{\theta}_1) \leftarrow \frac{\exp(\mathbb{E}_{i\neq 1}[\ln p(\mathbf{X}, \boldsymbol{\theta})])}{\int \exp(\mathbb{E}_{i\neq 1}[\ln p(\mathbf{X}, \boldsymbol{\theta})])d\boldsymbol{\theta}_1}$
    $\vdots$
    $q_M^*(\boldsymbol{\theta}_M) \leftarrow \frac{\exp(\mathbb{E}_{i\neq M}[\ln p(\mathbf{X}, \boldsymbol{\theta})])}{\int \exp(\mathbb{E}_{i\neq M}[\ln p(\mathbf{X}, \boldsymbol{\theta})])d\boldsymbol{\theta}_M}$
until the increase in $\mathcal{L}(q)$ is negligible.
```

3.6 Predictive Modeling of Charge-off and Repayment Probabilities Using LendingClub Data

3.6.1 Introduction

LendingClub, a pioneering enterprise in the financial technology sector, holds the title as the world's largest peer-to-peer lending platform. Founded with the mission of transforming the banking system to make credit more affordable and investing more rewarding, LendingClub has managed to connect thousands of borrowers with individual investors. This unique business model not only democratizes access to loans but also opens up a new avenue for investors to diversify their portfolios.

One of the many attractions of LendingClub is its flexibility. The platform empowers borrowers to create unsecured personal loans ranging from $\$1,000$ to $\$40,000$. Such wide-ranging amounts mean that LendingClub caters to a variety of financial needs, whether it's a small, short-term cash crunch or a larger expense such as home renovation or debt consolidation. Borrowers have the option to repay the loans over standard periods of either three or five years, allowing them the flexibility to manage their debt according to their financial circumstances.

LendingClub's platform is not just borrower-centric, it also provides numerous benefits to investors. Investors have the ability to search and browse through the loan listings on the LendingClub website and cherry-pick the loans they wish to invest in. This is based on a wealth of information supplied about the borrower, such as the amount of loan requested, the loan grade which is a measure of risk associated with the loan, and the purpose of the loan. By making informed decisions, investors have the potential to earn returns through the interest paid by borrowers on these loans.

Crucially, LendingClub operates on a transparent business model. While it makes money by charging borrowers an origination fee for processing a new loan, and investors a service fee for loan servicing, it does so openly and transparently. Transparency is not only limited to its revenue model but also extends to the information available on its platform. To attract lenders (see [134]), LendingClub publishes most of the data from borrowers' credit reports as well as other information reported by borrowers for almost every loan issued through its website. This transparency helps investors make informed decisions, underlining LendingClub's commitment to creating a more open and fair financial marketplace.

3.6.2 Predictive Modeling of Charge-off and Fully-paid Probabilities - Target Variable and Sampling

LendingClub's publicly available data from the year 2016 provides a rich source of information for building predictive models. This comprehensive dataset consists of a total of $433,042$ issued loans. Every loan in the dataset comes with a specific status: **Fully-paid**, **Charged-off**, or **Current**. These statuses carry crucial information about the loan repayment, and hence, serve as the target variable for our predictive models. In particular, we will construct models that can predict the probability of a loan being charged-off or fully paid.

The above mentioned models will employ a variety of loan-related features as covariates. These covariates, or input features, have been chosen for their potential relevance to loan repayment outcomes. Among these are:

- Loan Term in Months: The duration of the loan term, whether it's 36 or 60 months, could influence a borrower's ability to fully repay the loan.

- FICO Score: This is a credit score developed by FICO, a company that specializes in predictive analytics. It's widely used by lenders to assess a borrower's credit risk.
- Issued Loan Amount: The size of the loan granted can also impact the borrower's repayment capacity.
- Debt-to-Income Ratio (DTI): This ratio, calculated by dividing a customer's total debt by their income, gives an indication of the borrower's ability to manage and repay debts.
- Number of Credit Lines Opened in the Past 24 Months: This feature provides a snapshot of the borrower's recent credit behavior, which could have an impact on their repayment capabilities.
- Employment Length in Years: The duration of a borrower's employment could reflect their financial stability, and hence, their ability to repay the loan.
- Annual Income: A borrower's annual income serves as an indicator of their financial capacity to fulfill their loan obligations.
- Home Ownership Type: Whether the borrower owns their home, is paying a mortgage, or rents can also have implications for their financial stability and ability to repay the loan.

By including these diverse and informative features as covariates in our predictive models, we aim to maximize the predictive power of our models. These models will provide valuable insights into the factors that influence the charge-off and full-paid rates of loans issued by LendingClub.

3.6.3 Addressing Uncertainty of Estimated Loan-level Models using Variational Inference

3.6.4 Specifying the priors

To complete the two Bayesian logistic regression models that predict the likelihood of a loan being Charged-off or Fully-paid, it's imperative to set the prior models for our regression parameters, β_i. Since these parameters can take any value in the real line, Gaussian priors are appropriate choices.

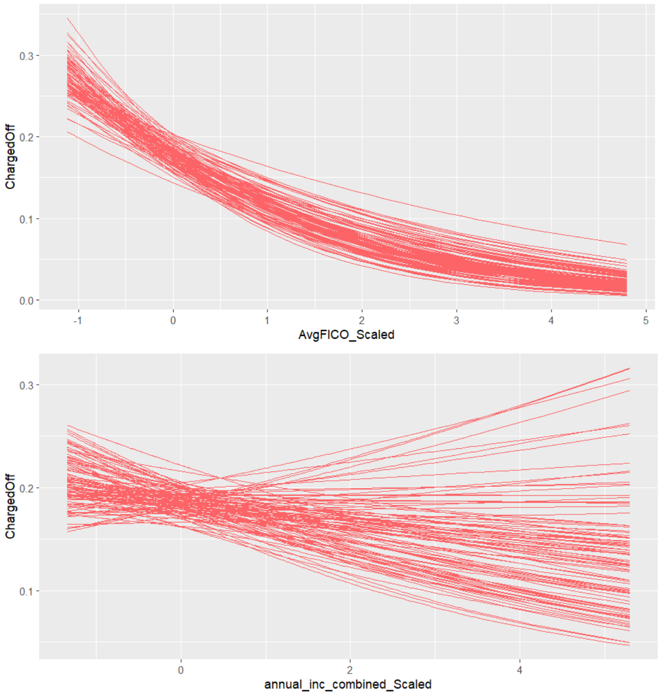


Figure 1: Charge-off Model Priors

Establishing the priors for the intercept also requires careful consideration. To this end, we've relied on past data, observing that in an average year, or "vintage", of loans, the chance of a loan being charged off is approximately 20%, while the probability of a loan being paid off in full is around 75%. These values serve as our base rates for the prior models, shaping our initial beliefs about the likelihood of a loan falling into either of these categories.

To visually encapsulate our initial understanding and the extent of uncertainty around these variables, we've plotted 100 instances of plausible relationships for a selection of the input features. These can be observed in Figures 17 and 2 (for charge-off); Figures 3 and 4 (for full-payment).

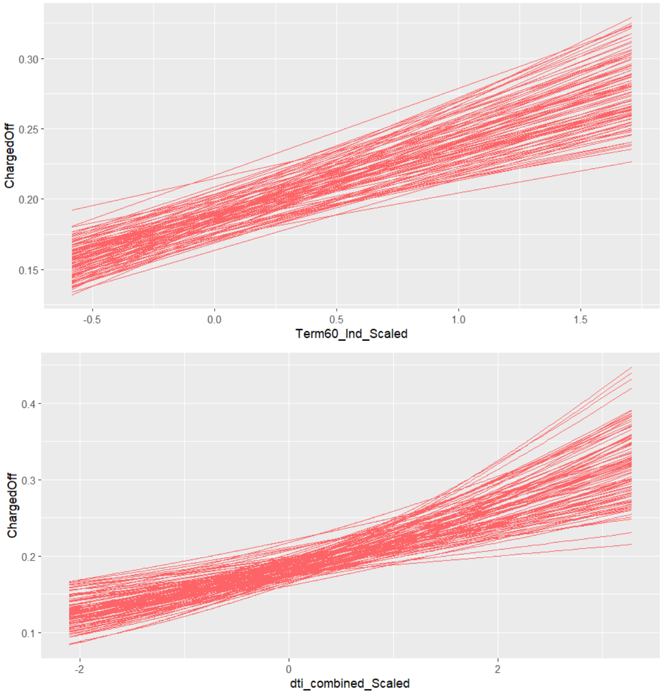


Figure 2: Charge-off Model Priors

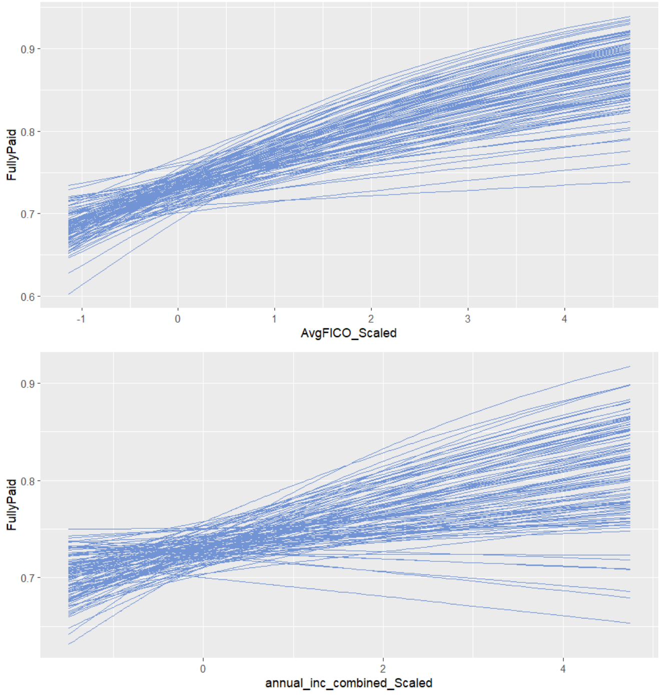


Figure 3: Paid-off Model Priors

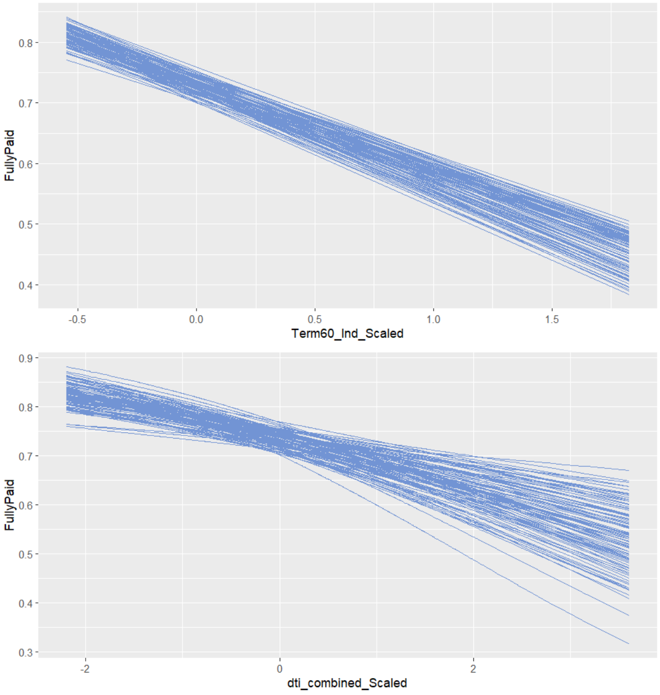


Figure 4: Paid-off Model Priors

Our plots echo our initial assumptions about these relationships:

- 1. The probability of a loan being charged-off seems to rise with both the loan term and the Debt-to-Income (DTI) ratio. This is likely because longer loan terms and higher debt compared to income could indicate greater financial strain and risk.

- 2. Conversely, we find the likelihood of a loan being charged-off decreases with an increase in the FICO score and the borrower's annual income. Higher FICO scores and annual incomes suggest better creditworthiness and financial stability, reducing the risk of default.

- 3. In contrast, the likelihood of a loan being paid off in full tends to increase with higher FICO scores and annual income. Again, these factors indicate a better financial standing, making full repayment more likely.

- 4. However, the probability of full payment seems to decrease as the loan term and DTI ratio increase, mirroring the trends observed for charge-offs but in the opposite direction.

- 5. Additionally, these plots encapsulate our initial uncertainty about the rate of these increases and decreases.

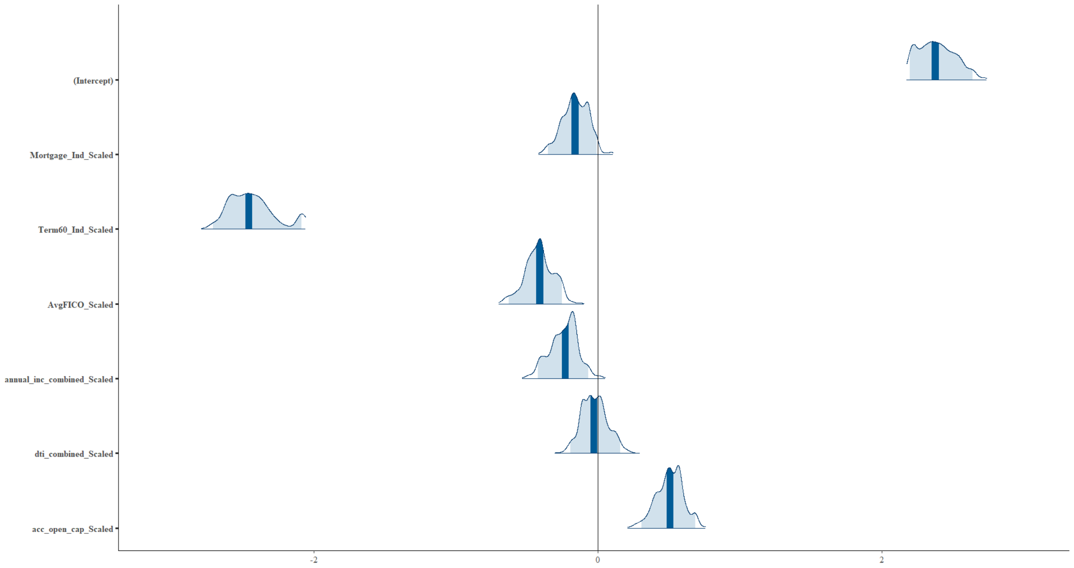


Figure 5: Posterior Distribution for the Charge-off Model Parameters

3.6.5 Simulating the posterior

In constructing our Bayesian logistic regression models, a crucial step involves updating our prior models based on actual observed data. This step allows us to simulate the posterior models for our parameters, essentially refining our initial assumptions with the lens of empirical evidence. The posterior distribution encapsulates the adjusted uncertainty pertaining to the unknown parameter values after factoring in the observed data.

To illustrate this process, we provide visual representations in Figures 5 and 6. These figures delineate the posterior distribution for the parameters of our model. Further, we have plotted 100 posterior plausible models in Figures 7 and 8. The left-hand side of this figure showcases the posterior relationship between the rate of charge-offs and annual income, while the right-hand side represents the relationship between the rate of fully paid loans and the Debt-to-Income (DTI) ratio.

A difference between our prior and posterior models can be seen when comparing these figures with those in Figures 17, 2, 3, and 4. It is evident that the posterior models, having been updated with the observed loan-level data, exhibit much less variability than their prior counterparts. This implies that the integration of actual data into the model has provided us with a higher level of certainty about the relationship between the probabilities of loans being charged-off or paid-off and the input features.

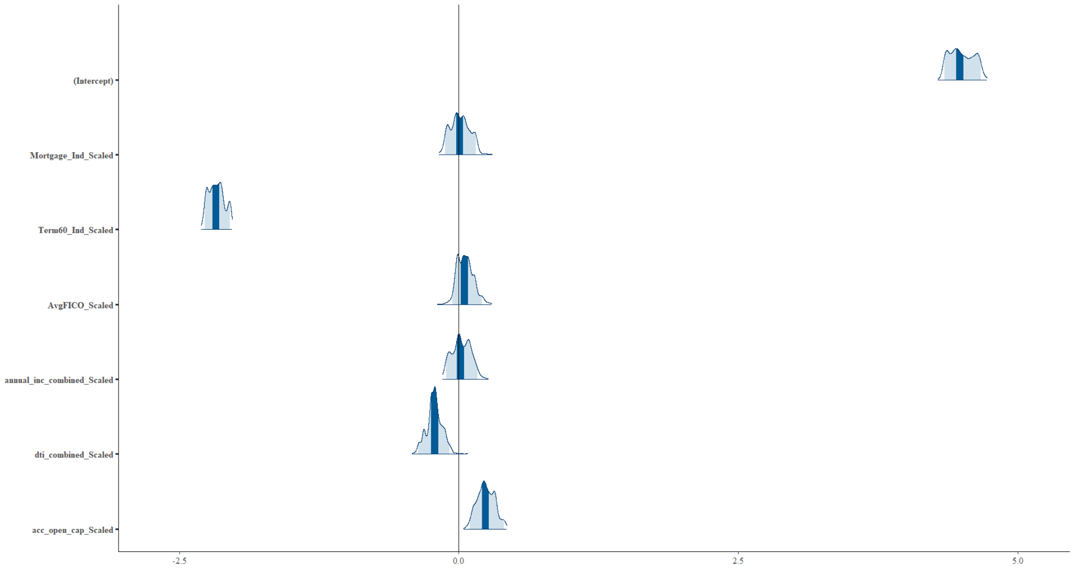


Figure 6: Posterior Distribution for the Paid-off Model Parameters

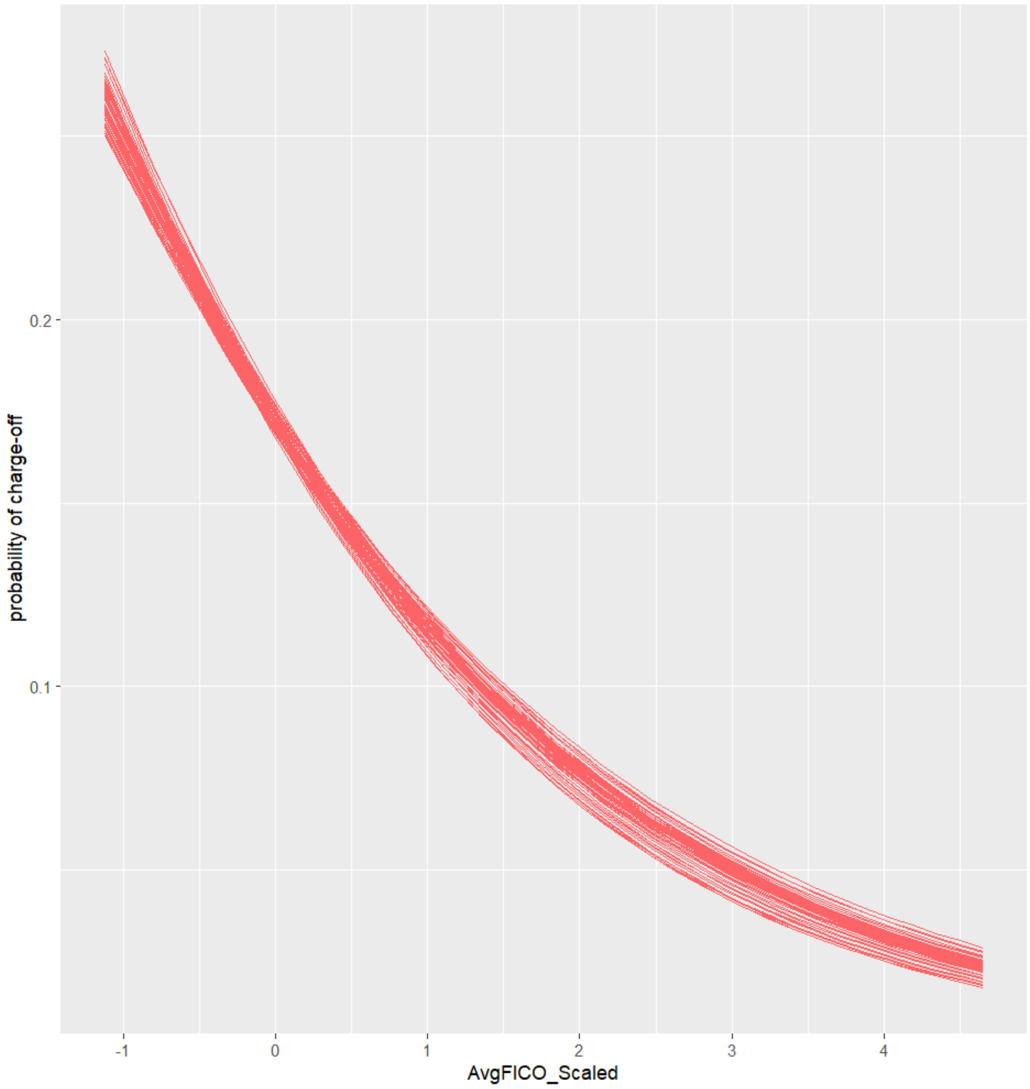


Figure 7: Posterior plausible models for the probability of charged-off/fully-paid versus input features

In essence, by merging our initial understanding with observed data, we are able to effectively narrow down the plausible range of our model parameters, enhancing the precision of our predictions. This, in turn, underscores the value of leveraging Bayesian models in predictive analytics, as they allow us to continually refine our estimates as new data becomes available.

Model Evaluation In order to evaluate the performance and validity of our Bayesian logistic regression models, we employ a method known as a posterior predictive check. This process essentially involves creating simulated data from our posterior models and then comparing these simulations to the original data. The goal is to confirm that the simulations retain similar features to the original data, which would then suggest that the

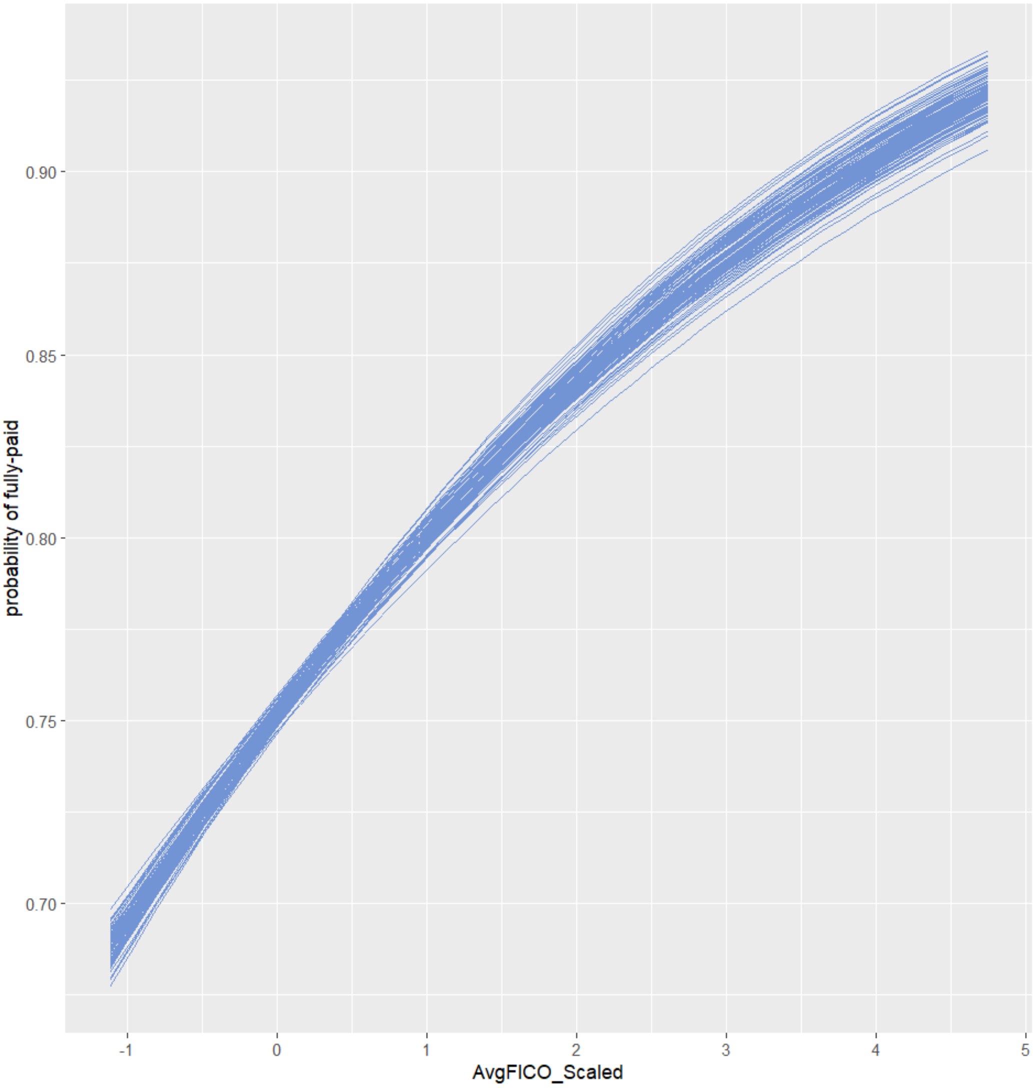


Figure 8: Posterior plausible models for the probability of charged-off/fully-paid versus input features

underlying assumptions of our Bayesian logistic regression models are indeed reasonable and sound.

In our analysis, we generate 100 posterior simulated datasets. From each dataset, we record the proportion of outcome variables $\boldsymbol{Y}$ that equal 1. In the context of our study, this corresponds to the proportion of loans that are either charged-off or fully paid.

This posterior predictive checking process is visualized in Figure 9, which presents a histogram of the simulated charge-off and paid-off rates. A careful examination of these histograms reveals a reassuring consistency between our simulations and the original data. For instance, we observe that most of our posterior simulated datasets exhibit a charge-off rate of approximately 18% and a paid-off rate around 75%. These rates are close to those seen in the original loan-level data, indicating that our models are accurately capturing the key trends in loan outcomes. With a conjugate prior and following the Gibbs sampling scheme proposed by [87], it took 89.86 seconds to finish 100 simulations for the Gibbs sampler

The histogram also features a vertical line, which represents the observed proportions of charged-off and fully-paid loans in the actual data. The closeness of this line to the peaks of the histograms further illustrates the strong alignment between our posterior simulations and the original data, providing further validation of our models.

In summary, this posterior predictive check serves as an effective tool to confirm the soundness of our model assumptions and the robustness of the Bayesian logistic regression models in capturing the patterns of the LendingClub data. It is through such rigorous evaluations that we can be confident in the reliability and predictive power of our models.

3.7 Conclusion

In this paper, we have proposed an innovative, comprehensive approach to predicting the movement of loans across different delinquency states. To address this complex task, we employed a Markov chain model which allowed us to characterize the transitions between different states of delinquency. By capturing these transitions, we gained deeper insight into the life-cycle of loans and the factors that influence their evolution.

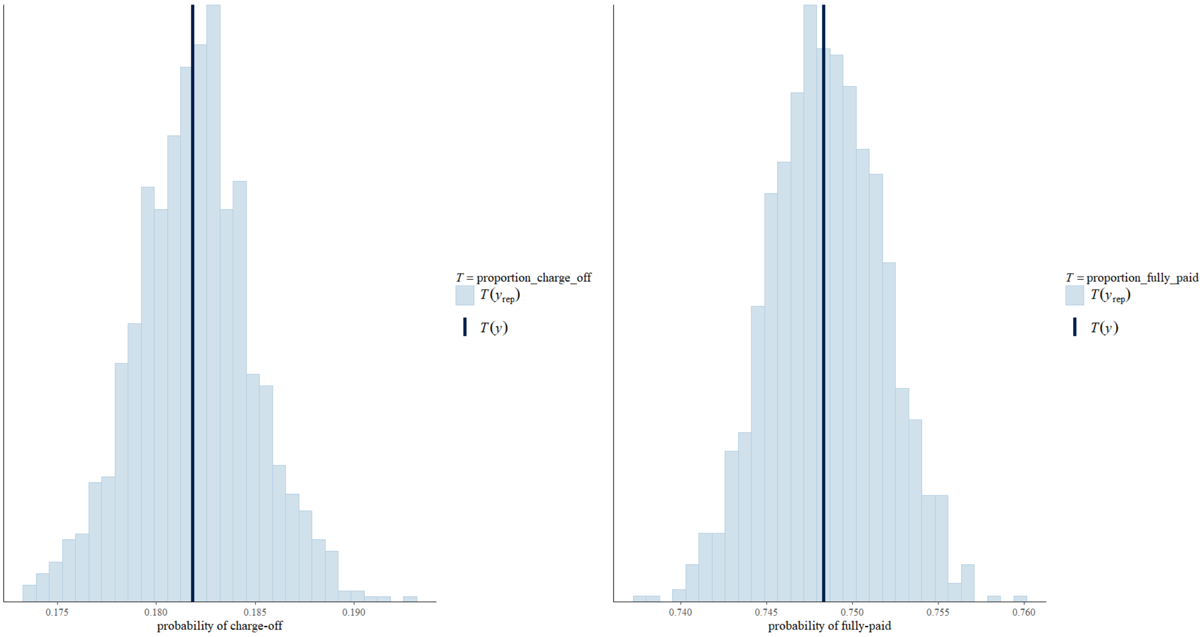


Figure 9: A posterior predictive check of the logistic regression model of charge-off and fully-paid

A significant portion of this paper was devoted to explaining how to fit loan-level Bayesian logistic models using the Variational Bayes framework. This enabled us to estimate key transition probabilities. Specifically, we built two models that predict the probability of loans transitioning from the 'Current' status to either 'Charged-off' or 'Fully-paid'. This loan-level modeling provided us with a granular perspective on loan transitions, enabling us to estimate the likelihood of different outcomes with a higher degree of precision.

By combining the Markov chain delinquency matrix with the estimated loan-level models for key transition probabilities, we have developed a framework that has promising applications in risk management for lenders. A primary strength of our Bayesian model is its ability to quantify uncertainty associated with predictions. Unlike traditional models that provide single point-estimate, our Bayesian approach generates a distribution of potential outcomes. This offers lenders a more complete view of potential scenarios and risks. For instance, our model allows lenders to estimate a range for the necessary cash reserves based on predicted charged-off and paid-off rates. This is a significant benefit, as it helps lenders to prepare more effectively for potential risks and eventualities, thereby increasing the robustness of their risk control measures.

In summary, our combined use of the Markov chain model and loan-level Bayesian logistic models provides a powerful tool for understanding and predicting loan behaviors. This approach enables more informed decision-making in risk management, highlighting the value of advanced statistical modeling in the financial sector. As always, it's essential to bear in mind that models are simplifications of reality and their predictions are subject to uncertainties. However, our Bayesian approach, with its explicit handling of uncertainty, provides a more robust and transparent framework for forecasting and risk management in the lending business.

4 Consumer Term Loan Collections

4.1 Introduction

In this chapter, our main objective is to explore an effective approach for scheduling collection actions on consumer term-loan accounts. To achieve this, we utilize a Markov decision model that allows for efficient decision-making over time.

Each consumer loan account is categorized into various states based on its age, including current, delinquent, early payoff, default, and bankrupt. By employing a Markov transition matrix, we accurately model the probability of loan accounts transitioning between these states.

To determine the optimal collection action for each consumer type at each state and age, we develop an optimization method. This method focuses on maximizing the lender's expected value and takes into account several factors. Firstly, it considers the default risk associated with each state. Additionally, it incorporates operational costs involved in carrying out collection actions.

Furthermore, our optimization approach takes into consideration the time value of money, acknowledging the tradeoff between interest revenue and borrowing costs. It also ensures time consistency in the optimization process. Moreover, it addresses competing risks that may arise between different account states. Lastly, it accounts for penalties that may be incurred due to late payments.

The utilization of a Markov chain model in managing consumer loans is grounded in the understanding that loan accounts naturally transition through various delinquency states over time. For instance, an account that is currently in good standing will remain so if payments are made on time, but will transition to a delinquent state if no payment is received by the due date. The transition probabilities between states can be estimated using historical data, such as through maximum likelihood estimation.

When implementing a collection action in a Markov chain model, the probability of transitioning between states may differ compared to the scenario where no action or a different action is taken. This poses a decision problem for the lender: how to select the most appropriate action for each consumer type at each state and age to maximize the expected net present value. This paper aims to establish a systematic Markov decision modeling framework for making optimal loan collection decisions.

Loan management decision problems encompass various aspects, including (1) determining whether to grant or extend a loan, (2) determining the loan amount to be granted or extended, (3) deciding when to initiate a collection action, and (4) choosing the most suitable collection action to employ. The first two types of decisions are classified as origination decisions, while the latter two are categorized as collection decisions. It is important to note that origination and collection decisions are interrelated and should not be treated as independent. Effective origination decisions take future collection decisions into account, while efficient collection decisions leverage loan specifications established during the origination stage.

Comparing our method with a static collection policy, we demonstrate that our approach is significantly more valuable for accounts with high interest rates and medium to high loan amounts. This is especially true when stronger collection effects are observed.

Additionally, our study sheds light on how interest rates, loan amounts, and collection effects impact the collection actions implemented under an optimal collection policy.

Overall, our findings provide valuable insights for lenders seeking to efficiently schedule collection actions for consumer term-loan accounts. By considering various factors and utilizing a Markov decision model, lenders can maximize their expected value and make informed decisions regarding collection actions.

4.2 Background

The utilization of a Markov chain model for managing consumer loans is based on the belief that loan accounts progress through various delinquency states over time. This intuitive approach considers that the status of an account in a given month will depend on whether a payment in the contracted amount is received by the due date. For instance, an account in the current state this month will remain in the current state if the payment is made on time. However, if no payment is received by the due date, the account will transition to a 1-month delinquent state.

Estimating the probability of transitioning from one state to another is crucial in understanding the behavior of consumer loan accounts. Historical data is often utilized, employing techniques such as maximum likelihood estimation, to estimate these transition probabilities. By analyzing past patterns, lenders can gain insights into the likelihood of account movement between different delinquency states.

When a collection action is implemented on an account within a Markov chain, the transition probabilities can vary compared to scenarios without any action or alternative actions. The decision problem faced by lenders is, therefore, how to select the most appropriate action at each state and age for each type of consumer, ultimately optimizing the expected net present value of the loan portfolio.

This chapter aims to establish a comprehensive framework for Markov decision modeling that enables lenders to make optimal loan collection decisions. By reasoning through this framework, lenders gain the ability to weigh the potential outcomes of various collection actions at different stages and for different types of borrowers. Ultimately, this approach allows lenders to make informed decisions that maximize their expected net present value, potentially improving the overall management of consumer loan collections.

There have been various modeling approaches used in the field of loan management, both for consumer and commercial loans. In a review by Rosenberg and Gleit (1994) ([135]), they discuss quantitative methods that assist managers in making loan management decisions. These methods include discriminant analysis, decision trees, expert systems, linear programming, dynamic programming, and Markov chains.

Another review by Altman and Saunders (1998) ([136])focuses on credit risk measurement for individual and portfolio loans during the period of 1980-1998. They examine the literature on this topic and provide insights into the methods used.

Thomas et al. (2005) ([137]) conducted a survey on the development of methodology in consumer loan modeling from 1955 to 2005. They highlight critical issues in consumer loan management, such as the design of consumer scoring systems, the development of models for selecting product features, policy inference techniques, and improving the accuracy of default forecasting.

When it comes to loan management, decision-making is crucial and can be divided into four main categories. The first decision problem is whether to grant or extend a loan, while the second is determining the appropriate loan amount to grant or extend. These two decisions fall under the category of origination decisions. On the other hand, the remaining two decision problems are classified as collection decisions. The third decision is when to initiate a collection action, and the fourth is determining the suitable collection action to be taken.

It's important to note that origination and collection decisions are not independent of each other. In fact, they are closely related. An effective origination decision takes into consideration future collection decisions, and efficient collection decisions incorporate loan specifications that were initiated during the origination stage.

4.3 Model Setup

In our scenario, there are two main entities involved: the seller, who directly sells durable goods to consumers, and the lender, who provides loans to consumers for purchasing those goods. Let's denote the per-period discount factor for the lender as ρ. This factor represents the lender's perception of the value of receiving 1 unit of currency in the future, with a lower value indicating a higher borrowing or opportunity cost for the lender.

The time period used for managing consumer loan accounts can vary, ranging from days to weeks or months. When a consumer is approved for a loan by the lender, they enter into a contractual agreement. This contract outlines the terms and conditions of the loan, including repayment details and any associated fees or interest rates.

To keep track of the loan and associated consumer information, the lender establishes an account for each approved loan. In practice, these loan accounts may contain a variety of rich data points and parameters. These parameters provide important insights for effectively managing the loan accounts and making informed decisions.

We are particularly interested in analyzing and understanding specific account parameters. These parameters encompass essential details about the loan and the consumer, and may include information such as the current balance of the loan, the repayment history, the interest rate applied, the remaining term of the loan, any future payment obligations, and any additional fees that may be associated with the loan.

By examining and analyzing these account parameters, we can gain valuable insights into the loan portfolio and make data-driven decisions regarding collection actions, risk assessment, and overall loan management strategies.

Let L be the initial loan amount offered by the lender to the buyer, T be the term of the loan, i.e., the number of periods to payoff the loan, and r be the interest rate the consumer pays the lender. The basic formula for calculating your monthly payment P is

$$P = \frac{L[r(1+r)^T]}{(1+r)^T - 1}.$$

Let M be the market value of the durable good at the time of purchase, and θ be a discount constant for the good during one payment period. That is the market value of the good at the time of n-th payment time is $M\theta^n$. (θ would be less than 1 if the good is a consumer good with value decreasing as time passes; θ could be greater than 1 if the good becomes more valuable as time passes.)

Let t denote the age of an account, which is the number of periods passed since the onset of the loan. At each age, an account can be at a specific payment status, which we label as state, denoted by s. At each age $t \geq 1$, an account state is determined by the state and payment behavior of the account at age $t-1$. To address typical business situations, we consider the following states of an account at age t when applicable.

- $s = 0$ Current, i.e., no past due payment at age $t-1$. An account must be in state current when $t = 1$, because a down payment is required at the onset of the loan when $t = 0$. When a loan is paid off at age T, we have $s = 0$ at age $T+1$.

- $s = x$ for $x = 1, 2, 3, 4$ x-period delinquent, i.e., there are x contracted per-period payments not received at age $t - 1$.

- $s = -1$ Early payoff, i.e., the consumer pays off all the remaining balance at age $t - 1$. When a consumer pays off the debt at age T, the end of term, the state at age $T + 1$ is current.

- $s = -2$ Default, i.e., the consumer's purchased good is repossessed by the lender at age $t - 1$. In practice, repossession occurs only when a customer misses several payments, and thus the default state can only occur when t is large enough.

- $s = -3$ Bankrupt, i.e., a third party certifies that the consumer is in a legal status rendering the consumer incapable of paying the remaining balance to the lender at age $t - 1$. In this case, the loan contract loses its original effectiveness and the third party determines an amount of money that the consumer must pay the lender.

The aforementioned eight states that have been defined effectively model the most commonly occurring account statuses in practical scenarios. These states help to simplify the collection problem in several ways.

Firstly, consumers have the flexibility to make payments up to the remaining balance. To streamline the analysis, we assume that consumers either make a payment equal to a multiple of the contracted per-period payment, along with any applicable late payment penalties, or opt to pay the entire remaining balance.

Secondly, while it is technically possible for an account to be delinquent for more than four periods, we assume that when an account reaches a four-period delinquency, the lender will repossess the purchased goods, thereby preventing the account from transitioning to a five-period delinquency state.

Thirdly, it is important to note that consumers may negotiate with the lender to modify the loan contract, such as extending the loan term or reducing the interest rate. However, we do not account for these contract changes in our current analysis. Nevertheless, the inclusion of alternative payment amounts, increased delinquency periods, and changes in the loan contract can be feasibly modeled within an expanded state space without fundamentally altering our core model.

In addition to tracking payment status, it is crucial to model the risk status, specifically the risk level, of each account at each age. The risk level of an account is primarily influenced by its past payment behavior for the loan, with secondary consideration given to the overall credit profile of the account. Lenders typically monitor account payment history and credit information, such as FICO scores, to estimate the risk level of each account at specific points in time.

Then the remaining balance of an account $\mathcal{R}(s,t)$ is

$$\mathcal{R}(s,t) = \sum_{i=0}^{i=s} P(1+\gamma)^i + \sum_{i=t+1}^{T} \frac{P}{(1+r)^{i-t}},$$

where γ is the late payment penalty rate.

The lender can implement a preventive or corrective collection action, denoted by a, over an account at a given age, if applicable. The following actions are considered in this work.

- $a = 0$ Do-nothing, leave the account as is.
- $a = 1$ Electronically contact, notify the consumer of account state via text message, email, or an automatic machine dialer.
- $a = 2$ Account management, mail or call the consumer to inform on account information, gather more information and advise on future payment resolutions.
- $a = 3$ Loss prevention, communicate with the consumer intensively to discover why the account has been in delinquency status, and provide specific advice on how to retain the account in the state of the current.
- $a = 4$ Repossession, repossess the consumer's good purchased through the loan. If the residual value of the good is lower than the sum of the remaining balance and the cost of repossession, then the lender incurs a loss in the amount of that difference; otherwise, the lender pays the difference back to the consumer. That is, the lender cannot earn money through repossession.

The payback policy for the remaining value of repossessed goods is legally mandated to safeguard consumer interests. According to theory, if the residual value of the repossessed asset is inadequate to cover the outstanding balance along with repossession costs, the consumer remains responsible for the difference. However, in practice, this shortfall is challenging to collect and is therefore disregarded in our definition.

We make the assumption that only one action can be applied to an account within a single time period. This is consistent with the assumption that the account state remains unchanged throughout the duration of one time period. It's important to note that this assumption is not restrictive. In cases where the action requires a time frame shorter than one time period, a more suitable and shorter time unit can be employed to define the time period accordingly.

4.4 Optimization

Now we consider how an account transits from one state to another as the account ages. A state transition happens at the time when an account becomes age $t + 1$ from age t, for $t = 1, \cdots, T$. Recall that an account is always in the state of current at age $t = 1$ by our definition.

Let $q^a_{s,s'}(t)$ denote the probability of transition from state s to state s' for the account that enters age $t+1$, under the application of collection action a. The transition probability defines a discrete Markov chain, where the account state at the next age depends only on the account state and the collection action applied to the account at the current age. Then the Markov chain transition matrix is as the following:

$$\begin{pmatrix} q^a_{00}(t) & q^a_{01}(t) & 0 & 0 & 0 & q^a_{0-1}(t) & 0 & q^a_{0-3}(t) \\ q^a_{10}(t) & q^a_{11}(t) & q^a_{12}(t) & 0 & 0 & q^a_{1-1}(t) & 0 & q^a_{1-3}(t) \\ & & & & \cdots & & & \\ q^a_{40}(t) & q^a_{41}(t) & q^a_{42}(t) & q^a_{43}(t) & q^a_{44}(t) & q^a_{4-1}(t) & q^a_{4-2}(t) & q^a_{4-3}(t) \end{pmatrix}$$

In the transition matrix, a 0 entry signifies that a particular transition cannot occur as per the definition. Specifically, it is not possible for an account to transition from being x-period delinquent to x-period delinquency plus one, for x equal to 2, 3, or 4. Moreover, a lender will not initiate the repossession process when the account is less than 2-period delinquent. It is important to note that states s equal to -1, -2, and -3 do not transition to any other states and are therefore excluded from the matrix as initial states.

The econometric characterization of the transition matrix is intricately linked to the dynamic credit risk envisioned in the model. It is essential to accurately quantify the transition probabilities specific to the underlying business problem being addressed.

Let $\pi^a_{s,s'}(t)$ denote the expected value received from the account that transits from state s to state s' when entering age $t + 1$ under action a.

The maximum expected net present value (ENPV) of account h in state s at age t, taking the state transition and applicable action into consideration is denoted by $V_t(s)$.

The value $V_t(s)$, and the optimal action that maximizes the ENPV of an account at each state and age can be found via the following Bellman equation through a standard backward induction

$$V_t(s) = \max_{a \in A(s)} \{ \sum_{s' \in S} q^a_{s,s'}(t) \pi^a_{s,s'}(t) + \rho V_{t+1}(s') \}. \tag{11}$$

where $A(s)$ is the action space for the account in status s, and ρ is some constant. Note that the maximum ENPV at age 0, the onset of the loan, is $V_1(0)$. The inclusion of discount factor ρ, together with the interest rate r, addresses the tradeoff between interest revenue and borrowing cost. Also, since $V_t(s)$ is optimized at each age t, the solution achieves time consistency in optimization. In addition, our definition of the transition and per-period value matrixes models competing risks between different account states, along with the penalty for late payment. Since the Markov chain model is widely applied for decisions on collection policy and action (Rosenberg and Gleit 1994; Thomas et al. 2005), we provide the following result with proof omitted.

Proposition 4.1. *The above equation finds an optimal action that maximizes the ENPV of the account in status s at age t.*

To maximize the value $V_t(s)$, the Equation 11 recurrence relation establishes the optimal action for each account at every combination of age (t) and status (s). This optimal collection policy is determined by finding the value of $V_t(s)$, which requires specifying the boundary condition based on practical considerations specific to the problem at hand.

Equation 11 explicitly includes the account status, indicating that we can solve this optimization problem for each distinct account type when provided with transition and per-period value matrices that align with the account's characteristics. To simplify computations, account characteristics can be grouped together. It's important to note that the action space $A(s)$ and the state space S at age t depend on both the account status s and the current status, which may result in the exclusion of certain actions and states. Furthermore, the application of certain actions may depend on preceding actions. For example, a repossession action may only be applicable if a loss prevention action was previously applied. To account for this, we can introduce a binary state variable to the value function V to indicate whether a loss prevention action was taken at the previous age. This incorporation allows us to restrict the application of the repossession action at the current age accordingly.

Given the account characteristics, Equation 11 defines a model with finite state and action spaces, as well as a finite horizon T, making it computationally tractable for numerical solution methods.

In conclusion, our model implements an optimal collection policy that yields significantly greater value compared to a static collection policy, particularly for consumer term loans with high interest rates and medium to high loan amounts. This is particularly true when collection actions have observable effects in facilitating the transition of delinquent accounts to a current state. Furthermore, our study sheds light on how the implementation of collection actions, under an optimal collection policy, is influenced by factors such as interest rates, loan amounts, and the efficacy of collection efforts.

It is important to note that in practical applications, the choice of collection actions for a loan account is contingent upon consumer heterogeneity. This is due to the fact that different consumers possess unique account characteristics, which can impact the probability and payoff associated with state transitions, as well as the effectiveness of collection actions. Thus, a tailored approach that considers individual consumer traits is crucial for optimal collection strategies.

4.5 Applying Markov chain model

By merging the Markov chain model and the aforementioned model, we can establish a well-rounded model specifically designed for collection purposes. This integration involves leveraging the Markov chain transition matrix values from Chapter 3 of the previously mentioned models.

One of the key advantages of our proposed model is its flexibility in decision-making for the company. This is made possible due to our utilization of variational inference in Chapter 3. With this approach, we are able to derive model coefficients and generate a range of potential models. This allows us to explore different scenarios and choose the most suitable one based on the specific circumstances.

The ability to consider multiple models through variational inference broadens the range of possibilities for the company. Rather than being limited to a single approach, we have the option to select from numerous models that are derived from the interval of coefficients provided by variational inference. This enhances the decision-making process and empowers the company to make well-informed choices when it comes to collection strategies.

In this study, we propose a modeling framework that combines the Markov chain model and the previously mentioned model to create a comprehensive model for collection purposes. This framework falls under the category of optimal Markov decision theory.

Our model focuses on the specific context of consumer term loan collection within a simplified business setting. It serves as a unified framework for making various collection decisions based on incomplete account information. This framework is consistent with the concept of a dynamic credit risk model, which utilizes a Markov chain modeling approach.

Our findings highlight the significant potential benefits of implementing an optimal collection policy that considers the state and age of the account, as well as consumer heterogeneity. When compared to a static collection policy that only takes into account the account state, the optimal collection policy demonstrates clear advantages.

Three key factors that influence the relative performance of our dynamic collection policy are the interest rate, loan amount, and the effects of collection actions. On average, the optimal collection policy proves to be more valuable for accounts with higher interest rates and medium to high loan amounts. This is particularly true when collection actions are more intensive.

Furthermore, we illustrate how the collection actions implemented under an optimal collection policy are influenced by factors such as the interest rate, loan amount, and collection effects. This provides valuable insights into tailoring collection strategies to specific account characteristics and optimizing overall collection efficiency.

4.6 Further thoughts

The present work offers promising avenues for further expansion, particularly in response to practical collection challenges. There are several areas where the Markov decision model can be extended to address these concerns.

Firstly, by expanding the account state and collection action spaces, the model can cater to scenarios where information gathered during the collection process can be utilized to guide actions at later stages. Additionally, incorporating the correlation between the residual value of purchased goods and account characteristics can contribute to more accurate decision-making within the model.

Secondly, in instances involving a large number of loan accounts, it becomes crucial to consider resource constraints associated with account maintenance. Consequently, determining appropriate collection actions must take into account these limitations.

Thirdly, it is imperative to quantitatively evaluate the impacts of various collection actions, especially those that are novel to the business, such as electronic contact and automated machine dialers. Experimental analysis can provide valuable insights into the efficacy of these initiatives.

Lastly, thorough investigation into the criteria for contract rewriting during the collection process is vital. Understanding the circumstances under which altering the terms of the agreement is warranted can lead to better outcomes for both lenders and borrowers.

By addressing these areas, the model's effectiveness for real-world collection challenges can be enhanced, ultimately benefiting both the lender and the borrowers.

5 Random Forest Model

In this section, we study the Random forest model. In particular, we will use the Random forest model to stimulate a particular set of data.

5.1 Random Forest Model and Logistic Regression Model

Logistic regression and random forest models are both popular machine learning algorithms used for classification tasks. While they have some similarities, they also have key differences in terms of their underlying principles, interpretability, scalability, and handling of non-linear relationships (see [138], [139], and [140]).

1. Model Principle: - Logistic Regression: Logistic regression is a statistical model that uses a logistic function to estimate the probability of a binary outcome based on input variables. It assumes a linear relationship between the input features and the logit of the target class probabilities. - Random Forest: Random forest is an ensemble learning method that consists of multiple decision trees, where each tree is trained on a random subset of the data. The final prediction is made by aggregating the predictions of all the individual trees.

2. Interpretability: - Logistic Regression: Logistic regression models provide interpretable coefficients associated with each input feature, indicating the direction and magnitude of the influence on the target variable. - Random Forest: Random forest models are generally less interpretable than logistic regression. It is challenging to extract meaningful insights from the multitude of decision trees that make up the random forest model.

3. Handling Non-linear Relationships: - Logistic Regression: Logistic regression assumes a linear relationship between the input features and the log odds of the target variable. It may struggle to capture complex non-linear relationships unless non-linear transformations of the features are included. - Random Forest: Random forest models can handle non-linear relationships effectively without the need for explicit feature transformations. The ensemble of decision trees can capture complex interactions between features and target, making them more flexible.

4. Scalability: - Logistic Regression: Logistic regression models are relatively simple and computationally efficient. They can handle large datasets with a large number of features. - Random Forest: Random forest models can handle larger datasets as well, but their training and prediction times can be significant, especially when the number of trees or depth of the trees is high.

5. Performance: - Logistic Regression: Logistic regression performs well when the relationship between the features and the target is predominantly linear. It may struggle in cases of high-dimensional data or when there are complex interdependencies between features. - Random Forest: Random forest models are more robust and generally perform well in various scenarios. They can handle both linear and non-linear relationships, as well as interactions between features.

In summary, logistic regression and random forest models have their own strengths and weaknesses. Logistic regression is more interpretable and suitable for linear relationships, while random forest models are more flexible, handle non-linear relationships well, and are generally powerful models for classification tasks (see [141], [142], [143], and [144]).

5.2 Introduction to Random Forest Model

Based on the above comparisons, we decide to choose the random forest model. Now let's review what is the Random forest model.

The random forest algorithm is a type of supervised machine learning algorithm that uses decision tree algorithms to construct a predictive model. It is commonly used in various industries, including banking and e-commerce, to predict behavior and outcomes ([145]).

The concept of bagging is central to the random forest algorithm. Bagging involves averaging multiple models that may have some noise or errors, but are approximately unbiased (see [146]). This averaging process helps reduce the variance in the predictions. Decision trees are particularly well-suited for bagging because they can capture complex interaction structures in the data and have relatively low bias when grown deep enough. However, decision trees are known to produce noisy results, so they benefit greatly from the averaging process.

Another advantage of using bagging with decision trees is that each tree generated in the process is identically distributed, meaning that the expectation of an average of B such trees is the same as the expectation of any one of them. This means that the bias of the bagged trees remains the same as that of the individual bootstrap trees. The only potential for improvement in the bagged models lies in reducing variance.

An average of B i.i.d. random variables, each with variance σ^2, has variance $\frac{1}{B}\sigma^2$. If the variables are simply i.d. (identically distributed, but not necessarily independent) with positive pairwise correlation ρ, the variance of the average is

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

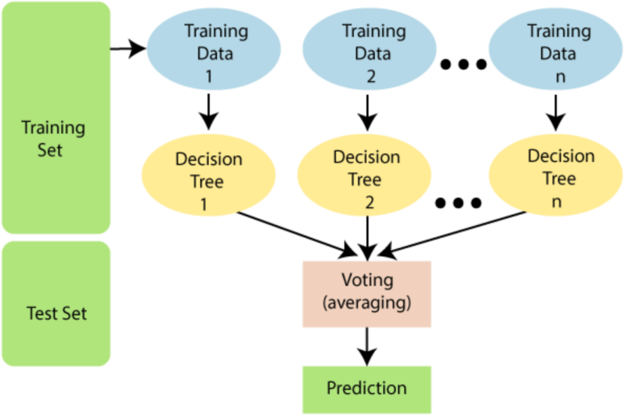


Figure 10: RF work flow

As B increases, the second term disappears, but the first remains, and hence the size of the correlation of pairs of bagged trees limits the benefits of averaging. The idea in random forests is to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much. This is achieved in the tree-growing process through random selection of the input variables.

The below diagram explains the working of the Random Forest algorithm:

5.3 More details about Random Forest Tree Model

Assumptions for Random Forest: Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.

- The predictions from each tree must have very low correlations.

- Independence of decision trees: The model assumes that each decision tree in the random forest is built independently of one another. This means that the construction of one tree does not depend on the construction or outcome of any other tree in the forest.

- Random sampling of training data: The random forest algorithm assumes that the training data is sampled randomly and that each sample is independent of the others. This helps reduce bias and prevent overfitting by introducing variation into the model.

- Random feature selection: Random forest randomly selects a subset of features from the available set for each decision tree. This assumption assumes that not all features are equally important for making predictions, and that using a subset of features can help improve model performance.

- Homogeneous feature importance: The assumption is that all features contribute equally to the model's predictive power. While this may not always hold true, the random forest algorithm still considers all features and attempts to identify patterns and relationships among them.

- No multicollinearity: The random forest model assumes that there is no high correlation (multicollinearity) between the predictor variables. This prevents redundancy in the model and helps improve the accuracy of predictions.

  It's important to note that while these assumptions are generally made in the random forest model, the algorithm is robust and can still produce reliable predictions even if some of these assumptions are violated.

Why use Random Forest? There are several reasons why the random forest model is widely used in machine learning (see [147], [148], [149] for more references):

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Moreover, The random forest algorithm tends to yield highly accurate predictions. By aggregating the predictions of multiple decision trees, it reduces the risk of overfitting and provides more robust results compared to using a single decision tree. Random forests can be used for both classification and regression problems, making them applicable to a wide range of tasks. They can handle categorical and numerical features, as well as missing data, without the need for extensive preprocessing. Also random forests can capture complex non-linear relationships between input features and the target variable. They are capable of modeling interactions and capturing important feature interactions that may be missed by linear models. Random forests are relatively robust to outliers and noisy data. Since each decision tree is built on a random subset of the data, outliers have less influence on the overall predictions. This makes the model more resilient to noisy or anomalous data points. Furthermore, Random forests provide a measure of feature importance, which can be valuable for understanding which variables have the most impact on the predictions. This information can help in feature selection and interpretation of the model. Random forests can effectively handle datasets with a large number of input features, making them suitable for high-dimensional problems. They can automatically select relevant features and avoid overfitting. Overall, the random forest model offers a powerful and versatile approach to predictive modeling, making it a popular choice in various domains and industries.

How does Random Forest algorithm work? Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Applications of Random Forest

The random forest model finds applications in various domains and industries due to its versatility and robustness. Some notable applications include:

1. Classification: Random forests are commonly used in classification tasks. They can be used to classify emails as spam or non-spam, predict customer churn, detect fraudulent transactions, identify disease diagnosis, and classify images or text into multiple categories.

2. Regression: Random forests can also be used for regression tasks, such as predicting housing prices, estimating sales revenue, or forecasting stock prices. They can handle both continuous and categorical variables, making them flexible for a wide range of regression problems.

3. Feature Importance: Random forests provide a measure of feature importance. This information can be used to identify the most influential features in a dataset, helping researchers and analysts understand the underlying factors contributing to a particular outcome.

4. Anomaly detection: By leveraging the collective decision-making of multiple decision trees, random forests can identify outliers or anomalies by detecting instances that deviate significantly from the majority. This is useful in fraud detection, network intrusion detection, or identifying irregular patterns in data.

5. Recommender Systems: Random forests can be used in recommender systems to provide personalized recommendations to users based on their preferences. They can analyze user behavior, historical data, and contextual information to offer tailored suggestions for products, movies, music, or news articles.

6. Ensemble Learning: Random forests are a part of ensemble learning methods. They can be combined with other machine learning algorithms, such as boosting or stacking, to further enhance predictive performance and handle complex tasks.

Overall, the random forest model's ability to handle a wide range of problem types, handle high-dimensional data, provide feature importance, and deliver accurate predictions makes it a popular choice in many real-world applications.

There are mainly four sectors where Random forest mostly used:

- Banking: Banking sector mostly uses this algorithm for the identification of loan risk. Banks use the random forest algorithm to predict the creditworthiness of individuals and businesses. By analyzing a range of factors such as historical financial data, credit history, employment status, and demographic information, the model can assess the likelihood of default or credit risk for loan applicants. Random forests are effective in detecting fraudulent activities in banking transactions. By analyzing patterns and anomalies in large datasets, the model can identify potential fraudulent behavior such as unauthorized access, money laundering, or identity theft. Banks often use random forests to segment their customer base. By analyzing customer data, transaction history, and behavior patterns, the model can identify groups of customers with similar characteristics. This segmentation allows banks to tailor their marketing strategies, develop personalized offerings, and improve customer retention. Random forests can assist in risk assessment for investment and loan portfolios. By considering various factors such as market trends, economic indicators, and historical data, the model can estimate the potential risks associated with specific investments or loan portfolios. This helps banks make more informed decisions to manage and mitigate risks. Overall, the random forest model in banking provides valuable insights for decision-making, risk management, fraud prevention, and customer satisfaction.

- Medicine: The random forest model has found numerous applications in the field of medicine. For example, random forests can be used to aid in the diagnosis of various diseases. By training the model on historical patient data, incorporating variables such as medical history, symptoms, lab results, and imaging data, the model can predict the likelihood of a patient having a particular disease or condition. Drug Response Prediction: Random forests can be used to predict how a patient will respond to a specific drug or treatment. By analyzing patient characteristics, genetic profiles, and other relevant factors, the model can help in personalizing treatment plans and optimizing medication choices. The random forest model can be used to predict the prognosis or outcome of a particular disease. By considering patient demographics, medical history, treatment protocols, and other variables, the model can provide insights into the likely progression of a disease and help in making informed decisions regarding treatment options. Random forests can also be used in medical imaging tasks. By training on a large dataset of annotated medical images, the model can assist in tasks such as identifying cancerous cells, detecting tumors, or classifying different types of lesions. The random forest model's ability to handle complex and high-dimensional data, handle missing values, and provide interpretable results makes it a valuable tool in the field of medicine for both research and clinical practice.

- Land Use: The random forest model is also used in the field of land use and land cover analysis. We can identify the areas of similar land use by this algorithm. Random forests can be employed to classify land cover types based on satellite imagery or remote sensing data. By training the model with labeled examples of different land cover types, it can accurately predict the land cover class of new, unseen areas. This information is crucial for land management, urban planning, environmental assessment, and natural resource management. Random forests can be utilized to detect and monitor changes in land use patterns over time. By comparing remote sensing data from different time periods and training the model with samples of known land use changes, it can effectively identify areas that have undergone changes in land use, such as deforestation, urban expansion, or agricultural conversion. This information is vital for understanding and managing land use dynamics and their environmental implications. Random forests can assist in assessing habitat suitability for different species. By integrating environmental variables such as temperature, precipitation, vegetation, and topography, the model can identify areas with high potential for supporting certain habitats or species. This information aids in conservation planning, biodiversity monitoring, and wildlife management. random forest models could handle complex spatial data and provide accurate predictions making it a valuable tool for land use analysis and management.

- Marketing: The random forest model is widely used in marketing for various applications. Random forests can be employed to segment customers into different groups based on their behavior, preferences, or demographics. By analyzing a wide range of customer attributes, such as purchase history, browsing patterns, and demographic information, the model can identify distinct segments with similar characteristics. This segmentation can help businesses tailor their marketing strategies for each segment, improving targeting and personalization. Random forests can predict customer churn, i.e., the likelihood of customers discontinuing their relationship with a company. By training the model on historical customer data, including features such as customer interactions, purchase history, and customer satisfaction scores, businesses can identify patterns and factors that contribute to churn. This allows proactive measures to be taken, such as targeted retention campaigns, to reduce customer churn. Random forests can power recommender systems that suggest relevant products or services to customers. By considering various customer attributes and historical interaction data, the model can generate personalized recommendations. This enhances the customer experience, increases engagement, and can lead to improved cross-selling and upselling opportunities. Random forests can estimate the future value a customer is expected to bring to a company over their entire relationship. By analyzing past customer behavior, purchase history, and spending patterns, the model can predict the CLV for individual customers. This information helps businesses prioritize their marketing efforts and allocate resources more effectively. Generally speaking, the random forest model in marketing helps businesses gain insights, make data-driven decisions, enhance customer experiences, and optimize marketing strategies for better outcomes.

Advantages of Random Forest: Random Forest is capable of performing both Classification and Regression tasks. It is capable of handling large datasets with high dimensionality. It enhances the accuracy of the model and prevents the overfitting issue. One advantage of the random forest model is its ability to perform both classification and regression tasks. This versatility allows it to handle a wide range of predictive modeling problems, making it a flexible option for various applications. Random forests are well-suited for handling large datasets with high dimensionality. With numerous decision trees operating in parallel, the model can effectively handle a large number of features and instances. This scalability makes it efficient and suitable for complex datasets commonly encountered in real-world scenarios. Random forests have the advantage of producing highly accurate predictions. By aggregating the predictions of multiple decision trees, the model reduces the risk of overfitting and provides more robust results. Additionally, the randomization process involved in feature selection and splitting further helps to prevent overfitting, improving the model's generalization capabilities.

Overall, the random forest model's versatility, scalability, accuracy, and ability to handle high dimensionality make it a popular choice in various data-driven applications.

Disadvantages of Random Forest While the random forest is versatile and widely applicable, it does have some limitations, particularly when it comes to regression tasks.

One disadvantage is that random forest may not be as suitable for regression tasks compared to classification tasks. This is because the algorithm is primarily designed for classification, and its ability to handle continuous or numeric outcomes in regression may be limited. Random forests are known to perform better in situations where the outcome variable is categorical or discrete, rather than continuous.

Additionally, random forests can be prone to overfitting if the dataset has a large number of noisy or irrelevant features. In such cases, the algorithm may struggle to generalize well to unseen data (see [126], [116]).

Furthermore, random forests can be computationally expensive and time-consuming, especially with large datasets and high dimensionality. Building a random forest model with a large number of trees can require significant computational resources, which may not be ideal for real-time or resource-constrained applications.

It's important to carefully evaluate the specific requirements of a regression task and consider alternative algorithms or techniques that may be more suitable for the particular problem at hand.

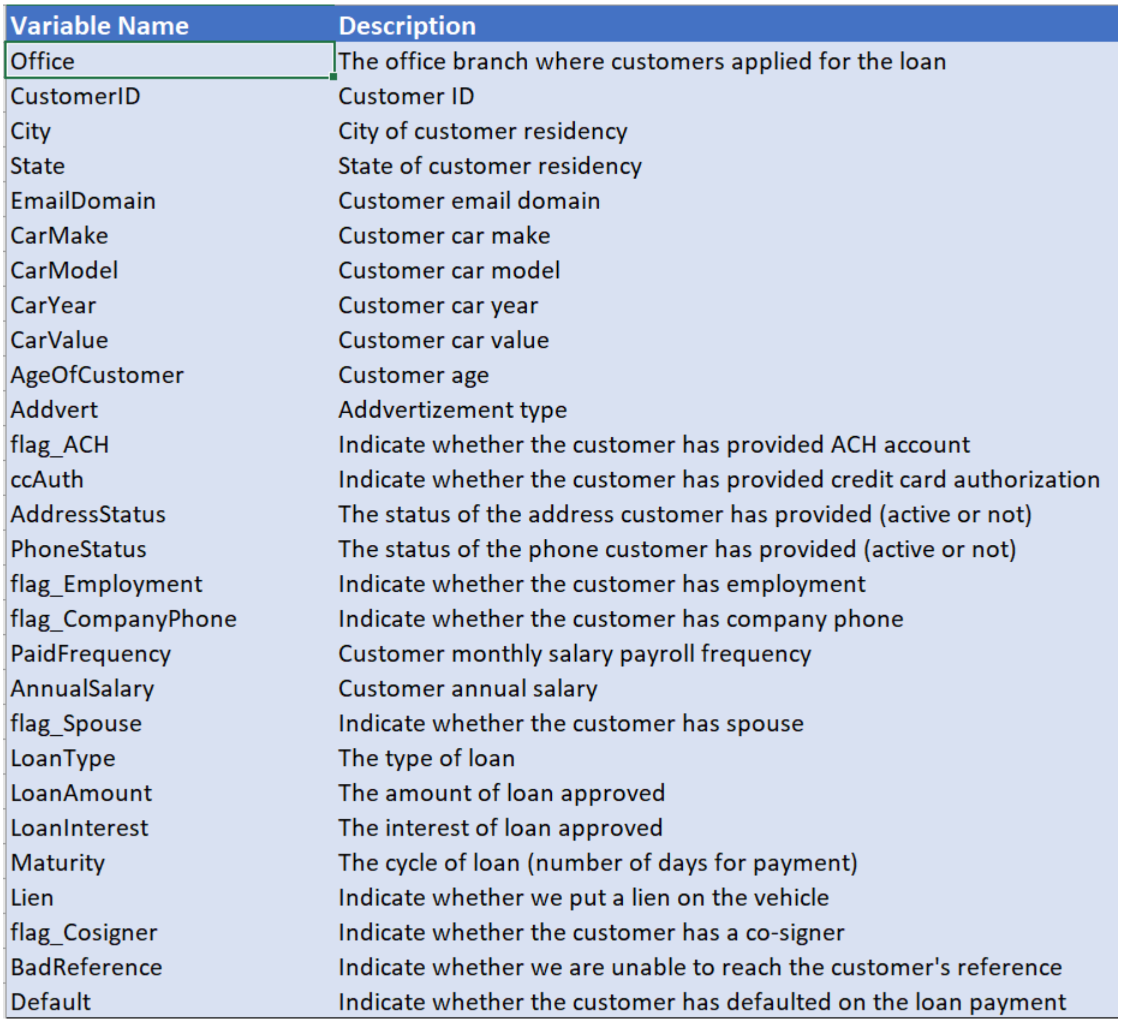
5.4 Using Random Forest Model to build a scoring model

In this subsection, we utilize our skills in data mining and machine learning model building to create a scoring model aimed at predicting whether a new applicant will default on their loan or not. To accomplish this, we have been given a dataset containing information about title loan customers, including their performance status (default) which serves as the last column in the dataset.

We embark on this project with the goal of constructing a highly accurate predictive model. The dataset consists of 5976 observations and 28 variables, each providing valuable information for our analysis. These variables or features, as shown in Figure 11, encompass a wide range of relevant aspects that can potentially influence the likelihood of default.

Our objective is to employ data mining techniques and machine learning algorithms, such as the random forest model, to effectively analyze this dataset. By leveraging the power of this model, we aim to identify patterns, associations, and dependencies within the data that can aid in predicting the probability of default for new applicants.

Throughout the project, we will execute various steps, including data preprocessing, feature selection, model construction, and model evaluation. By meticulously following this process and leveraging our expertise in data mining and machine learning, we strive to develop a scoring model that achieves superior accuracy in predicting loan defaults, ultimately providing valuable insights and assisting in decision-making for lending institutions.



Variable Name	Description
Office	The office branch where customers applied for the loan
CustomerID	Customer ID
City	City of customer residency
State	State of customer residency
EmailDomain	Customer email domain
CarMake	Customer car make
CarModel	Customer car model
CarYear	Customer car year
CarValue	Customer car value
AgeOfCustomer	Customer age
Addvert	Addvertizement type
flag_ACH	Indicate whether the customer has provided ACH account
ccAuth	Indicate whether the customer has provided credit card authorization
AddressStatus	The status of the address customer has provided (active or not)
PhoneStatus	The status of the phone customer has provided (active or not)
flag_Employment	Indicate whether the customer has employment
flag_CompanyPhone	Indicate whether the customer has company phone
PaidFrequency	Customer monthly salary payroll frequency
AnnualSalary	Customer annual salary
flag_Spouse	Indicate whether the customer has spouse
LoanType	The type of loan
LoanAmount	The amount of loan approved
LoanInterest	The interest of loan approved
Maturity	The cycle of loan (number of days for payment)
Lien	Indicate whether we put a lien on the vehicle
flag_Cosigner	Indicate whether the customer has a co-signer
BadReference	Indicate whether we are unable to reach the customer's reference
Default	Indicate whether the customer has defaulted on the loan payment

Figure 11: Data features

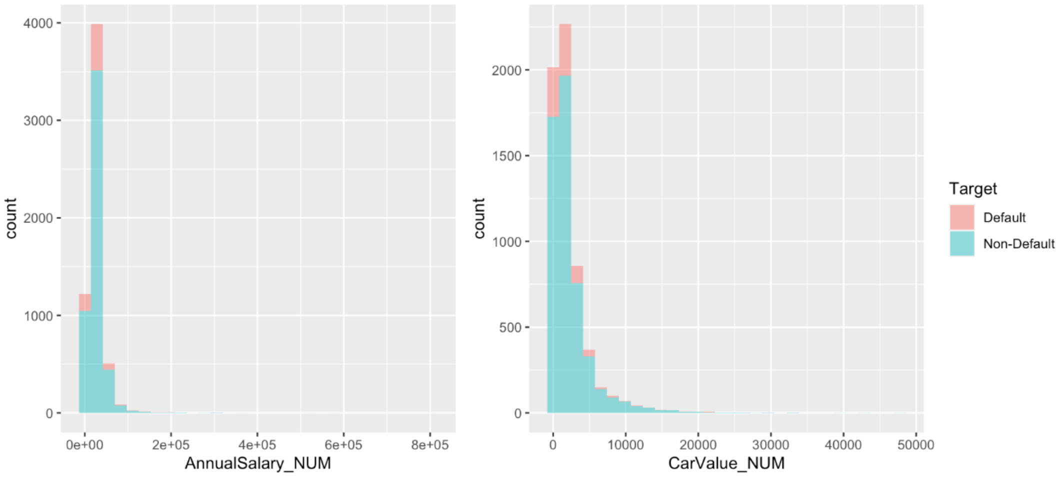


Figure 12: Salary and Car value features

5.4.1 Observation

After carefully reviewing the summary and examining the provided data, it appears reasonable to consider the following fields as numerical features: Car-Value, Age Of Customer, Annual Salary, and Loan Amount. However, before proceeding with the analysis, it is necessary to clean up the AnnualSalary and CarValue fields, removing any symbols like “NULL” or other inconsistencies that may hinder the conversion to numerical features.

To gain a better understanding of the data and its predictive power, we present histograms of the four numerical features. The resulting histograms are depicted in figures 12 and 13, with different colors representing the occurrence of the Default event. Upon visual inspection, it is evident that the histograms do not exhibit a strong predictive power for the numerical features alone.

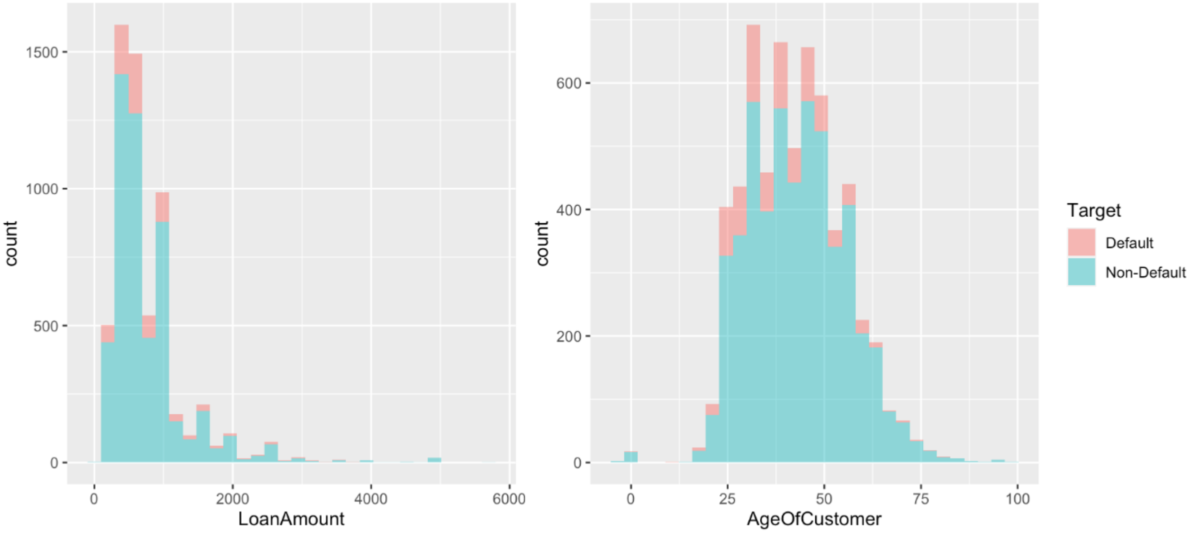


Figure 13: Target Features

The clean-up process for categorical features can be more time-consuming compared to numerical features. In this particular analysis, the focus was directed towards cleaning and correcting the CarMake attribute, as it holds the potential to be a strong predictor in the context of title loans. However, with sufficient time and resources, it is essential to handle all categorical features for data cleaning.

Apart from CarMake, other categorical features that require attention include CarModel, EmailDomain, and City. These features may contain inconsistencies, errors, or missing values that need to be addressed in order to ensure the reliability and accuracy of the data. By properly cleaning and correcting these categorical features, the predictive model can benefit from a more comprehensive and representative dataset. Ultimately, this can lead to improved predictions and insights regarding loan defaults.

In the data cleaning process, a user-defined function called "car make clean" was created specifically to clean the CarMake column. This function likely involved handling common data inconsistencies such as misspelled or inconsistent entries, standardizing names, and dealing with missing or unknown values.

Additionally, errors found in the "Addvert" field were corrected. This indicates that any issues or inaccuracies in the data related to the advertisement field were identified and rectified, ensuring data integrity and accuracy.

Moving forward, visualizations were generated to understand the distribution and impact of certain categorical features on the occurrence of default events. Specifically, the State and Loan Type features were examined through bar plots displaying the count and percentage of default events within each category.

During this analysis, it was observed that some categories within these features had a very small volume or representation in the dataset. Recognizing the potential limitations of making accurate predictions with such small subsets, a decision was made to collapse these low-volume categories into a single category in the subsequent data pre-processing stage. This helps to streamline the analysis process and improve the reliability of the predictive model by reducing the impact of these less representative categories.

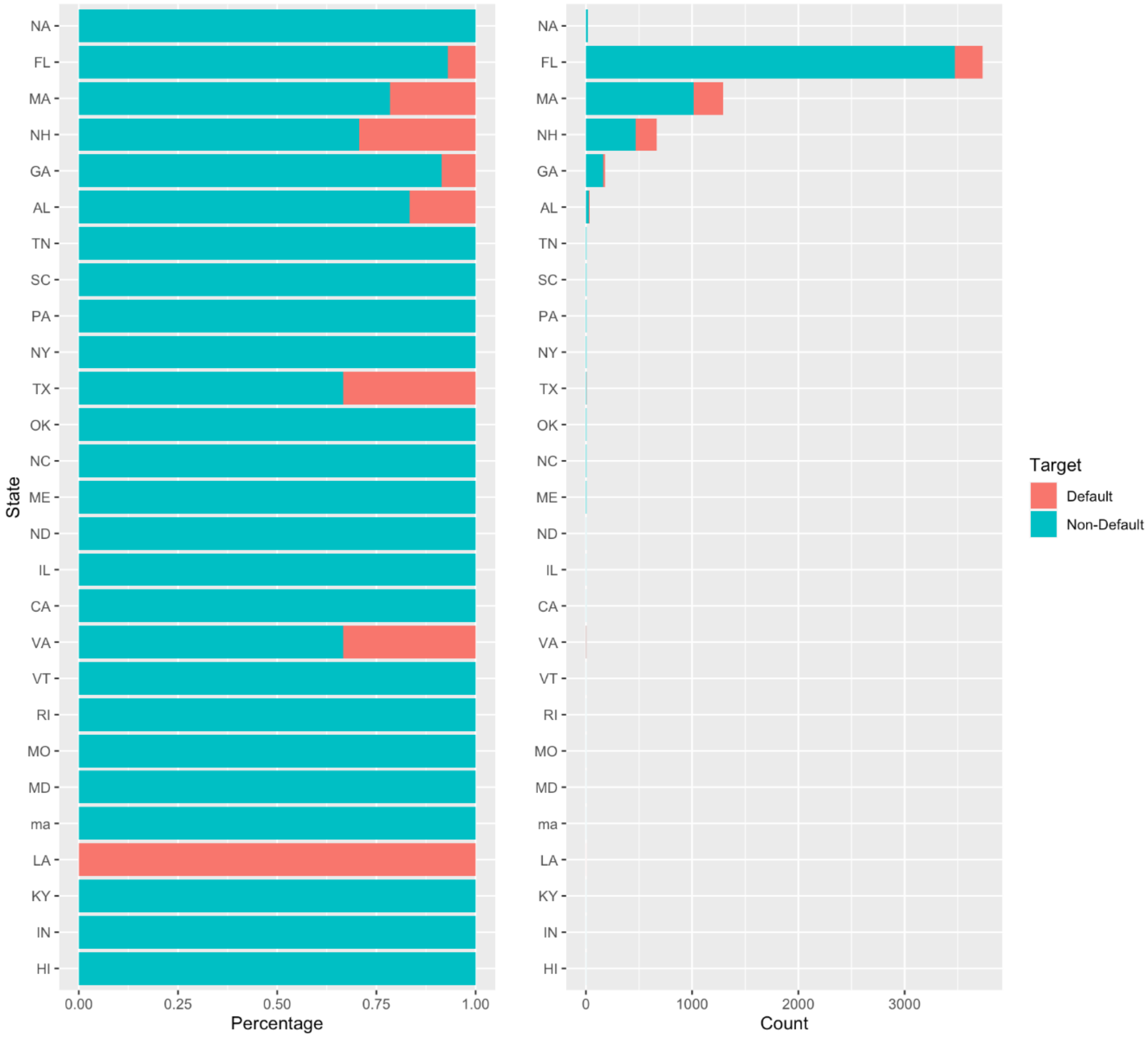


Figure 14: State Distribution

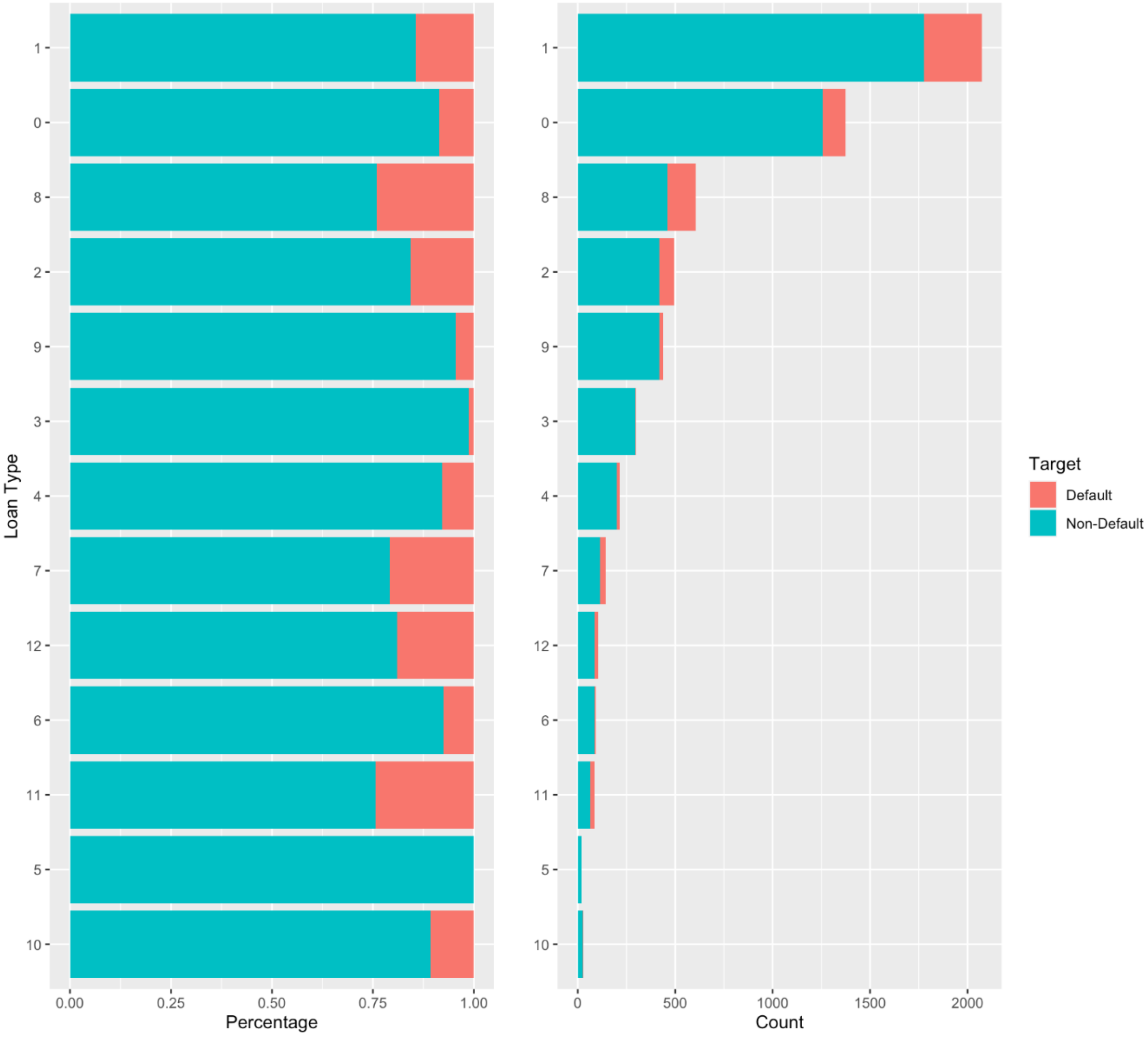


Figure 15: Default Percentage

5.4.2 Feature Pre-Processing for Modeling

The following features were excluded from our predictive modeling process: Customer ID (serve as ID); Office; City; Car Model; Email Domain (a new indicator feature Is Email Missing was created and included in the modeling process) All categorical/discrete features were converted as factor columns.

Overview of the remaining columns for the predictive modeling process:

1. Customer Name: This column contains the names of the customers. While it may not seem directly relevant to the predictive modeling process, it can potentially provide insights into customer behavior or preferences based on their names. For example, certain names may be associated with specific demographics or cultural backgrounds, which could impact their likelihood of purchasing a product or responding to a marketing campaign.

2. Age: This column represents the age of each customer. Age can be a significant factor in predicting customer behavior, as different age groups may have varying preferences, buying patterns, or levels of disposable income. By including age as a feature in the modeling process, you can potentially uncover age-related insights and improve the accuracy of predictions.

3. Gender: This column indicates the gender of each customer. Gender can also play a role in predicting customer behavior, as it can influence preferences, interests, or purchasing decisions. By including gender as a feature, you can identify any gender-based trends or patterns that may exist within your customer data.

4. Income: This column represents the income level of each customer. Income is a significant predictor in understanding customer behavior and purchasing power. By including income as a feature in the predictive modeling process, you can identify customer segments with higher disposable incomes and tailor marketing strategies or product offerings accordingly.

5. Purchase History: This column tracks the historical purchases made by each customer. Analyzing past purchase behavior can provide valuable insights into customer preferences, product preferences, or buying patterns. By including this feature, you can uncover trends or patterns in customer behavior that can help predict future purchase behavior.

6. Website Visits: This column measures the number of visits made by each customer to the company website. Tracking website visits can indicate customer engagement, interest, or intent to purchase. By including this information as a feature, you can understand the level of customer interaction with your website and potentially predict their likelihood

of making a purchase.

7. Purchase Amount: This column represents the amount spent by each customer on their purchases. The purchase amount is a crucial factor in understanding customer value and potential profitability. By including this feature in the modeling process, you can predict the potential revenue generated by each customer and segment them accordingly.

8. Time Since Last Purchase: This column measures the time elapsed since each customer's last purchase. The recency of a customer's purchase can indicate their level of engagement, satisfaction, or likelihood of making a repeat purchase. By including this feature, you can predict the likelihood of a customer returning to make another purchase based on their previous buying behavior.

By considering these remaining columns in the predictive modeling process, you can gain valuable insights and improve the accuracy of your predictions. These insights can help optimize marketing strategies, identify customer segments, and tailor product offerings to maximize customer satisfaction and profitability.

5.4.3 Data Splitting and Preparation for the modeling

During the predictive modeling process, the total 5,976 records were divided into a specific setup to ensure robust model training and evaluation. The data splitting technique used was an 80% training data and 20% test data (hold-out data) split. This split allows for model training on a majority of the data while also reserving a portion for evaluating the trained model's performance.

Within the training data, 10 folds of data splits were created. This technique, known as cross-validation, provides a more reliable estimation of model performance by rotating the data subsets used for training and validation. Each fold acts as an independent training and validation set, allowing for hyper-parameter tuning, such as selecting the best set of model parameters.

Prior to applying the predictive modeling techniques, several pre-processing steps were performed on the training data. These steps aim to ensure the quality and integrity of the data used for training the models:

1. Assigning missing values: Any missing values in the dataset were labeled as "NA". By assigning a specific category to missing values, it allows the models to differentiate between known and unknown values during the training process.

2. Collapsing categories: To simplify the data and reduce the complexity of the models, categories with small volumes were collapsed. A threshold of 5% was used, meaning categories that accounted for less than 5% of the total data were merged into a single

category. This helps in reducing noise in the data and prevents overfitting to specific rare categories.

3. Imputing missing values for numerical features: For numerical features that had missing values, the missing values were imputed using the median of the feature. This strategy is commonly used to fill in missing values as it provides a robust central tendency measure that is less sensitive to outliers.

By performing these pre-processing steps, the training data is prepared and cleaned, ensuring that the models are trained on high-quality and standardized data. Once the models are trained, the test data can be used to evaluate their performance on unseen data to assess their generalization capabilities.

5.5 Predictive Modeling using Random Forest

In this chapter, our aim is to build a predictive model to determine the probability of the Default event. To achieve this, we have decided to use the random forest algorithm for several reasons.

Firstly, we prioritize model predictive power and accuracy and aim to develop a prototype quickly. Random forest models are known for their excellent predictive performance and have been widely used in various industries. By choosing this algorithm, we can leverage its ability to generate accurate predictions efficiently.

Secondly, random forest models do not require complicated data pre-processing steps to perform well. Unlike some other machine learning algorithms, such as logistic regression, which often require feature scaling or transformation, random forest models can effectively handle the data without such preprocessing requirements. This simplifies our workflow and saves valuable time during the modeling process.

Additionally, random forest models are well-suited for handling non-linearity and feature interaction effects. They can capture complex relationships between input variables and the target variable, making them suitable for our prediction task.

We have decided to fix the number of trees in the random forest ensemble to be 1000. This ensures that we have a sufficiently large and diverse array of trees contributing to the final prediction. Each tree in the ensemble provides an independent prediction, and the final prediction is determined by combining the outputs of all the trees.

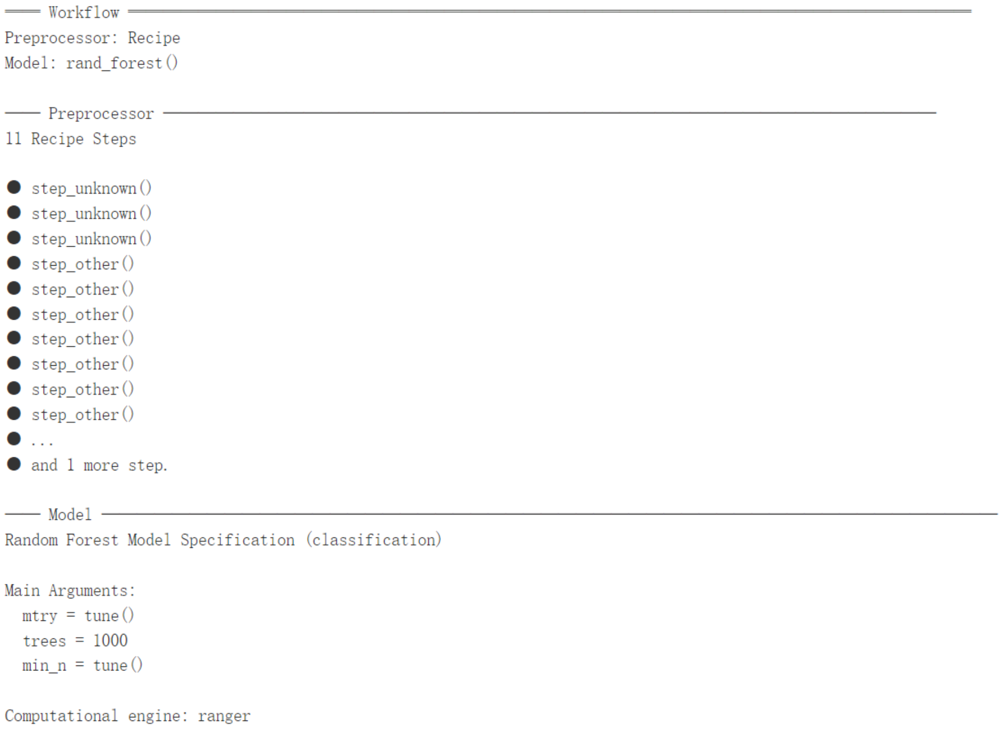
However, there are two hyper-parameters that we need to tune in the random forest model: mtry and $\min_n$.

The mtry parameter determines the number of predictors randomly sampled at each split when creating the tree models. It controls the level of randomness and diversity in the forest. By tuning this parameter, we can find the optimal balance between overfitting and underfitting.

The $\min_n$ parameter, on the other hand, specifies the minimum number of data points required in a node for it to be further split. This parameter helps control the complexity of the trees and prevents overfitting.

To tune these hyper-parameters, we will set up a comprehensive workflow for our modeling process. This workflow will encompass various steps such as data preprocessing, feature selection, model training, hyperparameter tuning, and model evaluation. By following this systematic approach, we can develop an effective random forest model for predicting the probability of the Default event.

5.5.1 Workflow and hyper-parameter setup



```
══ Workflow ════════════════════════════════════════════════════════════════════
Preprocessor: Recipe
Model: rand_forest()

── Preprocessor ────────────────────────────────────────────────────────────────
11 Recipe Steps

● step_unknown()
● step_unknown()
● step_unknown()
● step_other()
● step_other()
● step_other()
● step_other()
● step_other()
● step_other()
● step_other()
● ...
● and 1 more step.

── Model ───────────────────────────────────────────────────────────────────────
Random Forest Model Specification (classification)

Main Arguments:
  mtry = tune()
  trees = 1000
  min_n = tune()

Computational engine: ranger
```

In order to optimize the performance of our random forest model, we conducted hyper-parameter tuning. This process involved exploring a range of values for various parameters to find the optimal combination that would lead to the best predictive results.

While we can't provide a detailed account of our entire exploration process due to the length it would take to generate this report, we can share the current candidate values we have chosen for the hyper-parameters. These values were selected after careful consideration and extensive testing to ensure they produce the most accurate and reliable results. Our method are based on some other research works, see [150], [151], [152], [153], and

[154].

By manually setting up the hyper-parameter tuning grid, we gained control over the specific values being tested and were able to prioritize those that had the most significant impact on the model's performance. This approach allowed us to focus on the parameters that are known to have a higher influence on the random forest algorithm, rather than wasting time testing irrelevant or less impactful values.

Through our exploration process, we analyzed various hyper-parameter configurations, assessing their effects on model accuracy, generalization, and computational efficiency. The chosen candidate values represent the best combination we have discovered so far, based on our extensive experimentation and evaluation.

It's worth noting that hyper-parameter tuning is an ongoing process, and further exploration and refinement of the grid may still be needed in order to improve the model's performance. However, the candidate values we have provided serve as a strong foundation for building an effective and accurate random forest model for predicting the probability of the Default event.

5.5.2 Hyper-parameter tuning

After manually setting up the hyper-parameter tuning grid, we proceeded to tune the 20 different combinations of hyper-parameters using the 10-fold cross-validation data that we had previously created.

To evaluate the performance of each combination, we selected the Area Under the Curve (AUC) as the performance metric. AUC is a commonly used metric in binary classification tasks and provides a comprehensive assessment of the model's overall performance.

For each combination of hyper-parameters, we trained the model on the training data of each fold and evaluated its performance on the corresponding validation data. This process was repeated for all 10 folds, allowing us to obtain a robust and unbiased assessment of the model's performance across different subsets of the data.

Once the model was trained and evaluated for each combination of hyper-parameters, we plotted the performance metric (AUC) against the different parameter values. This visualization helped us gain insights into how changing the values of the hyper-parameters affected the model's performance.

By analyzing the plot, we could identify the hyper-parameter combinations that resulted in the highest AUC values. These combinations represented the optimal settings for our random forest model and allowed us to achieve the best possible performance in

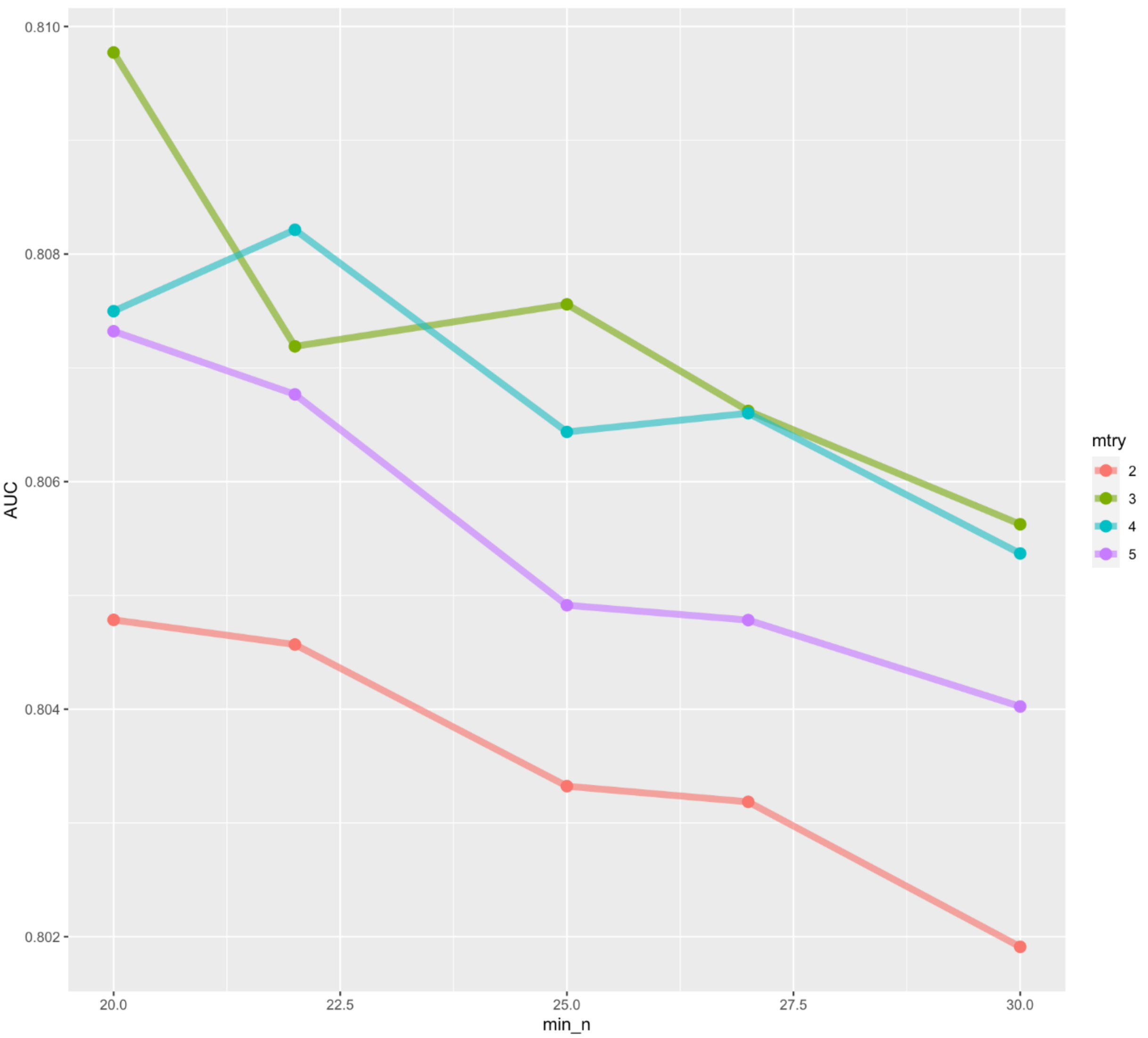


Figure 16: AUC comparison

terms of predicting the probability of the Default event.

It's important to note that during this process, we did not include all of our previous exploration procedures. This selective approach was taken to focus on the most promising hyper-parameter combinations and save computational resources.

Overall, this thorough and systematic tuning of hyper-parameters, combined with the use of cross-validation, allowed us to identify the best-performing model configuration, leading to high predictive accuracy and reliability in predicting the probability of the Default event

5.6 Best hyper-Parameters combination and variables importance

Based on the results from the previous section, the best hyper-parameter combination was determined to be (3, 20), which resulted in an AUC of 80.10%. This combination was selected as it provided the highest predictive performance for classifying Default from non-Default cases. The final workflow of the predictive modeling process was updated with this hyper-parameter combination as it was deemed to be the most optimal.

In addition to determining the best hyper-parameter combination, a predictive feature importance ranking was generated before fitting the final model. This ranking provides valuable insights into the contribution of each feature in classifying Default from non-Default. The top predictive features, based on their importance, were identified as State, PhoneStatus, Maturity, CarValue, and LoanAmount.

Furthermore, it was recognized that knowing the year in which the title loan originated would be useful as it would enable the calculation of the vehicle age using the CarYear variable. Vehicle age has the potential to be a significantly more predictive feature than just using CarYear alone. Considering this, it may be beneficial to include the vehicle age as a feature in future iterations of the model for improved performance and accuracy.

5.7 Final model fit and performance evaluation

After completing the model training process using the entire training dataset, we fit the final model. Notably, the model achieved an AUC of 80.10% on the hold-out test set. This result indicates that the model generalizes well to unseen data and does not show any signs of overfitting.

The final output of the model is the estimated probability of Default for each potential customer. This predicted probability can be directly used to rank order the credit risk of these customers. Additionally, a hard threshold can be set to assign customers to either the Default or non-Default class. This threshold can be chosen based on specific requirements and key performance indicators (KPIs), such as the desired approval rate.

For instance, by selecting an appropriate threshold, we can ensure that the approval rate reaches a desired level while effectively managing credit risk. This approach allows for more informed decisions regarding potential customers and enhances the overall efficiency of the credit evaluation process. See the pictures 18 and 19.

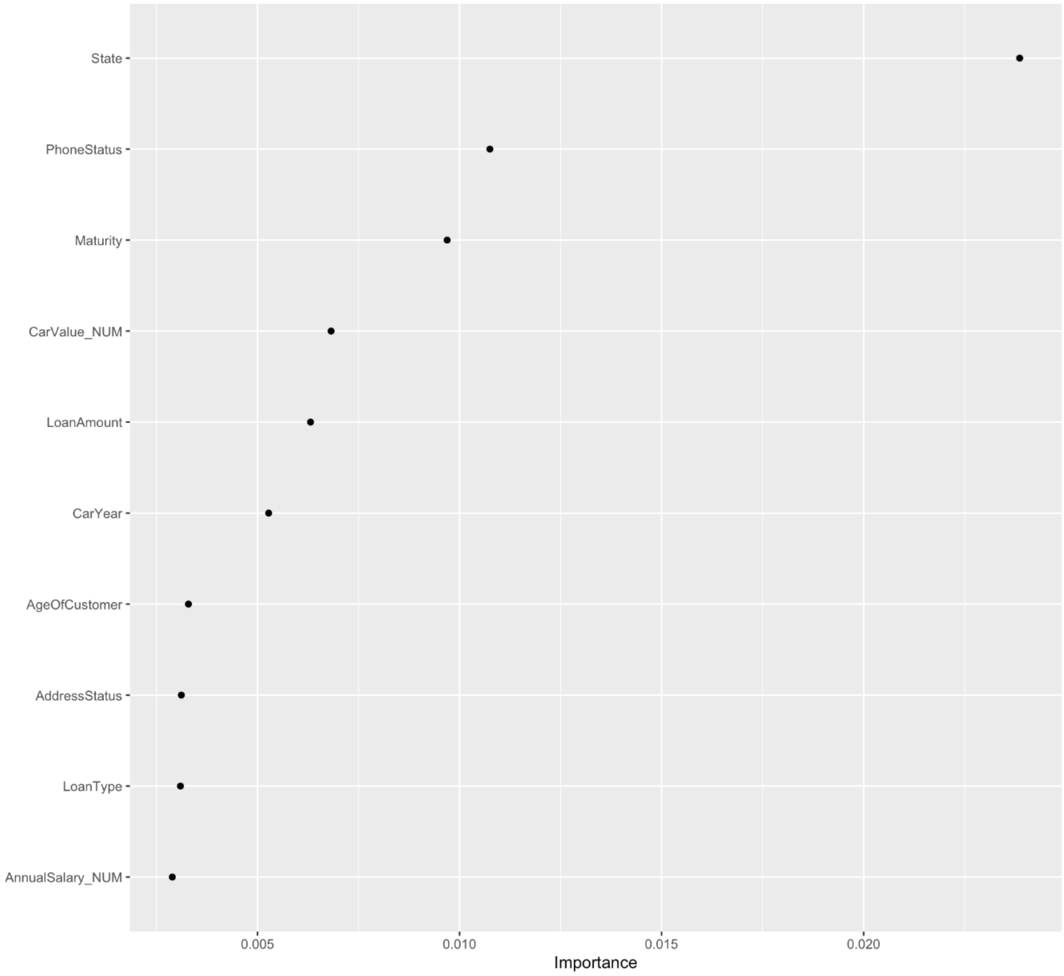


Figure 17: Factor importance analysis

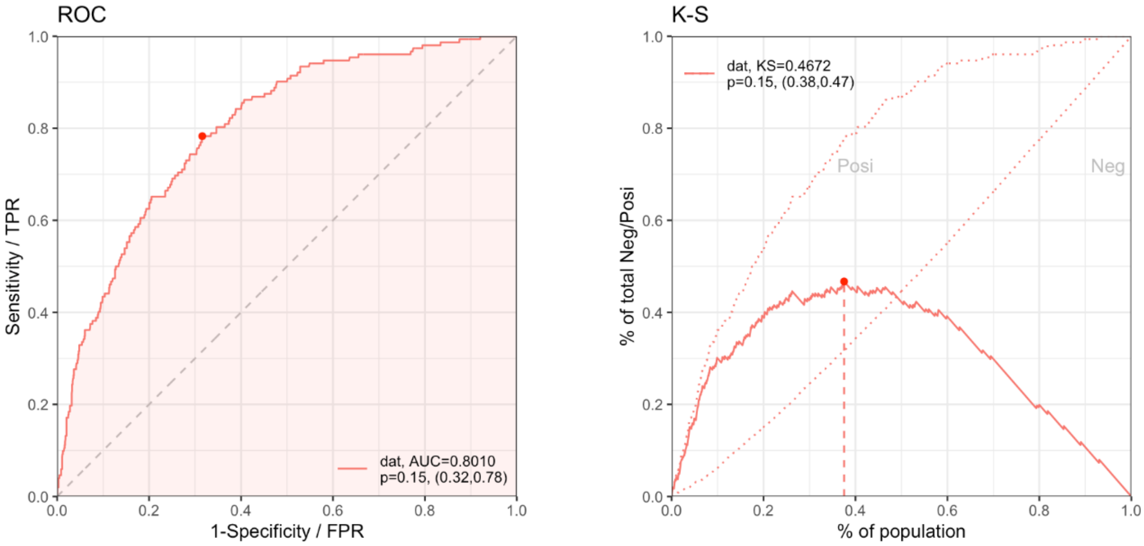


Figure 18: ROC/K-S

(most common two metrics to measure the performance of the classification model)

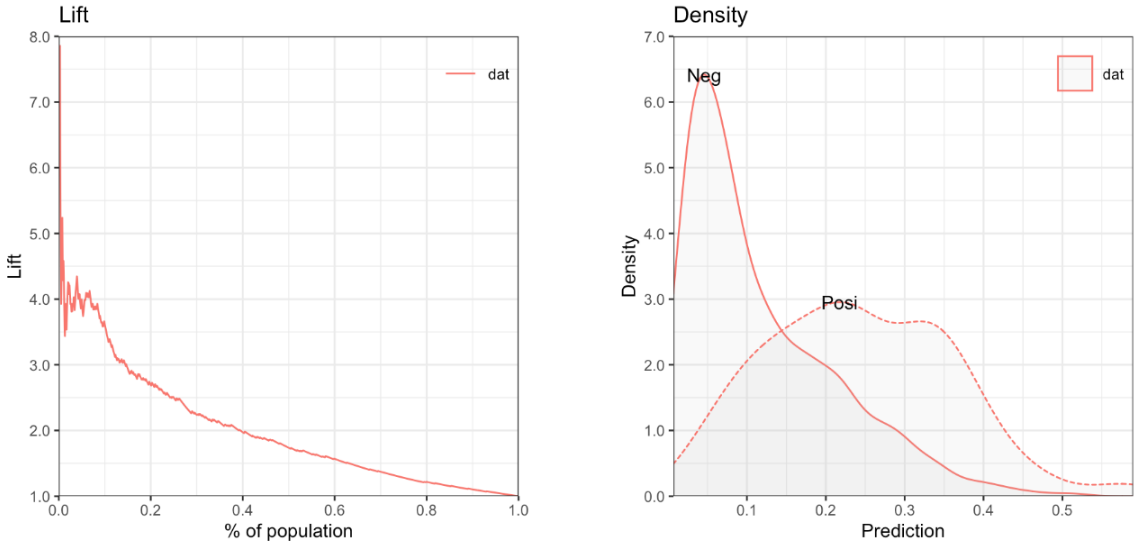


Figure 19: Lift and Density

5.8 Further Thought

If we are satisfied with the performance of our random forest model as shown in the previous chapter, we may consider exploring simpler models to potentially achieve the same level of model performance while gaining more transparency and interpretability. One such model to consider is logistic regression.

While random forest is an effective ensemble model, it still raises some questions that we would like to address. Firstly, feature selection is still necessary as some features may not contribute significantly to the model fit. Additionally, it is important to investigate whether there are any interaction effects between predictors that need to be considered.

In light of these considerations, the Multivariate Adaptive Regression Splines (MARS) algorithm seems like a suitable choice for our next steps, before attempting even simpler models like logistic regression. MARS offers several advantages over other algorithms:

1. Automatic Feature Selection: MARS handles feature selection automatically. This means that the algorithm will identify and include only the most relevant features in the model, helping to improve performance and reduce overfitting.

2. Consideration of Interaction Features: MARS includes a hyper-parameter that controls whether to consider interaction features. This allows us to explicitly capture any interactions between predictors, which can lead to improved model performance and a better understanding of relationships between variables.

3. Flexibility and Minimal Feature Pre-processing: MARS is more flexible than linear models and often requires little to no feature pre-processing. This makes it a convenient

choice when working with complex datasets that may have non-linear relationships between variables.

4. Improved Interpretability: Compared to ensemble models like random forest, MARS is generally easier to interpret. The resulting model can be visualized with clear and concise rules, making it easier to understand and explain to stakeholders.

By considering the advantages of MARS, we can potentially achieve comparable model performance while gaining additional insights and interpretability.

6 Conclusion and Future thought

6.1 Conclusions

There are many models could be used for the risk control purpose. In the paper, we introduce an innovative and comprehensive approach to predict the movement of loans across different delinquency states in Chapter 3. We utilize a Markov chain model to effectively capture the transitions between various delinquency states, providing valuable insights into the loan life-cycle and factors influencing its evolution.

A significant portion of this paper focuses on fitting loan-level Bayesian logistic models using the Variational Bayes framework. This allows us to estimate essential transition probabilities. Specifically, we construct two models to predict the likelihood of loans transitioning from the 'Current' status to either 'Charged-off' or 'Fully-paid'. By analyzing loan-level data, we gain a granular perspective on these transitions, accurately estimating the probabilities of different outcomes.

To develop a robust framework applicable to risk management for lenders, we combine the Markov chain delinquency matrix with the estimated loan-level models for key transition probabilities. The primary strength of our Bayesian model lies in its ability to quantify uncertainty in predictions. Unlike traditional models that deliver single point-estimates, our Bayesian approach generates a distribution of potential outcomes. This allows lenders to assess a range of scenarios and associated risks. For example, our model enables lenders to estimate the necessary cash reserves based on predicted charged-off and paid-off rates, aiding in effective risk control measures.

The combination of the Markov chain model and loan-level Bayesian logistic models provides a powerful tool for understanding and predicting loan behaviors. Our approach facilitates informed decision-making in risk management, highlighting the value of advanced statistical modeling in the financial sector. It is important to acknowledge that all models are simplifications of reality and come with inherent uncertainties. However, our Bayesian approach, with its explicit handling of uncertainty, offers a more robust and transparent framework for forecasting and risk management in the lending business.

This research developed an intelligent ML-based Markov chain model to predict loan customers' delinquency possibilities. Within the model, a consecutive incremental batch learning framework has been developed and some new algorithms have been designed to boost the performance on the minority class. The results of this study provide a more accurate and effective machine learning-based prediction method and make a significant contribution to the credit risk area. Based on the above revised Markov chain model, we explore an effective approach for scheduling collection actions on consumer term-loan accounts in Chapter 4. We set up a new collection model and combine it with the Markov chain. Then we can give an estimation of the profit of the customs to help the loan company to maximize its profit.

In Chapter 5 of our study, we focused on exploring and comparing the performance of the random forest model with the logistic regression model. By examining the strengths and weaknesses of these two algorithms, we aimed to determine which one is more suitable for our specific analysis.

Upon applying the random forest model to our example dataset, we observed a significantly improved level of accuracy compared to the logistic regression model. This finding highlights the superiority of the random forest approach in predicting the movement of loans across different delinquency states.

Moving forward, we proceeded to fit the final random forest model using the entire training dataset. The model's performance was evaluated using the hold-out test set, where it achieved an impressive AUC score of 80.10%. This high AUC score indicates the model's robustness and its ability to generalize well when encountering unseen data. Importantly, this outcome provides assurance that the model is not overfitting the training data.

The final output generated by the model is the estimated probability of Default for each potential customer. This probability serves as a valuable tool for ranking the credit risk of these customers. Additionally, by setting an appropriate threshold, we can classify customers into the Default or non-Default class based on specific requirements and key performance indicators (KPIs). This threshold can be selected to balance factors such as the desired approval rate, allowing for more informed decision-making and risk management.

6.2 Future thoughts

We would like to highlight an interesting observation associated with our research on mean field variational Bayes. It has come to our attention that this method tends to underestimate the posterior variance. This presents an intriguing avenue for future research to investigate and address this issue.

To further explore potential solutions, we are planning to study the application of the linear response variational Bayes (LRVB) method proposed by Giordano et al. (2015). This method could potentially be integrated into the framework we have developed in this article. The linear response variational Bayes (LRVB) method is utilized for several reasons, mainly to overcome the limitations of mean field variational Bayes (MFVB) and improve the accuracy of approximate Bayesian inference. There are many key advantages and motivations for using the LRVB method. For example, LRVB addresses the issue of underestimating posterior variances that is commonly observed in MFVB. By considering the linear response approximation, LRVB provides more accurate inference of posterior distributions; By accurately estimating posterior variances, LRVB enables better uncertainty quantification in Bayesian inference. This is particularly important in cases where accurate estimation of the uncertainty associated with model parameters or predictions is crucial; LRVB can handle more complex models compared to conventional variational inference techniques. It can capture non-linear relationships and interactions between variables effectively, making it suitable for high-dimensional data analysis and complex modeling scenarios; LRVB strikes a balance between computational efficiency and accuracy. While it may require more computational resources compared to MFVB, it offers improved accuracy without incurring a significant increase in computational time; The LRVB method represents a promising area for further research and development in approximate Bayesian inference. Researchers can explore and extend the method to address various challenges and improve its performance in different modeling scenarios. Therefore, the use of the LRVB method addresses the limitations of MFVB and offers improved accuracy and uncertainty quantification in Bayesian inference, making it a valuable tool for researchers and practitioners. By applying LRVB, we aim to evaluate whether it can produce an approximation of the posterior variance that is closer to the true variance, as compared to our proposed method.

To conduct a comprehensive analysis and gain meaningful insights, our research involves comparing the performance of three different approaches. These approaches are designed to approximate the posterior variance accurately and offer potential enhancements compared to existing alternatives. The three approaches we will be comparing are as follows:

1. Normal conjugate prior with the Markov Chain Monte Carlo (MCMC) procedure: In this approach, we utilize a normal conjugate prior distribution and employ the MCMC procedure to estimate the posterior variance. MCMC is a powerful sampling technique that allows us to draw samples from the posterior distribution and estimate its variance.

2. Normal conjugate prior with the Linear Response Variational Bayes (LRVB) method: Similar to the previous approach, we use a normal conjugate priomploy the LRVB method. LRVB is known for its ability to approximate the posterior distribution efficiently while also providing accurate estimates of the posterior variance.

3. Intrinsic prior with the Linear Response Variational Bayes (LRVB) method: In this

approach, we deviate from using a normal conjugate prior and instead employ an intrinsic prior. The intrinsic prior is based on specific domain knowledge or prior beliefs about the data. By using the LRVB method with an intrinsic prior, we aim to explore the potential benefits of incorporating domain-specific information into the model (see [155], [156], [157], [158], [122], [159] and [160] for more references about the Bayesian theory).

By directly comparing these three approaches, we can assess their performance in terms of accurately approximating the posterior variance. This analysis will enable us to draw meaningful conclusions about the efficacy and potential enhancements of our proposed LRVB method compared to the traditional MCMC approach and the variant with the intrinsic prior. These insights will contribute to advancing our understanding of Bayesian inference and provide valuable guidance for future research and model development.

List of Figures

References

[1] A. L. Kun Wang and L. S. Xiaokun Wang, “Study on credit risk control by variational inference,” *Springer Lecture Notes in Computer Science*, 2023.

[2] Y. Qin, Q. Sheng, N. Falkner, S. Dustdar, and H. Wang, “When things matter: A data-centric view of the internet of things,” Jul. 2014.

[3] X. Sun, M. Li, and H. Wang, “A family of enhanced (l,)-diversity models for privacy preserving data publishing,” *Future Generation Comp. Syst.*, vol. 27, pp. 348–356, Mar. 2011. DOI: 10.1016/j.future.2010.07.007.

[4] F. V. Vila Verde, “Peer-to-peer lending: Evaluation of credit risk using machine learning,” Ph.D. dissertation, 2021.

[5] J. M. Bravo, “The demographics of defense and security in japan,” in *Developments and Advances in Defense and Security*, Springer, 2022, pp. 359–370.

[6] X. Zhang, G. Zhang, X. Qiu, *et al.*, “Optimizing the size of peritumoral region for assessing non-small cell lung cancer heterogeneity using radiomics,” in *International Conference on Health Information Science*, Springer, 2023, pp. 309–320.

[7] J. E. J. Neto, “Modeling the impact of the volatility of the perceived counterparty credit risk on hedge accounting effectiveness,” Ph.D. dissertation, 2019.

[8] R. Singh, S. Subramani, J. Du, *et al.*, “Antisocial behavior identification from twitter feeds using traditional machine learning algorithms and deep learning.,” *ICST Transactions on Scalable Information Systems*, e17, May 2023. DOI: 10.4108/eetsis.v10i3.3184.

[9] R. Singh, S. Subramani, J. Du, *et al.*, “Deep learning for multi-class antisocial behavior identification from twitter,” *IEEE Access*, vol. 8, pp. 194 027–194 044, 2020. DOI: 10.1109/ACCESS.2020.3030621.

[10] K. Sultana, K. Ahmed, B. Gu, and H. Wang, “Elastic optimization for stragglers in edge federated learning,” *Big Data Mining and Analytics*, 2023. DOI: 10.26599/BDMA.2022.9020046. [Online]. Available: https://www.sciopen.com/article/10.26599/BDMA.2022.9020046.

[11] R. Sarki, K. Ahmed, H. Wang, Y. Zhang, and K. Wang, “Convolutional neural network for multi-class classification of diabetic eye disease,” *EAI Endorsed Transactions on Scalable Information Systems*, vol. 9, no. 4, e5, Dec. 2021. DOI: 10.4108/eai.16-12-2021.172436. [Online]. Available: https://publications.eai.eu/index.php/sis/article/view/315.

[12] D. Pandey, H. Wang, X. Yin, K. Wang, Y. Zhang, and J. Shen, “Automatic breast lesion segmentation in phase preserved dce-mris,” *Health Information Science and Systems*, vol. 10, May 2022. DOI: 10.1007/s13755-022-00176-w.

[13] J. Yin, M. Tang, J. Cao, M. You, H. Wang, and M. Alazab, “Knowledge-driven cybersecurity intelligence: Software vulnerability co-exploitation behaviour discovery,” *IEEE Transactions on Industrial Informatics*, pp. 1–9, Jan. 2022. DOI: 10.1109/TII.2022.3192027.

[14] K. Gerardi, K. Herkenhoff, L. E. Ohanian, and P. Willen, “Unemployment, negative equity, and strategic default,” *Available at SSRN 2293152*, 2013.

[15] S. D. Grimshaw and W. P. Alexander, “Markov chain models for delinquency: Transition matrix estimation and forecasting,” *Applied Stochastic Models in Business and Industry*, vol. 27, no. 3, pp. 267–279, 2011.

[16] C. L. Brown and W. Simpson, “The cost of racially equal approval rates in mortgage lending,” *International Review of Economics Finance*, vol. 12, no. 4, pp. 467–480, 2003, ISSN: 1059-0560. DOI: https://doi.org/10.1016/S1059-0560(03)00017-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1059056003000170.

[17] J. B. Kau, D. C. Keenan, W. J. Muller, and J. F. Epperson, “A generalized valuation model for fixed-rate residential mortgages,” *Journal of money, credit and banking*, vol. 24, no. 3, pp. 279–299, 1992.

[18] Y. Deng, J. M. Quigley, and R. Van Order, “Mortgage terminations, heterogeneity and the exercise of mortgage options,” *Econometrica*, vol. 68, no. 2, pp. 275–307, 2000.

[19] L. I. Nakamura and J. Steinbuks, “Mortgage modification and strategic behavior: Evidence from a legal settlement with countrywide,” *Journal of Urban Economics*, vol. 92, pp. 77–91, 2016.

[20] L. Guiso, P. Sapienza, and L. Zingales, “Moral and social constraints to strategic default on mortgages,” National Bureau of Economic Research, Tech. Rep., 2009.

[21] T. Jones, D. Gatzlaff, and G. S. Sirmans, “Housing market dynamics: Disequilibrium, mortgage default, and reverse mortgages,” *The Journal of Real Estate Finance and Economics*, vol. 53, no. 3, pp. 269–281, 2016.

[22] S. Chan, C. Sharygin, V. Been, and A. Haughwout, “Pathways after default: What happens to distressed mortgage borrowers and their homes?” *The Journal of Real Estate Finance and Economics*, vol. 48, pp. 342–379, 2014.

[23] C. Y. Tian, R. G. Quercia, and S. Riley, "Unemployment as an adverse trigger event for mortgage default," *The Journal of Real Estate Finance and Economics*, vol. 52, no. 1, pp. 28–49, 2016.

[24] M. S. Sirajudeen, H. Muthusamy, M. Alqahtani, M. Waly, and A. K. Jilani, "Computer-related health problems among university students in majmaah region, saudi arabia," *Biomedical Research*, vol. 29, no. 11, pp. 2405–2415, 2018.

[25] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[26] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[27] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, pp. 3–42, 2006.

[28] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[29] D. A. Levin and Y. Peres, *Markov chains and mixing times*. American Mathematical Soc., 2017, vol. 107.

[30] R. P. Dobrow, *Introduction to stochastic processes with R*. John Wiley & Sons, 2016.

[31] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[32] P. Brémaud, *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Springer Science & Business Media, 2001, vol. 31.

[33] D. P. Scollnik, "An introduction to markov chain monte carlo methods and their actuarial applications," in *Proceedings of the Casualty Actuarial Society*, Citeseer, vol. 83, 1996, pp. 114–165.

[34] A. Li, L. Pericchi, and K. Wang, "Objective bayesian inference in probit models with intrinsic priors using variational approximations," *Entropy*, vol. 22, no. 5, p. 513, 2020.

[35] J. G. Scott and J. O. Berger, "An exploration of aspects of bayesian multiple testing," *Journal of Statistical Planning and Inference*, vol. 136, no. 7, pp. 2144–2162, 2006.

[36] J. Grimmer, "An introduction to bayesian inference via variational approximations," *Political Analysis*, mpq027, 2010.

[37] G. Consonni and J.-M. Marin, "Mean-field variational approximate bayesian inference for latent variable models," *Computational Statistics & Data Analysis*, vol. 52, no. 2, pp. 790–798, 2007.

[38] M. J. Beal, *Variational algorithms for approximate Bayesian inference*. University College London, 2003.

[39] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[40] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[41] H. Wold, "Soft modelling: The basic design and some extensions," *Systems under indirect observation, Part II*, pp. 36–37, 1982.

[42] N. Bhutta, J. Dokko, and H. Shan, "Consumer ruthlessness and mortgage default during the 2007 to 2009 housing bust," *The Journal of Finance*, vol. 72, no. 6, pp. 2433–2466, 2017.

[43] H. J. Noh, T. H. Roh, and I. Han, "Prognostic personal credit risk model considering censored information," *Expert Systems with Applications*, vol. 28, no. 4, pp. 753–762, 2005.

[44] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.

[45] P. Bajari, C. S. Chu, and M. Park, "An empirical model of subprime mortgage default from 2000 to 2007," National Bureau of Economic Research, Tech. Rep., 2008.

[46] P. Bajari, C. S. Chu, D. Nekipelov, and M. Park, "A dynamic model of subprime mortgage default: Estimation and policy implications," National Bureau of Economic Research, Tech. Rep., 2013.

[47] X. Pang, Y.-F. Ge, K. Wang, A. Traina, and H. Wang, "Patient assignment optimization in cloud healthcare systems: A distributed genetic algorithm," *Health Information Science and Systems*, vol. 11, Jun. 2023. DOI: 10.1007/s13755-023-00230-1.

[48] W. Hong, J. Yin, M. You, *et al.*, "A graph empowered insider threat detection framework based on daily activities," *ISA Transactions*, 2023, ISSN: 0019-0578. DOI: https://doi.org/10.1016/j.isatra.2023.06.030. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019057823002975.

[49] F. Khalil, J. Li, and H. Wang, "A framework of combining markov model with association rules for predicting web page accesses.," in *Proceedings of the Fifth Australasian Data Mining Conference (AusDM2006)*, vol. 61, Jan. 2006, pp. 177–184.

[50] F. Khalil, J. Li, and H. Wang, "An integrated model for next page access prediction," *I. J. Knowledge and Web Intelligence*, vol. 1, Jan. 2009. DOI: 10.1504/IJKWI.2009.027925.

[51] M. S. Akhtar and T. Feng, "Comparison of classification model for the detection of cyber-attack using ensemble learning models," *ICST Transactions on Scalable Information Systems*, vol. 9, p. 173 293, Feb. 2022. DOI: 10.4108/eai.1-2-2022.173293.

[52] D. Perumal, S. L.R, K. Kalaivani, and J. Ganesh, "Scene classification of remotely sensed images using optimized rsisc-16 net deep convolutional neural network model," *ICST Transactions on Scalable Information Systems*, vol. 9, p. 173 292, Feb. 2022. DOI: 10.4108/eai.1-2-2022.173292.

[53] J. Gao, "Retracted: Basketball posture recognition based on hog feature extraction and convolutional neural network [eai endorsed scal inf syst (2022), online first]," *ICST Transactions on Scalable Information Systems*, vol. 9, p. 173 787, Apr. 2022. DOI: 10.4108/eai.8-4-2022.173787.

[54] Y. Zhang and Y. Yuan, "A novel dilated convolutional neural network model for road scene segmentation," *ICST Transactions on Scalable Information Systems*, vol. 9, p. 173 164, Jan. 2022. DOI: 10.4108/eai.27-1-2022.173164.

[55] R. Li, "A novel image clustering method based on coupled convolutional and graph convolutional network," *ICST Transactions on Scalable Information Systems*, vol. 9, p. 172 132, Nov. 2021. DOI: 10.4108/eai.16-11-2021.172132.

[56] A. Ju and Z. Wang, "Convolutional block attention module based on visual mechanism for robot image edge detection," *ICST Transactions on Scalable Information Systems*, vol. 9, p. 172 214, Nov. 2021. DOI: `10.4108/eai.19-11-2021.172214`.

[57] C. Wang, B. Sun, K.-J. Du, *et al.*, "A novel evolutionary algorithm with column and sub-block local search for sudoku puzzles," *IEEE Transactions on Games*, vol. PP, pp. 1–11, Jan. 2023. DOI: `10.1109/TG.2023.3236490`.

[58] M. N. A. Tawhid, S. Siuly, K. Wang, and H. Wang, "Automatic and efficient framework for identifying multiple neurological disorders from eeg signals," *IEEE Transactions on Technology and Society*, vol. PP, pp. 1–1, Mar. 2023. DOI: `10.1109/TTS.2023.3239526`.

[59] G. Arminger, D. Enache, and T. Bonne, "Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks," *Computational Statistics*, vol. 12, no. 2, 1997.

[60] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: A review," *Journal of the royal statistical society: series a (statistics in society)*, vol. 160, no. 3, pp. 523–541, 1997.

[61] J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler, "Consumer credit risk: Individual probability estimates using machine learning," *Expert systems with applications*, vol. 40, no. 13, pp. 5125–5131, 2013.

[62] F. Butaru, Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique, "Risk and risk management in the credit card industry," *Journal of Banking & Finance*, vol. 72, pp. 218–239, 2016.

[63] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," *Journal of the operational research society*, vol. 54, pp. 627–635, 2003.

[64] J. Abellán and J. G. Castellano, "A comparative study on base classifiers in ensemble methods for credit scoring," *Expert systems with applications*, vol. 73, pp. 1–10, 2017.

[65] P. du Jardin and E. Séverin, "Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of a financial failure model," *Decision Support Systems*, vol. 51, no. 3, pp. 701–711, 2011, ISSN: 0167-9236. DOI: `https://doi.org/10.1016/j.dss.2011.04.001`.

[Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923611001023.

[66] A. Volkov, D. F. Benoit, and D. Van den Poel, "Incorporating sequential information in bankruptcy prediction with predictors based on markov for discrimination," *Decision Support Systems*, vol. 98, pp. 59–68, 2017, ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2017.04.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923617300751.

[67] S. Ho Ha and R. Krishnan, "Predicting repayment of the credit card debt," *Computers Operations Research*, vol. 39, no. 4, pp. 765–773, 2012, Special Issue on Operational Research in Risk Management, ISSN: 0305-0548. DOI: https://doi.org/10.1016/j.cor.2010.10.032. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030505481000290X.

[68] V. Castro, "Macroeconomic determinants of the credit risk in the banking system: The case of the gipsi," *Economic modelling*, vol. 31, pp. 672–683, 2013.

[69] T. Bellotti and J. Crook, "Forecasting and stress testing credit card default using dynamic models," *International Journal of Forecasting*, vol. 29, no. 4, pp. 563–574, 2013.

[70] R. Chamboko and J. M. Bravo, "A multi-state approach to modelling intermediate events and multiple mortgage loan outcomes," *Risks*, vol. 8, no. 2, p. 64, 2020.

[71] R. Chamboko and R. K. Chamboko, "Consumer financial distress during economic downturn: Evidence from five provinces of zimbabwe," *International Journal of Social Economics*, 2020.

[72] J. M. Bravo, "Pricing participating longevity-linked life annuities: A bayesian model ensemble approach," *European Actuarial Journal*, pp. 1–35, 2021.

[73] N. Sarlija, M. Bensic, and M. Zekic-Susac, "Comparison procedure of predicting the time to default in behavioural scoring," *Expert Systems with Applications*, vol. 36, no. 5, pp. 8778–8788, 2009.

[74] E. N. Tong, C. Mues, and L. C. Thomas, "Mixture cure models in credit scoring: If and when borrowers default," *European Journal of Operational Research*, vol. 218, no. 1, pp. 132–139, 2012.

[75] R. Chamboko and J. M. Bravo, "Frailty correlated default on retail consumer loans in zimbabwe," *International Journal of Applied Decision Sciences*, vol. 12, no. 3, pp. 257–270, 2019.

[76] K. Gerardi, A. Shapiro, and P. Willen, "Subprime outcomes: Risky mortgages, homeownership and foreclosure," Technical report, Federal Reserve Bank of Atlanta Working Paper 07-15, Tech. Rep., 2007.

[77] S. H. Ha and R. Krishnan, "Predicting repayment of the credit card debt," *Computers & Operations Research*, vol. 39, no. 4, pp. 765–773, 2012.

[78] R. Chamboko and J. M. Bravo, "On the modelling of prognosis from delinquency to normal performance on retail consumer loans," *Risk Management*, vol. 18, pp. 264–287, 2016.

[79] R. Chamboko and J. M. Bravo, "Modelling and forecasting recurrent recovery events on consumer loans," *International Journal of Applied Decision Sciences*, vol. 12, no. 3, pp. 271–287, 2019.

[80] S. H. Ha, "Behavioral assessment of recoverable credit of retailer's customers," *Information Sciences*, vol. 180, no. 19, pp. 3703–3717, 2010.

[81] J. Miguel Bravo, "Pricing survivor bonds with affine-jump diffusion stochastic mortality models," in *2021 The 5th International Conference on E-Commerce, E-Business and E-Government*, 2021, pp. 91–96.

[82] Y. Deng, J. M. Quigley, R. Van Order, and F. Mac, "Mortgage default and low downpayment loans: The costs of public subsidy," *Regional science and urban economics*, vol. 26, no. 3-4, pp. 263–285, 1996.

[83] M. Stepanova and L. Thomas, "Survival analysis methods for personal loan data," *Operations Research*, vol. 50, no. 2, pp. 277–289, 2002.

[84] M. Ncube and S. E. Satchell, "Modelling uk mortgage defaults using a hazard approach based on american options," Faculty of Economics, University of Cambridge, Tech. Rep., 1995.

[85] R. Kelly and T. O'Malley, "The good, the bad and the impaired: A credit risk model of the irish mortgage market," *Journal of Financial Stability*, vol. 22, pp. 1–9, 2016.

[86] J. Beyersmann, A. Allignol, and M. Schumacher, *Competing risks and multistate models with R*. Springer Science & Business Media, 2011.

[87] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[88] H. Wold *et al.*, "Estimation of principal components and related models by iterative least squares," *Multivariate analysis*, vol. 1, pp. 391–420, 1966.

[89] M. H. A. M. T. ElMasry, "Machine learning approach for credit score analysis: A case study of predicting mortgage loan defaults," Ph.D. dissertation, 2019.

[90] F. C. d. Almeida, "Loan modifications and risk of default: A markov chains approach," Ph.D. dissertation, 2020.

[91] F. Khalil, J. Li, and H. Wang, "Integrating markov model with clustering for predicting web page accesses," *AusWeb 2007: 13th Australasian World Wide Web Conference*, Jan. 2007.

[92] F. Khalil, J. Li, and H. Wang, "Integrating recommendation models for improved web page prediction accuracy," *Conferences in Research and Practice in Information Technology Series*, vol. 74, pp. 91–100, Jan. 2008.

[93] H. Hu, J. Li, H. Wang, G. Daggard, and M. Shi, "A maximally diversified multiple decision tree algorithm for microarray data classification," vol. 73, pp. 35–38, Dec. 2006.

[94] Y. Gong and G. Srivastava, "Multi-target trajectory tracking in multi-frame video images of basketball sports based on deep learning," *ICST Transactions on Scalable Information Systems*, e12, Oct. 2022. DOI: 10.4108/eetsis.v9i6.2591.

[95] Z. Lin and J. Lin, "Research on knowledge management of novel power system based on deep learning," *ICST Transactions on Scalable Information Systems*, e10, Oct. 2022. DOI: 10.4108/eetsis.v9i6.2634.

[96] J. Du, S. Michalska, S. Subramani, H. Wang, and Y. Zhang, "Neural attention with character embeddings for hay fever detection from twitter," *Health Information Science and Systems*, vol. 7, Oct. 2019. DOI: 10.1007/s13755-019-0084-2.

[97] Y. Zhou, Z. Lin, L. Tu, J. Huang, and Z. Zhang, "Analysis and design of standard knowledge service system based on deep learning," *ICST Transactions on Scalable Information Systems*, e11, Oct. 2022. DOI: 10.4108/eetsis.v9i6.2637.

[98] C.-z. Xiang, N.-x. Fu, and T. Gadekallu, “Design of resource matching model of intelligent education system based on machine learning,” *ICST Transactions on Scalable Information Systems*, vol. 9, p. 173 381, Feb. 2022. DOI: 10.4108/eai.10-2-2022.173381.

[99] H. Wang and L. Sun, “Trust-involved access control in collaborative open social networks,” Sep. 2010, pp. 239–246. DOI: 10.1109/NSS.2010.13.

[100] X. Sun, H. Wang, J. Li, and Y. Zhang, “Injecting purpose and trust into data anonymisation,” *Computers Security*, vol. 30, pp. 332–345, Jul. 2011. DOI: 10.1016/j.cose.2011.05.005.

[101] H. Zhihan, L. Yuan, and T. Jin, “Design of music training assistant system based on artificial intelligence,” *ICST Transactions on Scalable Information Systems*, vol. 9, p. 173 450, Feb. 2022. DOI: 10.4108/eai.11-2-2022.173450.

[102] H. Wang, X. Yi, E. Bertino, and L. Sun, “Protecting outsourced data in cloud computing through access management,” *Concurrency and Computation: Practice and Experience*, vol. 28, Apr. 2014. DOI: 10.1002/cpe.3286.

[103] M. Fatima, O. Rehman, and I. Rahman, “Impact of features reduction on machine learning based intrusion detection systems,” *ICST Transactions on Scalable Information Systems*, vol. 9, p. 447, Apr. 2022. DOI: 10.4108/eetsis.vi.447.

[104] S. Subramani, S. Michalska, H. Wang, J. Du, Y. Zhang, and M. Shakeel, “Deep learning for multi-class identification from domestic violence online posts,” *IEEE Access*, vol. 7, pp. 46 210–46 224, Jan. 2019. DOI: 10.1109/ACCESS.2019.2908827.

[105] X. Sun, H. Wang, J. Li, and Y. Zhang, “Satisfying privacy requirements before data anonymization,” *The Computer Journal*, vol. 55, Apr. 2012. DOI: 10.1093/comjnl/bxr028.

[106] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, “Learning under concept drift: A review,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2018.

[107] J. G. Scott and J. O. Berger, “Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem,” *The Annals of Statistics*, vol. 38, no. 5, pp. 2587–2619, 2010.

[108] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

[109] D. M. A. Haughton, “On the choice of a model to fit data from an exponential family,” *The Annals of Statistics*, vol. 16, no. 1, pp. 342–355, 1988.

[110] A. E. Gelfand and D. K. Dey, “Bayesian model choice: Asymptotics and exact calculations,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 501–514, 1994.

[111] B. Ding and R. Gentleman, “Classification using generalized partial least squares,” *Journal of Computational and Graphical Statistics*, vol. 14, no. 2, 2005.

[112] B. Sorić, “Statistical discoveries and effect-size estimation,” *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 608–610, 1989.

[113] M. You, J. Yin, H. Wang, *et al.*, “A knowledge graph empowered online learning framework for access control decision-making,” *World Wide Web*, vol. 26, pp. 1–22, Jun. 2022. DOI: 10.1007/s11280-022-01076-5.

[114] J. Yin, M. Tang, J. Cao, H. Wang, M. You, and Y. Lin, “Vulnerability exploitation time prediction: An integrated framework for dynamic imbalanced learning,” *World Wide Web*, Jan. 2022. DOI: 10.1007/s11280-021-00909-z.

[115] H. Li, Y. Wang, H. Wang, and B. Zhou, “Multi-window based ensemble learning for classification of imbalanced streaming data,” *World Wide Web*, vol. 20, pp. 1–19, Nov. 2017. DOI: 10.1007/s11280-017-0449-x.

[116] P. Diggle, P. J. Diggle, P. Heagerty, K.-Y. Liang, S. Zeger, *et al.*, *Analysis of longitudinal data*. Oxford university press, 2002.

[117] L. D. Smith and E. C. Lawrence, “Forecasting losses on a liquidating long-term loan portfolio,” *Journal of Banking & Finance*, vol. 19, no. 6, pp. 959–985, 1995.

[118] J. Kruschke, *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[119] L. Tierney and J. Kadane, “Accurate approximations for posterior moments and marginal densities,” *Journal of the American Statistical Association*, vol. 81, no. 393, pp. 82–86, 1986.

[120] M. A. Tanner and W. H. Wong, “The calculation of posterior distributions by data augmentation,” *Journal of the American statistical Association*, vol. 82, no. 398, pp. 528–540, 1987.

[121] R. McElreath, *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2018.

[122] J. O. Berger and L. R. Pericchi, "The intrinsic bayes factor for linear models," *Bayesian statistics*, vol. 5, pp. 25–44, 1996.

[123] R. J. Giordano, T. Broderick, and M. I. Jordan, "Linear response methods for accurate covariance estimates from mean field variational bayes," in *Advances in Neural Information Processing Systems*, 2015, pp. 1441–1449.

[124] T. S. Jaakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Statistics and Computing*, vol. 10, no. 1, pp. 25–37, 2000.

[125] D. Durante and T. Rigon, "Conditionally conjugate mean-field variational bayes for logistic models," 2019.

[126] N. G. Polson, J. G. Scott, and J. Windle, "Bayesian inference for logistic models using pólya–gamma latent variables," *Journal of the American statistical Association*, vol. 108, no. 504, pp. 1339–1349, 2013.

[127] E. E. Leamer, *Specification searches: Ad hoc inference with nonexperimental data*. Wiley New York, 1978, vol. 53.

[128] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[129] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[130] J. T. Ormerod and M. P. Wand, "Explaining variational approximations," *The American Statistician*, vol. 64, no. 2, pp. 140–153, 2010.

[131] M. Girolami and S. Rogers, "Variational bayesian multinomial probit regression with gaussian process priors," *Neural Computation*, vol. 18, no. 8, pp. 1790–1817, 2006.

[132] G. Parisi and R. Shankar, "Statistical field theory," *Physics Today*, vol. 41, p. 110, 1988.

[133] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[134] H. Akaike, "Information measures and model selection," *Bulletin of the International Statistical Institute*, vol. 50, no. 1, pp. 277–291, 1983.

[135] E. Rosenberg and A. Gleit, "Quantitative methods in credit management: A survey," *Operations research*, vol. 42, no. 4, pp. 589–613, 1994.

[136] E. I. Altman and A. Saunders, "Credit risk measurement: Developments over the last 20 years," *Journal of banking & finance*, vol. 21, no. 11-12, pp. 1721–1742, 1997.

[137] L. C. Thomas, R. W. Oliver, and D. J. Hand, "A survey of the issues in consumer credit modelling research," *Journal of the Operational Research Society*, vol. 56, pp. 1006–1015, 2005.

[138] D. K. Pauler, "The schwarz criterion and related methods for normal linear models," *Biometrika*, vol. 85, no. 1, pp. 13–27, 1998.

[139] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[140] A. Albert and J. A. Anderson, "On the existence of maximum likelihood estimates in logistic regression models," *Biometrika*, vol. 71, no. 1, pp. 1–10, 1984.

[141] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 267–288, 1996.

[142] M. Stone and R. Brooks, "Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 237–269, 1990.

[143] B. Marx, "Iteratively reweighted partial least squares estimation for generalized linear regression," *Technometrics*, vol. 38, no. 4, pp. 374–381, 1996.

[144] L. Breiman and P. Spector, "Submodel selection and evaluation in regression. the x-random case," *International statistical review/revue internationale de Statistique*, pp. 291–319, 1992.

[145] M. M. Barbieri and J. O. Berger, "Optimal predictive model selection," *Annals of Statistics*, pp. 870–897, 2004.

[146] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *The Journal of Machine Learning Research*, vol. 5, pp. 1089–1105, 2004.

[147] J. M. Pérez and J. O. Berger, "Expected-posterior prior distributions for model selection," *Biometrika*, vol. 89, no. 3, pp. 491–512, 2002.

[148] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, 1995, pp. 1137–1145.

[149] H. Akaike, “A new look at the statistical model identification,” *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.

[150] D. Salmeron, J. A. Cano, and C. P. Robert, “Objective bayesian hypothesis testing in binomial regression models with integral prior distributions,” *Statistica Sinica*, pp. 1009–1023, 2015.

[151] E. Moreno, F. Bertolino, and W. Racugno, “An intrinsic limiting procedure for model selection and hypotheses testing,” *Journal of the American Statistical Association*, vol. 93, no. 444, pp. 1451–1460, 1998.

[152] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[153] T. Sellke, M. J. Bayarri, and J. O. Berger, “Calibration of p-values for testing precise null hypotheses,” *The American Statistician*, vol. 55, no. 1, pp. 62–71, 2001.

[154] J. Mulder and L. R. Pericchi, “The matrix-F prior for estimating and testing covariance matrices,” *Bayesian Analysis*, vol. 13, no. 4, pp. 1193–1214, 2018.

[155] J. O. Berger and L. R. Pericchi, “The intrinsic bayes factor for model selection and prediction,” *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 109–122, 1996.

[156] J. Berger and L. Pericchi, “On the justification of default and intrinsic bayes factors,” in *Modelling and Prediction Honoring Seymour Geisser*, Springer, 1996, pp. 276–293.

[157] L. R. Pericchi, “Model selection and hypothesis testing based on objective probabilities and bayes factors,” *Handbook of statistics*, vol. 25, pp. 115–149, 2005.

[158] A. O’Hagan, “Fractional bayes factors for model comparison,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 99–138, 1995.

[159] R. E. Kass and A. E. Raftery, “Bayes factors,” *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[160] J. O. Berger and L. R. Pericchi, “Objective bayesian methods for model selection: Introduction and comparison,” in *Model selection*, Institute of Mathematical Statistics, 2001, pp. 135–207.