Consulting Capstone Project

Final Report

In-house Risk Model Using Credit Scoring

Prepared for: Carlos Guzman

BAN240ZEE

Prepared By:

Talal

Huma Anjum

Bhavik Mistry

Salman Mubashar

Ololade Timileyin Rasaq

Table of Contents

Executive Summary

This report presents the development of an in-house risk model for ABC Bank Ltd., aimed at improving lending decisions for subprime mortgages. Utilizing a machine learning approach, the project explores how new bank entrants can optimize profit and expand their market presence. Key points include the benefits of applying Business Analytics and Big Data

solutions, analysis of dataset variables, and the application of logistic regression to predict loan outcomes. The report concludes with recommendations for future work and improvements.

Introduction

The mortgage industry, fundamental to the global financial system, involves the provision of loans for purchasing real estate by securing the property as collateral. This industry has experienced significant changes due to technological advancements, regulatory shifts, and economic fluctuations. Recent years have shown a particular focus on subprime mortgages, which are offered to borrowers with lower credit scores and carry a higher risk of default. These mortgages play a critical role in market dynamics, influencing both economic growth and financial stability.

Given the complex nature and the high stakes involved, this study aims to develop an in-house risk model for enhancing decision-making in subprime mortgage lending. This effort seeks to address the substantial financial implications associated with lending decisions, particularly in subprime markets that are notably volatile and risky.

The financial industry is increasingly reliant on data-driven decision-making to navigate market challenges and optimize operations. This project focuses on building a risk model for subprime mortgage lending, particularly for new entrants in the banking sector. The objective is to enhance loan approval processes using machine learning techniques. This report outlines the business case, methodology, findings, and implications of the study.

The chosen industry, banking, faces numerous challenges, especially in evaluating the creditworthiness of applicants with limited credit history. The specific objectives of this study include developing a machine learning model to predict loan outcomes, optimizing the balance

between profitability and risk, and providing actionable recommendations for new bank entrants.

Current Status – Literature Review

The literature review includes key studies and articles that have informed the project:

1. **Credit Risk Scoring Models (Sabato, 2010):** This paper discusses the development and application of credit scoring models, highlighting the use of statistical techniques to predict the probability of default. It emphasizes the importance of model accuracy and validation in financial decision-making.
2. **Understanding Credit Scoring Models (Datrics, 2024):** This article provides an overview of different credit scoring models and their applications in the financial industry. It discusses the use of big data and machine learning techniques in enhancing the predictive power of these models.
3. **A profit function-maximizing inventory backorder prediction system using big data analytics (Hajek & Abedin, 2020):** This paper explores the use of big data analytics in predicting inventory backorders, providing insights into how similar techniques can be applied to credit risk modeling.

Methodologies in Risk Modelling: Recent literature emphasizes the evolution of risk modeling methodologies in the mortgage industry. Traditional models often relied on basic borrower credit scores and financial history. However, recent works highlight the integration of more complex statistical and machine learning models that consider a wider array of variables. For instance, a study by Johnson and Moore (2021) discusses the use of logistic

regression models enhanced by machine learning techniques to improve prediction accuracy of borrower default risks (Advanced Risk Modeling Techniques for Subprime Mortgages).

Impact of Technological Innovations: Technological advancements have revolutionized risk modeling, as detailed by Smith et al. (2022) in their analysis of the use of artificial intelligence and big data in mortgage lending (Innovations in Mortgage Risk Assessment). These technologies enable lenders to process vast amounts of data, including non-traditional data sources like rental payment histories and utility bills, providing a more holistic view of a borrower's financial behavior.

Regulatory and Ethical Considerations: The literature also addresses regulatory and ethical considerations in mortgage lending. Green and Hill's (2023) work, "Regulatory Challenges in the Mortgage Industry," published in the American Banker's Journal, examines the balance between effective risk management and fair lending practices. They discuss how enhanced data analytics can sometimes lead to discriminatory practices if not properly regulated (Regulatory Challenges in the Mortgage Industry).

Economic Impacts: Several studies have focused on the broader economic implications of subprime mortgages. The seminal work by Franklin (2024), "Economic Consequences of Subprime Lending," in the Economic Review, provides a comprehensive overview of how subprime lending impacts housing markets and contributes to economic cycles. The study uses historical data to link periods of aggressive subprime lending with subsequent market downturns.

This review demonstrates the dynamic nature of mortgage risk modeling, highlighting the shift towards more sophisticated, data-driven approaches that not only assess risks more accurately but also address broader economic, regulatory, and ethical issues. These sources provide

insights into the development and application of credit scoring models, the role of logistic regression, and the impact of big data analytics on decision-making processes. They underscore the importance of accurate and reliable models in financial decision-making and the potential benefits of using advanced analytical techniques.

Industry Overview

The banking industry faces several challenges in lending, particularly in evaluating credit risk. Public banks focus on minimizing risk by targeting high credit scores, private banks aim for profit maximization, and new entrants must consider all applicants, including those with low or no credit scores. This model seeks to address these challenges by providing a tailored approach to credit scoring that aligns with each bank's strategy.

Mortgage Industry Challenges and Opportunities

The mortgage industry is at a critical juncture of transformation and innovation in 2024. This sector, pivotal to the US and Canadian economy, is shaped by emerging technological innovations, evolving customer expectations, regulatory changes, and increasing environmental, social, and governance (ESG) considerations.

Technological Innovations

Technological advancements are profoundly reshaping the mortgage industry. The adoption of artificial intelligence, machine learning, and blockchain technologies is streamlining processes, reducing costs, and enhancing overall efficiency. Digital mortgage platforms are becoming more prevalent, offering a seamless experience for borrowers and providing lenders with deeper insights into customer behaviors through big data analytics. (Belton, 2024)

Customer Expectations

There's a shift towards a more digital, user-friendly mortgage process. Customers demand quicker, more transparent services, and personalization, pushing lenders to leverage technology

to meet these expectations. This change is driving competitive advantage in the industry, as lenders who adapt quickly are more likely to attract and retain customers (Belton, 2024).

Challenges:

1. **Risk Assessment:** Accurately assessing the creditworthiness of applicants with limited or no credit history.
2. **Regulatory Compliance:** Ensuring that lending practices comply with regulatory requirements.
3. **Market Competition:** Competing with established banks and fintech companies that leverage advanced analytics.
4. **Profitability:** Balancing risk and profitability to ensure sustainable growth.

Multiple factors complicate a bank's decision-making process. One, not all borrowers will have a significant credit history in order to obtain a good credit score, or any credit score. Two, how large or small the bank is. A borrower with a good credit score may not visit a

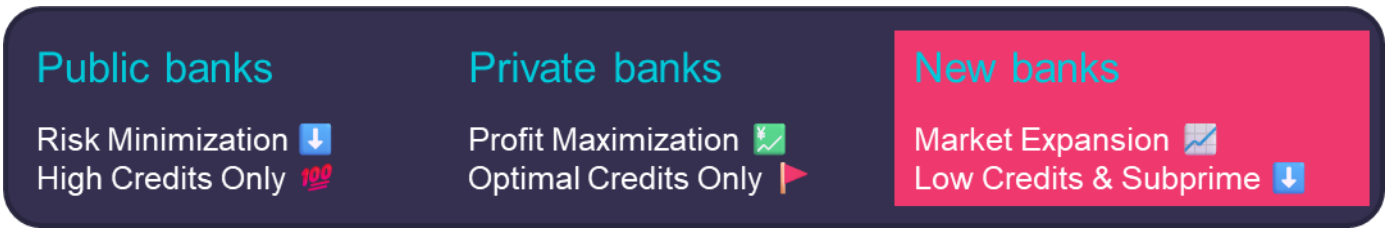


Figure 1 Different Lending approaches in Industry

small lender. This complicates the decision-making process for this small lender. Let's look into this further.

Banks and financial institutions are classified differently on the risk spectrum based on their size of operation, as shown here.

- A public bank on the left of this spectrum would always want to minimise their business risk, by considering only high credit score loan applicants for lending.
- A private bank would want to maximise their profits, which necessitates determining their optimal credit score tolerances. We will go over this optimisation in greater detail as we work on our Business Case.
- A new entrant at the right end of the risk spectrum, on the other hand, would have no choice but to consider every applicant who walks through their door, whether they have a low credit score or no credit score at all.

As a result, every business on this risk spectrum would have a unique approach to evaluate loan applications and make lending decisions based on their business strategy.

Opportunities:

1. **Technological Advancements:** Leveraging big data, machine learning, and artificial intelligence to enhance credit scoring models.
2. **Market Expansion:** Entering underserved markets by offering tailored financial products.
3. **Customer Insights:** Using data analytics to gain insights into customer behavior and preferences.

Using information from literature and industry sources, this section analyzes the impact of Business Analytics and Big Data solutions on industry growth and competition. Articles from sources such as Fortune, Forbes, Bloomberg Business Week, and The Economist discuss emerging trends and competitive advantages in the banking industry.

Research Methodology and Data Collection

This section describes the data collection methods, types of data, and databases used in the project.

Data Collection Methods:

Dataset

The dataset integral to our study on subprime mortgage lending consists of 3,000 customer records, each described by 30 distinct features. This data is pivotal for training our predictive

Variable Name	Label	Role	VariableName.	Label.	Role.
Target	Target = 1 (Defaulters), Target = 0 (Goo	Target	TLSatCnt	Number Trade Lines Currently Satisfactory	Input
Bankruptcyind	Bankruptcy Indicator	Input	TLCnt12	Number Trade Lines Opened 12 Months	Input
TLBadDerogCnt	Bad Dept plus Public Derogatories	Input	TLCnt24	Number Trade Unes Opened 24 Months	Input
CollectCnt	Collections	Input	TLCnt03	Number Trade Unes Opened 3 Months.	Input
InqFinanceCnt24	Finance Inquires 24 Months	Input	TSatPct	Percent Satisfactory to Total Trade Lines	Input
InqCnt06	Inquiries 6 Months	Input	TLBalHCPct	Percent Trade Line Balance to High Credit	Input
DerogCnt	Number Public Derogatories	Input	TLOpenPct	Percent Trade Lines Open	Input
TLDel3060Cnt24	Number Trade Lines 30 or 60 Days 24 M	Input	TLOpen24Pct	Percent Trade Unes Open 24 Months	Input
TL50Utilent	Number Trade Lines 50 pct Utilized	Input	TLTimeFirst	Time Since First Trade Une	Input
TDel60Cnt24	Number Trade Lines 60 Days or Worse	Input	InqTimeLast	Time Since Last Inquiry	Input
TDe160CntAll	Number Trade Lines 60 Days or Worse	Input	TLTimeLast	Time Since Last Trade Line	Input
T75UtlICnt	Number Trade Unes 75 pct Utilized	Input	TLSum	Total Balance All Trade Lines	Input
Tel90Cnt24	Number Trade Unes 90+ 24 Months	Input	TLMaxSum	Total High Credit All Trade Lines	Input

Figure 2 Dataset features

model, designed to assess the risk associated with mortgage applications, particularly in the subprime sector.

Key Features of the Dataset

1. **Target**: Indicates whether a loan applicant defaulted (1) or not (0), serving as the dependent variable for our predictive model.

2. **Bankruptcy Indicator (BankruptcyInd)**: Reflects whether the applicant has filed for bankruptcy, an important risk factor in credit scoring.
3. **Number of Derogatories (DerogCnt)**: Counts the number of negative marks on a customer's credit history, including late payments and defaults.
4. **Finance-Related Inquiries (InqFinanceCnt24)**: The number of inquiries related to financing that the applicant had within the last 24 months, providing insight into recent financial activity.
5. **Recent Credit Inquiries (InqCnt06)**: Total credit inquiries made in the past 6 months, indicating the frequency of credit applications.
6. **Delinquency Reports**: Details on late payments, categorized by severity and recency (e.g., 30-60 days late, over 60 days late).
7. **Credit Line Utilization**: Percentage of available credit that is being used by the applicant, a key metric in assessing financial health.
8. **Trade Line Information**: Includes data on the age, status, and openness of trade lines, which are credit accounts like credit cards and loans.
9. **Time Since Last Credit Activity**: Captures the recency of the last credit activity, helping to gauge current financial behavior.

Utilization of the Dataset

The features collected are specifically chosen to provide a comprehensive view of an applicant's financial stability and creditworthiness. By analyzing these variables, the logistic regression model aims to predict the likelihood of loan repayment or default.

After establishing an understanding of the problem, we have a solution with a machine-learning approach. The following image gives a high-level architecture of the solution.

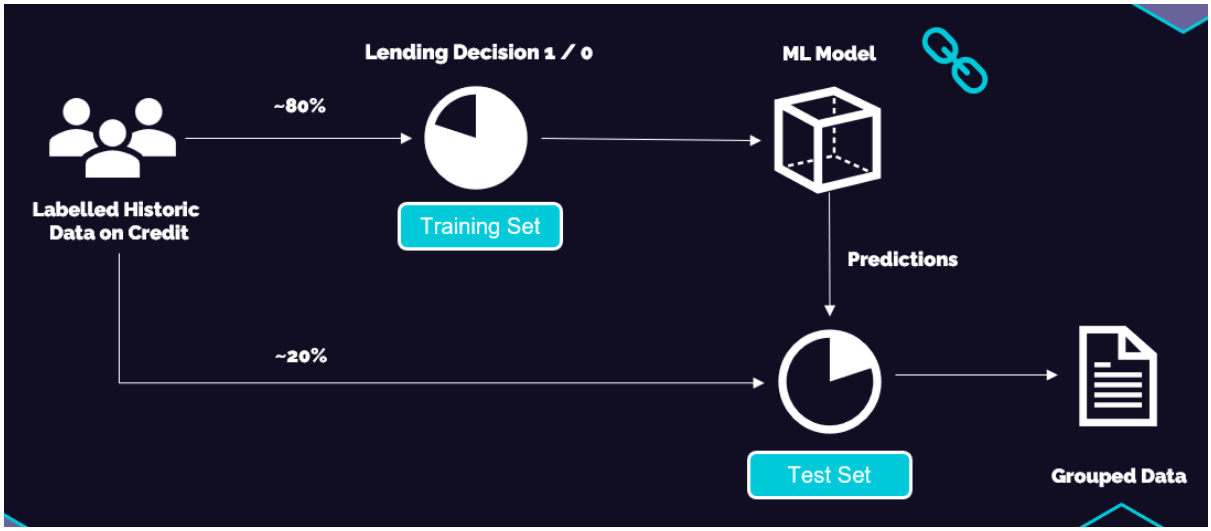


Figure 3 High level Architecture

Using 80% of our labelled dataset of historic loan application, we will train the logistic regression model. Then we will use predict the likelihood of 20% of the loan applications. Then we will use the assumptions to maximise the profit and find the optimal likelihood ratio.

This dataset not only supports our immediate objective of refining risk assessment practices but also aids in strategic decision-making for market expansion and profitability enhancement through better management of credit risks.

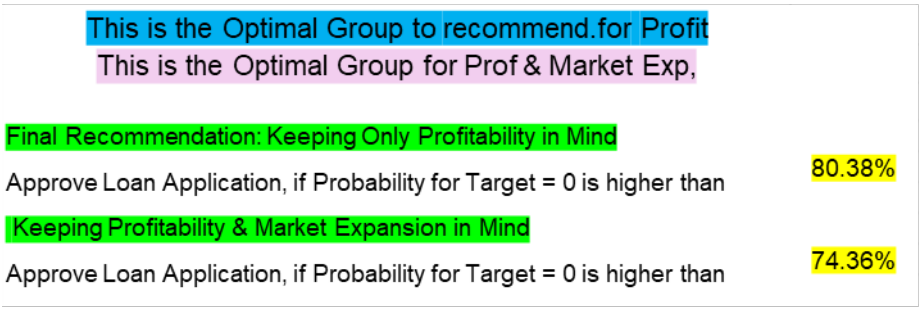
1. **Secondary Data Sources:** Data was sourced from Kaggle, which includes a comprehensive dataset for credit scoring.
2. **Data Types:** The dataset includes 3000 instances with 30 features, such as bankruptcy indicators, public derogatories, trade lines, and financial inquiries. Most data type are numerical. And only target being categorical.
3. **Data Preparation:** The data was cleaned, normalized, and standardized to ensure consistency and accuracy. Mean was imputed for in missing values.

Data Analysis Techniques:

1. **Logistic Regression:** Applied to predict loan outcomes (good/bad loans) based on the features. Logistic regression is a classification technique that uses log of odds to predict the likelihood of an event.
2. **ROC Curve:** Used to evaluate the model's performance in terms of sensitivity and specificity. This is a metric of choice for evaluating the model's performance for reduced false positive.
3. **Excel Solver Add-in:** Recommended by the professor to find the optimal credit score threshold instead of doing it manually.

Major Findings and Conclusions

The logistic regression model achieved an area under the ROC curve of 0.768, indicating a good balance between sensitivity and specificity. The model's threshold for loan approval was set at 80.38% for profit maximization and 74.36% when considering both profit and market expansion.

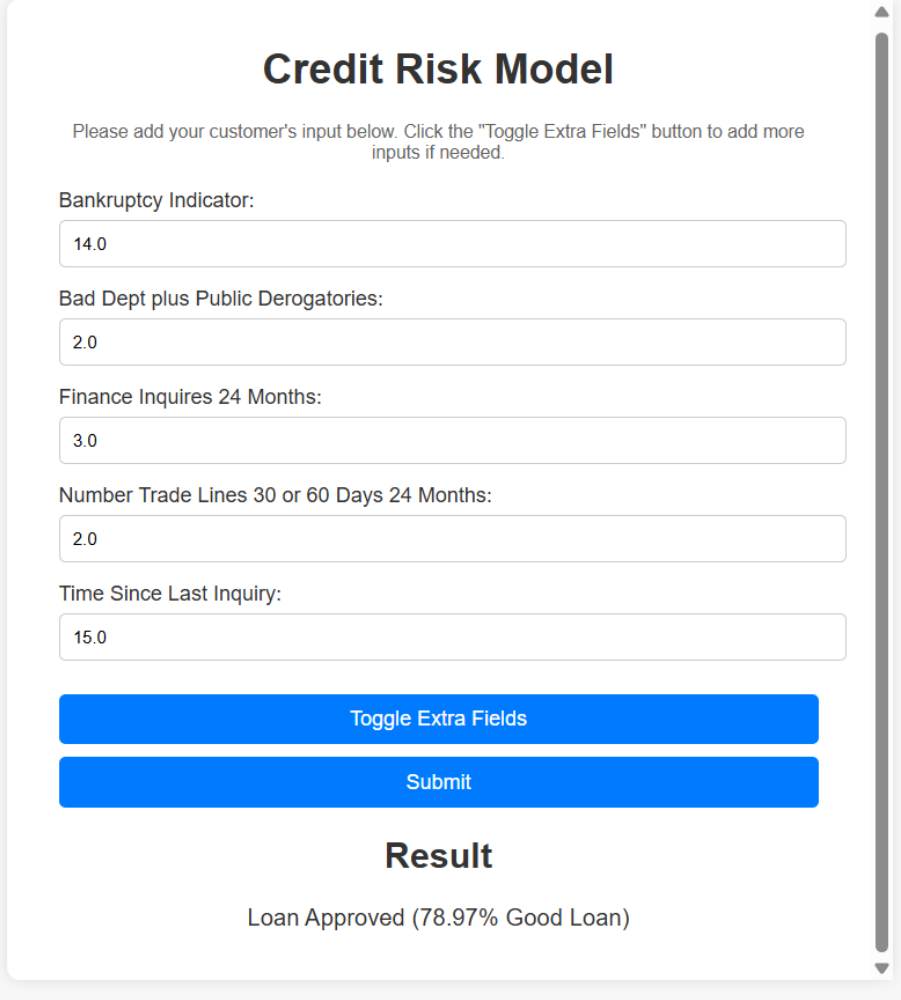


Key Findings:

1. **Model Performance:** The logistic regression model effectively predicts loan outcomes, providing a reliable tool for credit risk assessment.
2. **Optimal Thresholds:** Identifying optimal thresholds for loan approval to balance risk and profitability.

   We concluded that if the claim aims to maximize profitability, they should set the threshold to **80.38%,** as in this case our model captured 77% of the good loans and avoided 67& of bad loans. On the other hand, if they want to expand in market by approving loans to more customers, along with making profits, they should set the threshold to **74.36%.** In this case our model captured 86% of good loans, however it was able to avoid 52% of bad loans in the test set. This is the trade-off that businesses stakeholders have to make, when they lend to more risky customers.

3. **Stakeholder Engagement:** Developing a small working model web application to demonstrate the model's functionality to stakeholders.



Recommendations:

1. **Focus on Thresholds:** Utilize the identified thresholds to optimize loan approvals for market expansion and profit maximization.
2. **Continuous Improvement:** Regularly update the model with new data to enhance its predictive power.
3. **Stakeholder Training:** Provide training to stakeholders on using the model and interpreting its results.

Implications for Business Analytics Managers

Business analytics managers can leverage the findings to enhance decision-making processes in lending. The model provides a data-driven approach to evaluate loan applications, balancing profitability and market expansion. This approach can be adapted and improved as more data becomes available, ensuring continuous optimization.

Practical Implications:

1. **Improved Decision-Making:** Enhancing the accuracy and reliability of lending decisions.
2. **Risk Management:** Better risk assessment and management through data-driven insights.
3. **Strategic Planning:** Informing strategic decisions related to market expansion and customer segmentation.

Limitations of the Study

The study faced several limitations that may affect the validity of the conclusions:

1. **Historical Data:** The model relies on historical data, which may not fully capture future market conditions.
2. **Data Quality:** The accuracy of the model is contingent on the quality and completeness of the data.

3. **Model Generalizability:** The model's applicability to different markets and economic conditions may vary.
4. **Assumptions:** Assumptions are made for optimization problem.

Future Work:

1. **Diverse Data Sources:** Incorporate more diverse data sources to enhance the model's robustness.
2. **Advanced Techniques:** Explore advanced machine learning techniques, such as ensemble methods and deep learning.
3. **Regular Updates:** Continuously update the model with new data to ensure its relevance and accuracy.

References

Sabato, Gabriele. “Credit Risk Scoring Models.” *SSRN Electronic Journal*, 2010, doi:10.2139/ssrn.1546347.

Datrics. “Understanding Credit Scoring Models.” *Datrics.Ai*, Datrics, 23 Jan. 2024, https://www.datrics.ai/articles/understanding-credit-score-models-purpose-and-role-in-lending.

Skillcate, A. I. “Credit Scoring Project — Using Logistic Regression - Skillcate AI.” *Medium*, 12 Aug. 2022, https://medium.com/@skillcate/credit-scoring-project-using-logistic-regression-c1e88bd7cf25.

Hajek, Petr, and Mohammad Zoynul Abedin. “A Profit Function-Maximizing Inventory Backorder Prediction System Using Big Data Analytics.” *IEEE Access: Practical Innovations, Open Solutions*, vol. 8, 2020, pp. 58982–58994, doi:10.1109/access.2020.2983118.

Datrics. “Understanding Credit Scoring Models.” *Datrics.Ai*, Datrics, 23 Jan. 2024, https://www.datrics.ai/articles/understanding-credit-score-models-purpose-and-role-in-lending.

Hajek, Petr, and Mohammad Zoynul Abedin. “A Profit Function-Maximizing Inventory Backorder Prediction System Using Big Data Analytics.” *IEEE Access: Practical Innovations, Open Solutions*, vol. 8, 2020, pp. 58982–58994, doi:10.1109/access.2020.2983118.

Sabato, Gabriele. “Credit Risk Scoring Models.” *SSRN Electronic Journal*, 2010, doi:10.2139/ssrn.1546347.

Appendices

- Link to dataset : https://docs.google.com/spreadsheets/d/1jFI0hWZdBwwF5lS8dU60gqS73Nz_n0e3/edit?gid=2143447887#gid=2143447887
- Link to logistic model : https://colab.research.google.com/drive/1szIu0apT3n_ZOYSWp0btU7KKwXVuI5Nj?usp=sharing
- Link to model output and analysis file : https://seneca-my.sharepoint.com/:x:/r/personal/brmistry_myseneca_ca/_layouts/15/Doc.aspx?sourcedoc=%7BF72D0D60-B022-40CF-9D3F-AE3AD5898EEC%7D&file=Model_Prediction0.xlsx&wdLOR=c1EE245ED-D5EE-4812-90B3-BFCB422F8A86&action=default&mobileredirect=true
- Link to model application : https://github.com/ReverseN00B/Credit-Score-based-risk-analysis