Loan Portfolio Prediction Capstone Project

Final Paper
NYU MSBA Class of 2024

Team Members:
Yen Fu (Michael) Lin
Juan Pena
Alfa Wesley Sirra
Pattamon Taltatsirapob
Mingi Um

Table of Contents

Executive Summary

The Loan Portfolio Prediction project involves an automotive loan lender in the Malaysian market that leverages the use of machine learning models to screen applicants to determine their qualification for an automotive loan. This involves the gathering of several critical pieces of information from each applicant to be processed and identify their probability of defaulting on their loan based on the information provided. This project evaluates a series of machine learning models using historical customer data to predict the probability of default for future applicants to aid the client in making lending decisions, while mitigating the risk of a financial loss due to borrower defaults. Additionally, our project investigates the timeframe that it would take for a loan applicant to default on their loan, if they are classified as a defaulter by our algorithms.

Furthermore, the Loan Portfolio Prediction project also consists of a causal analysis component to identify the characteristics of those applicants who default on their automotive loans compared to those that do not default over the same time period. This investigation consists of a natural experiment and counterfactual analysis to understand the underlying factors that may contribute to why an applicant is more likely to default.

As a result of our investigation and work on improving the predictive capability of our client, we have developed two machine learning models that will predict if an applicant will default as well as the timeframe in which they are expected to default as well. The causal analysis aspect of our project has also uncovered potential hypotheses for what behavior is exhibited by those who default on their loans as opposed to those that do not default.

Project Objective & Scope

The objective of the Loan Portfolio Prediction project is to improve the decision-making process of our client when it comes to quantifying the risk of each loan applicant. Tactically speaking, we aim to improve the predictive performance of the client's existing machine learning model that quantifies the risk of an applicant upon the completion of their motorcycle loan application. Additionally, to help inform

the client of the timing it may take for the applicant to default, a secondary model has been developed to predict when an applicant will default. Both of these models have been developed with the goal to minimize the financial loss that can stem from defaulted loans by providing the client information with who may default and when they will default.

Additionally, our primary focus is to improve the predictive capabilities of our clients to aid their business. However, we have conducted a causal analysis to build a potential understanding of what is driving defaults using a combination of causal inference techniques to determine why an applicant defaults on their motorcycle loan. The purpose of this analysis is to help our client understand how they should approach specific applicants who display certain characteristics before moving forward when they submit a loan request for a new motorcycle.

It should also be noted that the scope of this project will be limited to the Malaysia market and will only consider motorcycle loans. While the overarching process to submit and process a loan application is the same, motorcycle loans have subtle differences (e.g. loan term) than a typical automobile loan. The details of these subtle differences will be described in the appendix of this report.

Project Drivers

The key driver for this project is to minimize the financial loss our client experiences on an annual basis. On an annual basis, our client experiences an approximately 17.4% default rate, which results in a significant financial loss. By being able to eliminate defaulters from our client's portfolio would maximize the return on their investment.

While there are known difficulties with perfectly predicting loan defaulters, improving the accuracy in our client's model by a fraction of a percentage could result in significant upside, which we have considered through a cost/benefit analysis across all models that were considered.

Key Data Sources

The dataset for this project was provided directly by our client. The dataset our team received spans multiple years and contains applicant information that was provided at the time they applied for their automotive loan. This includes information such as income, housing, and other pieces of information that can be used to determine whether our client should provide motorcycle financing to the applicant. Additionally, this data set has information on whether the applicant defaulted on their loan, where a default is classified as a borrower who has missed at least 3 payments during the life of the loan. This dataset is the primary source of how each of our models have been trained and tested as part of our work as well as the generation of any descriptive analytics.

In addition to the dataset, we have also been provided information related to the average profit and loss per loan from the client. This information has been used to conduct the cost/benefit analysis of each model to determine the estimated value of each model we have tested to this point. This information was used to finalize our recommendation as to which model should be used moving forward.

Lastly, our team has conducted independent research to find information related to the Malaysia motorcycle market. This information has been used to build our understanding of business processes in the market and inform our decision making as we formulated recommendations for the client.

Methodology

Our approach for our project was to first understand how our client currently conducts their lending of funds to borrowers today. This required us to learn the business and how loans were being currently processed. After we had an initial understanding of the business and regulatory requirements in the Malaysia market, we began to formulate a problem statement on what exactly we were going to achieve with this project. Ultimately, we settled on our objectives for the project and began our discovery phase.

During the discovery phase, we began gathering the necessary data to begin the development of our models by training and testing them with the data we were provided by our client. Our team then focused on efforts on a cost/benefit analysis to identify the optimal model to maximize the financial return on loans provided to borrowers while limiting the number of defaulted loans.

As we completed the development of our model, we then shifted our focus from prescriptive to causal inference. This phase allows us to explore beyond the correlations in our data set and identify potential reasons behind customer defaults in Malaysia over our two-year dataset by employing a mix of natural and counterfactual methodologies.

Summary Results

In summary, the Loan Portfolio Prediction project has focused on improving the default rate for loans provided by our client to customers. A combination of skills that we developed over the time in the NYU MSBA program have contributed to how this project was approached from a business and analytical standpoint.

We have been able to use a combination of prescriptive and causal inference approaches to help our client address their business needs so that they may improve their business moving forward with our recommendations. Ultimately, we recommend that our client utilize the combination of our models to identify defaulters as well as understanding when those borrowers will default to make informed decisions regarding whether or not an applicant should be extended a motorcycle loan. Lastly, we recommend that our client evaluate opportunities to leverage information regarding when a borrower will default on their loan to see if there are options to maintain profitability despite the fact that they will default.

The remaining sections of this paper provide information regarding the business of our client, the methodology we used to approach this project, and ultimately how we have come to our conclusion and recommendations.

Background

Market Overview

Market : Our client is a financial company in Malaysia that provides financing options for individuals looking to purchase motorcycles. Malaysia's motorcycle market ranks 12th globally—and 10th in Asia—in terms of unit sales, and the market has been experiencing rapid growth recently. From 2014 to 2023, motorcycle sales in Malaysia skyrocketed from 426,600 units to 667,800 units, marking a significant growth rate over nine years.[1]

Figure 1.1 Transportation Breakdown in Malaysia (Source: Statistia)

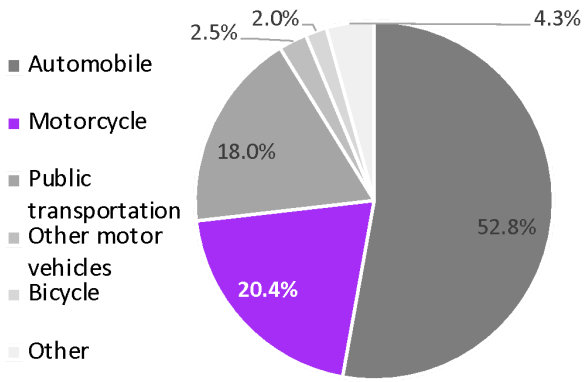
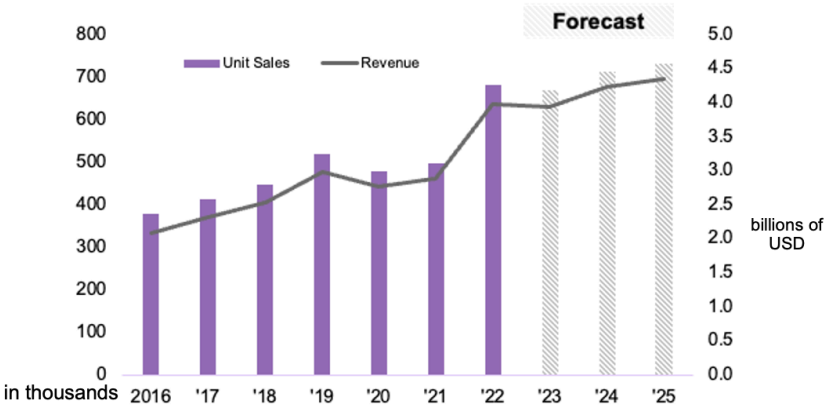


Figure 1.2 Motorcycle Market Trend in Malaysia (Source: Statistia)



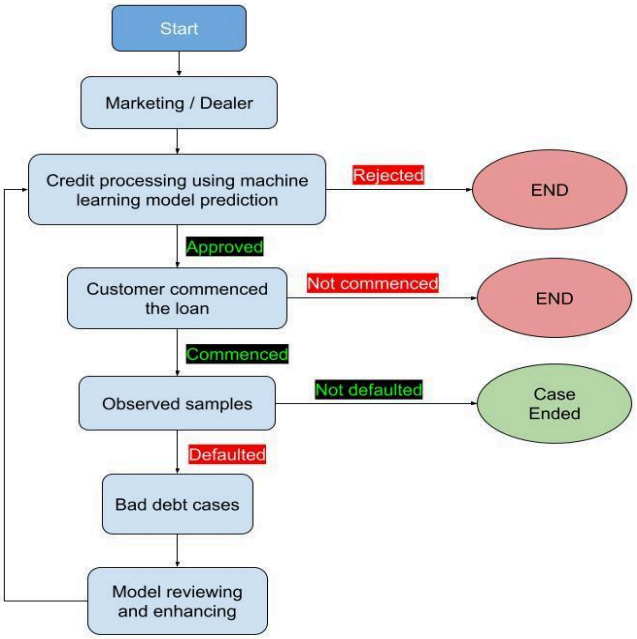
[1] Statistia Market Insights, “Motorcycles: market data & analysis”, Statistia, December 2023

A January 2023 survey revealed that 20.35% of Malaysians prefer private motorcycles as their primary mode of transportation, second only to private automobiles (52.84%) and surpassing public transportation (18.0%)[2]. This growing preference underscores the motorcycle market's potential and, by extension, the financing sector's opportunities and challenges.

Business Overview

Our client is one of the active financial companies in the Malaysian motorcycle financial market. Our client offers both motorcycle and car loan products (our loan prediction model is specifically focused on motorcycle loans). Our client has maintained stable financial health so far. Moving forward, as the motorcycle market continues to grow rapidly, our client aims to achieve market leadership as a financial provider.

Figure 1.3 Customer Journey Flowchart



The flow chart above illustrates the loan application process. The process begins at the start point, where a customer becomes interested in a loan through Marketing / Dealer initiatives. Once the customer applies for a loan, their application

[2] Standard Insights, "Leading Modes of Transportation Among Consumers in Malaysia as of January 2023." Statista, April 2023

undergoes Credit processing using a machine learning model prediction to evaluate the creditworthiness. If the application is Rejected, the process ends there. However, if it's approved, the customer proceeds to commence the loan. At this juncture, if the customer decides not to commence with the loan, the process will end. If the customer does commence the loan, they move to the next phase, where their payment behavior is monitored. In this monitoring phase, if the customer does not default, which means they continue to make timely repayments, the case is considered closed, marked as case ended. If a default occurs, it is categorized under Bad debt cases, which indicates the customer failed to repay the loan as agreed. The final step involves model reviewing and enhancing, where the predictive performance of the machine learning model is assessed and improved upon based on the observed outcomes, especially from the defaulted loans, to refine future credit assessments.

Project Overview

Challenges

As the Malaysian motorcycle market grows, more people are looking to finance their purchases, leading to a boom in the loan sector. This growth has sparked fierce competition among finance companies, all competing to lend money to customers. But with this competition comes a downside—more people are failing to pay back their loans on time, and the number of risky loans is on the rise. This problem is partly because loans are being approved for people who may not have the means to pay them back. As a result, lending companies are facing financial pressures, and the cost of lending is going up because of these higher risks. This creates a tough environment for finance companies, who must be careful to manage these challenges.

Purpose and Objective

The main purpose of our project is to enhance our client's ability to predict and manage risks associated with motorcycle loans. This is critically important as the business grows and operational assets increase while maintaining a minimized delinquency rate and stable profitability.

A decrease in the delinquency rate can directly reduce financial losses and ensure a stable and predictable revenue flow, allowing the company to be more proactive in business expansion. In other words, if the delinquency rate increases even as operational assets grow, the company’s profit may not increase proportionally. We will closely identify the risk factors leading to loan defaults and incorporate them into our predictive model to help our client maintain a stable asset portfolio and lay the groundwork for business expansion.

Through our predictive model, our client will be able to establish various strategies such as adjusting loan conditions for specific customer groups. It will be possible to determine how much certain independent variables influence our target variable (Defaulted or not). By doing so, our client can minimize cases where they approve loans to customers who lack the ability to repay (False Negatives) and reject loan requests from customers who are capable of repayment (False Positives).

Data Understanding

We used data provided by a client. This data has 55,373 records of clients from January 2020 to July 2022 and includes 28 different features or variables. It is made up of numeric, binary, and categorical type variables. It is typical financial loan data showing each unique client's income level, living situation, work experience, and loan details like amount and term, whether they defaulted or not, and the month of default.

Dependent Variable (Target Variable)

Figure 2.1 Loan Status in the Dataset

Non defaulted		**45,737**
	On paying	36,790
	Completed	8,947
Defaulted		**9,636**
	On paying	1,844
	Completed	333
	True Loss	7,459
Total Number of Observations		**55,373**

The dependent variable is 'whether a customer defaults,' which is a binary variable. Out of a total of 55,373 loans, 45,737 are non-defaulted, and 9,636 are defaulted loans (delinquent for more than three months). Within the defaulted loans, 7,459 cases represent those that have been delinquent for over six months continuously and are considered write-offs. 1,844 cases are where customers are still trying to pay the remaining balance and interest after being three months delinquent, and 333 cases are those where the principal has been fully repaid after having defaulted.

Figure 2.2 Defaulted Loan Cases by Year

Defaulted in the 1st year	3,783
Defaulted in the 2nd year	3,725
Defaulted in the 3rd year	1,671
Defaulted after the 3rd year	457
Total Number of Defaulted Cases	**9,636**

Defaulted loans are also categorized based on the timing of the default occurrence. There are 3,783 cases that defaulted within the first year after loan origination, 3,725 cases that defaulted between the first and second year, 1,671 cases that defaulted between the second and third year, and 457 cases that defaulted after the third year. The reason for this reclassification by the timing of default is that the recoverable amount varies depending on the residual value of the loans. Typically, loans that default more quickly are considered to have a higher loss rate.

Independent Variables

The dataset's independent variables number 28 in total, with each variable's type being diverse, including Integer, Binary, Categorical, and Boolean. The independent variables describe clients' financial situations, such as the number of existing loans, living arrangements, work experience, income, education level, and credit score. They also include loan terms, like borrowing amount, term length, and vehicle brand. Some variables were removed during the data preparation phase to enhance the accuracy of the analysis.

Data Preparation

We made our data processing efficient by preparing our data first. For example, out of 28 independent variables, we chose only 7 that were important — that is, they had a meaningful impact on whether there would be a default. We picked these 7 by looking at the Information Value of each variable. The selected data are as follows.

Figure 2.3 Independent variables

Variable name	Description	Type
Risk grade	Credit range of the applicant form by 3rd party credit agency	Category
Net income	Applicant's net income per month in Malaysian currency (Ringgits)	Category
Loan amount	Total initial borrowing amount	Numeric
Trade bureau	Count of how many other financial obligations this customer has with Companies like Telephone bill, and etc	Category
House status	Where the applicant lives. The value options are "Own", "Relatives' house (local)" , and "Rent and Quarter (out state)"	Category
Employment period	How long has the applicant been employed	Category
Enquiry count	Count of how many times this customer requested for a loan in past 12 months	Category

Feature selection and Variable Reduction

SelectKBest: It is a feature selection technique used to select the top k features that have the strongest relationship with the target variable. It works by scoring each feature using a specific statistical test and then selecting the top k features based on these scores.

In this project, we are using SelectKBest with two different statistical tests:

- **Chi-Squared test** for categorical variables: It measures the dependence between categorical variables. For each categorical feature, the test measures the dependence between the feature and the target variable, and selects the features with the strongest relationship.
- **ANOVA (Analysis of Variance) test** for numeric variables: It measures the difference in means between multiple groups. For each numeric feature, the test measures the variance between different groups defined by the target variable, and selects the features with the highest variance.

By using SelectKBest with these two different tests, we are able to select the top features from both categorical and numeric variables, which are most relevant for predicting the target variable (loan defaulters) in this project.

Regarding T-Test: T-test is generally used to compare the means of two groups. In our case, if we want to test whether there's a significant difference in means between defaulters and non-defaulters for a particular numeric variable, we could have used a T-test. However, T-test alone may not be the best method for feature selection when we have both numeric and categorical variables. It's better to use techniques like ANOVA (for numeric) and Chi-Squared test (for categorical) for a more comprehensive feature selection.

Correlation: Initially, our dataset included twenty eight independent variables, each potentially offering unique insights into the phenomenon under investigation. However, as we delved deeper into our analysis, we encountered the challenge of multicollinearity—a situation where predictor variables exhibit strong correlations with each other, potentially undermining the predictive accuracy and interpretability of our model.

To address this challenge, we turned to correlation analysis, a powerful tool for uncovering relationships between variables. By carefully examining the pairwise correlations among our independent variables, we aimed to identify those with high correlations. Variables with high correlations often convey redundant information, and retaining them in our model could lead to instability in parameter estimates and compromised inference.

After thorough examination, we found that several variables shared strong correlations, suggesting redundancy in our feature set. Consequently, we made the strategic decision to trim our variable set, retaining only those that offered unique insights while discarding those that added little beyond what was already captured by other variables. Through this process of elimination, we were able to distill our initial set of twenty eight independent variables down to a more streamlined set of eight.

With our reduced set of independent variables at hand, we turned our attention to preparing categorical variables for inclusion in a logistic regression model. Since logistic regression requires numerical input, we employed a technique known as dummy variable encoding to transform our categorical variables into a format suitable for regression analysis.
This transformation resulted in the creation of twenty seven dummy variables, each representing a distinct category within our original categorical variables. However, even in this transformed state, the challenge of multicollinearity persisted. To mitigate its adverse effects, we once again turned to correlation analysis, this time focusing on the relationships among our dummy variables.

Our goal was clear: to select a subset of dummy variables with minimal correlations with each other, thereby ensuring the robustness and reliability of our logistic regression model. To achieve this, we established stringent criteria, retaining only those variables whose correlations fell within the narrow range of -0.69 to 0.69. These correlation coefficients are considered indicative of moderate to low linear relationships.

Figure 4.1 *Heatmap exhibiting the correlation of all the variables (dummies) from the main dataset of 28 variables*

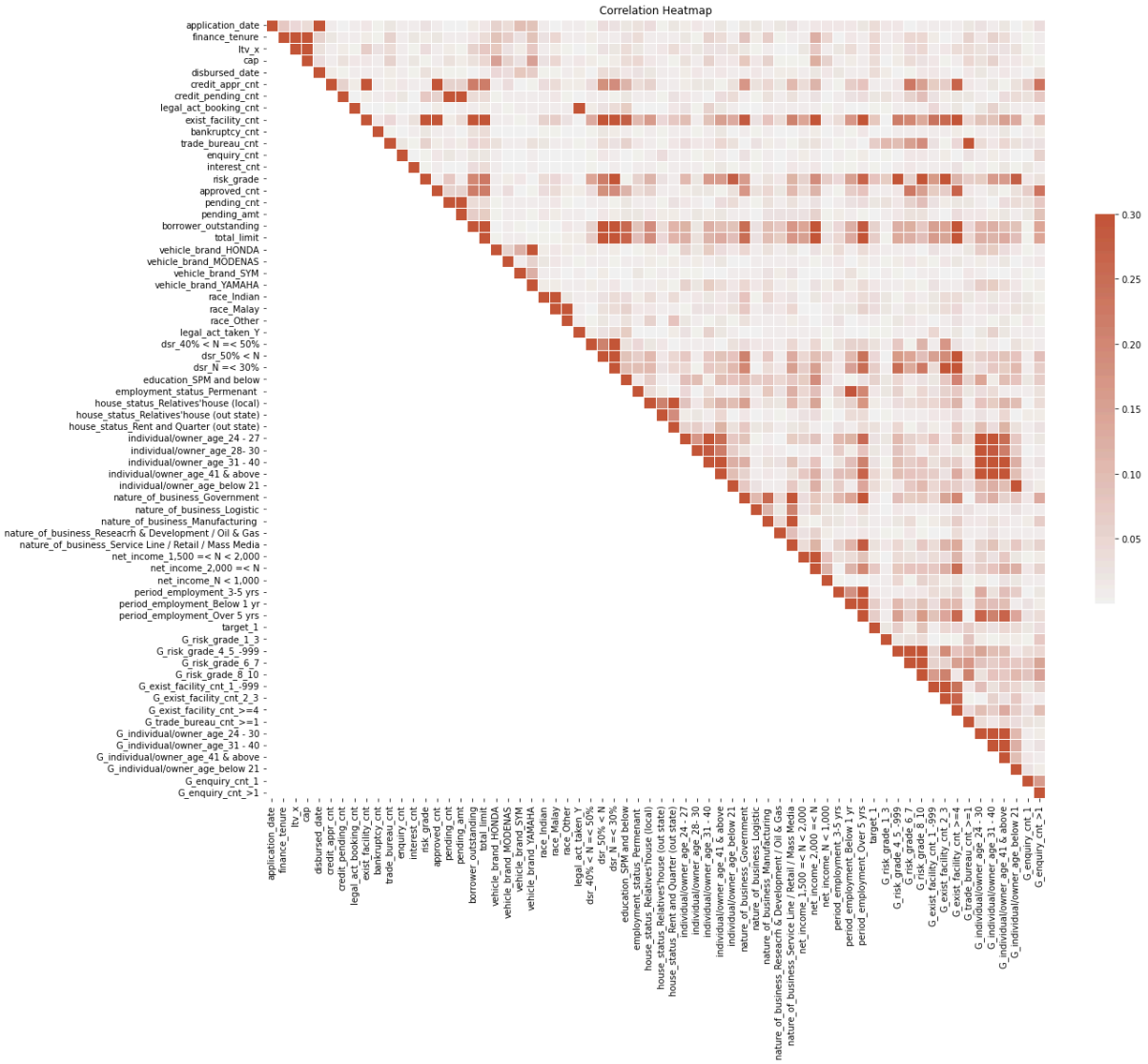
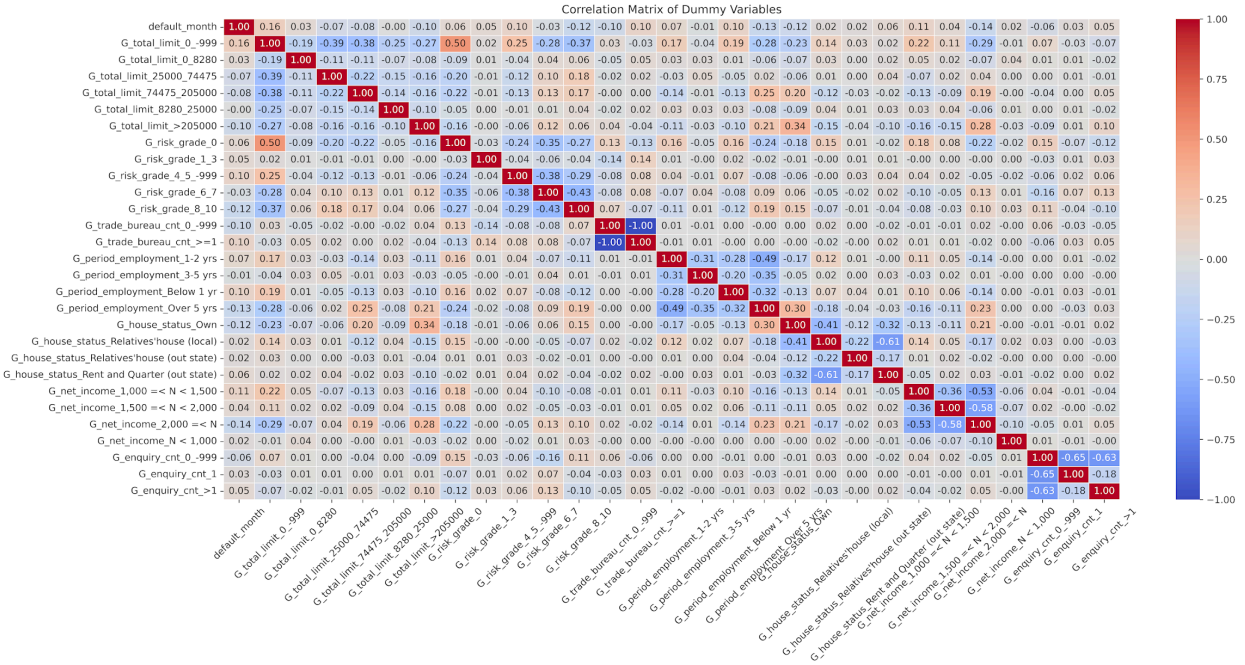


Figure 4.2 *Heatmap exhibiting the correlation of all the variables (dummies) from the dataset of 8 variables after feature selection*



We set up all our independent variables as categories. For example, for net income, we divided it into groups: income less than USD 1,500, income between USD 1,500 and 2,000, and income more than USD 2,000. And for risk grade, we split it into categories: 0, 1-3, 4-5, 6-7, and 8-10.

Working with categorical variables has given us several advantages. First, it made our model simpler, which helps a lot with certain algorithms like decision trees and random forests - they can work more efficiently now.

Additionally, it's more straightforward to look at net income in categories such as 'below USD 1,500', 'between USD 1,500 and 2,000', and 'above USD 2,000', rather than dealing with exact amounts.

Lastly, categories help us avoid making our model too complex - this means it's not just memorizing our data but really learning from it. They also make it easier to play around with the data to find new insights, which helps us get the most out of the information we have. All these advantages come together to make our prediction model not just simpler, but also more powerful.

Model Development

Model selection

After thorough evaluation of several machine learning models including XGBoost, decision tree, random forest, and logistic regression, we selected logistic regression as the primary model for our Loan Portfolio Prediction project. Our decision was based on several factors.

Interpretability: Logistic regression offers transparency, allowing us to understand the relationship between input features and the probability of loan default. This interpretability is crucial for our client to comprehend the reasoning behind the model's predictions.

Performance: Despite its simplicity compared to other models tested, logistic regression demonstrated competitive performance and generalization capability on our dataset.

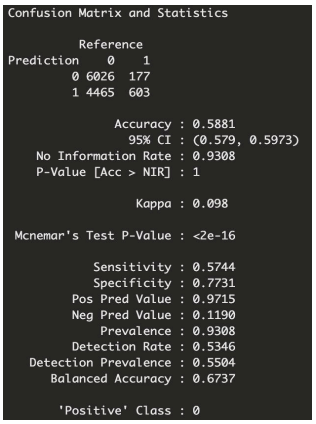
Robustness: Logistic regression is less prone to overfitting, making it a reliable choice for our project, especially considering the limited size of our dataset.

Computational efficiency: Logistic regression provides the advantage of computational efficiency, making it feasible for real-time application within our client's lending process.

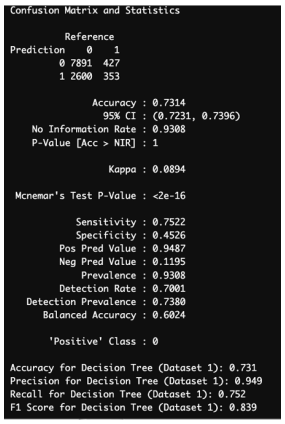
Overall, logistic regression emerged as the optimal choice due to its interpretability, performance, robustness, and computational efficiency, aligning perfectly with the objectives of our project.

Figure 5.1 *Results of our models*

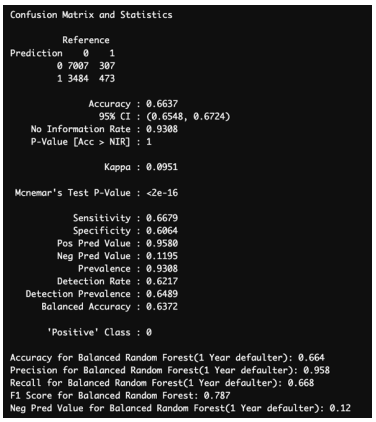
XG Boost



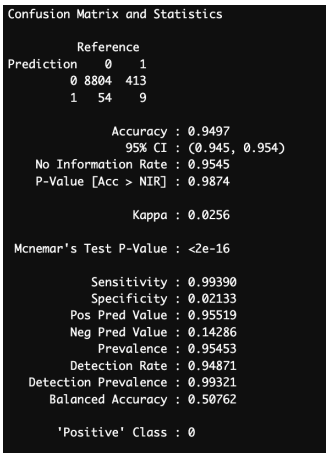
Decision tree



Random forest



Logistic regression



In addition to the factors discussed above that led us to choose logistic regression, the performance of the other models was a significant consideration. Upon evaluation, it became evident that the accuracy and other performance metrics of all other models were notably weaker compared to the logistic regression model. Therefore, logistic regression appeared to be the most logical choice.

Considering the current default rate at the automobile loan institution is approximately 83%, achieving a 95% accuracy in predicting non-defaulters marks a significant advancement. Implementing this project in practice could greatly contribute to improved success rates.

Logistic Regression

We conducted logistic regression analysis to examine the characteristics of each independent variable. Through logistic regression, we were able to analyze how changes in the values of each independent variable affect the likelihood of default, which is our target variable. Initially, we established a baseline for each independent variable, aiming to determine how much they increase or decrease the probability of default relative to this baseline. The variables we analyzed include G_total_limit(), G_risk_grade, G_trade_bureau_cnt, G_period_employment, G_house_status, G_net_income, and G_enquiry_cnt, making a total of seven variables.

Figure 5.2 Independent Variables in Logistic Regression

```
Call:
glm(formula = target_variable ~ ., family = "binomial", data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1250  -0.4287  -0.3051  -0.2087   3.1681

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -1.91862    0.15233 -12.595  < 2e-16 ***
G_total_limit0_-999                     0.73609    0.11622   6.334 2.39e-10 ***
G_total_limit0_8280                     0.66302    0.13435   4.935 8.02e-07 ***
G_total_limit25000_74475                0.20994    0.12002   1.749 0.080255 .
G_total_limit74475_205000               0.20599    0.12022   1.713 0.086633 .
G_total_limit8280_25000                 0.28203    0.13010   2.168 0.030174 *
G_risk_grade1_3                         0.17073    0.18695   0.913 0.361130
G_risk_grade4_5_-999                   -0.10896    0.05487  -1.986 0.047039 *
G_risk_grade6_7                        -0.23038    0.06090  -3.783 0.000155 ***
G_risk_grade8_10                       -0.77927    0.08263  -9.431  < 2e-16 ***
G_trade_bureau_cnt0_-999               -0.74900    0.05630 -13.303  < 2e-16 ***
G_period_employment3-5 yrs             -0.18267    0.05881  -3.106 0.001895 **
G_period_employmentBelow 1 yr           0.36116    0.04872   7.413 1.24e-13 ***
G_period_employmentOver 5 yrs          -0.39325    0.05825  -6.751 1.47e-11 ***
G_house_statusRelatives'house (local)   0.43759    0.08622   5.075 3.87e-07 ***
G_house_statusRelatives'house (out state)  0.50249    0.11121   4.519 6.23e-06 ***
G_house_statusRent and Quarter (out state)  0.76260    0.08589   8.879  < 2e-16 ***
G_net_income1,500 =< N < 2,000         -0.23641    0.04685  -5.046 4.51e-07 ***
G_net_income2,000 =< N                 -0.52148    0.05128 -10.168  < 2e-16 ***
G_net_incomeN < 1,000                  -0.08554    0.17492  -0.489 0.624837
G_enquiry_cnt0_-999                    -0.65030    0.05090 -12.777  < 2e-16 ***
G_enquiry_cnt1                         -0.33702    0.06281  -5.366 8.05e-08 ***
---
```

The Odds Ratio is a statistical measure commonly seen in the results of logistic regression analysis. It represents the ratio of the odds of an event occurring to the odds of it not occurring within a selected category, and indicates the relative probability of that event happening in one category compared to a reference category. For example, in the case of G_total_limit0_8280, the odds of defaulting increase by approximately 1.94 times compared to the baseline.

1) G_total_limit: It refers to the total initial borrowing amount.

Baseline: G_total_limit > 205000	
Category	Odds Ratio
G_total_limit0_-999	2.09
G_total_limit0_8280	1.94
G_total_limit8280_25000	1.33
G_total_limit25000_74475	1.23
G_total_limit74475_205000	1.23

2) G_risk_grade: This indicates the credit range of the applicant as assessed by a third-party credit agency. The baseline is risk_grade = 0, which represents the highest level of risk.

Baseline: G_risk_grade0	
Category	Odds Ratio
G_risk_grade1_3	1.19
G_risk_grade4_5_-999	0.9
G_risk_grade6_7	0.79
G_risk_grade8_10	0.46

3) G_trade_bureau_cnt: It represents the count of the customer's financial obligations with companies, including, but not limited to, telephone bills.

Baseline: G_trade_bureau_cnt >= 1	
Category	Odds Ratio
G_trade_bureau_cnt0_-999	0.47

4) G_period_employment: This measures the length of the loan applicant's employment period.

Baseline: G_period_employment1-2 yrs	
Category	Odds Ratio
G_period_employment Below 1 yr	1.43
G_period_employment 3-5 yrs	0.83
G_period_employment Over 5 yrs	0.67

5) G_house_status: This indicates the ownership status of the residence in which the loan applicant lives.

Baseline: G_house_statusOwn	
Category	Odds Ratio
G_house_statusRelatives'house (local)	1.55
G_house_statusRelatives'house (out state)	1.65
G_house_statusRent and Quarter (out state)	2.14

6) G_net_income: This refers to the customer's average monthly income, expressed in USD.

Baseline: G_net_income1,000 =< N < 1,500	
Category	Odds Ratio
G_net_income1,500 =< N < 2,000	0.79
G_net_income2,000 =< N	0.59
G_net_incomeN < 1,000	0.92

7) G_enquiry_cnt: This denotes the number of loan inquiries made by the customer in the past 12 months.

Baseline: G_enquiry_cnt > 1	
Category	Odds Ratio
G_enquiry_cnt0_-999	0.52
G_enquiry_cnt1	0.71

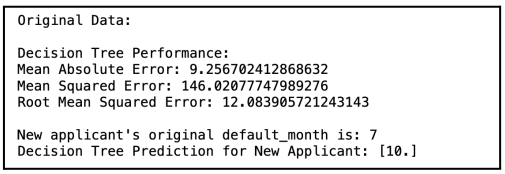
Predicting default month

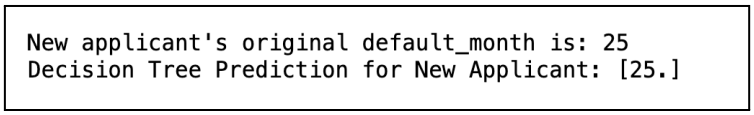
In addition to developing and comparing several models to predict whether a potential customer will default on a loan in the future, we have also worked on creating another model to predict when this default might occur—specifically, in which month after the loan was issued.

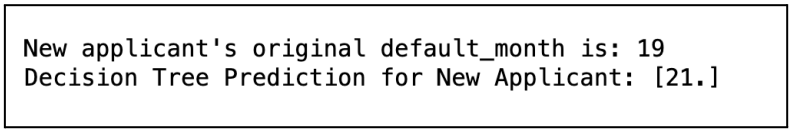
The aim is to increase our confidence when approving loans and to better plan for the future of our existing customers. This approach allows us to calculate more accurate interest rates for new loan approvals. Moreover, knowing precisely when a loan will default enables us to plan more effectively to prevent defaults by implementing strategies such as loyalty programs.

While logistic regression appears to be the better choice for predicting whether someone will default on a loan, decision trees have proven to be more effective for predicting when a default might occur.

5.3 Results of the model







Predicting the default month can be a valuable tool for improving our business strategies. Here are some suggestions on how we can leverage this prediction to enhance various aspects of our business:

1. Risk Assessment and Loan Approval: Utilize the predicted default month as a risk indicator during the loan approval process. Offer lower interest rates or more

favorable terms to applicants predicted to have a later default month, indicating lower risk of default.

2. Customized Loan Products: Develop tailored loan products based on predicted default months. For customers predicted to have a longer default window, offer longer loan durations or larger loan amounts. Conversely, for customers with shorter default windows, consider offering shorter loan durations or smaller loan amounts.

3. Customer Retention and Loyalty Programs: Identify existing customers with predicted early default months and proactively engage with them to mitigate potential default risks. Offer incentives, such as refinancing options with reduced interest rates or personalized financial counseling, to help customers manage their loans effectively and maintain loyalty to your company.

4. Collection Strategies: Prioritize collection efforts based on predicted default months. Allocate resources towards customers with earlier predicted default months to mitigate losses and improve recovery rates. Implement targeted communication strategies, such as reminders or financial assistance programs, to support customers at higher risk of default.

5. Credit Scoring and Underwriting: Incorporate predicted default months into our credit scoring and underwriting models to refine risk assessment processes. Combine traditional credit metrics with predicted default months to develop more accurate risk profiles and make informed lending decisions.

6. Portfolio Management: Monitor and adjust our loan portfolio composition based on predicted default months to optimize risk-return trade-offs. Diversify our portfolio by balancing exposure to customers with different predicted default months to mitigate overall default risk.

7. Marketing and Customer Acquisition: Tailor marketing campaigns and customer acquisition strategies to target segments of the population with favorable predicted default months. Highlight the benefits of our loan products, such as competitive

interest rates or flexible repayment options, to attract customers with lower default risk.

8. Continuous Improvement: Continuously evaluate the performance of our default month prediction model and refine it based on feedback and new data. Incorporate additional features or advanced modeling techniques to enhance prediction accuracy and stay ahead of evolving market dynamics.

By incorporating predicted default months into various aspects of our business operations, we can optimize risk management strategies, enhance customer relationships, and drive sustainable business growth. It's essential to maintain a customer-centric approach and prioritize transparency, fairness, and ethical lending practices in all business decisions.

Cost/Benefit Matrix Analysis

After selecting logistic regression as our primary model, we further compared three different techniques for handling class imbalance: Logistic Regression Model with a 0.3 threshold, ROSE (Random Over-Sampling Examples), and SMOTE (Synthetic Minority Over-sampling Technique).

GLM with a 0.3 threshold: This technique involves setting a threshold for classification. In this case, a threshold of 0.3 was used with the Generalized Linear Model.

ROSE (Random Over-Sampling Examples): ROSE is a technique used to handle class imbalance by randomly oversampling examples from the minority class.

SMOTE (Synthetic Minority Over-sampling Technique): SMOTE is another technique used to handle class imbalance. It works by generating synthetic examples from the minority class based on feature space similarity.

We compared these techniques to determine which would yield the maximum total revenue using confusion matrices and cost-benefit matrices. This analysis helped us identify the most suitable model for predicting defaulters in our loan portfolio.

Figure 6.1 Cost/Benefit Matrix for Logistic Regression Models

Confusion Matrices

GLM 1 0.3

Prediction/Actual	Non-Default	Default
Non-Default	**0.928**	**0.068**
Default	**0.002**	**0.002**

GLM 2 ROSE 0.85

Prediction/Actual	Non-Default	Default
Non-Default	**0.927**	**0.068**
Default	**0.004**	**0.002**

GLM 3 SMOTE 0.9

Prediction/Actual	Non-Default	Default
Non-Default	**0.925**	**0.067**
Default	**0.006**	**0.002**

Cost/Benefit Matrix

Prediction/Actual	Non-Default	Default
Non-Default	**$23.65**	**($73.95)**
Default	**$0**	**$0**

Results

Model	EV/App	Total Payoff
GLM 1	**$16.948**	**$191,041**
GLM 2	$16.913	$190,639
GLM 3	$16.912	$190,633

Furthermore, we completed a similar analysis across each of our prior models to determine how the selected logistic model would compare to those models from a cost/benefit perspective. Ultimately, we were able to validate that our selected logistic regression model was still the ideal model from the ones we had developed.

Figure 6.2 Cost/Benefit Matrix for All Models Developed

Cost/Benefit Matrix

For every $100

Prediction/Actual	Non-Default	Default
Non-Default	$23.65	($73.95)
Default	$0	$0

Model	Accuracy	Expected Value
Random Forest	66%	$134,360
Decision Tree	83%	$166,770
Logistic Regression	93%	$191,040
XGBoost	59%	$148,345

Causal Inference

Causal inference is the process of drawing conclusions about causal relationships between variables based on observed data. It aims to understand whether changes in one variable directly cause changes in another variable. In other words, causal inference seeks to establish a cause-and-effect relationship between variables in a dataset, allowing researchers to make predictions about how changes in one variable will affect another.

When conducting causal analysis in a predictive project like predicting loan defaulters, several options are available. Some of the common approaches include:

Experimental Studies: Conducting randomized controlled trials (RCTs) where the intervention (e.g., offering financial education) is randomly assigned to borrowers, and the outcome (loan default) is measured.

Quasi-Experimental Studies: Utilizing natural experiments or quasi-experimental designs where the intervention is not randomly assigned but occurs naturally (e.g., economic shocks like the Covid-19 pandemic).

Observational Studies: Using observational data and statistical methods such as regression analysis, propensity score matching, instrumental variable analysis, and difference-in-differences analysis to estimate causal effects.

Counterfactual Analysis: Comparing what actually happened to what could have happened under different scenarios (e.g., changing income levels in the dataset to observe its effect on loan default).

Instrumental Variable Analysis: Identifying an instrument (a variable that affects the treatment variable but not the outcome directly) to estimate causal effects.

For our project, we have utilized both counterfactual analysis and a natural experiment, which are appropriate and effective methods for understanding the causal factors contributing to loan default. It is also important to note that the structure of our dataset limits our ability to conduct a thorough causal analysis as it is not necessarily a time-series dataset. We are limited to one observation per borrower so we are limited in our ability to evaluate that borrower with multiple observations over a period of time. Therefore, we have focused on a counterfactual and natural experiment approach to provide us with some insight as to potentially why some borrowers default on their loans.

Counterfactual Analysis

In our project, we employed counterfactual analysis as a method of causal inference. Counterfactual analysis allows us to assess the causal impact of a specific variable, such as total limit, on the likelihood of loan default. To recap, the 'Total limit' is a categorical variable indicating the total amount of funds a customer owes to other financial institutions (excluding the automobile loan issued by us).

To implement this, we first built a logistic regression model to predict loan default rate based on various customer attributes. Then, we systematically modified one variable at a time, such as total limit, in our dataset and observed the impact on the default prediction. For example, by incrementally increasing all customers' total limit levels by one level and reapplying the original default prediction model, we were able to determine whether higher total limit levels resulted in a lower default rate. This process helped us identify and understand the causal relationship between total limit and loan default within our loan portfolio.

Here are the comparative results.

Figure 7.1 *comparing the change in performance of non-defaulters*

	Percentage change in True-negative	Percentage change in False - Positive
Origianl dataset - 0_8280 to 8280_25000	17.70053476	-96.08127721
Origianl dataset - 8280_25000 to 25000_74475	-24.47860963	132.87373
Origianl dataset - 25000_74475 to 74475_205000	13.26203209	-71.98838897
Origianl dataset - 74475_205000 to >205000	13.27540107	-72.06095791

The table above illustrates the impact on the default rate as we sequentially adjust the levels of the total limit variable. For instance, when we transitioned the total limit of customers from the range 0_8280 to 8280_25000, the non-default rate of the dataset increased by 17.7 percent. However, not all adjustments resulted in the same magnitude or direction of change. For instance, when we moved from the range 8280_25000 to 25000_74475, the non-default rate decreased by 24.47 percent.

This analysis suggests that approving more loans to individuals within the total limit range associated with a 17 percent higher non-default rate could significantly enhance the overall non-default rate of our portfolio, particularly when viewed in its entirety. Conversely, issuing loans to individuals with a total limit between 25000 and 74475 could prove counterproductive, as evidenced by its negative impact on the total non-default rate, which undermines our overall success.

This observation is consistent with our intuition. When we encounter customers with no other debts with other institutions, it suggests that these individuals may have been evaluated by other institutions and denied a loan. Consequently, it's logical to infer that we were the only institution to grant them a loan, which raises the possibility that our decision might not have been optimal. Therefore, the absence of any other debts at all may not necessarily be a positive indicator; rather, it could indicate the opposite.

Given that our objective is to enhance our ability to predict non-defaulters (and defaulters) and thereby identify the profiles of individuals to whom we should approve loans, this insight into which 'total limit' customers have the lowest default rates enables us to streamline and refine our loan approval process.

Figure 7.2 exhibiting the results of all the counterfactual datasets

total_limit Counter factual analysis - confusion matrix

Origianl dataset		
Accuracy = 82.24	Reference	
Prediction	0	1
0	7480	270
1	1378	152
0_8280 to 8280_25000		
Accuracy = 94.97	Reference	
Prediction	0	1
0	8804	413
1	54	9
8280_25000 to 25000_74475		
Accuracy = 63.55	Reference	
Prediction	0	1
0	5649	174
1	3209	248
25000_74475 to 74475_205000		
Accuracy = 92.52	Reference	
Prediction	0	1
0	8472	308
1	386	114
74475_205000 to >205000		
Accuracy = 92.53	Reference	
Prediction	0	1
0	8473	308
1	385	114

A similar analysis was conducted on the "income level" variable, revealing that customers within the income levels of 1500 to 2000 are the most favorable to us. Transitioning customers from the income range of 1000 - 1500 to 1500 - 2000, as well as from incomes greater than 2000 to the 1500 - 2000 range, resulted in an improvement in the non-default rate. Just like the total limit variable, income level appears to be a significant factor influencing the default rate. Therefore, income level should be carefully considered when approving loans.

How did we determine which variables to include in our analysis? We relied on the **Information Value (IV)** as a valuable tool for variable selection during the model-building process. IV is commonly used to assess the predictive power of variables by measuring the strength of the relationship between a predictor variable

and the target variable. Below is a heatmap displaying the Information Values of the selected variables.

Figure 7.3 Information Values (IV)

		IV
1	G_total_limit_dummy1	0.18
2	G_net_income_dummy3	0.14
3	G_period_employment_dummy4	0.14
4	G_house_status_dummy1	0.12
5	G_risk_grade_dummy5	0.12
6	G_total_limit_dummy6	0.09
7	G_net_income_dummy1	0.08
8	G_risk_grade_dummy3	0.06
9	G_period_employment_dummy3	0.06
10	G_total_limit_dummy4	0.05
11	G_trade_bureau_cnt_dummy1	0.05
12	G_trade_bureau_cnt_dummy2	0.05
13	G_total_limit_dummy3	0.03
14	G_period_employment_dummy1	0.03
15	G_enquiry_cnt_dummy1	0.03
16	G_house_status_dummy4	0.02
17	G_risk_grade_dummy1	0.02
18	G_enquiry_cnt_dummy3	0.02
19	G_net_income_dummy2	0.01
20	G_total_limit_dummy2	0.01
21	G_risk_grade_dummy4	0.01
22	G_enquiry_cnt_dummy2	0.01
23	G_house_status_dummy2	0.00
24	G_house_status_dummy3	0.00
25	G_period_employment_dummy2	0.00
26	G_total_limit_dummy5	0.00
27	G_risk_grade_dummy2	0.00
28	G_net_income_dummy4	0.00

Variables with IV values above a certain threshold are considered to be informative. Typically, IV values are interpreted as follows:

- < 0.02: No predictive power
- 0.02 to 0.1: Weak predictive power
- 0.1 to 0.3: Moderate predictive power
- 0.3 to 0.5: Strong predictive power
- > 0.5: Suspicious, possibly too good to be true

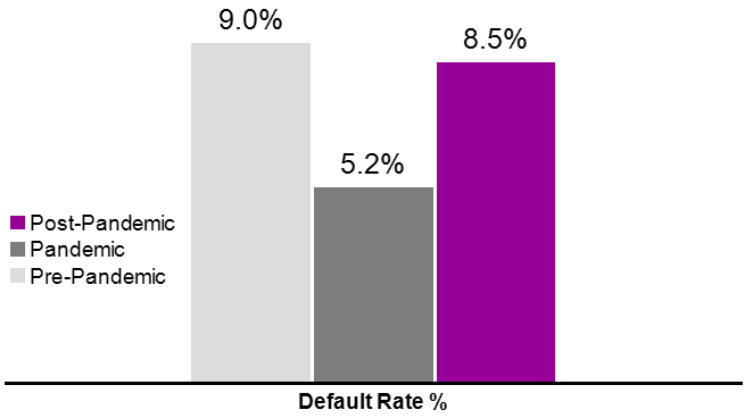
Once we have the IV values for all variables, we selected the ones with IV above a certain threshold for the model building process.

Natural Experiment

An alternative methodology that was used to understand the reason for why certain borrowers default on their loan was to evaluate loans before, during, and after the Covid-19 pandemic in 2020. This analysis was conducted to determine if economic conditions around these time periods could be attributing to defaults that are client was experiencing with borrowers. The approach for the natural experiment analysis was to conduct a statistical analysis as well as a regression model across our dataset to understand how a loan that was dispersed during each of these time periods would affect the probability of an applicant defaulting on their loan.

The definition of a pandemic loan was defined as a loan that was disbursed from March 2020 through November 2021. This definition was defined using the timeframe that contained lockdowns in Malaysia to combat the spread of Covid-19. Loans disbursed prior to March 2020 are considered pre-pandemic and loans after November 2021 are considered to be post-pandemic. The figures below show default rates our client experienced based on each of the loan types and it can be seen that default rates for loans disbursed during the pandemic actually dropped from 9% to 5.2% and eventually increased to 8.5% after the pandemic period.

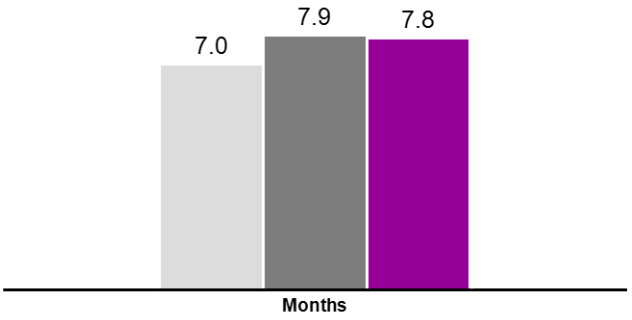
Figure 7.4 Default Rate by Period



Despite the decline in default rates, we also evaluated the time to default for the loans that defaulted in terms of months. We were able to notice that once the

pandemic started, the time to default increased to 7.9 months from the average of 7 months during the pre-pandemic.

Figure 7.5 Average time to default



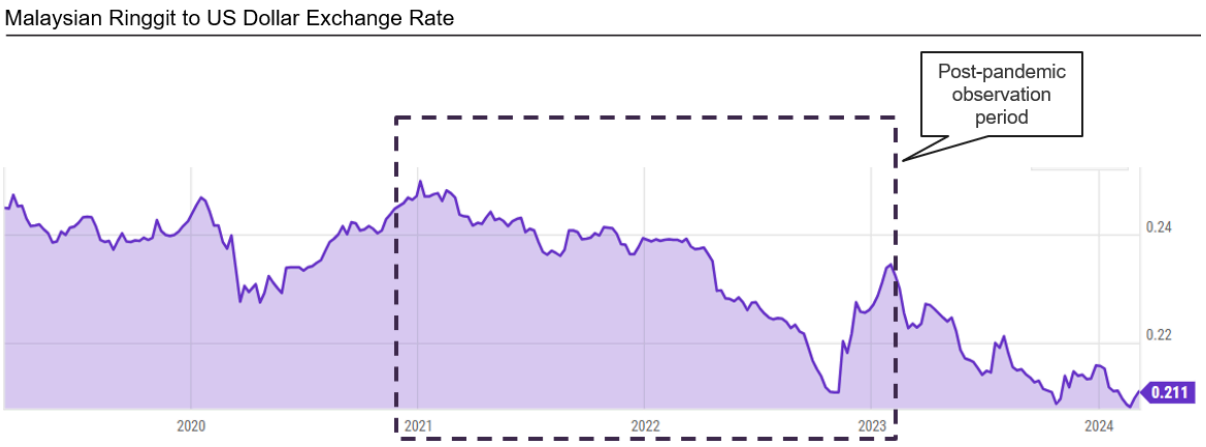
Initially, we believed that this information showed that borrowers who received motorcycle loans during and after the pandemic were less likely to default and would take longer to default on their loans, hence repaying back most of their loan. However, we decided to run a logistic regression to understand if loans disbursed during these time periods significantly influenced whether an applicant was more likely to default. Below is a figure of the logistic regression output.

Figure 7.5 Logistic regression outputs

```
Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                               -1.97923    0.12718 -15.562  < 2e-16 ***
G_total_limit0_8280                        0.86625    0.09555   9.066  < 2e-16 ***
G_total_limit74475_205000                  0.25494    0.10077   2.530  0.01141 *
G_total_limit8280_74475                    0.28895    0.09701   2.979  0.00289 **
G_risk_grade4_5_-999                       0.03662    0.04182   0.876  0.38121
G_risk_grade6_7                           -0.13162    0.04601  -2.861  0.00422 **
G_risk_grade8_10                          -0.60330    0.06446  -9.359  < 2e-16 ***
G_trade_bureau_cnt0_-999                  -0.78226    0.04412 -17.731  < 2e-16 ***
G_period_employment3-5 yrs                -0.21212    0.04723  -4.491 7.08e-06 ***
G_period_employmentBelow 1 yr              0.28380    0.03838   7.394 1.43e-13 ***
G_period_employmentOver 5 yrs             -0.41724    0.04759  -8.768  < 2e-16 ***
G_house_statusRelatives'house (local)      0.34738    0.06819   5.095 3.49e-07 ***
G_house_statusRelatives'house (out state)  0.47141    0.08588   5.489 4.03e-08 ***
G_house_statusRent and Quarter (out state) 0.67433    0.06809   9.903  < 2e-16 ***
G_net_income2,000 =< N                    -0.31386    0.04107  -7.642 2.13e-14 ***
G_net_incomeN < 1,500                      0.22201    0.03695   6.008 1.88e-09 ***
G_enquiry_cnt0_-999                       -0.67475    0.04292 -15.721  < 2e-16 ***
G_enquiry_cnt1                            -0.39469    0.05263  -7.499 6.42e-14 ***
disbur_post_pandemic1                      0.54194    0.03772  14.366  < 2e-16 ***
disbur_pre_pandemic0                      -0.49392    0.04191 -11.784  < 2e-16 ***
```

From this output we are able to see that loans that were disbursed during or after the pandemic had a higher likelihood of default than those that were disbursed prior to the pandemic. Therefore, we could reasonably assume that perhaps macroeconomic factors in the Malaysian market during and after the pandemic could have been attributed to loan defaults. While we cannot say this is an exact reason for the change in behavior, our hypothesis is supported by an evaluation of the Malaysian Ringgit after the pandemic, which can be seen below.

Figure 7.6 MYR/USD Exchange Rate



Key Findings and Recommendations

Dynamic Interest Rate for Motorcycle Loan

Introducing a dynamic interest motorcycle loan product in Malaysia could revolutionize the lending landscape, offering numerous benefits for both the financial institution and its clients. By leveraging advanced predictive analytics models to assess the creditworthiness of applicants, the dynamic interest rate mechanism allows for personalized loan pricing tailored to individual risk profiles.

One of the primary advantages of dynamic interest rates is their ability to optimize revenue while minimizing risk. By utilizing predictive algorithms to assess the likelihood of default, the financial institution can adjust interest rates accordingly, ensuring that clients with higher creditworthiness are offered more competitive rates, while those deemed higher risk are charged higher interest to mitigate potential

losses. This personalized approach not only maximizes profitability for the company but also promotes responsible lending practices by aligning interest rates with borrowers' risk levels.

Furthermore, dynamic interest rates provide flexibility and transparency to clients, empowering them to access financing at rates that accurately reflect their creditworthiness. Unlike traditional fixed-rate loans, which offer standardized terms to all borrowers regardless of risk profile, dynamic interest loans offer greater fairness and affordability for clients with strong credit histories, potentially saving them money over the loan term.

Additionally, by leveraging automated loan prediction systems, financial institutions can streamline the lending process, reducing administrative overhead and accelerating loan approvals. The ability to rapidly assess credit risk and calculate personalized interest rates in real-time enhances operational efficiency and improves the overall customer experience.

Offering dynamic interest motorcycle loans aligns with the evolving needs and expectations of Malaysian consumers, who increasingly seek tailored financial solutions that cater to their individual circumstances. By embracing innovation and leveraging predictive analytics technology, financial institutions can differentiate themselves in the market, attract new clients, and foster long-term customer loyalty.

In summary, dynamic interest motorcycle loans represent a forward-thinking approach to lending that combines data-driven risk assessment with personalized pricing, benefiting both financial institutions and clients alike. By harnessing the power of predictive analytics, companies can optimize revenue, mitigate risk, and enhance customer satisfaction, positioning themselves as leaders in the dynamic and competitive Malaysian lending market.

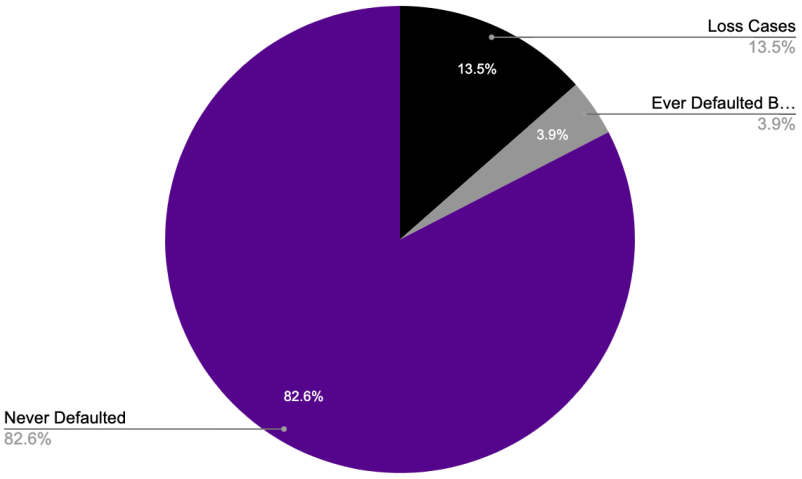
Optimizing Asset Recovery: Selling Motorcycle Loan Defaults

Selling off motorcycle loan defaulters to debt management companies presents a strategic opportunity for the company, particularly in navigating the challenging economic environment that Malaysia currently faces. With the lifting of

the loan moratorium imposed by Bank Negara, individuals and businesses are increasingly negotiating debt restructuring with banks. However, despite these efforts, many customers continue to struggle with loan repayments amidst the ongoing economic uncertainties.

The trend of rising non-performing loans (NPLs) underscores the urgency for proactive measures to manage risk and mitigate losses. By outsourcing the collection of NPLs to specialized debt management firms, motorcycle loan provider companies can efficiently offload distressed assets while recovering a portion of the outstanding balances. This approach not only helps in optimizing the company's balance sheet but also frees up resources that can be redirected towards core business activities.

Figure 8.1 Distribution of Clients Across Each Type



Moreover, partnering with established debt management firms offers access to expertise and resources in debt recovery, enhancing the effectiveness of collection efforts. These companies, with their extensive footprint and track record in debt management services across multiple markets in Asia, provide a valuable opportunity to leverage their capabilities in maximizing recovery rates and minimizing losses.

The composition of bad debts being sold or outsourced, particularly in the motorcycle leasing segment, underscores the relevance of this strategy. As motorbike leasing companies face increasing loan defaults amid stricter credit

conditions and challenging economic conditions, selling off these distressed debts provides a viable avenue to alleviate liquidity constraints and mitigate financial risks.

Looking ahead, the outlook for NPLs remains uncertain, with economic indicators signaling potential challenges in credit growth and asset quality deterioration. In such a volatile environment, proactive measures to manage NPLs and optimize asset performance are imperative for sustained business resilience.

Overall, selling off motorcycle loan defaulters to debt management companies aligns with other strategic objectives of risk management, liquidity optimization, and operational efficiency. By leveraging the expertise and resources of specialized partners, the company can navigate the complexities of the current economic landscape while safeguarding long-term profitability and sustainability.

User Acquisition Adjustments

Since we have insights into the traits of both defaulters and non-defaulters, we can identify the attributes of high-quality clients. This understanding enables us to refine our acquisition strategies to target potential clients more effectively. For instance, individuals earning above RM2,000 are less likely to default. Thus, we may opt to attract more individuals within this income bracket by focusing our advertising efforts on platforms frequented by this demographic. Additionally, variables with strong predictive capabilities, such as Initial Loan Amount, can be leveraged in pre screening applications to identify and exclude groups with a higher propensity for default.

Figure 8.2 Google Ads Set Up Page

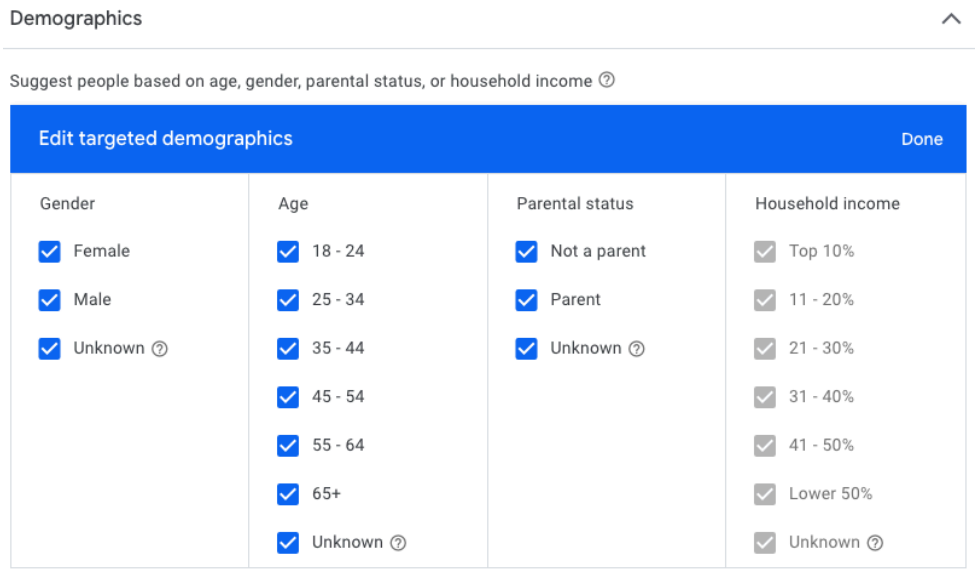


Figure 8.4 Comparison of Observation Percentage by House Status

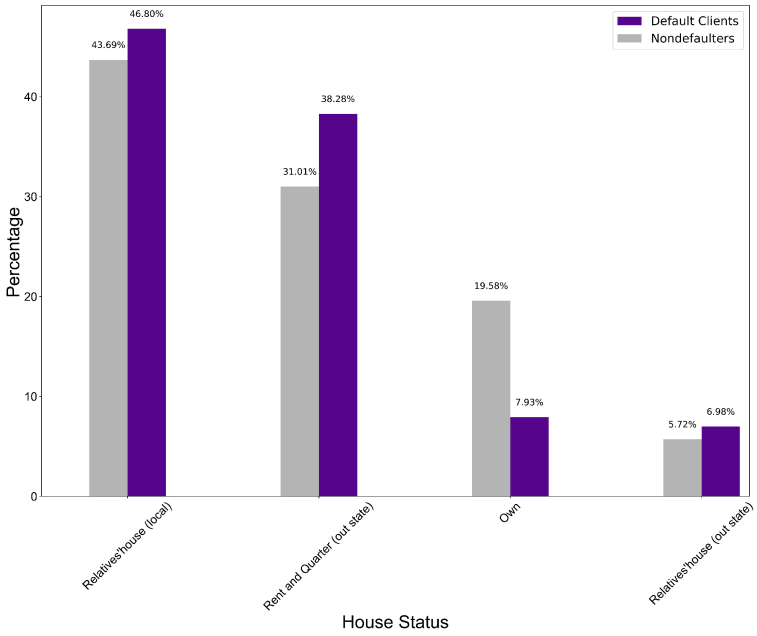
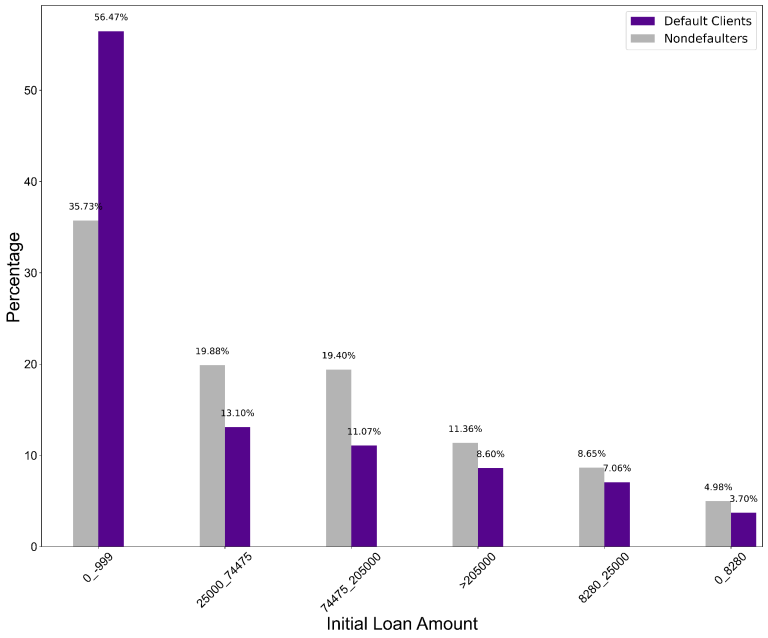


Figure 8.5 Comparison of Observation Percentage by Initial Loan Amount



Appendices

Appx.1 Dictionary of variables

Variable name	Definition
target	Normal case or defaulted case
total_limit	Total initial loan amount
risk_grade	Credit range of the applicant form 3rd party credit agency
trade_bureau_cnt	Count of how many other financial obligations this customer has with other companies like telephone bill etc.
period_employment	How long the applicant has been employed
house_status	Where the applicant lives and if the applicant owns a house
net_income	Applicant's Net income per month in Malaysian Ringgits
enquiry_cnt	Count of how many times customer requested for a loan in past 12 months

References

Author name, “Title of the document”, Publisher, Publication date, URL

Websites:

- Standard Insights, "Leading Modes of Transportation Among Consumers in Malaysia as of January 2023." Statista, April 2023, https://www-statista-com/statistics/1385981/malaysia-preferred-modes-of-transportation/
- Statistia Market Insights, “Motorcycles: market data & analysis”, Statistia, December 2023, https://www-statista-com/outlook/mmo/motorcycles/worldwide
- MYR and USD exchange rate, Investing.com https://www.investing.com/currencies/myr-usd-chart