NATIONAL ECONOMICS UNIVERSITY

APPLICATION OF MACHINE LEARNING IN PREDICTING THE DEFAULT PROBABILITY OF INDIVIDUAL CUSTOMERS AT CREDIT INSTITUTIONS

Instructor: Mrs. Nguyen Thi Quynh Giang

Prepared by:

1. **Bui Thi Mai Luong (Leader) - 11193191**
2. **Nguyen Thi Thuy Duong – 11191271**
3. **Ho Duc Duy – 11191318**

Hanoi, June 8th, 2022

ACKNOWLEDGEMENTS

Our team would like to express our deep respect and gratitude to Mrs. Nguyen Thi Quynh Giang, lecturer at the Department of Mathematical Economics, who directly guided and helped with knowledge, materials and methods for the group to complete this scientific research project. At the same time, the group also sent Mr. Nguyen Thanh Tuan deep thanks for the advice and help on knowledge so that the group can have the right orientation on the topic.

Our team would like to thank the leadership of the Department of Mathematics and Economics, the Board of Directors of the National Economics University, and the teachers who supported and accompanied the group.

The team would also like to thank the relatives, family and friends who have always supported and helped the group to complete this report.

Hanoi, June 8th, 2022

Implementation Group

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

List of Tables:

List of Figures:

INTRODUCTION

1. Overview

In 1999, Dunn & Kim conducted a study on the probability of customer default based on a data set by the Survey Research Center at Ohio State University that conducted a monthly random telephone survey of 500 households in this state from February 1998 to May 1999. The two researchers focused primarily on the relationship between customers' likelihood of default and their demographic's financial characteristics. If they initially hypothesized that factors such as education, income and home ownership had a strong influence on a customer's likelihood of default, the results of the study suggested the opposite, that it is demographic variables such as age, marital status and number of children that are the factors that strongly influence the probability of customers defaulting.

According to Peter's research in 2006, he predicted default for a dataset of 3431 households based on Australian Bureau of Statistics data, ABS 2001. Accordingly, the age of the holder had a strong impact. on household default. Also, income and demographic variables such as education, marital situation have a significant impact on the probability of default.

Another study conducted by Jacobson & Roszbach with a data set of 13,338 personal loan applications at a bank in Sweden in 2003 showed that, out of 57 initial recognition variables, only 16 remained. The variable is applied to the model after dealing with the high correlation problem. In the 16 variables, variables such as age, income, annual income growth, and some eligibility for credit waivers are the ones that have the most impact on the likelihood of default.

A broader study in Southeast Asian countries, specifically the prediction of creditworthiness by Logistic Regression model in two countries, Thailand and Vietnam by Dufhues et al was carried out in 2011. Results forecast on 467 households in Thailand show that in this country only two factors, namely the education level of the household head and the value of household investment, have a positive effect on the probability of household default and found no influence from other factors. The same study was conducted in Vietnam with 198 households, two factors were found to have an impact on the probability of default: loan of the household head and ethnicity of the household head.

Developing more than the above study, in addition to the Logistic regression model, Kocenda & Vojtek (2011) used Decision Trees to build credit risk models for Czech retail lending banks. The study was carried out with 3403 observations and 21 explanatory variables. Both models show that variables such as customer assets, education level, personnel status, target loan and number of years with a bank account are the most important variables. insolvent goods. Since then, the authors have commented on the importance of academic societies and made recommendations for new members of the EU.

Using the same Logistic model, Ojiaki & Ogbukwa conducted a study on the repayment results of farmers when borrowing loans at Nigerian banks in 2012. The results of the study indicated that in the variables of characteristics for the individual households included in the model, there are only three variables that affect the household's debt repayment capacity: household size, agricultural land use area and loan amount.

Research paper on this issue was conducted in 2018 by Abid et al. The authors have compared the efficiency of two logistic regression models and analyzed the discriminant analysis model when forecasting customer default based on data at a commercial bank from 2010 to 2012 The results show that when the index of the logistic regression model reaches 99%, the number of Discriminant Analysis only reaches 68.49%, the ie prediction error rate is up to 31.51%. From that, the authors concluded that the Logistic Regression model has a better ability to predict default than the Discriminant Analysis.

2. Rationale

In this period of global integration and opening of the economy, commercial banks or credit institutions are likened to the lifeblood of the economy on both a micro and macro perspective, when it plays the role of financial intermediaries that transfer and store currency. It conducts lending, investment and almost every activity of the economy, as well as the operation of businesses. In fact, a commercial bank or a credit institution (non-bank) also operates as a business, but it does not deal in ordinary items on the market, but it does business in "currency". Therefore, the main goal of banks is still to maximize profits and minimize risks. But it can be seen that, with this special product, commercial banks will gain huge profits from credit activities, but at the same time, will have to face a lot of risks. Bad debts and defaults of borrowers due to insolvency cause heavy losses, even leading to insolvency if the investment is too large. Especially during the pandemic, the situation is more bewildering than ever and the market is in a panic, thus making it a perilous time for financial consultation.

Faced with that reality, minimizing financial risks, reducing debt and increasing risk identification of borrowers has become a necessary and urgent task. Until now, building a system to help evaluate credit risk optimally has always been a difficult problem faced by commercial banks, there is a lot of research with many different approaches and aspects.

However, most of the research is based on social practice as well as accumulated experience from old losses combined with conclusions about credit risk in a rather subjective way. Read first the overview, well-defined researchers, the majority of the study focuses on Researching and explaining the constituent factors to break the debt of borrowers at banks or financial institutions. Although there are also studies that directly suggest customer default such as using some Logistic Regression or Random Forest models, there are still few studies that apply scientific sets in this field. analysis and quantification. Additionally, previous studies mostly used balance data while the current customer data set is mostly unbalanced data. Moreover, related data to customers is in increasing abundance, along with the rise of effectiveness of models and estimation methods, resulting in projects with objective customers and binding rationale.

With the urgency of the topic as well as the status quo, the research team decided to choose the topic: "Application of machine learning to predict the probability of customer default at a credit institution". The team hopes that this will be a highly applicable topic and that banks or credit institutions can apply it in their process of reducing their lending risks. From there, credit institutions can map out appropriate lending and investment strategies.

3. Research Purpose

The study was carried out with the aim of developing a model to help commercial banks assess the possibility of customer default; segment customers according to the probability of default, classify the signals affecting the customer's default. based on customer signals and information. From there, credit institutions can map out appropriate lending and investment strategies. Specific goals include:

- Build a highly applicable model in predicting default.
- Evaluate and analyze the influence of such signals on the customer's default.
- Propose a number of solutions and recommendations in the lending strategy of credit institutions to each customer group after segmenting customers according to the possibility of default.

4. Research subjects and research scope

- Research subjects: Factors affecting the probability of default of customers at credit institutions.
- Research scope: Customers who have had transactions with credit institutions.

5. Outline of the research

The project report has five main parts:

- Part 1: Theoretical basis.
- Part 2: Research Methods.
- Part 3: Research Data
- Part 4: Insight and model results.
- Part 5: Conclusion

PART 1: THEORETICAL BASIS

1.1 Bank credit

1.1.1 Definition

Bank credit is a credit relationship between a bank and other entities in the economy (be it individuals or businesses), with the transfer of assets between the participating parties. In particular, with the function of financial intermediaries, banks both act as lenders providing credit products to individuals and businesses, as well as borrowers receiving deposits from entrepreneurs, businesses and individuals or issuing certificates of deposit and bonds to mobilize capital in society.

According to Article 20, Law on Credit Institutions 2010 has shown that: “Credit extension means an agreement allowing an organization or individual to use a sum of money or a commitment allowing the use of a sum of money on the repayment principle by such professional operations as lending, discount, financial leasing, factoring, bank guarantee and other credit extension operations”.

Therefore, it can be understood that the credit of banks is the relationship of transferring the right to use capital or assets between commercial banks and economic individuals within a certain agreed period based on the principle of refund. pay both the principal and interest of the borrower at maturity.

Based on the principle of "borrowing to lend", commercial banks are both capital mobilization organizations and major investors for most subjects in society from individuals, private enterprises to state enterprises. More specifically, based on the specific nature of capital flows, when there is a difference in time and quantity between income and expenditure in the process of social reproduction, it often forms "idle" cash flows, and the bank is a bridge to mobilize idle cash flows. cash flow to "pour" back, but the organization needs to borrow capital at the same time to maintain costs and generate profits for the government. The bank credit relationship is based on three main relationships:

- Relationship between banks and businesses.
- Relationship between the bank and individual customers.
- Relationships between banks (both domestic and foreign banks).

1.1.2 Characteristics and roles of bank credit

a. Characteristics of bank credit:

Firstly, the subjects in the bank's credit relationship include the lender who assigns the right to use capital and the borrower who receives the right to use the capital. In some cases, there is also a third party in this relationship, acting as a guarantor for the borrower as a means of hedging.

Secondly, bank credit is built on the basis of trust and voluntariness between the parties involved. This credit relationship is only really established when the bank believes in the customer's willingness to repay and the ability to repay. On the other hand, the borrower also believes that he can use the loan capital effectively. However, the transaction process between the bank and the customer must also be based on legal documents/contracts such as credit contracts, debt receipts, guarantees, etc.

Thirdly, the nature of a bank is a business operating on the basis of profit, "borrowing to lend", so for each credit transaction, the bank must receive a return value greater than the value of the loan (including principal, interest and fees).

Fourthly, based on the asymmetry in information between banks and borrowers, credit loans often contain many risks. Since then, one of the causes of credit risk comes from the customer appraisal process of commercial banks. Besides, in addition to the subjective reasons from within the bank, there are also objective causes of credit risk such as market volatility and changes in economic policy, cycle economic, or even random risks such as natural disasters, epidemics, etc.

Fifthly, bank credit can meet all customers in the national economy in the form of currency loans, which is very popular and flexible. Lenders mainly mobilize capital from members of society, not entirely from personal capital like other forms. Cash capital can meet the needs of many borrowers. Flexible loan terms: short, medium and long term. Banks can adjust capital sources with each other to meet the needs of the term of customers. Satisfying the capital needs of individuals and organizations in the economy to the fullest extent because they can mobilize capital in many forms and in large volumes.

b. The roles of the bank credit.

It cannot be denied that banks are an important link in promoting production and development, regulating and circulating capital, and at the same time increasing the efficiency of capital flows in the market. businesses and individuals. Besides, now that the operation model of the bank is gradually improving, the actors in the economy can access loans more easily and quickly, creating a rhythmic harmony for the economy.

For businesses, bank credit pulls future demand to the present, production scale increases rapidly, innovation level and diversity in quality are increasing. This has made the entire process of production, exchange, distribution and consumption take place quickly and efficiently, which is the foundation for economic growth.

For banks, this is considered a business that mainly brings high profits to credit institutions. Having abundant capital helps to expand scale, invest in technological innovation, etc. to create strong and sustainable development.

For the economy, in the current global integration economy, foreign borrowing has become an objective need for countries around the world.

1.1.3 Forms of credit by commercial bank

a. Classification by credit period:

Short-term credit: term of less than 12 months, are documents with a term of less than 1 year for the purpose of making up for short-term capital shortages or short-term expenses.

Medium-term credit: Loan term from 12 to 60 months; Possibly with goals that have the ability to quickly recover capital such as equipment purchases, fixed assets (assets with rapid wear and tear); business expansions; or implementing projects with a small and medium scale;…

Long-term credit: The loan term is greater than 60 months, even up to 30 years; aims to provide capital for customers' long-term goals such as building and repairing factories, starting a business, …

b. Classification by the purpose of using capital:

Credit for production and circulation of goods: For enterprises and business entities to use for the purpose of production and circulation of goods.

Consumer credit: For individuals to meet consumption needs such as shopping, building houses, vehicles, studying... to meet daily spending needs.

c. Classification by the credit object:

Working capital credit: Used to form working capital for business organizations.

Fixed capital credit: used to form fixed assets.

d. Classification by the degree of credit to customers:

Secured credit: These are loans that are lent by the bank with the customer's collateral/ pledge or the guarantee of a third party. This type of loan allows banks to hedge risks with customers with low creditworthiness.

Unsecured Credit: A loan without a mortgage, pledge or guarantee from a third party. It is only based on the credit of commercial banks to customers, usually applied to honest and capable customers.

e. Classification by customer’s debt repayment method:

One-time repayment loans: are loans that pay both principal and interest once upon the maturity date agreed upon by the bank and the customer.

Installment loans: These are loans that customers will pay principal and interest in installments at an agreed period. This loan usually applies to loans for real estate, residential, consumer, commercial.

f. Classification by economic sectors:

Loans to financial institutions such as credit contracts between banks, insurance companies and some other financial institutions.

Loans for corporate customers: These are loans granted to businesses to use as working capital, project investment, production expansion, procurement of materials, etc.

Loans for individual customers: These are loans to individuals to meet the needs of life through the issuance of credit cards.

1.2 Credit to individual customers of the bank.

1.2.1 Definition

In the banking business, credit is defined as a transaction of assets between a lender and a borrower, in which the lender transfers the property to the borrower for use within a certain period of time. According to the agreement, the borrower is responsible for unconditionally repaying the principal and interest to the borrower when the payment is due (Ho Dieu 2003). Credit activities are considered as the main profitable business activities for banks, usually for most banks, credit outstanding accounts for more than of total assets and income from credit activities. accounts for about ½ to more than 70% of the bank's total income. Since credit is the main profitable item in banking activities, it is also the main risk item of commercial banks (Le Van Te 2009).

Personal customer credit at a bank is an economic category that reflects the relationship between two objects: the bank and its individual customers. This relationship is recognized based on a credit agreement, where the bank transfers a certain amount of currency to an individual customer to use within a certain period of time, and promises to repay it at the same time. agreed period.

1.2.2 Characteristics and roles of individual customer credit

a. Characteristics of individual customer credit:

The first is that the loan size is usually small, but the number of loans is large. The main purposes of personal loans are: (i) Supplementing business capital. These production and business activities are not large in scale, besides, most customers come to the bank when they have a relatively large amount of capital, so they only supplement a relatively small gap; (ii) Serving daily consumption needs. Personal loans for this purpose are directly for living expenses such as buying a house, car, shopping for household items, construction, home repairs, studying abroad. These are all small loans, but because customers are all individuals in society, from high- and middle-income people to low-income people, credit needs are quite diverse with no lack of variety.

Secondly, individual customer credit reduces many costs. Due to the special features of individual customers, which are large in number and widely dispersed, it will reduce a lot of costs for jobs such as network expansion, advertising, and marketing to facilitate work. Next, individual customers in each location and area, develop the entire staff to serve customers quickly and accurately from the person receiving the application, appraising it to deciding on the loan, disbursement and debt collection as well as related expenses such as administrative costs, stationery, etc. Therefore, credit interest rates of individual customers are often higher than that of corporate customers. For individual accounts, making information about the customer's identity, background and financial situation is often incomplete and difficult to collect, so the budget usually has less costs for determining loan appraisal and approval.

Thirdly, individual customers are more risky than corporate customers. In addition to the general management of the credit bank, personal lending banks also have risks arising from the customers themselves, such as the financial condition of an individual or a household may change rapidly.

Fourthly, the quality of the borrower's financial information is often not high. When assessing a loan, this information is an extremely important factor for the bank to make a loan decision, besides the soundness and legitimacy of capital needs, debt repayment capacity and collateral. For institutional customers, capturing customer information is relatively convenient because there are many publicly available sources of information such as financial statements, tax payment status, reputation and relationships with partners... In contrast, for individual customers, it is often difficult to fully, clearly and accurately assess their identity, repayment source, and loan use purpose, leading to asymmetric information risk, making the appraisal process difficult. inaccurate customer identification.

b. The role of individual customers:

For individual customers, personal customer credit is a safe and quick source of funds for daily spending requirements or business capital shortages. Currently, the bank has also been developing this format more and more diversely to easily help individual customers access capital more conveniently and easily.

For credit institutions, personal credit is akin to a “media darling” in service to promote its brand. Along with the service of providing signals to a large number of individual customers are sales tactics through products such as: providing e-banking services, issuing signals and saving deposits, etc. help improve position and competitiveness in the field.

For economic development, individual customers credit as a bridge to exploit leisure capital flows from individual customers. Thereby, helping the capital circulation process be circulated more effectively from the place of excess capital to the place of lack of capital.

1.3 Several factors influence the probability of default of individual customers

1.3.1 Credit risk of individual customers

Credit risk is simply the possibility that a bank borrower or counterparty would fail to satisfy its obligations under agreed-upon terms. Credit risk management is to increase a bank's risk-adjusted rate of return while keeping credit risk exposure below acceptable limits. Banks must manage both the overall credit risk and the risk associated with individual credits or transactions. Banks should also think about how credit risk interacts with other hazards. Credit risk management is a vital component of a comprehensive risk management strategy and is critical to any banking organization's long-term success.

Loans are the greatest and most visible source of credit risk for most banks; nevertheless, other sources of credit risk can be found throughout a bank's operations, including in the banking and trading books, as well as on and off the balance sheet. Acceptances, interbank transactions, trade financing, foreign exchange transactions, financial futures, swaps, bonds, equities, options, and the extension of commitments and guarantees, as well as the settlement of transactions, are all examples of financial instruments in which banks are increasingly exposed to credit risk (or counterparty risk).

1.3.2 The probability of default of individual customers

Currently, there is no unified definition of the customer's default, but mainly focuses on whether the customer is able to repay the loan or not.

However, the Basel Committee has also provided a definition of "default" - default (inability to repay) in the Basel Committee on Banking Supervision - 2006 document as customers who have one or more of the symptoms in the following conditions:

- The customer is not able to fulfill the payment obligation in full upon maturity without taking into account the sale of assets (if any) to repay.
- Customers have bad debts with overdue time of more than 90 days.

For Vietnam alone, bad debts are debts from group 3 to group 5: debts that are assessed as likely to lose a part of capital and interest (group 3 debt), debts that are likely to cause loss of money. high losses (group 4) and irrecoverable debts (group 5 debt). Group 2 debts receive less attention than those with poor repayment ability, but those loans only need more attention because customers still have the ability to repay.

Thus, it can be understood that the default ability of an individual customer is the possibility that this customer will not be able to pay the loan and interest when it is due or the possibility of bad debt at the bank.

From the bank's perspective, credit risk is the risk that the bank may face when it fails to collect debts, causing problems with capital turnover and liquidity.

In fact, the bank's decision to grant credit must be based on assumptions about the customer's ability to repay. However, any prediction or test is subject to error, and validation results can be faulty for many different reasons. Therefore, from the bank's management point of view, credit risk is unavoidable and it is like a "companion" in business activities and the best course of action of a bank is to reduce the risk to a minimum.

Credit risks from individual customers can be divided into the following 4 cases: principal and interest cannot be recovered; not collect enough principal and interest.

Accordingly, the failure to collect interest on time only causes low damage to the bank. In this case, the bank will incur additional frozen interest unless the bank exempts or reduces interest rates for customers. In case the risk occurs with the principal when the customer does not pay the debt on time, then the bank will have an overdue debt arising. However, it has not been completely lost, but it is possible that customers will still pay the debt. The case that causes the highest credit risk is the unrecoverable customer's debt (possibly due to default).

1.4 Status of assessing the possibility of default of individual customers at commercial banks

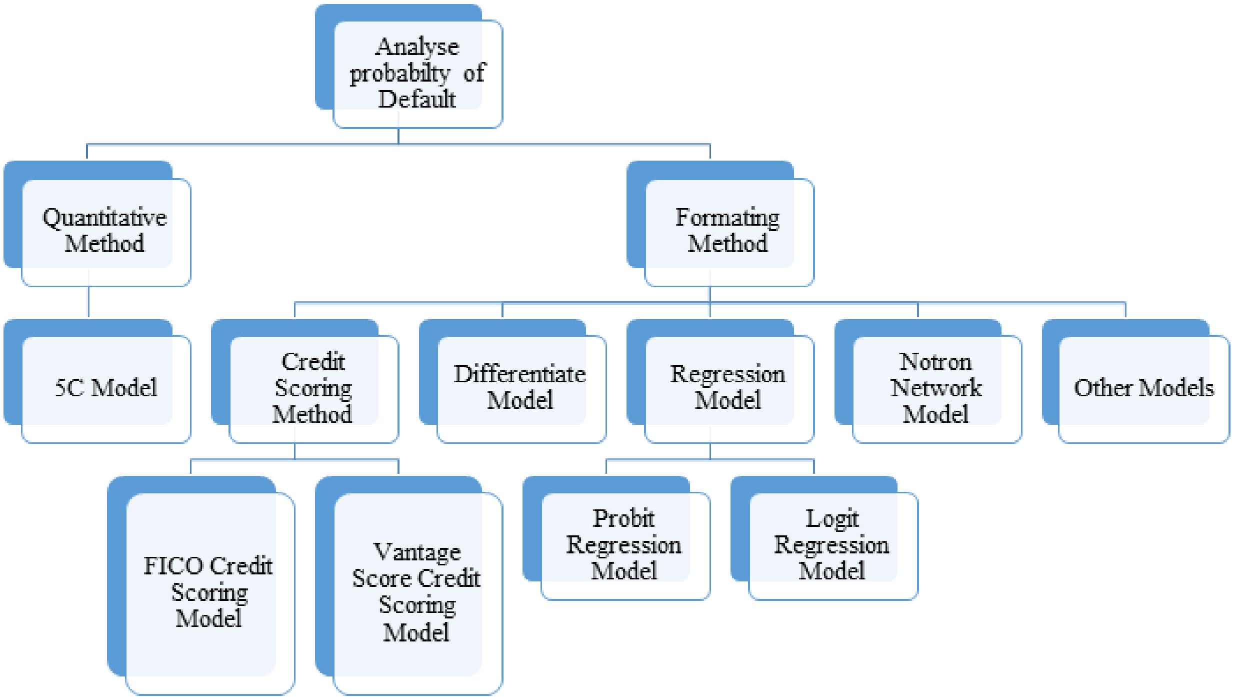


Figure 1: Analyse probability of Default

1.4.1 Qualitative Model (5C Model)

This is a model that considers the intention as well as the ability of the customer to repay the loan when it comes to maturity. According to the Commercial Banking textbook (Tran Thi Xuan Huong and Hoang Thi Minh Ngoc, 2012) and research documents, the specific 5C model includes the following five elements:

- Reputation and attitude of customers (Character): This is a factor that is quite subjectively assessed from the customer's side based on favorable factors such as customer transaction attitude, accuracy in information, etc. There are also a number of other qualitative factors to be considered such as professional seniority, personal qualities, education level, etc.
- Capacity: The borrower's capacity in this case is considered on factors such as civil act capacity, civil liability capacity and financial capacity. This is one of the key factors to evaluate whether a customer can repay the debt or not.
- Borrower's income (Cash): This is a factor reflecting the customer's ability to repay loans. Here, the borrower's income is determined based on factors such as salary, income from property rental, business revenue, interest from shares/dividend, etc.
- Collateral: In addition to income, this is the second source for banks to ensure the repayment ability of customers and is also a factor for banks to consider a reasonable loan rate.
- Conditions: Depending on the credit policy of the bank from time to time, the bank may issue specific regulations on preferential policies for credit lines, interest rates and collaterals. for each customer. For example, for customers borrowing capital for export purposes, the condition may be that revenue and interest must be paid through a bank.

Advantages and disadvantages of the 5C model:

- Advantages: The 5C model is quite simple, easy to implement and evaluate
- Disadvantages: The results of this model depend too much on the accuracy of information provided by customers and subjective factors from the predictive and analytical ability of credit appraisers.

1.4.2 Quantitative model

- **FICO personal credit score model:**

Fair Isaac Corp has built a FICO personal credit score model based on a mathematical equation with the input of customers' credit information collected from credit reports of financial institutions. provided. Then, FICO compares the collected information with the standards drawn through the huge credit report store in the past to assess the credit risk level of the bank. Accordingly, with each customer having their own credit score, the lower the score, the higher the credit risk of the customer. In the convention of the FICO model, 300 is the lowest score and 850 is the highest score. Specifically, the grading is based on the following criteria:

Weight	Evaluation Criteria
35%	Payment history: the number of late payments, bankruptcy, etc., the more credit score, the lower the credit
30%	Amount owed: According to FICO, there are six different metrics in a debt portfolio including debt-to-limit ratio, number of accounts with balances, and amount owed across categories. Different accounts and installment loans. More debt will lower your credit score
15%	Length of credit history: The longer the history of the information is, the more reliable the credit score will be. There are two metrics in this category: the average age of accounts on a report and the age of the oldest account.
10%	New credit: frequent borrowing by customers is considered a sign of financial difficulty, so it will have a negative impact on credit scores.
10%	Types of credit used: Different types of debt such as installments, revolving, consumer finance and mortgages will be scored differently.

Table 1: Scoring criteria according to the FICO credit model

Accordingly, customers with credit scores of 700 or higher are customers with high credit ratings. In contrast, banks need to be careful when lending to customers with credit scores of 620 or less. The FICO credit score model is a simple and easy-to-implement model, but the model still lacks the important factor that is the factors related to the customer's relatives - an important cause affecting the creditworthiness of the customer.

- **Personal credit score model Vantage Score:**

Along with the FICO model, the VantageScore model is widely used in the US. With the input of credit data from three companies: Equifax, Experian, and TransUnion, the Vantage Score model was built with 5 ratings descending from A to F, corresponding to a score from 501 to 990. Specifically, A: 900 – 990, B: 800 – 899, C: 700 – 799, D: 600 – 699, F: 501 – 599.

Advantages: The credit scoring models are more complete than the 5C model when incorporating a variety of factors into the model affecting the customer's ability to repay loans such as: personal information, information/ Debt history, financial level and non-financial indicators, etc.

Disadvantage: The results are still quite subjective based on subjective opinions from the evaluator. The model's criteria have not been quantified to measure customer default and identify important factors.

- **Multiple Discriminant Analysis – MDA:**

The MDA model allows to classify individual customers into groups and analyze differences between groups. The MDA method is in fact capable of discriminating into more than 2 groups, the number of discriminant functions will be less than the number of discriminant groups by 1 unit. Here, discriminant analysis finds a linear function (discriminant function) of financial and market variables to distinguish between two groups of high and low default risk (Durand, 1941). The difference between the two groups was measured by the mean of the discriminant variables - the z-score.

- **Neural network model:**

One of the most popular quantitative models used to assess default in the world is the neural network model. This model uses the principle of parallel computing, consisting of many simple interconnected processes. In each of these processes, each calculation is performed simply, by one neuron. These neurons connect, organize together logically to solve complex tasks. The neural network model can perceive and process in real states for incomplete input data or process for large variable scale data. This model is especially suitable for predictive models where there is no general mathematical formula to express the relationship between input and output variables.

According to Rosenberg and Gleit (1994), this model, which has fewer assumptions than other models, is best suited for heterogeneous credit grants. However, at present, the application of the neural network method is still quite complicated and not guaranteed to be much more effective than traditional methods such as discriminant analysis and regression models (Altman et al. events, 1994).

In addition, there are a number of other models such as Markov chain model, recursive partitioning (Breiman, 1984), mathematical programming (Hand, 1981), ... which are also used in the analysis of customer default.

PART 2: PROJECT’S RESEARCH METHOD

2.1. Traditional Credit Scoring Method

At some banks, to evaluate the bank's default ability, the technical side will usually process the data using the WOE - IV method due to its good classification to then create a new data set with the weighted features and finally processed into a training logistic regression model.

2.1.1. WOE – IV

WOE (weight of evidence) is a feature engineering and feature selection approach that is frequently used in the scorecard model. Based on the capacity and strength to predict bad debt, this technique will rate factors as strong, medium, weak, no influence, and so on. The information value index IV (information value) generated using the WOE approach will be used as the ranking criterion. In addition, the model generates feature values for each variable. This value will represent the difference in the distribution of good and bad. Here are several examples:

For continuous and categorical data, the WOE approach will use several processing techniques:

- When dealing with continuous variables, the WOE will label each observation based on the bins value label to which it belongs. The bins will be calculated from the continuous variable in such a way that the number of observations in each bin is equal. We must first identify the number of bins in order to determine the bins. The endpoints of bin intervals can be thought of as quantiles.

- When dealing with categorical data, the WOE may treat each class as a separate bin, or it may combine numerous groups with a limited number of observations into a single bin. Furthermore, the degree of difference between the excellent and bad distributions as evaluated by the WOE index may be utilized to identify groups with similar category traits. If their WOE values are near to one other, they may be assigned to the same group. Furthermore, if its numbers are considerable, the Null case can be regarded as a separate group, or it can be placed into other groups if it is a minority.

WOE is calculated by dividing the percent of good by the percent of bad by the natural logarithm (log to base e).

$$WOE = ln\left(\frac{\% \, Good}{\% \, Bad} \right)$$

Where:

- %Good: Distribution of good records across all bins. The sum of the column is 1.
- %Bad: Distribution of bad records across all bins. Similar to %Good, it also sums to 1

Properties of WOE:

- The larger the WOE value at a bin is a sign that the feature is very good in identifying good records and conversely, the smaller the WOE value, the better the bin feature will be in identifying bad records.
- WOE > 1 then the distribution of Good records is dominant over Bad and vice versa.

IV (Information Value): An index of information value that determines whether or not a variable has the ability to classify bad debt. Formula IV is:

$$IV = \sum_{i=1}^{n} (\%Good_i - \%Bad_i) * WOE_i$$

We see that IV (Information Value) always gets a positive value because WOE_i and $(\%Good_i - \%Bad_i)$ are covariates. The IV value will tell us how large of a difference there is between the percentages of Good and Bad in each bin. If the IV is high, the difference in distribution between percent Good and percent Bad will be considerable, and the variable will be more useful in categorizing records; if the IV is low, the variable will be less useful in classifying records. Some publications also include the following criteria for determining the power of variables based on their IV value:

- If the IV value is less than 0.02, the variable has no effect in classifying Good/Bad records.

- If the IV value is from 0.02 to 0.1, the predictor only has a shaky link with the Goods/Bads odds ratio.
- If the IV value is from 0.1 to 0.3, the predictor has a medium strength link with the Goods/Bads chances ratio.
- If the IV value is between 0.3 and 0.5, the predictor has a high link with the Goods/Bads odds ratio.
- If the IV value is over 0.5, the variable is very strong, however, must be checked again to avoid the case where the variable has a direct relationship to the definition of Good/Bad records.

2.1.2. Logistic Regression Model

Logistic Regression is a sort of supervised learning technique that computes the connection between input and output characteristics using a logistic/sigmoid function. Despite the fact that the name contains the term "Regression," this is a Classification algorithm. As a result, unlike other Regression methods, Logistic Regression is used to predict a binary outcome (with a value of 0/1 or -1/1 or True/False) based on the data it receives.The result of a normal regression problem is a continuous quantity, whereas the result of a logistic regression problem is a discrete quantity. For example, the price of a property after knowing its location, and the return result with Logistic Regression will be a discrete amount like 0.1 signifying an event that is projected to happen or not. It is, in fact, one of the most widely used Machine Learning algorithms. Typical issues for this algorithm include: predicting whether a message is spam or not; determining whether a person is at risk for stroke based on the findings of a physical exam, etc

The estimated value of y is $\hat{y} = \sigma\,(w_0 + w_1x_1 + w_2x_2 + \ldots + w_kx_k\,)$ after regressing the Logistic model. The customer's repayment probabilities is then estimated using the formula below:

$$P = \frac{1}{1 + e^{-y}}$$

The generated P-value in the range (0,1) is compared to the bank's customer rating criterion. The threshold value for customer classification was set at 0.5 in this study to make it easier to compare the models' efficacy. This means that if the P-value is less than

0.5, the client is likely to default; conversely, if the P-value is greater than 0.5, the customer is likely to return the loan.

2.2. Machine Learning Approach

2.2.1 Theory of the techniques

a. Correlation

To determine the strength of a relationship between data, correlation coefficient formulas are utilized. The formulas produce a number between -1 and 1, with 1 denoting a strong positive relationship, -1 denoting a strong negative association, and zero denoting no relationship at all.

A correlation value of 1 indicates that for every positive increase in one measure, a set proportion of the other increases positively. For instance,shoe sizes increase in (nearly) perfect proportion to foot length. A correlation coefficient of -1 indicates that for every positive rise in one measure, a set proportion of the other decreases. The amount of gas in a tank, for example, diminishes in (nearly) perfect proportion to speed. There is no positive or negative rise for every increase of zero. The two aren't related in any way.

The correlation coefficient's absolute value indicates the strength of the link. The stronger the association, the higher the number. For instance, |-0.75| = 0.75 has a stronger association than 0.64.

b. Encoding

Encoding is the process of converting data into a different format for use in a different device or system. The approach utilized in this transformation has a public content that anyone can use. After encoding, the data can be reversed or utilized to encode to another representation system.

For instance, base64 encoding is a popular encoding. This is one of the forms of encoding used to transform binary data to text. Consider the situation where we need to email a photo. Because binary data cannot be sent over email, we must base64 encoding it to convert it to text.

A few things to keep in mind about encoding:

- The scenario of data exchange between distinct systems or contexts is addressed by encoding.
- The objective of encoding is to ensure that the data is usable.
- We can deduce from the first statement that encoding has nothing to do with data security.
- All encoded actions are reversible, and the encoding technique is open, due to the need to interchange data.
- The encoder's input does not have to be text.

c. Sampling:

Sampling is a crucial approach in statistics, and it plays a significant role in establishing the correctness of a research/survey. Any inaccuracy in the sample procedure will have a direct impact on the final outcome. There are several strategies that may be used to collect samples based on our demands and circumstances. We may divide them into two categories:

- Probability Sampling: Algorithms in this category utilize a "random" function to ensure that every element has an equal probability of being chosen, assuring impartiality. Although there is still a danger that the sample chosen does not contain the primary features of the population, this amount of risk is quantified using statistical methods. This procedure is also known as random sampling. Simple Random Sampling, Stratified Sampling, Systematic Sampling, Cluster Sampling, and Multistage Sampling are examples of approaches in this category.
- Non-probability sampling: Techniques that do not employ a random function come under this category. To choose individuals for the sample set, this approach relies on the researchers' awareness of the current population. As a result, the findings of this sampling are highly likely to be biased. The reason for selecting this strategy is because, while random sampling will not be favorable if the material error is concentrated in the distribution of specific populations or if the population is small, it will be advantageous if the population is large. This strategy will produce better findings than the probability sampling method if the researcher has strong judgment and

expertise. Convenience Sampling, Purposive Sampling, Quota Sampling, and Referral / Snowball Sampling are some of the methods included in this category.

One of the most popular methods for dealing with uneven data sets is data resampling. There are two types of methods for this in general:

- Under sampling refers to the process of reducing the number of observations of the majority group to the same number as the minority group. Under sampling has the advantage of being able to make sample balance quickly and readily without the use of a simulation algorithm. However, it has the disadvantage of reducing the sample size greatly.
- Over sampling is a technique for resolving sample imbalance by enlarging the sample size of the minority group using various strategies. There are two primary strategies for performing over sampling: Samples that can be replicated; Simulate the new model using the old samples' synthesis.

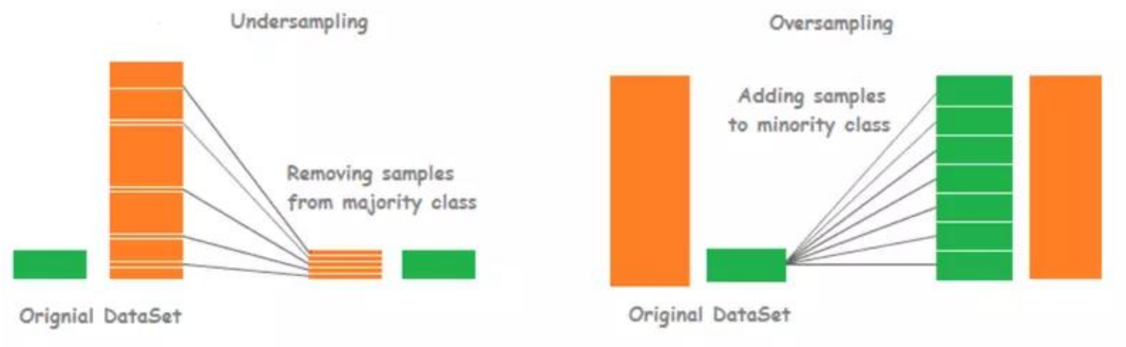


Figure 2: Undersampling and Oversampling

Oversampling techniques usually take precedence over undersampling techniques. The reason for this is that in the latter, we tend to discard versions of the data that may include critical information. I specifically mention an oversampling technique in this study, specifically SMOTE (Synthetic Minority Oversampling Technique).

SMOTE is an oversampling technique in which composite samples for the minority class are generated. This approach aids in overcoming the problem of overfitting caused by random oversampling. It focuses on feature space in order to produce new examples via interpolation between positive cases that are close together.

The total number of oversampled observations, N, is set initially. In general, the binary class distribution is selected to be 1:1. However, this can be changed depending on the situation. The iterative process then begins by selecting an active class instance at random.

The KNNs for that instance (by default 5) are then obtained. Finally, to interpolate the new composite individuals, N of these K people are chosen.

The KNNs for that instance (by default 5) are then obtained. Finally, N people are chosen from among these K to interpolate the new composite individuals. The difference in distance between the feature vector and its neighborhoods will be determined using any distance measure.

This difference is now multiplied by any random value in the range (0,1) and added to the feature vector. The illustration below depicts this:

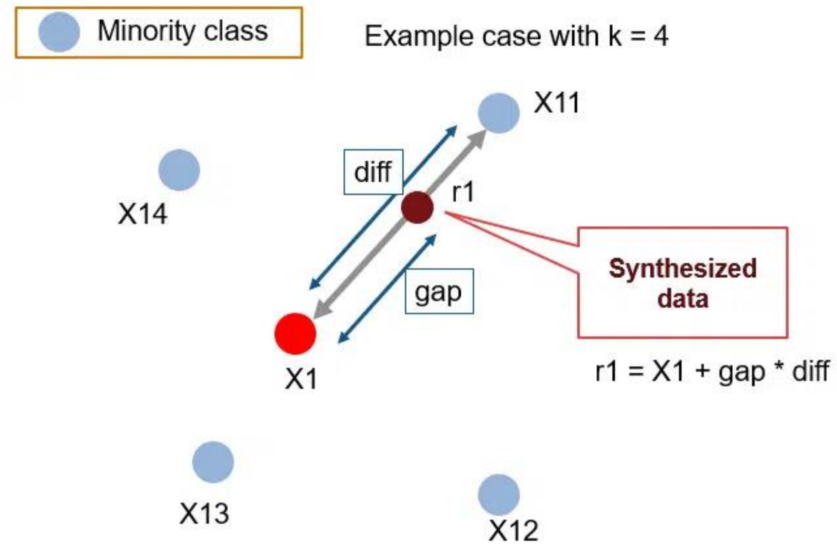


Figure 3: The difference in distance between the feature vector and its neighborhoods

While this technique is quite useful, it does have certain drawbacks.

- The composite instances are all generated in the same direction, with an artificial line connecting them. As a result, the decision surface generated by some categorization algorithms becomes more complicated.
- SMOTE has a tendency to generate a lot of zeros in the object space.

2.2.2 Theory of the model Machine Learning

a. Random Forest:

The Random Forests method is a supervised learning system. It may be used for classification as well as regression. It is also the most adaptable and user-friendly algorithm. Its name indicates how it works in part: A forest is made up of trees, and the more trees there are, the stronger the forest. It operates in four steps:

- Choose at random from the supplied data collection.
- Create a decision tree for each sample and compute prediction results for each decision tree.
- Vote for each predicted result.
- Choose The last forecast is the most likely outcome.

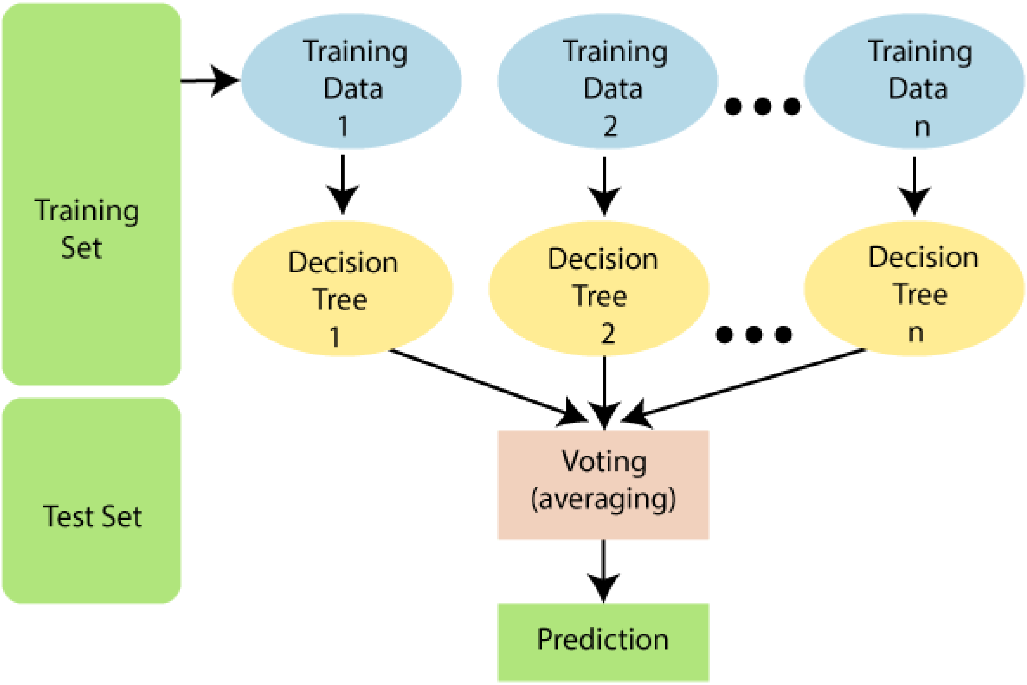


Figure 4: Steps of Random Forest

Pros: Because of the large number of decision trees involved in the process, random forests are considered an accurate and powerful method. It doesn't have any overfitting issues. The fundamental reason for this is that it averages all predictions, canceling out any bias. Both classification and regression problems can benefit from the approach. Missing values can also be handled using random forests. The mean values can be used to substitute continuous variables, or the approximate average of the missing values can be calculated. You can acquire relative feature importance, which might help you pick the features that will aid the classifier the most.

Cons: Because there are so many decision trees in random forests, it takes a long time to make forecasts. Every time it makes a prediction, all of the trees in the forest must make a prediction for the same input and vote on it. The entire procedure takes time. Models are more perplexing than decision trees, which allow you to make judgments quickly by following a path through the tree.

b. LightGBM:

In January 2016, Microsoft originally released an experimental version of LightGBM, and LightGBM swiftly supplanted XGBoost as the most popular ensemble method. What distinguishes LightGBM is:

- LightGBM finds split sites during tree building using "histogram-based algorithms" rather than "pre-sort-based techniques" as is usual in other boosting tools. This enhancement allows LightGBM to boost training speed while decreasing memory utilization. Both xgboost and lightgbm employ histogram-based methods; however, lightgbm outperforms xgboost in two techniques: GOSS (Gradient Based One Side Sampling) and EFB (Exclusive Feature Bundling), which considerably accelerate the computation process.

- LightGBM builds trees leaf-by-leaf, unlike most other boosting programs (including xgboost) that grow trees level (depth)-by-level. Leaf-wise chooses nodes to expand the tree based on the optimization of the entire tree, but level-wise optimizes on the branch under consideration, hence for a limited number of nodes, trees produced from leaf-wise typically outperform trees created from level-wise.

2.3 Metrics for evaluating model

2.3.1 ROC - AUC curve:

An assessment measure for binary classification issues is the receiver operator characteristic (ROC) curve. This is a probability curve that shows TPR vs FPR at various threshold levels, separating the "signal" from the "noise." The area under the curve (AUC) is a summary of the ROC curve that measures a classifier's ability to differentiate between classes.

The AUC measures how well the model distinguishes between positive and negative classes. The greater the AUC, the better the model's performance.

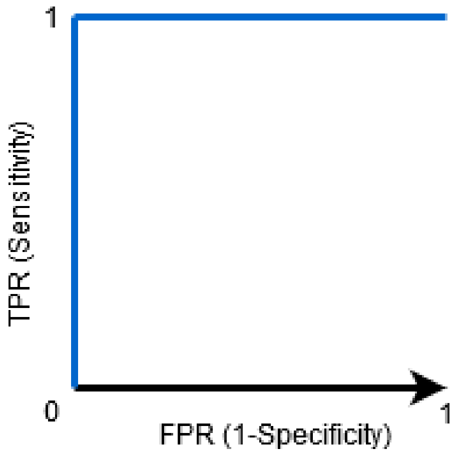


Figure 5: The classifier when AUC =1

When AUC = 1, the classifier is capable of successfully distinguishing between all Positive and Negative class points. If the AUC was 0, however, the classifier would expect all Negatives to be Positives and all Positives to be Negatives.

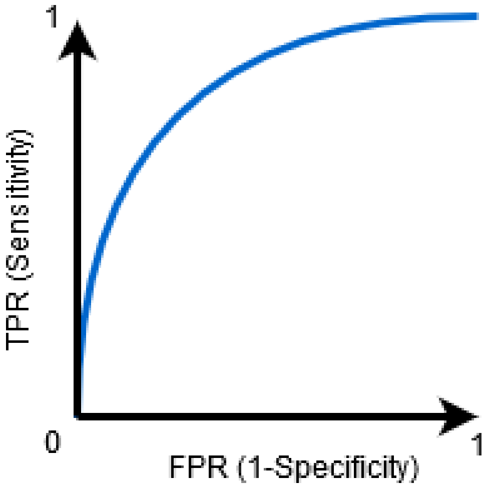


Figure 6: The classifier when 0.5 < AUC < 1

When the 0.5 < AUC < 1, there is a good likelihood that the classifier will be able to tell the difference between positive and negative class values. Because the classifier can recognize more True positives and True negatives than False negatives and False positives, this is the case. When the AUC is less than 0.5, the classifier is unable to distinguish between Positive and Negative ratings. That is, for all data points, the classifier predicts a random or constant class.

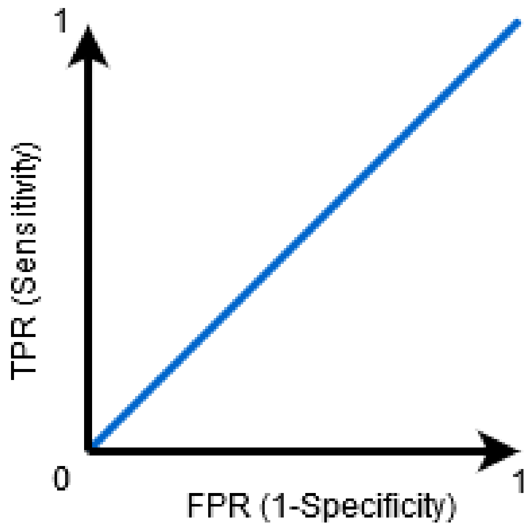


Figure 7: The classifier when AUC = 0.5

When the AUC = 0.5, the classifier is unable to differentiate between Positive and Negative class points. That is, the classifier predicts either a random or a constant class for all data points.

As a result, the higher a classifier's AUC score is, the better it is at distinguishing between positive and negative classes.

2.3.2. Precision & Recall:

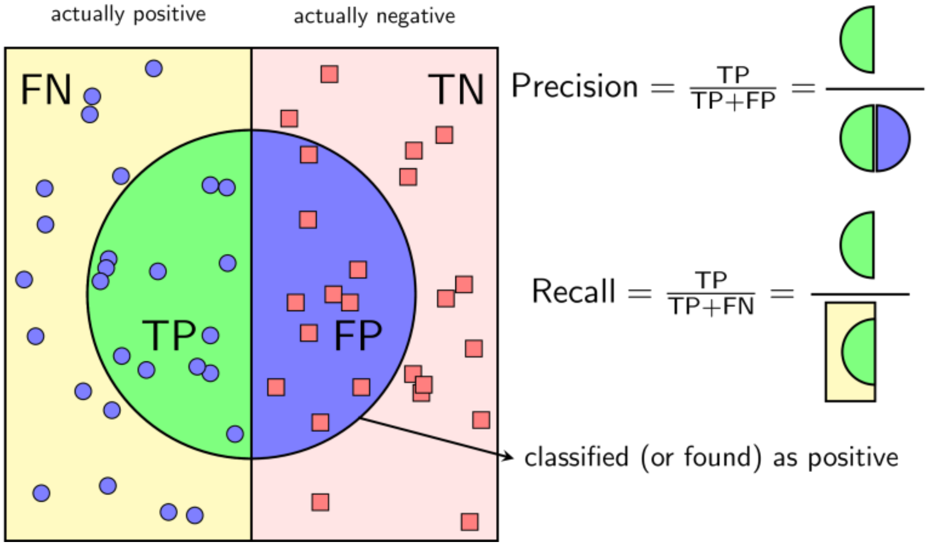

actually positive
actually negative
FN
TN
TP
FP
Precision = $\frac{TP}{TP+FP}$ =
Recall = $\frac{TP}{TP+FN}$ =
classified (or found) as positive

Figure 8: Precision and Recall

The ratio of accurately anticipated positive observations to total predicted positive observations is known as **Precision**. The ratio of accurately predicted positive observations to all observations in the actual class is known as **Recall**.

The greater the precision, the greater the quantity of positive prediction model scores. Precision = 1, that is, the model properly predicts all Positive scores, or there are no Negative points that the model wrongly forecasts as Positive.

The lesser the amount of false positives, the higher the recall. The model recognizes all points labeled as Positive with a recall of one.

2.3.3. F1 score:

However, the model's quality cannot be determined solely by Precision or Recall.

The model can only generate predictions for the point where it is most certain if it is only using Precision. Then Precision = 1, however this model isn't really good. If the model predicts that all points will be positive, only use Recall. Then Recall = 1, however this model isn't very good.

F1-score is then applied. **F1-score** is the harmonic mean of precision and recall (assuming these two quantities are different from zero).

F1-score is calculated according to the following formula:

$$\text{F1-Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

PART 3: RESEARCH DATA

3.1 Meaning of Features

The data set used to describe the relationship between creditworthiness of a loan and personal financial and demographic information has been recorded through the credit records of individual customers. of the Home Credit organization. Home Credit is a non-banking financial institution - a type of business organization in the financial - monetary field. The dataset is publicly available on Kaggle by Home Credit, the service is offered to a less privileged group of customers and the question asked was to predict whether customers will be able to pay back their loans or if there is a problem. This question is very important for any bank or lending institution because it directly affects the bottom line of the organization's business. This dataset contains eight different datasets: application train, bureau, bureau balance, credit card balance, installments payments, previous payments and POS CASH balance. The dataset is completely unbalanced (skewed labels), with the target variable being able to repay the debt or not, the value 0 represents the customer who repays the debt accounted for 91.9% of the total sample, where the value 1 represents customers who have not yet paid their debts account for only 8.07%. In addition, the data has a lot of missing values, which can greatly affect the quality of the training model used.

Figure 9 - a diagrammatic structure of data sets that shows the relationship between the data sets. The next section is an explanation of each dataset.

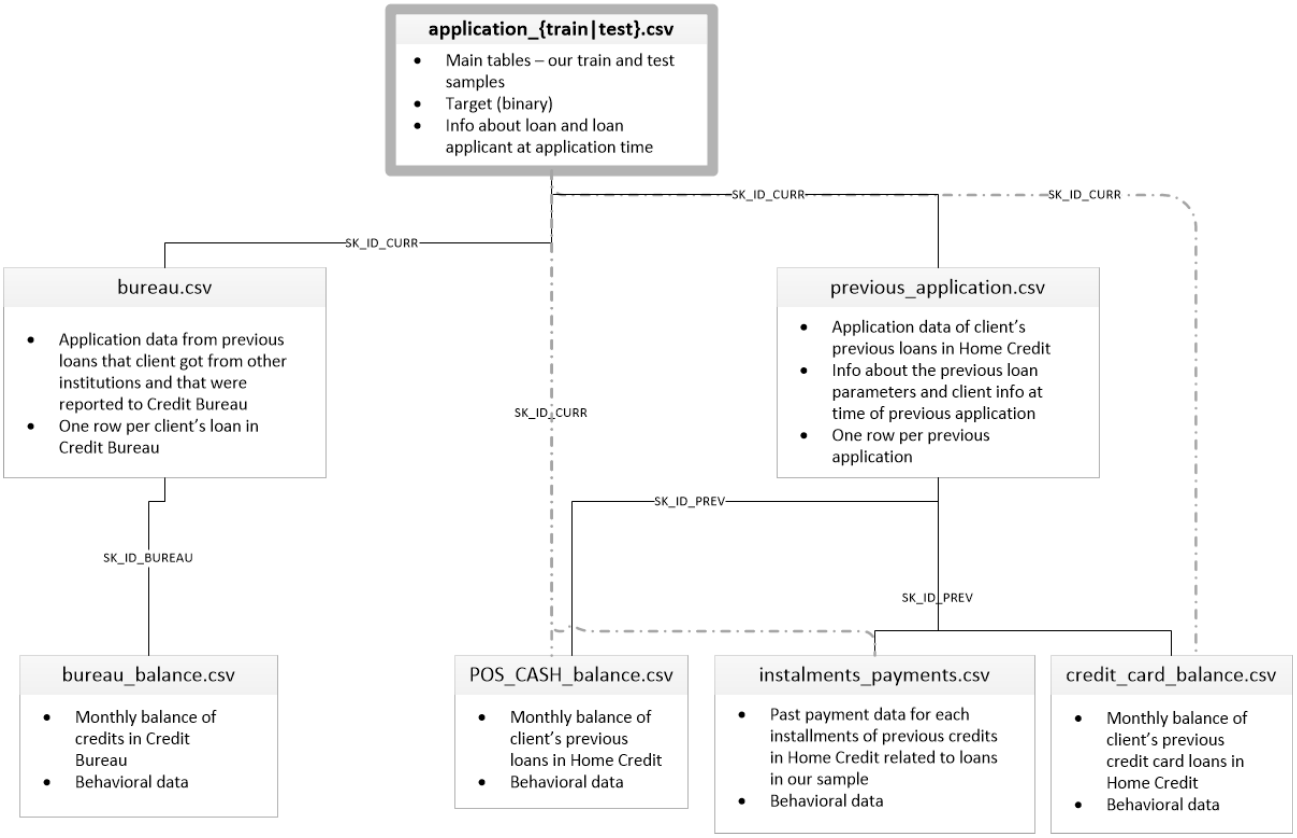

application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_PREV
SK_ID_BUREAU
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

Figure 9: The structure of data sets

- **Application train/test:**

The application train/test dataset is the main table that includes training and testing data with key information about each loan application at Home Credit for each debt. This information is shown line-by-line, and each customer is identified by the SK_ID_CURR variable. In the training dataset, there is an additional TARGET column with a value of 0 representing a successfully paid debt and a value of 1 unpaid debt.

The dataframe application train has 307511 rows and 122 columns (including the TARGET feature). The dataset has 32/74 features containing missing values; contains 16 features of type categorical and 106 features of type numeric.

The dataset is divided into 5 main information: customer's personal information, customer's debt information, customer's debt-related information, customer's property information and other information.

Type	Features	Type (Categorical/ Numeric)	Meaning	Number of Unique values	% Null	Examples
Customer's personal information	CODE_ GENDER	Categorical	Gender of the client	3	0	M,F,XNA
	CNT_ CHILDREN	Numeric	Number of children the client has	12	0	
	AMT_ INCOME_ TOTAL	Numeric	Income of the client	2548	0	
	NAME_ INCOME_ TYPE	Categorical	Income sources of Applicant	5	0	Working, Commercial associate, Pensioner, State servant, Unemployed
	NAME_ EDUCATION_ TYPE	Categorical	Client's education level	5	0	Secondary/ secondary special, Higher education, Incomplete higher, Lower secondary, Academic degree

	NAME_FAMILY_STATUS	Categorical	Family Status of Applicant's	5	0	Married, Single / not married, Civil marriage, Widow, Separated
	OWN_CAR_AGE	Numeric	Total number of years from Applicant had car	62	64.97	
Customer's debt information	TARGET	Numeric	Customer's debt status	2	0	0, 1
	NAME_CONTRACT_TYPE	Categorical	Types of loan	2	0	Cash loans, Revolving loans
	AMT_CREDIT	Numeric	Credit amount of the loan	5603	0	
	AMT_ANNUITY	Numeric	Loan annuity	13672	0	
	NAME_HOUSING_TYPE	Categorical	Types of house which applicants applied for loan	6	0	House/apartment, With parents, Municipal apartment, Rented apartment, Office apartment, Co-

						op apartment
Customer's debt-related information,	NAME_TYPE_SUITE	Categorical	Who accompanied applicant when applying for the previous application	7	0.39	Unaccompanied, Family, 'Spouse, partner', Children, Other_B, Other_A, Group of people
	DAYS_REGISTRATION	Numeric	Total number of days from date of Applicant registration to now		0	Negative Number
	DAYS_ID_PUBLISH	Numeric	Total number of days applicant published ID		0	Negative Number
	EXT_SOURCE_1 (2,3)	Numeric	Normalized score from external data source		1-56.39 2-0.22 3-19.86	
Customer's property information	FLAG_OWN_CAR	Categorical	Flag if the client owns a car	2	0	N, Y

	FLAG_OWN_REALTY	Categorical	Flag if client owns a house or flat	2	0	N, Y
Other information	OCCUPATION_TYPE	Categorical	Occupation of applicant's who applied of loan	18	31.32	Laborers, Sales staff, Core staff, Drivers, Managers, High skill tech staff, Accountants, Medicine staff, Security staff, Cooking staff, Cleaning staff
	REGION_RATING_CLIENT	Numeric	The rating of clients in the city		0	
	REGION_RATING_CLIENT_W_CITY	Numeric	The rating of clients in the city		0	
	WEEKDAY_APPR_PROCESS_START	Numeric	Days of the APPR(Annual Professional Performance Review) process		0	
	HOUR_APPR_PROCESS_START	Numeric	Approximately at what hour did the client apply for the loan		0	

Table 2: Information of the application train/test dataset

***Descriptive statistics (Numeric Feature):**

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
AMT_ INCOME_ TOTAL	308000	169000	237000	25600	113000	147000	203000	117000000
AMT_ CREDIT	308000	599000	402000	45000	270000	514000	809000	4050000
AMT_ ANNUITY	308000	27100	14500	1620	16500	25000	34600	258000

Table 3: Information about the numeric features

- **Bureau:**

The Dataset Bureau contains previous loan applications. All customer credits previously provided by other financial institutions have been reported to the Credit Bureau.

For the loan in the dataset sample, the number of credits the customer had in the Credit Department prior to the application date is shown by each row in the table. In addition, a loan in the application data can have multiple previous credits.

The dataframe bureau has 1716428 rows and 17 columns (17 features); the dataset has 6 features containing missing data (of which 2 features: AMT_ANNUITY, AMT_CREDIT_MAX_OVERDUE contain more than 50% missing data); there are 3 features whose value is categorical and 14 features of type numeric.

Type	Features	Type (Categorical/ Numeric)	Meaning	Number of Unique values	% Null	Examples
Basic infomation	CREDIT_ ACTIVE	Categorical	Status of the Credit Bureau	4	0	Closed, Active

			(CB) reported credits			
	CREDIT_ TYPE	Categorical	Type of Credit Bureau credit (Car, cash, ...)	4	0	currency 1, currency 2
Time	DAYS_ CREDIT	Numeric	How many days before current application did client apply for Credit Bureau credit		0	min: -2922 max: 0
	CREDIT_ DAY_ OVERDUE	Numeric	Number of days past due on CB credit at the time of application for related loan in our sample		0	min: 0, max: 2792
Money	AMT_CRED IT_MAX_ OVERDUE	Numeric	Maximal amount overdue on the Credit Bureau credit so far (at application date of loan in our sample)		65	min: 0 max:116,000,000 mean: 3,830

	AMT_ CREDIT_ SUM	Numeric	Current credit amount for the Credit Bureau credit		0	min: 0 max: 585.000.000 mean: 355.000
	AMT_ CREDIT_ SUM_DEBT	Numeric	Current debt on Credit Bureau credit		15	min: -4.710.000 max:170.000.000 mean: 137.000

Table 4: Information of the Bureau dataset

- **Bureau balance:**

This data set describes the goods balance through the previous account credit in the Credit Bureau. Each row represents information for each historical month of all previous signals reported to the Credit Bureau, and a previous signal can have multiple rows, one row for each month of the signal (the table contains multiple rows). of the previous loan application (previous loan application). It has many relationships.

The dataset has 27299925 rows and 3 columns (3 features); the dataset has no features that contain missing values; there is 1 feature whose value is categorical and 2 features of type number.

Type	**Features**	**Type (Categorical/ Numeric)**	**Meaning**	**Number of Unique values**	**%Null**	**Examples**
Basic information	STATUS	Categorical	Status of Credit Bureau loan during the month	8	0	C, 0

Time	MONTHS_ BALANCE	Numeric	Month of balance relative to application date (-1 means the information to the freshest monthly snapshot, 0 means the information at application - often it will be the same as -1 as many banks are not updating the information to Credit Bureau regularly)		0	min: -96 max: 0

Table 5: Information of the Bureau balance dataset

- **Previous Application:**

Dataset Previous applications contain loans received from the same institution (Home Credit Default Risk). The customer's previous Home Credit loan records are included in the application data. Each current loan in the application data can have multiple previous loans. Each previous application is represented by a commodity and is identified using the SK_ID_PREV feature.

The dataset has 1670214 rows and 37 columns (37 features); of which 15 features contain missing values: 4 features contain more than 50% missing values); the dataset has 16 features whose value is categorical and 21 features that are numeric.

Type	**Features**	**Type (Categorical/ Numeric)**	**Meaning**	**Number of Unique values**	**% Null**	**Examples**
Basic Information	NAME_ CONTRACT_ TYPE	Categorical	Contract product type (Cash loan, consumer loan [POS] ,...) of the previous application	4	0	Cash loans, Consumer loans
	FLAG_ LAST_APPL_ PER_ CONTRACT	Categorical	Flag if it was the last application for the previous contract. Sometimes by mistake of client or our clerk there could be more applications for one single contract	2	0	True, False
Money	AMT_ ANNUITY	Numeric	Annuity of previous application		22	min: 0 max: 418.000 mean: 16.000
	AMT_ APPLICATION	Numeric	For how much credit did client ask on the previous application		0	min: 0 max:6.910.000 mean:175.000

	AMT_DOWN_PAYMENT	Numeric	Down payment on the previous application		0	min: -0.9 max: 3.060.000 mean: 6.700
Time	WEEKDAY_APPR_PROCESS_START	Categorical	On which day of the week did the client apply for previous application	7	54	Mondey, Tuesday
	HOUR_APPR_PROCESS_START	Numeric	Approximately at what day hour did the client apply for the previous application	0 - 23	0	0, 23

Table 6: Information of the Previous Application dataset

- **POS CASH balance:**

The dataset describes the monthly balance of previous POS (point of sale) and cash loans that the subscriber has with Home Credit. This table each row can current history each month of all account credit before that in Home Credit relating to the loans in the data.

The dataset has 10001358 rows and 8 columns (8 features); there are 2 features that contain missing values but the percentage missing values is very small (0.26%); dataset has 1 feature whose value is categorical and 7 features of type number.

Type	Features	Type (Categorical/ Numeric)	Meaning	Number of Unique values	% Null	Examples
Basic infomation	NAME_ CONTRACT_ STATUS	Categorical	Contract status during the month	9	0	Active, Completed
Time	MONTHS_ BALANCE	Numeric	Month of balance relative to application date (-1 means the information to the freshest monthly snapshot, 0 means the information at application - often it will be the same as -1 as many banks are not updating the information to Credit Bureau regularly)		0	min: -96 max: -1 mean: -35
	CNT_ INSTALMENT	Numeric	Term of previous credit (can change over time)		0	min: 1 max: 92 mean: 17.1
Money	CNT_ INSTALMENT _FUTURE	Numeric	Installments left to pay on the previous credit		0	min: 0 max: 85 mean: 10.5

Table 7: Information of the POS CASH balance dataset

- **Credit card balance:**

The dataset includes monthly data about previous card signal customers. Each row is a card balance signal from one month ago.

The dataset has 3840312 lines and 23 columns (23 features); there are 9 features containing missing values (the highest percentage with missing values is 20%); the data set has 1 feature whose value is categorical and 22 features of numeric type.

Type	**Features**	**Type (Categorical/ Numeric)**	**Meaning**	**Number of Unique values**	**% Null**	**Examples**
Money	AMT_ BALANCE	Numeric	Balance during the month of previous credit		0	min: -420.000 max:1.510.000 mean: 58.300
	AMT_ CREDIT_ LIMIT_ ACTUAL	Numeric	Credit card limit during the month of the previous credit		0	min: 0 max:1.350.000 mean: 154.000
	AMT_ DRAWINGS_ATM_ CURRENT	Numeric	Amount drawing at ATM during the month of the previous credit		19	min: -6.830 max: 2.120.000 mean: 5.960
Information basic	NAME_ CONTRACT_STATUS	Categorical	Contract status (active signed,...) on the previous credit	7	0	Active - Completed
Time	SK_DPD	Numeric	DPD (Days past due) during the month on the previous credit		0	min: 0 max: 3.260

	SK_ DPD_DEF	Numeric	DPD (Days past due) during the month with tolerance (debts with low loan amounts are ignored) of the previous credit		0	min: 0 max: 3.260

Table 8: Information of the Credit card balance dataset

- **Installment payment:**

The data describes the payment history of previous loans at Home Credit group. Each row corresponds to a one-time payment or an installment payment corresponding to a previous Home Credit bank payment.

The dataset has 13605401 rows and 8 columns (8 features); there are 2 features containing missing values but the missing percentage is very small (about 0.021%); all features in this dataset are numeric by value.

Type	**Features**	**Type (Categorical/ Numeric)**	**Meaning**	**% Null**	**Examples**
Basic infomation	NUM_ INSTALMENT _VERSION	Numeric	Version of installment calendar (0 is for credit card) of previous credit. Change of installment version from month to month signifies that some parameter of payment calendar has changed	0	min: 0 max: 178
Money	AMT_ INSTALMENT	Numeric	What was the prescribed installment amount of previous credit on this installment	0	min: 0 max: 3.770.000 mean: 17.100
	AMT_ PAYMENT	Numeric	What the client actually paid on previous credit on this installment	0	min: 0 max: 3.770.000 mean: 17.200

Time	DAYS_INSTALMENT	Numeric	When the installment of previous credit was supposed to be paid (relative to application date of current loan)	0	min: -2.920 max: -1
	DAYS_ENTRY_PAYMENT	Numeric	When was the installments of previous credit paid actually (relative to application date of current loan)	0	min: -4.920 max: -1

Table 9: Information of the Installment payment dataset

3.2 Data preprocessing

3.2.1 Cleaning data

First, process the data of each table and then combine them into a complete dataset.

When processing data, it is found that there are some mismatches in model building as well as label coding to be included in the model, so it will be removed from the data set such as:“SK_ID_CURR”, “SK_BUREAU_ID”, “SK_ID_CURR”, “SK_ID_PREV”, “SK_DPD”, “SK_DPD_DEF”,...

Next, process data containing missing values, we do this by removing features with a percentage of missing data up to more than 75%.

Similar to drop feature, there will be some features created to train the model better (based on data discovery):

- DAY EMPLOYED = $\frac{DAYS\ EMPLOYED}{DAY\ BIRTH}$

- INCOME CREDIT PERC = $\frac{AMT\ INCOME\ TOTAL}{AMT\ CREDIT}$

- INCOME PER PERSON = $\frac{AMT\ INCOME\ TOTAL}{CNT\ FAM\ MEMBERS}$

- ANNUITY INCOME PERC = $\frac{AMT\ ANNUITY}{AMT\ INCOME\ TOTAL}$

- PAYMENT RATE = $\frac{AMT\ ANNUITY}{AMT\ CREDIT}$

3.2.2 Test for correlation

The higher the correlation, the greater the chance of multicollinearity. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

3.2.3 WOE - IV

WOE - IV, this method will help select the features that have the best ability to classify the target variable. As mentioned in section 2 – Part 2, features with IVs less than 0.02 are unlikely to be able to classify variables. Therefore, consider removing features with IV index < 0.02.

After removing features with value IV < 0.02, the dataset has 67 features.

3.2.4 Imbalance data

Before putting data into the training model, we will handle imbalance data.In this project, three different methods were used: Undersampling, Oversampling and SMOTE . Each method when training the model will give different results (The results will be presented in the following section). In these unbalanced data processing techniques, the research team will bring class 0 and class 1 to 1:1 ratio, the processed data sets all apply this ratio.

PART 4: INSIGHT AND RESULT MODEL

4.1 Insight

4.1.1. What is the customer's debt status?

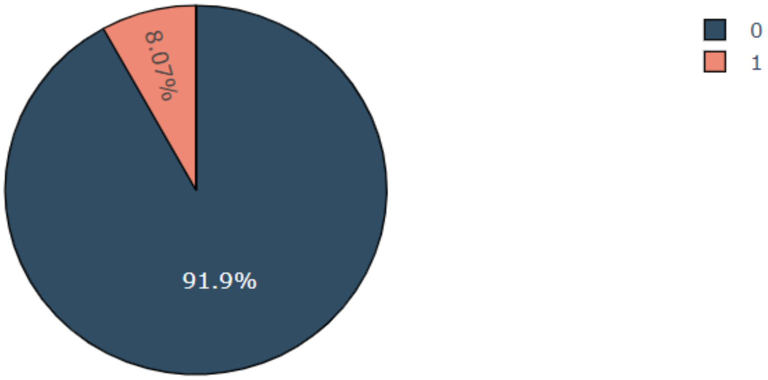


Figure 10: Customer's debt status

1 represents the number of customers having difficulty paying their loans, and only 24825 were unable to pay, equivalent to 8.07%. According to the State Bank of Vietnam in 2021, the on-balance sheet bad debt ratio reached 1.92% and the bad debt ratio including potential and restructured debt reached 7.31% so the figure of 8.07 is considered relatively high compared to the debt ratio of the market.

4.1.2. What are the types of loans that customers often use?

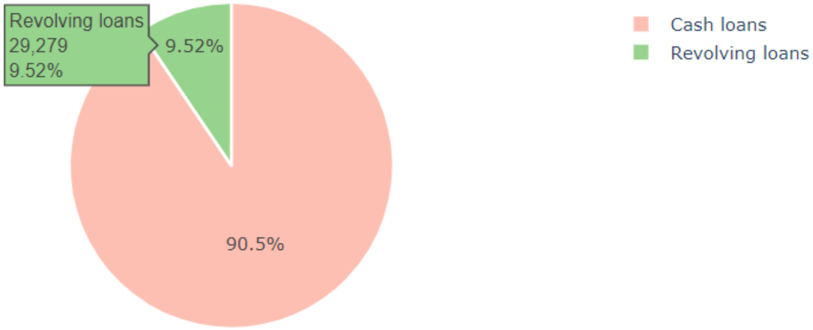


Figure 11: Type of loans

The number of cash loans is 278232, accounting for more than 90% while the revolving loan has a much smaller number of 29279. Customers often borrow cash (cash loans) more than revolving loans (revolving loans) because cash loans do not need collateral or guarantee. *(Revolving loans: A loan that can be withdrawn, returned, and redrawn in any* *manner and for any number of times until the agreement expires. Revolving loans include credit card debt and overdrafts. It's also known as an evergreen loan.)*

4.1.3. What gender are the customers applying for a loan?

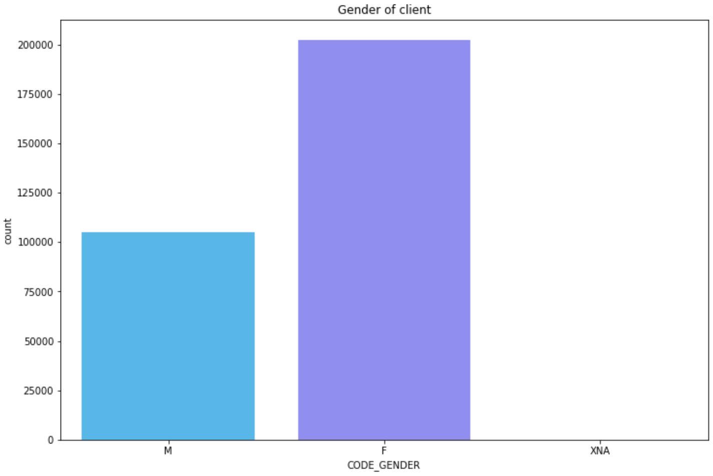


Figure 12: Gender of customers

From the chart above it can be seen that women have applied for the majority of loans, almost twice as many as men. In total, there were about 202,448 loan applications submitted by women, while only 105059 loan applications were submitted by men.

- **Gender of Customer in terms of loan is repaid or not**

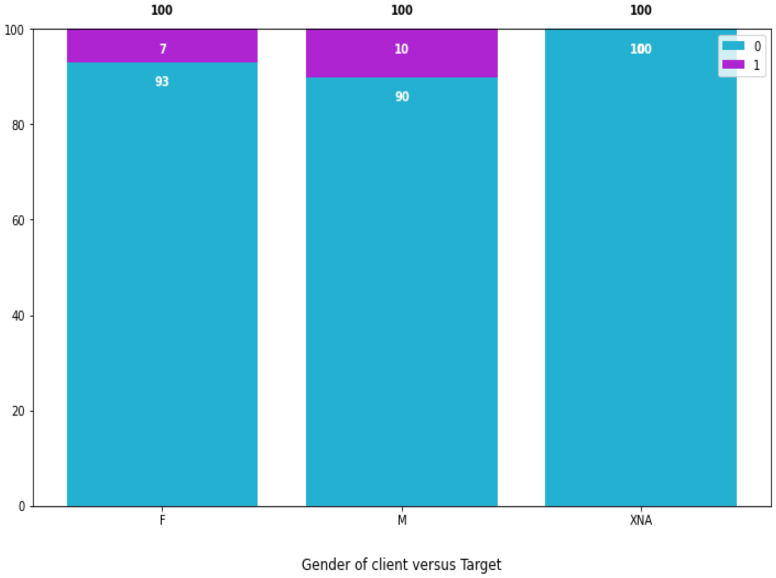


Figure 13: Gender of customer versus Target

Although the number of applications submitted by female customers is almost twice as high as that of male customers, the default rate of male customers is higher than that of female customers, specifically 3%.

4.1.4. What is the family status of the customer who applied for the loan?

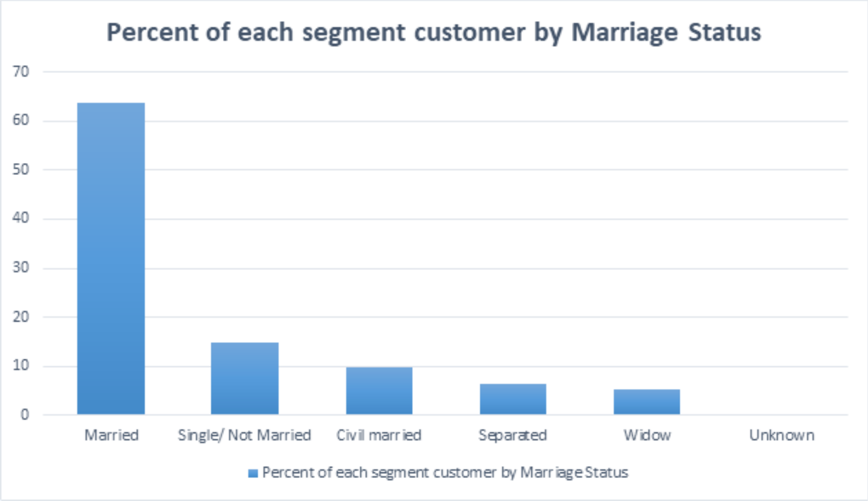


Figure 14: Family Status of the Customers

Borrowers are mainly married people (63.87%). Those who are unmarried/single are about a quarter less than those who are married (14.77%). Married people often have more family spending plans and plans, and often they will take a bank loan to pay for those expenses.

- **Family Status of Customers in terms of loan is repaid or not**

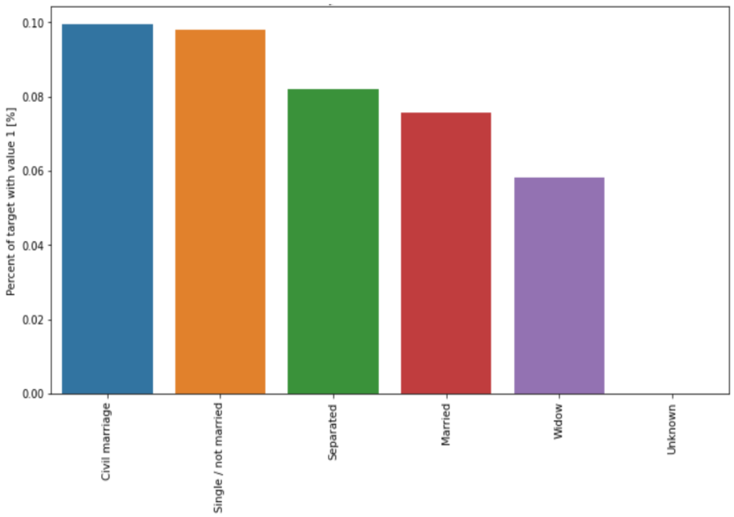


Figure 15: Family Status versus Target

Considering the default rate, we see that the default percentage is the highest in the group of Single/not married and civil married customers, and the lowest is in the widow customer group. With the high lending rate of these two groups of customers, the default rate is high.

4.1.5. How many family's members do the Customers have ?

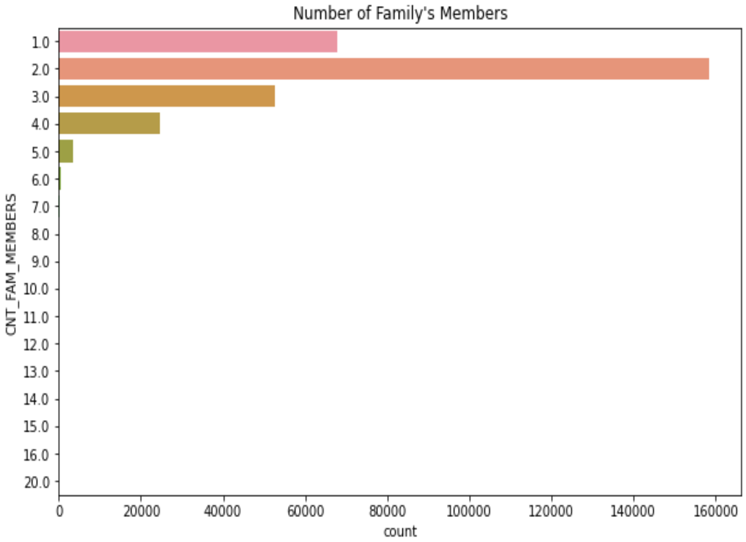


Figure 16: Number of family's members

The number of family members is mainly 2 people. With a relatively high number of married customers (accounting for more than 63% of total customers). There are many customers who do not have children.

4.1.6. How many Children do the Customers have ?

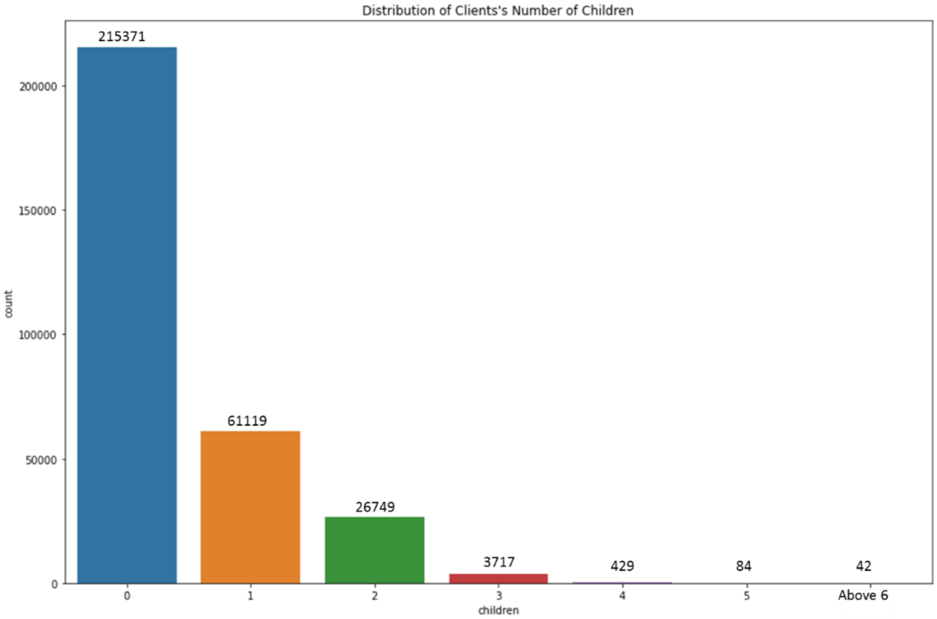


Figure 17: Customer's Number of Children

Due to the relatively rare number of customers with 6 or more children, we grouped them into a group called "Above 6". From the chart above, we can see that the majority of customers applying for loans are childless with the number up to 215371. Customers with 1 and 2 are 4 and 8 times more than customers without children, respectively. Customers with 5 or more children are much rarer. The reason for this is that these are newlyweds, young or have many plans. This financial situation is not stable, so the demand for loans is high.

- **Customer's Number of Children in terms of loan is repaid or not?**

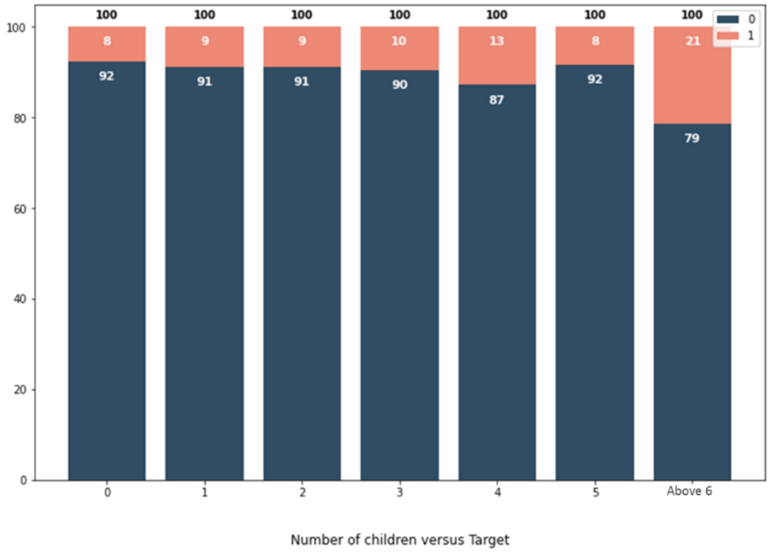


Figure 18: Number of children versus Target

For debt repayment, the group of customers with 0 to 3 children has an average default rate (from 8 to 10%) and the group of customers with 6 or more children, although the number of loan applications is small, the default rate of these customer groups are quite high (up to 21%). This shows that the economy from customers with many children is not stable, so there should be a more suitable plan or plans for this customer group.

- **Number of Family's members of Customers in terms of loan is repaid or not**

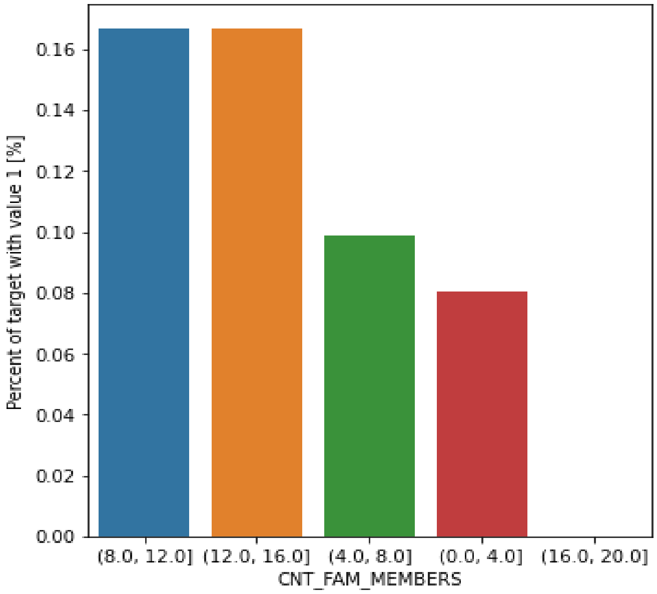


Figure 19: Number of Family's members versus Target

Most of the customer's family have 1-4 members (98.69% of customers in this customer group), which is reasonable because currently the number of family members is almost like only 1-4 people. Number of customers with the number of family members from 4 to 8 people ranked second (3987 customers - accounting for 1.3% of total customers) These customers often live with their parents, have a number of children from 2-3 people. The customer group of 8-20 people has the smallest number, in today's society very few families have such a large number of people. There are two cases for high values in the family member feature. It is an outlier value or the value is faulty.

Next we view it to be the default rate of the group customer. As we can see, although there is a group of customers with 1 - 4 family members, the default rate is quite low compared to the other groups (accounting for 8%). In families with a large number of members from 8 people or more, the default rate is very high (16.67%), maybe because of the large number of people, the economic conditions are not good, and they have to pay a lot of expenses. Therefore, banks must have a private policy for this group of customers.

4.1.7. Which type of house does the applicant apply for a loan?

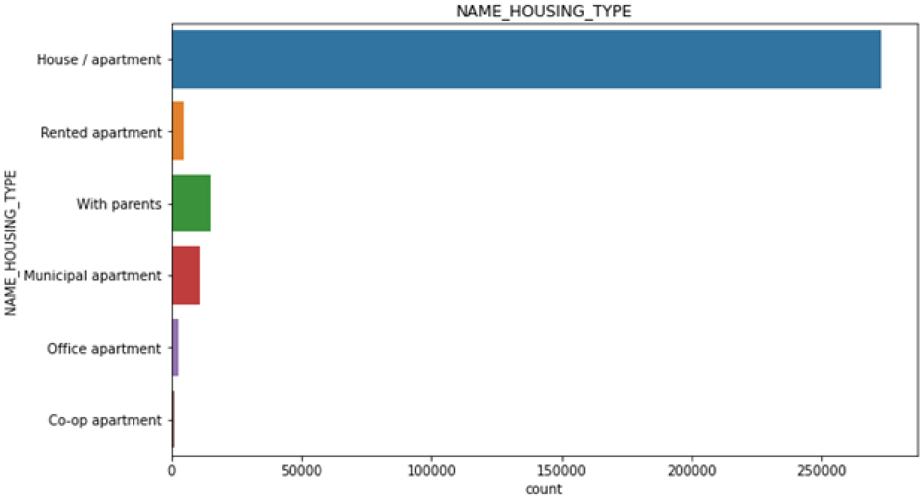


Figure 20: Type of House

Customers mainly live in Houses/apartments with 88.73%. Meanwhile, the proportion of customers living in other forms only accounts for 11.27%. From here, we realize that most of the customers applying for a loan have their own apartment/house. And this is reasonable because 63.87% of customers are married.

- **For which types of house higher customer applied for loan in terms of loan is repaid or not:**

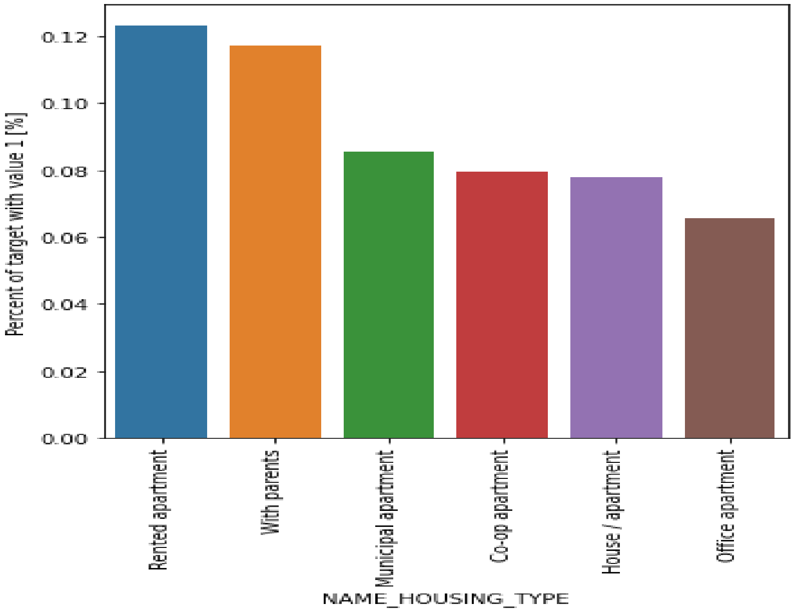


Figure 21: Type of house versus Target

The number of customers with Houses/apartments is quite large, but this group of customers default rate is only fourth in placements. In contrast, the total number of customers in the group of customers who live with their parents (with parents), and live in rented houses (rented apartment) is less than the group of customers who have their own houses or apartments, but the rate is high. The default is highest in groups, especially the group of customers living in rented houses. The reason is that the group of customers who have their own houses and apartments have more stable finances. The group of tenants has financial difficulties, because they have to rent a house, so in the month, the cost of housing is quite large, creating a high default rate in this group of customers.

4.1.8. What is the source of income of customers applying for a loan?

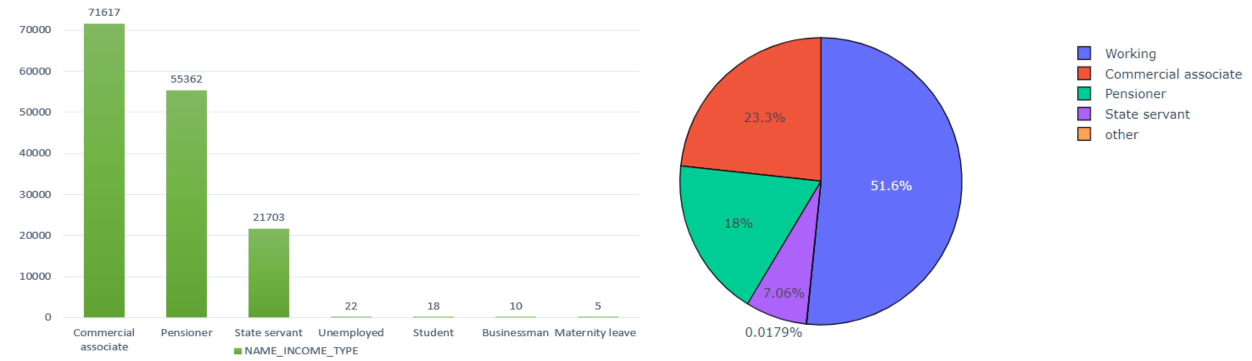


Figure 22: Income type of Customers

Customers' income comes from many different sources such as: Working, Commercial associate or Pensioner, ... of which 51.6% of customers have income from Working, followed by income from Commercial associate and Pensioner with 23.3% and 18%.

- **Income sources of Customer's in terms of loan is repaid or not**

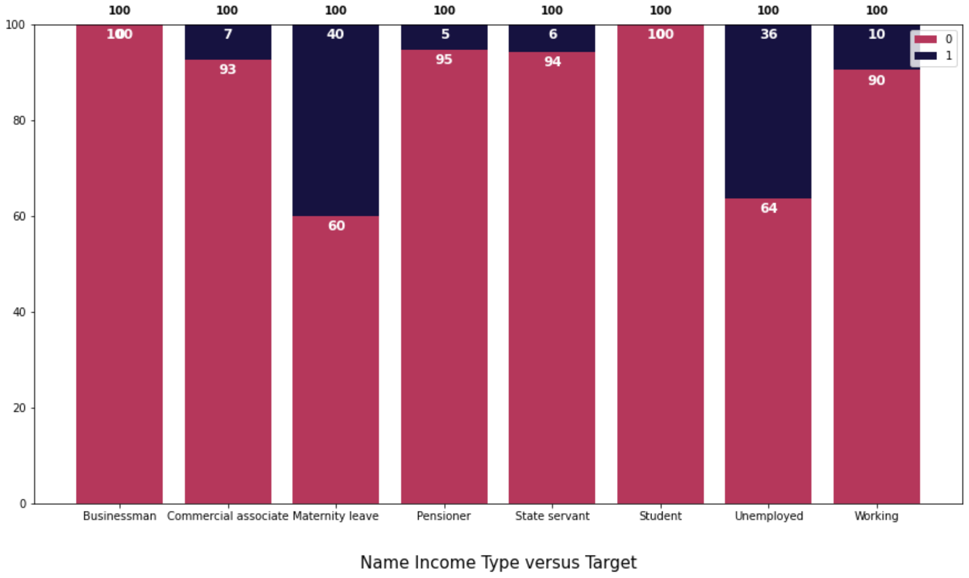


Figure 23: Income Sources versus Target

Maternity leave income type applicants have a 40% loan default rate, followed by the Unemployed with 36% and the Working group with 10 %. The remaining types of income all have default rates below the average, specifically from 5-7% or 0% like the Businessman group or the Student group. The reason why these two groups have a default rate of 0% could be that the Businessman group is a group of people who have stable careers or are successful, so they can afford to repay loans on time; As for the Student group, most of them are still dependent on their parents so the default rate is low (0%).

4.1.9. What is the education level of the client applying for the loan?

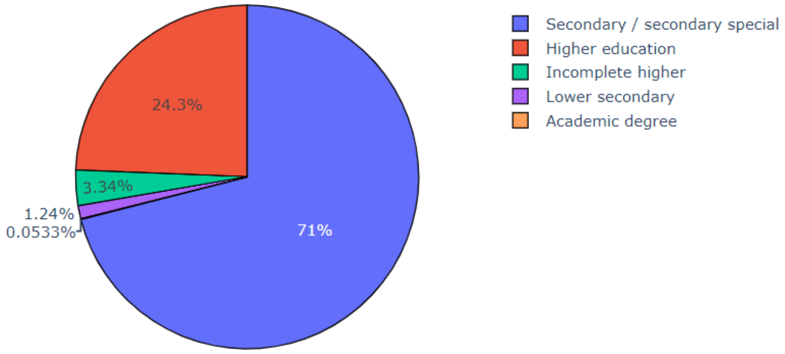


Figure 24: Customer's Education Level

The majority of customers have secondary / high school qualifications with 71% (more than 218 thousand applicants have Secondary / secondary special qualifications), followed by customers with education University with 24.3%. From the chart above, we see that only a very small number of customers have degrees.

- **Education of Customers in terms of loan is repaid or not**

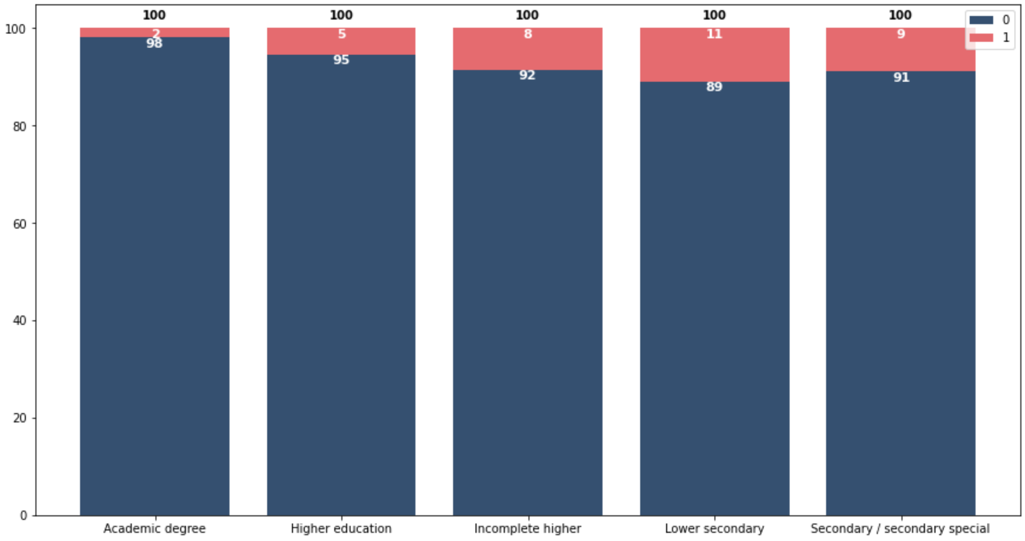


Figure 25: Education versus Target

Lower secondary group, although accounting for only a small part, has the largest loan default rate of 11%. In contrast, those in the Academic degree group had the lowest non-return rate at less than 2% (1.83% to be exact). The remaining groups have an average default rate of 5-9%.

4.1.10. What is the age group of the customer applying for the loan?

The bank does not have an age division for loan products except for those aged over 64. The age group should be divided into 2 ranges: under 65 years old and over 65 years old.

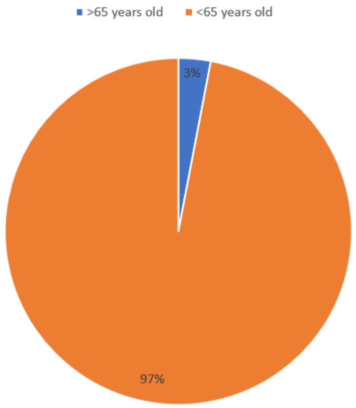


Figure 26: Customer's Age

It can be clearly seen on the data set that customers over 65 years old account for a very small number (3%) of the total number of customers. In addition, to see more insight, we have broken down the age of customers into smaller age groups.

- **Age group of Customers in terms of loan is repaid or not**

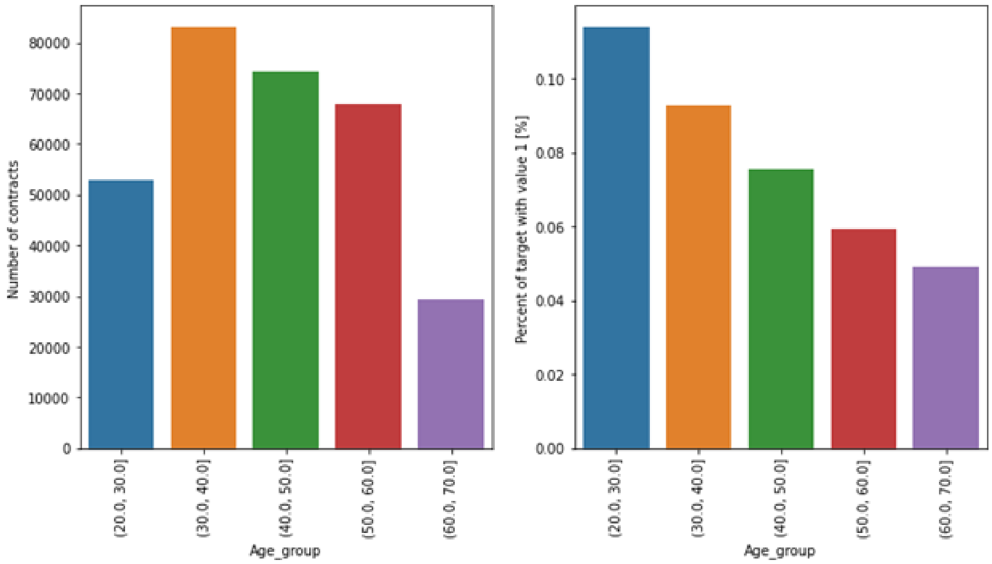


Figure 27: Age group versus Target

Looking at the graph, the trend is quite clear, younger people are more likely to default (higher default rate than older people). The default rate is approximately 10% for the two youngest age groups, and the default rate is 4.6% for the oldest age group. Through two charts, it is clear that the age group from 20-30 has a small number but the default rate is the highest. This is information that banks can use directly: because younger customers are less likely to borrow. The reason why the group of young people can't pay their debts is probably because they don't have enough experience or knowledge to manage or identify risks in finance.

4.1.11. What is the Occupation of a customer who applied for a loan ?

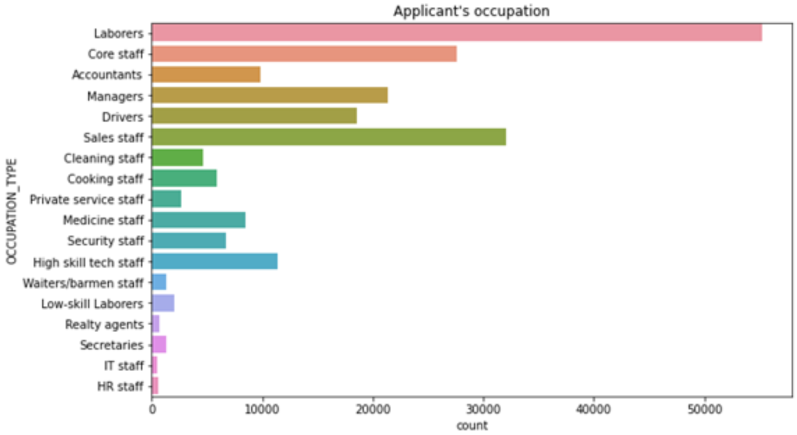
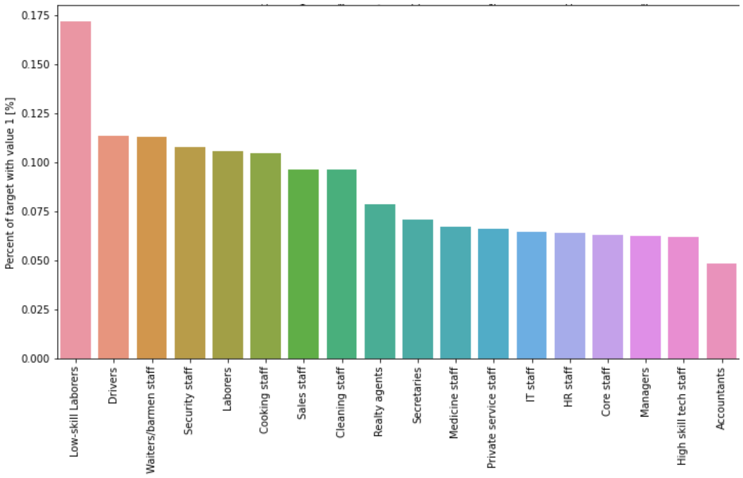


Figure 28: Customer's Occupation

Most of the loans are taken by Laborers, followed by Sales staff. IT staff take the lowest amount of loans. The reason is that IT staff often have a relatively high income compared to the common folk, so the need for debt is at least understandable, while workers with a middle or low income should have a higher need for debt.

- **Occupation of Customers in terms of loan is repaid or not**



From the chart, we can clearly see the division of jobs: group of low-skilled workers, group of manual workers and group of high-skilled workers. Hence, dividing occupations into those 3 groups.

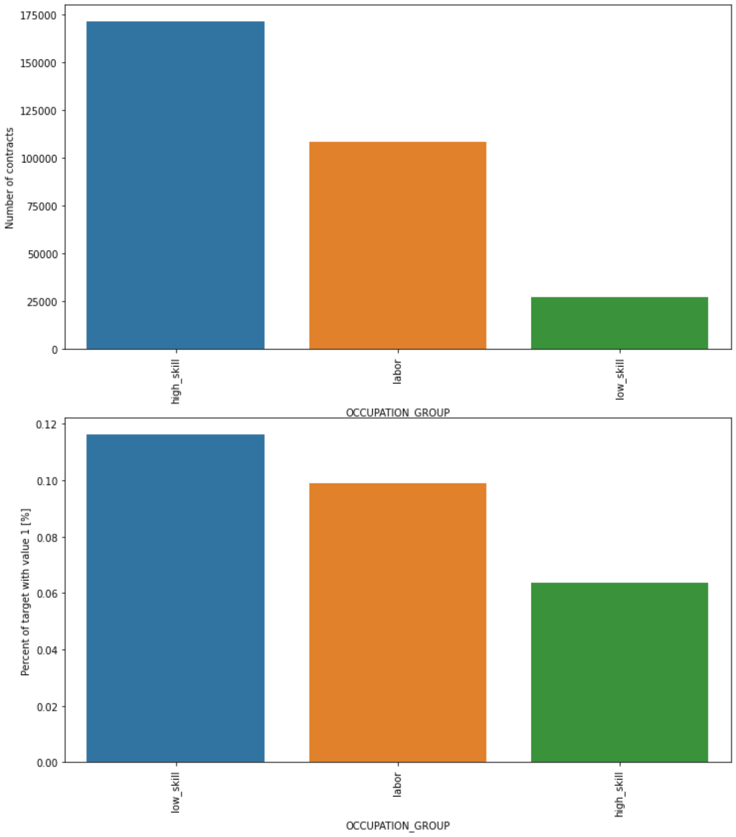


Figure 29: Occupation versus Target

Looking at the first chart, we see that borrowers belonging to the high-skill group account for the largest proportion (171614 customers - accounting for 54.81% of the total number of customers), but the default rate of this group is the lowest of the three groups. In contrast to the low-skill group, the number of borrowers is very low (only 8.9% of the total number of customers), but the default rate is the highest among customers (the default rate is 11.62%). The group of manual labor customers has the middle number of borrowers in the group and the default rate is 9.9%.

As the chart shows, there is a trend, with low-level people, the need to borrow money is high and the default rate is higher. The reason is that the second group will usually have a more stable income, and have large expenditures such as buying a house, buying a car, but because of the stable income, the default rate is low. Low-skill Laborers because of their low qualifications, they won't have jobs with high and stable salaries, because their finances are low, so when borrowing, the default rate is much higher. A group of highly skilled people (IT staff, ...) often have a fairly high income compared to the average level of society, they learn and invest in many large projects, but also because this customer group is stable in terms of income. finance, so the ability to default is low.

4.1.12. Years Employed of Customers in terms of loan is repaid or not

Customers work for at least half a month and 1000 years at the most. Data has people who worked up to 1000 years is ridiculous. And there are also some incidents where the number of people doing this work in 1000 years accounts for 18% (55374 customers) of the total number of borrowers. Therefore, this value needs to be processed.

Since the oldest client is 68 years old, the 50 year old mark should be taken to consider outliers. As expected, only 55374 of these people have over 1000 years of service or the data in this column is faulty.

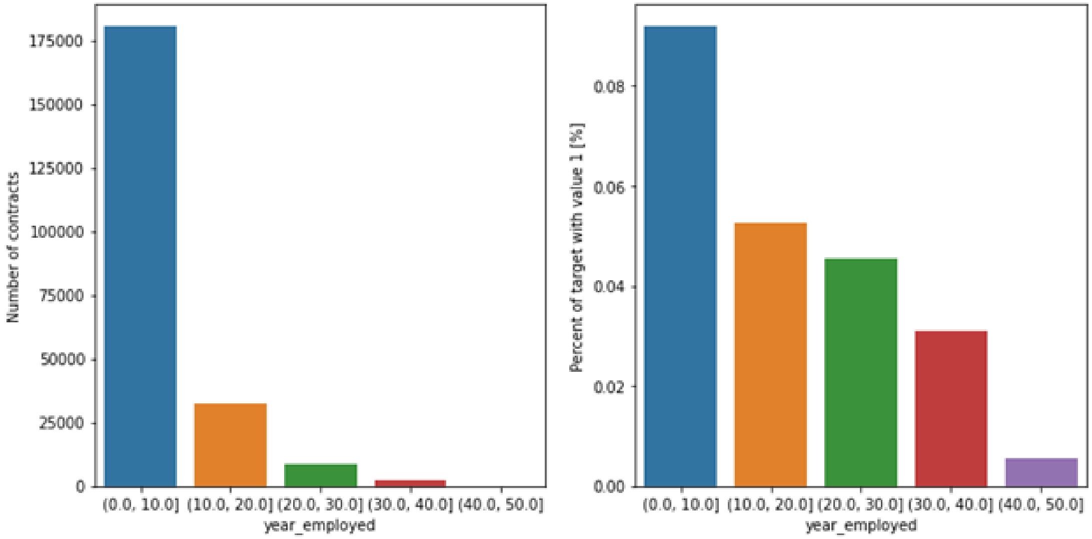


Figure 30: Customer's year-employed versus Target

From the chart, customers' commute time is trending. Borrowers are mainly working people for 10 years (accounting for 58.73%) of customers. The group of borrowers is at least the group of customers who have worked for 30-50 years. This is quite reasonable given the characteristics of the customer's age, the short working time means that the younger age will borrow more. Along with that, customers who have been working for about 10 years have a much higher default rate than the rest (the default rate of this group is 9.18%), that rate is almost double. the second group (working for 10-20 years), and many times higher than the last group of customers with a long working time (from 30-50 years).

4.2. Scoring Model

4.2.1. Method 1: Using WOE/IV and Logistic Regression Method.

Imbalanced data	**AUC index**	0.740			
		precision	**recall**	**f1-score**	**support**
	Class 0	0.92	1.00	0.96	50928
	Class 1	0.53	0.00	0.01	4429
	accuracy			0.92	55357
	weighted avg	0.89	0.92	0.88	55357

Table 10: Result of method 1 with Imbalanced data

Oversampling data	**AUC index**	0.741			
		precision	**recall**	**f1-score**	**support**
	Class 0	0.96	0.68	0.79	50928
	Class 1	0.16	0.68	0.25	4429
	accuracy			0.68	55357
	weighted avg	0.90	0.68	0.75	55357

Table 11: Result of method 1 with Oversampling data

Undersampling data	**AUC index**	0.741			
		precision	**recall**	**f1-score**	**support**
	Class 0	0.96	0.68	0.79	50928
	Class 1	0.15	0.68	0.25	4429
	accuracy			0.68	55357
	weighted avg	0.90	0.68	0.75	55357

Table 12: Result of method 1 with Undersampling data

SMOTE data	**AUC index**	0.735			
		precision	**recall**	**f1-score**	**support**
	Class 0	0.96	0.69	0.8	50928
	Class 1	0.16	0.66	0.25	4429
	accuracy			0.68	55357
	weighted avg	0.89	0.68	0.76	55357

Table 13: Result of method 1 with SMOTE data

In the traditional method, the research team used unbalanced and balanced data sets to train the Logistic Regression model. As mentioned above, because the dataset has a lot of missing data, the accuracy when training the model may not be high. Therefore, when the project is training the model, the research team expects the AUC index to reach 0.75.

Looking at 4 tables of results from 4 different ways of data processing to be included in the Logistic Regression model, it is clear that the AUC index of the different data processing methods is approximately 0.74 - not yet met the expected threshold. of the research group. Regarding the indexes of precision, recall, f1-score with weight avg, it is clear that the indexes when the data are not processed, the skewed label situation has a much higher value than the alternatives when dealing with skewed labels (Oversampling, Undersampling, SMOTE). In the unbalanced sample processing technique in this project, the team uses Sampling technique so that the samples in class 0 and class 1 have a 1:1 ratio and all later models, the team also handles the loss of samples. balance the data in such proportions.

However, when looking at the classifier efficiency 0, 1 shows a clear result, the unprocessed data result is imbalanced data showing that the recall in class 0 has a value of 1 but the recall value in class 1. again equal to 0. Similarly, with the value f1-score, the f1-score index at class 0 is 0.96 but for class 1, that value is only 0.01. Therefore, this model is bad and unusable.

In terms of training results when handling imbalanced data, tables 2, 3, and 4 show that there is not much difference between the results of these different treatment options. In these three options, the recall index in class 1 has been improved (approximately 0.68) but the recall in class 0 is lower (approximately 0.68), so the weight average of the recall is also not high (about 0.68). Regarding the f1-score, the values of the indexes in these three options are approximately the same and equal to about 0.75, the results are quite good.

In summary, in method 1 - the traditional method when classifying features by WOE - IV method and training with Logistic Regression model, the model results are not really as good as expected.

4.2.2. Method 2: Machine Learning

In the second method, the research team uses three machine learning models, Logistic Regression, Random Forest and LightGBM. Simultaneously train with four different datasets: unlabeled skewed data, and three skewed label processed datasets with three different methods: Oversampling, Undersampling and SMOTE.

a. Logistic Regression.

Dataset	Imbalanced data	Oversampling data	Undersampling data	SMOTE data
AUC score	0.583	0.576	0.58	0.576576

Table 14: Result of model Logistic Regression.

In method 2, the research team does not use WOE - IV to classify variables or filter variables anymore, but uses Scaling Standard technique to scale data points.

The research team's results when using all 4 different data sets (with skewed labels and not with skewed labels) are not good. The AUC index only reached the threshold of 0.58, excluding the remaining metrics (precision, recall, f1-score), in the AUC metrics, the obtained results were too low compared to the expected threshold of the research group (0.75). Therefore, the Logistic Regression model with all 4 different data sets cannot be used.

b. Random Forest.

Imbalanced data	**AUC index**	**0.769**			
		precision	**recall**	**f1-score**	**support**
	Class 0	0.93	1.00	0.96	66285
	Class 1	0.00	0.00	0.00	4966
	accuracy			0.93	55357
	weighted avg	0.87	0.93	0.9	55357

Table 15: Result of method 1 with Imbalanced data Random Forest model

Oversampling data	**AUC index**	**0.785**			
		precision	**recall**	**f1-score**	**support**
	Class 0	0.93	1.00	0.96	66285
	Class 1	0.46	0.04	0.08	4966
	accuracy			0.68	71251
	weighted avg	0.90	0.93	0.9	71251

Table 16: Result of method 1 with Oversampling data Random Forest model

Undersampling data	**AUC index**	**0.787**			
		precision	**recall**	**f1-score**	**support**
	Class 0	0.97	0.69	0.81	66285
	Class 1	0.15	0.73	0.25	4966
	accuracy			0.69	71251
	weighted avg	0.91	0.69	0.77	71251

Table 17: Result of method 1 with Undersampling data Random Forest model

SMOTE data	**AUC index**	**0.747**			
		precision	**recall**	**f1-score**	**support**
	Class 0	0.93	1.00	0.96	66285
	Class 1	0.38	0.01	0.02	4966
	accuracy			0.93	71251
	weighted avg	0.89	0.93	0.9	71251

Table 18: Result of method 1 with SMOTE data Random Forest model

In the section using the Random Forest model, the research team continues to use four datasets to train with the Random Forest model. The results are much better than that of the Logistic Regression model, the AUC indexes are all improved, and the research team's expected results are achieved (AUC-score > 0.75). Especially with two data sets that have been processed with imbalanced data through Oversampling and Undersampling techniques with AUC indexes of 0.785 and 0.789 (~0.79).

Next, consider the metrics precision, recall and f1-score in each class (class 0 and class 1). For datasets treated with imbalanced data by Oversampling technique, although the f1-score at class 0 is very good (0.96) or the precision and recall indexes are both high (0.93, 1.00), which means that the model has a very good classification. with customers in group 0; but when observing the metric indexes in class 1, we see the opposite result compared to class 0, with the f1-score of 0.08, similar to the recall or precision index of 0.04 and 0.46, respectively, this result is quite bad. when classifying customers in class 1. In this project, the purpose of building a model is to find customers who are insolvent, but the Random Forest model with this dataset has bad metrics in class 1. Therefore, Random Forest with these Oversampling treatments is not a good way to classify customers.

For datasets treated with imbalanced data by Undersampling technique, the precision, recall and f1-score indexes of class 0 are not good with 0.97, 0.69 and 0.81, respectively, but this is also a good result for analysis. class 0. With the classification level of class 1 and the above metrics are somewhat improved with the dataset processed with imbalanced data with Oversampling technique. Especially, the recall index is 0.73, which means that the recognition of defaulted customers has been much better. In the second method, the dataset is processed by Undersampling technique and building the Random Forest model is a pretty good method to classify customers who can repay their debts to credit institutions or not.

c. LightGBM

- **Hyperparameter:**

boosting_type	goss
learning_rate	0.005134
max_depth	10
min_split_gain	0.024766

Table 19: Hyperparameter of LightGBM model

- **Result:**

Imbalanced data	**AUC index**	0.82			
		precision	**recall**	**f1-score**	**support**
	Class 0	0.93	1.00	0.96	66285
	Class 1	0.56	0.05	0.10	4966
	accuracy			0.93	55357
	weighted avg	0.9191	0.93	0.90	55357

Table 20: Result of method 2 with Imbalanced data LightGBM model

Undersampling data	**AUC index**	**0.81**			
		precision	**recall**	**f1-score**	**support**
	Class 0	0.97	0.71	0.82	66285
	Class 1	0.16	0.75	0.27	4966
	accuracy			0.71	71251
	weighted avg	0.92	0.71	0.78	71251

Table 21: Result of method 2 with UnderUndersampling data LightGBMGBM model

The final model in the second method is the dataset trained with the LightGBM tuning hyperparameter model. In this method, the model is trained with two datasets: the unbalanced set and the processed imbalanced data set, specifically dealing with Undersampling technique.

For the LightGBM model, it is clear that the AUC metric has been much improved, for both datasets it is larger than 0.8 - a good result for the models above with the dataset, specifically with the imbalanced dataset. The AUC score is 0.82 and the data set processed with Undersampling technique has an AUC score of 0.81. The above two AUC results are good and do not differ much.

Considering the metrics precision, recall and f1-score, similar to the results in the above models with the imbalance data set, in class 0 the indexes are all high and very good (recall index is 1.00 and f1-score is 0.96), however the metrics are all very bad in class 1 (recall index is 0.05 and f1-score is 0.1). It shows that with this dataset, the LightGBM model performs well or classifies defaulted customers well.

With the dataset being processed with imbalanced data by Undersampling technique, the metrics are better, especially the recall score in two classes 0 and 1 is 0.71 and 0.75 respectively. This shows that the model recognition of customers in default or not is quite good. The metrics when using the LightGBM model are better than other models, and the dataset when processing Undersampling data works well on the LightGBM tuning hyperparameter model.

- Limit:

The study has some resource limitations to build and train the LightGBM model when the data is too large. As observed above, the research team does not show the results of the dataset processed by Oversampling and SMOTE techniques because there are some limitations due to the large size of data.

PART 5: CONCLUSION

In this report, the research team has visualized outstanding insights of the data set such as the customer's debt status or the types of loans that customers often apply for, ... Through the Insight section, we can understand that the needs and status of customers in taking out loans is that nowadays, many people have difficulty in getting loans due to insufficient or non-existent credit history. Not only that, the research team also have a more general view of the debt repayment capacity of different customer groups (For example, the female customer group has a high debt ratio that is nearly twice as high as that of the male customer group, but the default rate of male customer group is higher than female customer group,...).

In this paper, the process of forecasting bad debt of credit institutions has been implemented. Thus, a robust model was built for the intended purpose. The data has been extracted from the bank with the objective of bad debt forecasting, the sample data is used to train the dataset (with 7 interconnected tables of information and the main data table is application_train) based on the stated criteria to apply the built model. The project has applied the technique of processing unbalanced data Undersampling, Oversampling and SMOTE in the ratio 1:1 with class 0 and class 1; and train those data sets in two different methods: the traditional method of classifying the data using WOE - IV and train the dataset with Logistic Regression and the second method is to use machine learning models to build an efficient model: Logistic Regression, Random Forest and LightGBM. With previous studies, most of the authors used logit regression - a traditional method to make predictions, but this paper shows more clearly that the Random Forest algorithm and the LightGBM algorithm are capable of making better predictions, especially with the LightGBM model tuning hyperparameter. The prediction performance is much better even though the training time of the model is longer than that of the Logistic Regression model. In addition, depending on the actual bad debt data; predictions seem to be in line with reality. These are reasonable outcomes for specific predictions and for the use of real-life data mining techniques; this gives a good indication that using data mining techniques can help banks make different decisions when using LightGBM for data analysis. Therefore, the best forecasting model for this dataset is LightGBM tuning.

LINK GITHUB

https://github.com/btmluogg01/Home-Credit-Default-Risk

REFERENCES

1. Abid, L., Masmoudi, A., & Zouari-Ghorbel, S. (2018), “The Consumer Loan’s Payment Default Predictive Model: An Application of the Logistic Regression and the Discriminant Analysis in a Tunisian Commercial Bank”, *Journal of the Knowledge Economy,* 9(3), 948-962.

2. Dufhues, T., Buchenrieder, G., Quoc, H. D., & Munkung, N. (2011), “Social capital and loan repayment performance in Southeast Asia", *The Journal of Socio-Economics,* 40(5), 679-691.

3. Dunn, L., & Kim, T. (1990), *Empirical Investigation of Credit Card Default [Working Paper],* Ohio State University, Department of Economic.

4. Jacobson, T., & Roszbach, K. (2003), “Bank lending policy, credit scoring and value-at-risk”, *Journal of Banking & Finance,* 27(4), 615-633.

5. Kocenda, E., & Vojek, M. 92011), *Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data,* Emerging Markets Finance and Trade, 47(6), 80-98.

6. Ojiako, I. A., & Ogbukwa, B. C. (2012), *Economic analysis of loan repayment capacity of smallholder cooperative farmers in Yewa North Local Government Area of Ogun State, Nigeria.*