**TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN**

**KHOA TOÁN KINH TẾ**

**-----\*\*-----**



# BÀI TẬP LỚN

MÔN RISK MANAGEMENT

***Đề tài: Predicting Bank Loan Defaults.***

**Sinh viên: Bùi Thị Mai Lương – 11193191**

**Lớp : DSEB**

Table of content

Part 1: Introduction

1. Bank Loans
2. Bad Loans
3. Predict Bad Loans.

Part 2: Method

1. Correlation.
2. WOE - IV.
3. Imbalance data.
4. Scaling.
5. Descriptive statistics.
6. Exploratory Data Analysis (EDA).
7. Model.

7.1. Logistic Regression.

7.2. Random Forest.

7.3. K-Nearest Neighbor (kNN)

7.4. LightGBM

1. Metrics for evaluating model.

8.1. Precision and Recall

8.2. F1 - Score.

8.3. ROC - AUC Curve.

Part 3: Data

1. Data Description.

1.1. Dataset.

1.2. Data Analysis.

1. Data processing
2. Model and Results

Part 4: Conclusion.

Part 1: Introduction

1. Bank Loans.

Bank loans are one of the most popular forms of financial assistance available to small and medium-sized enterprises (SMEs). This is the fastest and simplest way to secure the necessary funding and is usually provided within a fixed period of time.

Bank loans will likely require payment of both capital and principal or can be structured to meet your business needs. For businesses looking to buy business premises, they can choose a commercial mortgage with more flexible terms.

Bank loans can be short-term or long-term, depending on the purpose of the loan.

1. Bad Loans.

Bad loan is the total amount borrowed for which the customer has not made payments within a given time period.

Although each debt is different in terms of loan terms, loan types and payment terms. However, in general, this period usually falls between 90 and 180 days. (According to Investopedia).

According to Makri, Tsagkanos, and Bellas (2014), ineffective debt collection is the cause of the increase in bad debt as well as the difficulties encountered in dealing with bad debts. In addition, the bad debts outstanding from previous years to the present have not been completely resolved, which will increase bad debts in the current year.

Credit risk is by far the most significant risk that banks face, and their economic success is dependent on accurately measuring and effectively managing this risk to some extent (Gieseche, 2004). It is the risk of financial loss if a borrower or counterparty fails to meet their agreed-upon obligations, and this failure has a negative impact on the bank's financial performance.

1. Predict Bad Loans.

Monitoring liquidity and credit risk is one of the main issues in banking operations. Liquidity risk due to unmarketable investments, where the underlying asset cannot be bought or sold quickly enough to hedge or minimize loss. Credit risk arising from holding bonds, or any delay in making loan payments as they come due, is one of the main concerns of investors. As a result, they want to do a credit analysis before approving any loan, while also assessing a borrower's likelihood of default.

Credit scoring has numerous advantages for all sectors of the economy. Credit ratings enable lenders to reach out to previously untapped markets. Furthermore, judgments are made more promptly and objectively now, with the majority of loan applicants receiving responses in minutes rather than days. Finally, lenders have been able to lower prices for important services such as mortgages and consumer loans by better predicting risk using credit ratings. as well as credit cards. Despite the fact that it has been extended to previously neglected regions, the real risk ratio is still lower when credit scores are used, because lenders can more actively manage risk and keep risk at an appropriate level.

Credit scoring helps the national economy control consumer activities during periods of cyclical unemployment and minimizes business cycle volatility. Credit scoring connects clients to the secondary capital market and increases the volume of funds available to issue credit or invest in economic growth by allowing lending and credit products to be divided into risk-based groupings and selling securitized derivatives.

Part 2: Method

1. Correlation.

Correlation are utilized in a wide scope of uses, frequently related to likelihood thickness capacities, data on the distribution's arbitrary signal. Application analysis of electrophysiological estimation and which give of amplitudes of reach from the signs to stream weariness examination.

The correlation coefficient between two vectors of equal lengths with corresponding elements is a widely used statistical measure. Though a correlation value between two such vectors with non-zero variances can always be computed and is in the range [-1, 1], it is difficult to assign a meaning to a specific value of correlation, say, 0.8 or -0.3. Correlation values above 0.8 are deemed to indicate a strong positive linear relationship between the variables. Values between 0 and 0.3 indicate a weak relationship or none. Interpretations such as these come handy, but they are subjective and tell very little about specific correlation values.

1. WOE – IV.

The WOE calculates how effective each property, or collection of attributes, is at distinguishing between good and bad accounts. It's a metric for determining the proportion of good and bad in each feature (i.e., the odds of a person with that attribute being good or bad). The WOE depends on the log of odds calculation:

Which measures odds of being good.

A more user-friendly way to calculate WOE:

Multiplication by 100 is done to make the numbers easier to work with. Negative numbers imply that the particular attribute is isolating a higher proportion of bads than goods. Information Value, or total strength of the characteristic, comes from information theory, and is measured using the formula:

Based on this methodology, one rule of thumb regarding IV is:

* Less than 0.02: unless.
* 0.02 to 0.1: weak.
* to 0.3: medium.
* 0.3 to 0.5: strong.
* ≥ 0.5: very strong variable. However, this case needs to be reconsidered to avoid the case where the variable has a direct relationship to the definition of the good/bad profile.

1. Imbalance data.

In many constant applications enormous measure of information is created with slanted dissemination. An informational collection said to be exceptionally slanted assuming example from one class is in larger number than other. In an imbalance dataset, the class with the most cases is known as the significant class, while the one with the fewest is known as the minor class.

There are many methods to deal with data imbalance, but in this project two main methods were used, under-sampling and over-sampling.

* **Under-sampling**

**Under-sampling balances the dataset by reducing the size of the abundant class. This method is used when quantity of data is sufficient. By keeping all samples in the rare class and randomly selecting an equal number of samples in the abundant class, a balanced new dataset can be retrieved for further modelling.**

* **Over-sampling**

**On the contrary, oversampling is used when the quantity of data is insufficient. It tries to balance dataset by increasing the size of rare samples. Rather than getting rid of abundant samples, new rare samples are generated by using e.g. repetition, bootstrapping or SMOTE (Synthetic Minority Over-Sampling Technique).**

**Note that there is no absolute advantage of one resampling method over another. Application of these two methods depends on the use case it applies to and the dataset itself. A combination of over- and under-sampling is often successful as well.**

1. Scaling.

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.

For this project, the dataset was used two scaling methods: Standard Scaler and MinMaxScale.

1. Descriptive statistics

Descriptive statistics are used to describe the basic characteristics of data collected from experimental studies in different ways.

Descriptive statistics of the observed variables, the research team found that the answers in the questionnaire were quite diverse.

1. Exploratory Data Analysis (EDA)

EDA of observed and target variables helps the research team have an overview of how the observed variable affects the target variable.

How do the observed variables tend, what values of the observed variables are unreasonable or out of the ordinary.

1. Model
   1. Logistic Regression.

Logit model is a regression model with auxiliary variable belongs (Y) is a binary variable, taking only two values are 0 and 1; The independent variables can be binary integral, discrete or continuous variables. In tissue credit rating model, dependent variable Y received value 0 when the customer fails to pay the debt and Once the customer repays the debt (Lee et al.,2000). The independent variables represent the information qualitative and quantitative information of customers such as income, age, gender, education level...

After regression Logit model, obtained is the estimated value of Y. Then, the customer's debt repayment probabilities

is calculated by the following formula:

P value is obtained in the range (0,1) get compared with the threshold set by the bank for customer ratings. However, in the article this study, to make it easier to compare

Compare the effectiveness of the models, the threshold value to categorize selected customers as 0.5. This means that if the P value < 0.5, the customer

goods will be forecast to default, and vice versa if P ≥ 0.5, customers will be expected to repay the debt (good loan).

* 1. Random Forest.

Decision trees can be used for different machine learning applications. But when the data noise can cause the tree to grow in a completely different way. This is due to the fact that decision trees have very low bias and high variance. Random Forests overcome this problem by training multiple decision trees on different subspaces of the object space at a slightly increased cost. This means that no tree in the forest sees the entire training data. The data is recursively separated into partitions. At a particular node, separation is performed by questioning an attribute. The choice for separation criteria is based on some impurity measure such as Shannon Entropy or Gini impurity. Gini impurity was used as a function to measure the separation quality in each node. The gini impurity at node N is given by

where is the proportion of the population labeled as i. Another function that can be used to evaluate separation quality is Shannon Entropy. It measures disorder in Content information. In Decision trees, Shannon entropy is used to measure unpredictability in the information contained in a particular node of the tree (In this context, it measures how mixed the population in a node is). The entropy in node N can be calculated as follows:

(13)

Where d is the number of classes to be considered and P() is the proportion of the population labeled as i. Entropy is highest when all layers are contained in equal proportions in the node. It is lowest when only one class is present in a node (when node is pure). The obvious heuristic approach for choosing the best split decision at a node is to reduce impurities as much as possible. In order, the best separation is characterized by the highest information acquisition or the highest impurity reduction. The information obtained as a result of a split can be calculated as follows:

(14)

where I(N) is the impurity measure (Gini or Shannon Entropy) of the node N, is the proportion of the population in node N that moves to the left subsection of N after splitting and so on, is the proportion of the population in node N that moves to the right child after splitting . and are the left and right child nodes of N, respectively.

* 1. K-Nearest Neighbor (kNN)

KNN is a machine learning method to classify objects based on the closest distance between the object to be classified and all the objects in the training data. Class of a new data point (or customer classifier) ​​is directly inferred from the nearest K data points in the training data. This class can be decided according to the class with the most number of points (within the nearest K points).

In fact, the data to be classified has many attributes in which each attribute corresponds to a spatial dimension, so when calculating the closest distance, it is necessary to calculate the vector distance in multidimensional space with the Euclidean distance formula:

The prediction of bad loan is computed using kNN as follows:

a) Determine the number of nearest neighbors, k.

b) Compute the distance between the training samples and the query record.

c) Sort all training records according to the distance values. d) Use a majority vote for the class labels of k nearest neighbors, and assign it as a prediction value of the query record.

* 1. LightGBM

LightGBM uses "histogram-based algorithms" as an alternative to "pre-sort-based algorithms" commonly used in other boosting tools to find split points during tree construction. This improvement helps LightGBM increase training speed and reduce memory usage. LightGBM both use histogram-based algorithms, the optimal point of lightgbm is in 2 algorithms: GOSS (Gradient Based One Side Sampling) and EFB (Exclusive Feature Bundling) significantly speeds up the computation process.

LightGBM grows trees based on leaf-wise, while most other boosting tools are level (depth)-wise. Leaf-wise selects nodes to grow the tree based on the optimization of the whole tree, while level-wise optimizes on the branch under consideration, so for a small number of nodes, trees built from leaf-wise are often out-perform level-wise.

According to the levelwise growth strategy, the leaves on the same layer are simultaneously split. It is favorable to optimize with multiple threads, and control model complexity. However, leaves on same layer are indiscriminately treated, whereas they have different information gain. Information gain indicates the expected reduction in entropy caused by splitting the nodes based on attributes.

where En(B) is the information entropy of the collection B, pd is the ratio of B pertaining to category d, D is the number of categories, v is the value of attribute V, and Bv is the subset of B for which attribute has value v.

Actually, many leaves with low information gain are unnecessary to be searched and split, which increases a great deal of extra memory consumption and causes this method to be ineffective. In contrast, leaf-wise growth strategy is more efficient because it only split the leaf that has the largest information gain on the same layer. Furthermore, considering this strategy may cause trees with high depth, resulting in overfitting, a maximum depth limitation is adopted during the growth of trees.

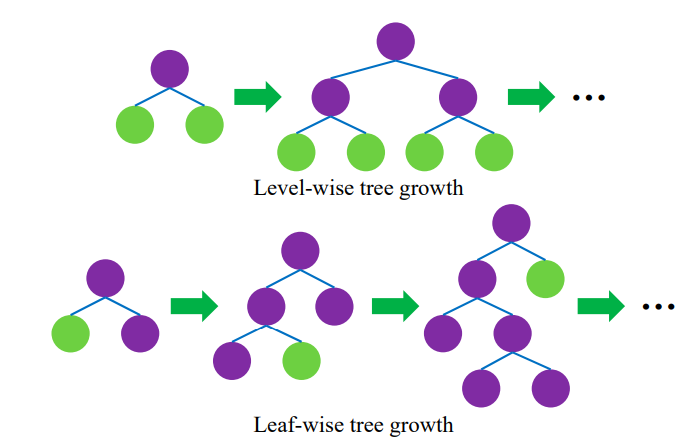


Figure. Level – wise and leaf – wise tree construction

1. Metrics for evaluating model.
   1. Precision and Recall

For classification problems where the data sets of the classes are very different, there is an effective operation that is often used, which is Precision-Recall. First consider the binary classification problem, consider one of the two classes to be positive, the other to be negative.



Figure: Calculate Recall and Precision

* 1. F1 – Score.

Precision and Recall are the two building blocks of the F1 score. The goal of the F1 score is to **combine the precision and recall metrics into a single metric.** It is primarily used to compare the performance of two classifiers. At the same time, the F1 score has been designed to **work well on imbalanced data**.

The F1-score of a classification model is calculated as follows:

* A model will obtain a high F1 score if both Precision and Recall are high.
* A model will obtain a low F1 score if both Precision and Recall are low.
* A model will obtain a medium F1 score if one of Precision and Recall is low and the other is high.
  1. ROC – AUC Curve.

AUC - ROC is a method of calculating the performance of a classifier model against different classification thresholds.

ROC curve, (Receiver Operating Characteristics Curve), is a metric used to measure the performance of a classifier model. The ROC curve depicts the rate of true positives with respect to the rate of false positives, therefore highlighting the sensitivity of the classifier model. The ROC is also known as a relative operating characteristic curve, as it is a comparison of two operating characteristics, the True Positive Rate and the False Positive Rate, as the criterion changes. An ideal classifier will have a ROC where the graph would hit a true positive rate of 100% with zero false positives. ROC curve can be used to select a threshold for a classifier, which maximizes the true positives and in turn minimizes the false positives. ROC Curves help determine the exact trade-off between the true positive rate and false-positive rate for a model using different measures of probability thresholds. ROC curves are more appropriate to be used when the observations present are balanced between each class. This method was first used in signal detection but is now also being used in many other areas such as medicine, radiology, natural hazards other than machine learning. A discrete classifier returns only the predicted class and gives a single point on the ROC space. But for probabilistic classifiers, which give a probability or score that reflects the degree to which an instance belongs to one class rather than another, we can create a curve by changing the threshold for the score.

Area Under Curve or AUC is one of the most widely used metrics for model evaluation. It is generally used for binary classification problems. AUC measures the entire two-dimensional area present underneath the entire ROC curve. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than that of a randomly chosen negative example. The Area Under the Curve provides the ability for a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, it is assumed that the better the performance of the model at distinguishing between the positive and negative classes.

One of the good approaches to measure the model's accuracy is to look at the area under the curve. An excellent model has an AUC close to 1, indicating that it has a high level of separability. An AUC approaching 0 indicates that a model is bad, indicating that it has the lowest measure of separability. In fact, it implies that it is reversing the result, anticipating 0s as 1s and 1s as 0s. An AUC of 0.5 indicates that the model has no class separation capacity at all.

Part 3: Data

1. Data Description.
   1. Dataset.

* Dataset: 30 features (including feature target) and 887379 rows.
* Dataset have no null datapoint.
* Type: 10 categorical features, 20 numeric features.
* % Null: Dataset does not null feature (Feature’s Percentage = 0%)
* Source: Loan Data for Dummy Bank (Kaggle).
* Dataset divided two main information: Customer's personal information, Customer's debt information. Variable types are shown in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Information** | **Feature** | **Type Feature** | **Meaning** | **The Number of Unique values** | **Values** |
| Customer’s personal information | ID | Numeric | Customer's ID |  |  |
| emp\_length\_int | Numeric | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year | 11 | 0.5-10 |
| home\_ownership | Categorical | The home ownership status provided by the borrower during registration. | 6 | Mortgage Rent Own Other None Any |
| income\_category | Categorical | Customer's income category | 3 | Low Medium High |
| annual\_inc | Numeric | The self-reported annual income provided by the borrower during registration |  |  |
| grade | Categorical | Loan Customer (LC) assigned loan grade | 7 | A, B, C, D, E, F, G |
| region | Categorical | region of Loan being executed |  |  |
| Customer's debt information | Year | Numeric |  | 9 | 2007-2015 |
| issue\_d | Day |  |  |  |
| final\_d | Numeric |  |  |  |
| loan\_amount | Numeric | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |  |  |
| term | Categorical | The number of payments on the loan. | 2 | 36 months 60 months |
| application\_type | Categorical | Indicates whether the loan is an individual application or a joint application with two co-borrowers | 2 | Individual Joint |
| purpose | Categorical | Customer's purpose loan | 14 | Debt consolidation Credit card Home improvement Major Purchase Smaill business Car Medical Moving Vacation House Wedding Renewable enery Educational. |
| interest\_payments | Categorical |  | 2 | Low, High |
| interest\_rate | Numeric | Interest Rate on the loan |  |  |
| dti | Numeric | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, - - - excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income. |  |  |
| total\_pymnt | Numeric | Payments received to date for total amount funded |  |  |
| total\_rec\_prncp | Numeric | Principal received to date |  |  |
| recoveries | Numeric | post charge off gross recovery |  |  |
| installment | Numeric | The monthly payment owed by the borrower if the loan originates. |  |  |
| Customer's debt information | target (loan\_condition) | Categorical | Condition of Loan | 2 | Good Loan Bad Loan |

Table 1. Information Feature

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Count** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **Max** |
| annual\_inc | 887379 | 75027.59 | 64698 | 0 | 45000 | 65000 | 90000 | 95000000 |
| loan\_amount | 887379 | 14755.26 | 8435.455 | 500 | 8000 | 13000 | 20000 | 35000 |
| interest\_rate | 887379 | 13.24 | 4.38 | 5.32 | 9.99 | 12.99 | 16.2 | 28.99 |
| dti | 887379 | 18.15 | 17.19 | 0 | 11.91 | 17.65 | 23.95 | 9999 |
| total\_pymnt | 887379 | 7558.82 | 7871.24 | 0 | 1914.59 | 4894.99 | 10616.81 | 57777.58 |
| total\_rec\_prncp | 887379 | 5757.71 | 6625.44 | 0 | 1200.57 | 3215.32 | 8000 | 35000.03 |
| recoveries | 887379 | 45.92 | 409.69 | 0 | 0 | 0 | 0 | 33520.27 |
| installment | 887379 | 436.71 | 244.18 | 15.67 | 260.71 | 382.55 | 572.6 | 1445.46 |

(Source: Quote from python).

Table 2. Descriptive statistics (Numeric Feature)

* 1. Data Analysis.
     1. Loan condition (Target Feature).

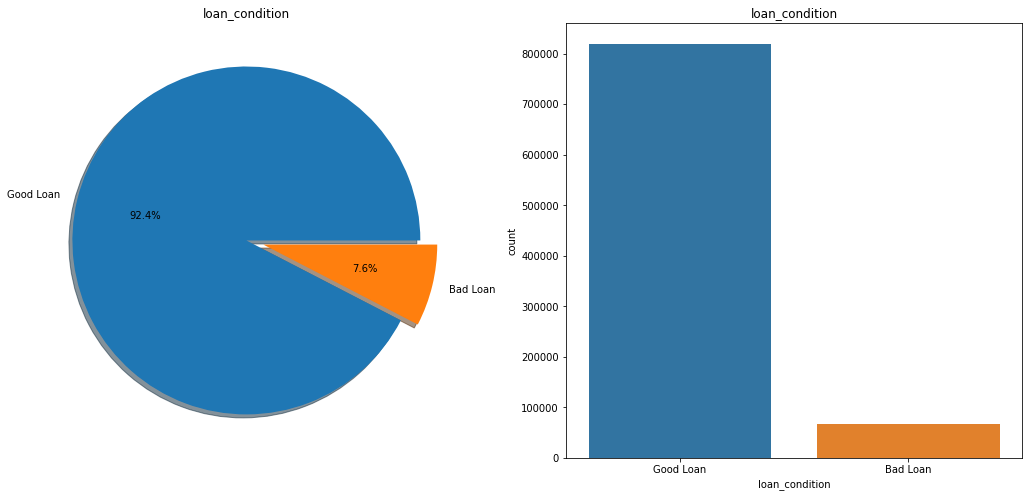


Figure 1.2.1. Analysis Target Feature (Loan Condition)

The customer's debt condition dataset has 92.4% Good Loans and only 7.6% Bad Loans.

=> Dataset is imbalanced data.

* + 1. Year of customer's debt.

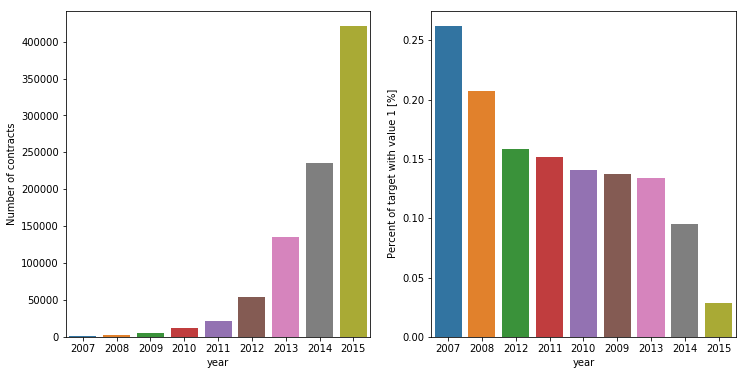


Figure 1.2.2. Analysis Year of customer’s debt (With Target)

The borrowers were mainly in 2013 and beyond, when the economy had stabilized. In the years 2007-2011, very few people applied for loans, especially in the period 2007-2009 - the period when the economic crisis occurred.

It can be seen the effect of that period on the customer's debt status more clearly in the chart of the customer's bad debt ratio year by year. Bad Loan rate was highest in 2007 (more than 25%), then in 2008 (>20%), and Bad Loan rate will decrease gradually in the following years.

* + 1. Home Ownership.

Figure 1.2.3. Analysis Home Ownership (with Target)

The home ownership status feature provided by the borrower upon registration has six different status categories: rent, own, mortgage, other, none, any. None and any account for quite a small amount, this can happen while filling in information, customers or employees have made this data error. Thus processing: merging two data into one.

As can be seen the chart Home Ownership (%), it is clear that customers are mainly living in rented houses (40.13%) or in installments (49.98%), and the number of customers owning their own houses is much less with 9.86% of the total number of customers.

As can be seen the chart Home Ownership (% Bad Loan), although the bad loan rate of other and none is not high, compared to other values in the feature, it is much higher (0.2% and 0.19%).

* + 1. Customer’s Income

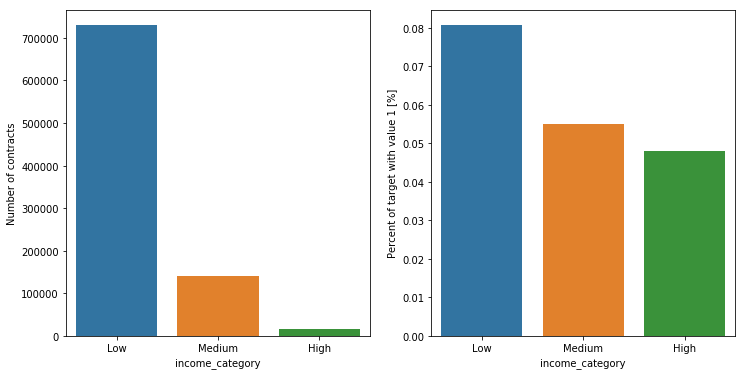


Figure 1.2.4. Analysis Customer’s Income (with Target)

The chart shows that almost customers who borrow from banks have low income (accounting for 82.22%) of the total number of customers. Middle-income customers account for 15.89% and the number of high-income customers is very low (only 1.89% ~ 16786 customers).

* Customers with low income from 100000 down.
* Customers have an average income of 200000 or less.
* Customers with high income over 200000 - 950000.

The second chart shows the trends in customer income and Bad Loan rates. The higher the customer's income, the lower the Bad Loan rate.

* + 1. Loan term of customers (Term).

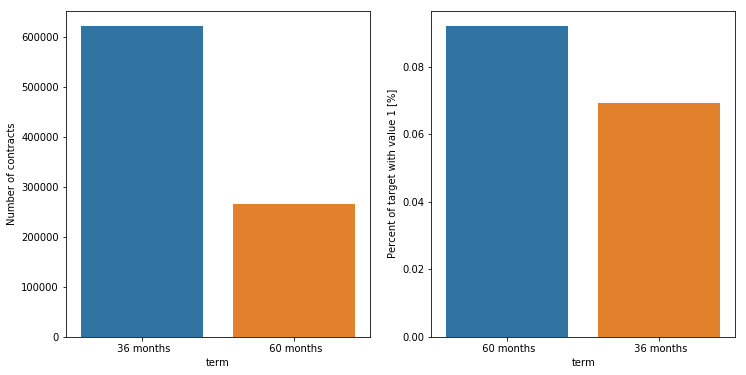


Figure 1.2.5. Analysis Loan term of customers (with Target)

When borrowing from a bank, there are two repayment periods: 60 months and 36 months.

Customers mainly borrow with a term of 36 months, the number of borrowers with a term of 36 months (accounting for 70% of total customers) is more than double the number of borrowers with a term of 60 months (accounting for 30% of total customers). However, the Bad Loan rate for customers who borrow within 60 months is higher than that for customers who borrow within 36 months.

It is possible that a longer loan period results in a lower repayment capacity.

* + 1. Customer’s Purpose.

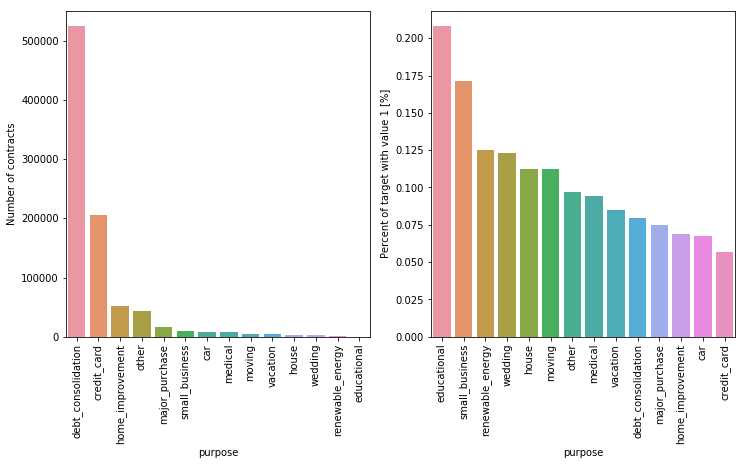


Figure 1.2.6. Analysis Customer’s Purpose (with Target)

There are 14 different reasons (including other options) customers apply for a loan, in which customers apply with the most for the purpose of debt consolidation (Up to 59% of the total customers). Next, Credit Card is also the purpose of many customers to register (accounting for 23.23% of customers).

The two purposes that customers mention the least are renewable energy (0.065%) and educational (0.047%).

Although the fewest people borrow money for educational spending, the percentage of bad loans is the largest (0.2%), which is much larger than for debt consolidation borrowers (although the top number of borrowers - more 59% but Bad Loan rate is only 0.08%).

Bad Loan rate ranks second when customers borrow for small businesses (0.17%), it is possible to hypothesize that the business is not profitable, difficult times => high Bad Loan rate (considering the loan purpose with five years). get a loan).

Customers who use credit cards have the lowest Bad Loan rate (0.06%), because when issuing a credit card the bank has managed how much monthly income, how much income they can open an account with much => lowest Bad Loan rate.

* + 1. Interest Rate.

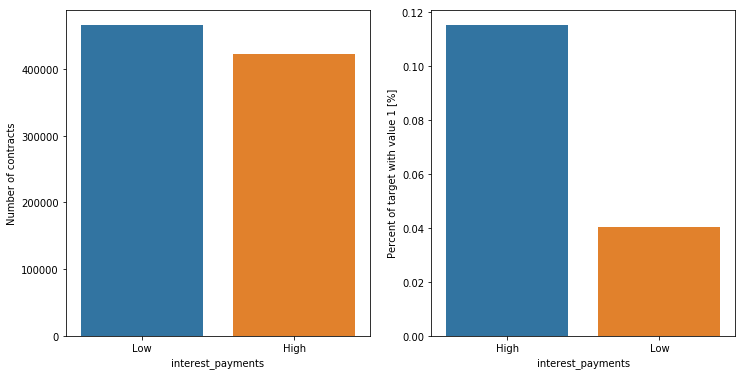


Figure 1.2.7. Analysis Interest Rate (with Target)

The number of customers borrowing with high and low interest rates is not much different.

But the Bad Loan rate of borrowers with high interest rates is much higher (0.11) than customers with low interest rates (0.04) (This is reasonable, because when borrowing with high interest rates, the amount must be pay a lot more => customers can't afford to pay => Bad Loan rate is high).

* + 1. Customer’s Grade

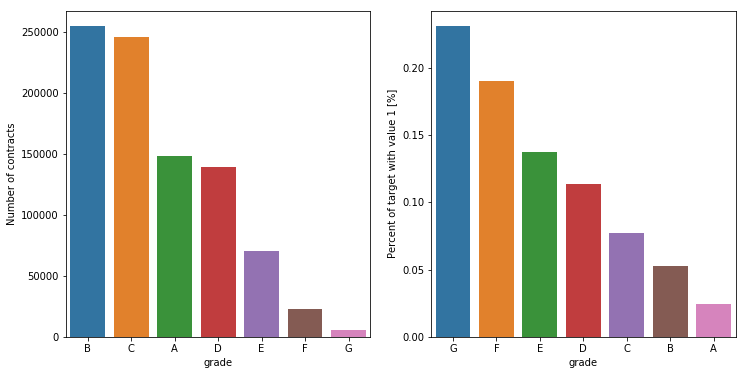


Figure 1.2.8. Analysis Customer’s Grade (with Target)

Customers are mainly in Grade B (28.68%), and Grade C (27%), customers of Grade G and F are very few (accounting for 3%).

Looking at the chart of Bad Loan rate of customers in different grades follows the trend: the higher the customer rank, the lower the rate of Bad Loan. (Grade A Bad Loan rate: 0.024%, Grade G Bad Loan rate: 0.23%).

* + 1. Region.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Leinster | Ulster | Northern-Irl | Cannught | Munster |
| % Percentage | 24.18 | 23.52 | 23.03 | 17.47 | 11.78 |
| % Bad Loan | 0.077 | 0.078 | 0.078 | 0.069 | 0.07 |

Table 3: Information Customer’s Region (with Target)

Customers are almost evenly distributed in 5 regions: Leinster, Ulster, Northern-Iri, Cannught, Munster. The rate of Bad Loan in these areas is almost the same and is very small (<0.08%). Therefore, the living area does not affect the Debt Status of the customer.

1. Data processing.
   1. Clean data.

When processing the data, it is found that there are some features that are not suitable in building the model as well as the label encoding to include in the model, so it will be removed from the data set: 'issue\_d', 'id', 'grade'.

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

* 1. Test for correlation.

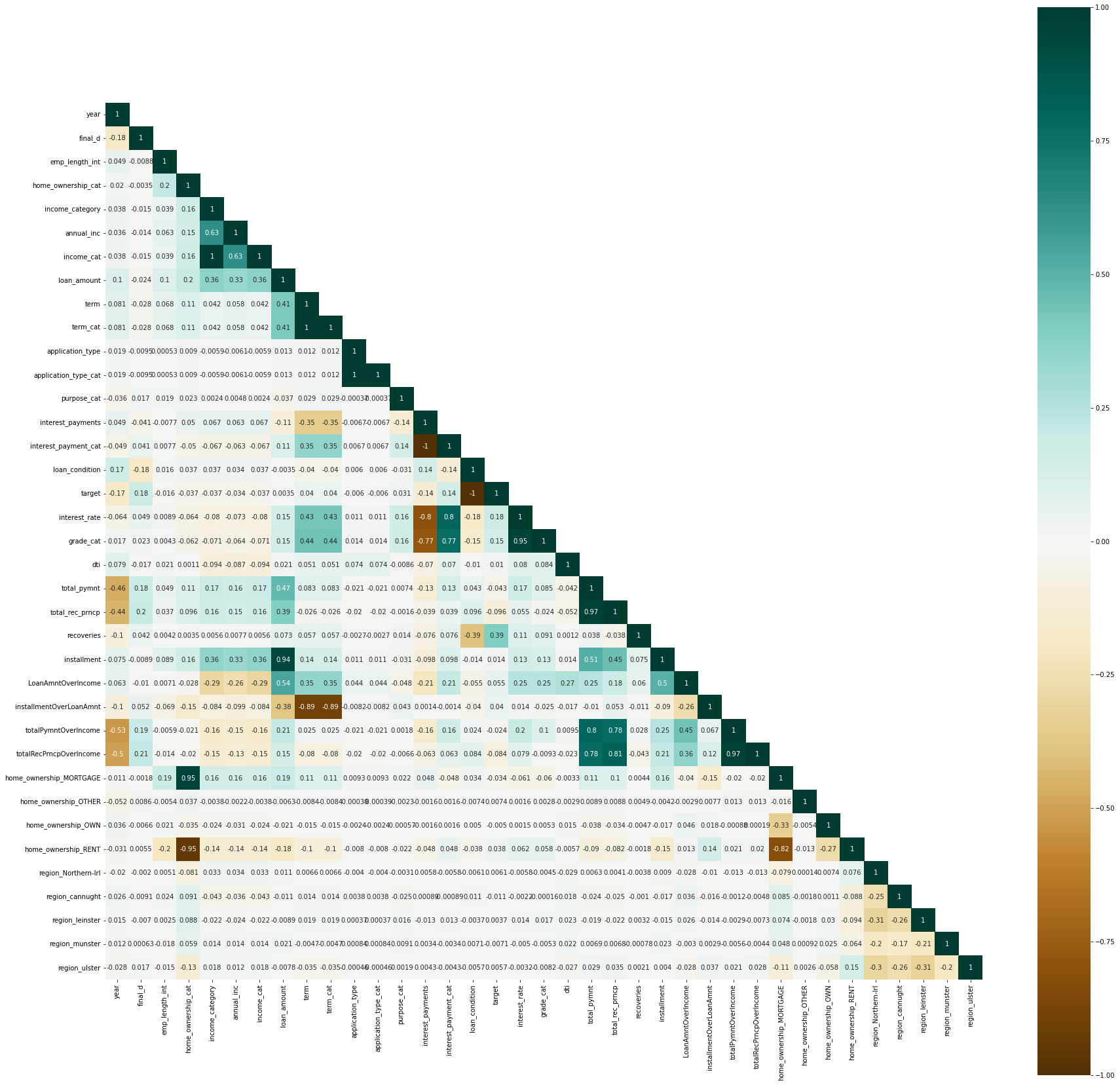


Figure 2.2. Correlation Feature

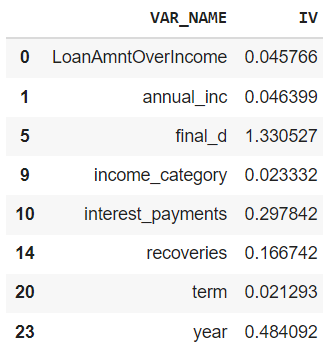
Figure 2.2, green indicates positive correlation, darker green indicates higher correlation; Yellow indicates negative correlation, the darker the yellow, the higher the correlation.

The higher the correlation, the greater the correlation between the features, such features included in the model will be wrong, in this project will remove the features with correlation greater than 0.8.

In this dataset, there are 12 features that need to be deleted because the correlation value with other features is greater than 0.8: ‘income\_cat’, ‘term\_cat’, ‘ application\_type\_cat’, ‘interest\_payment\_cat’, ‘interest\_rate’, 'grade\_cat', 'total\_rec\_prncp', 'installment', 'installmentOverLoanAmnt', 'totalRecPrncpOverIncome', 'home\_ownership\_MORTGAGE', 'home\_ownership\_RENT'.

* 1. 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 9 variables to include in the model: 'LoanAmntOverIncome','annual\_inc','final\_d','income\_category','interest\_payments','recoveries','term','year','totalPymntOverIncome'

* 1. 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).

1. Model and Results.

This project use model for training data: logistic regression, random forest, K- Nearest Neighbors (KNN) and LightGBM.

* 1. Logistic Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Metrics Performance** | | | |
| **Method** | **Precision** | **Recall** | **F1 - Score** | **AUC** |
| Oversampling | 0.92 | 0.82 | 0.86 | 0.89 |
| Undersampling | 0.92 | 0.82 | 0.86 | 0.88 |
| SMOTE | 0.92 | 0.82 | 0.86 | 0.89 |

**Table 5. Model Logistic Regression ‘s Result**

The goal of this project is to expect the model's metrics performance to all be greater than 0.8 (the Precision, Recall, F1-Score, and AUC metrics). Looking at the results of Table 5, we can see that no matter which method is used, the metric scores are all greater than 0.8.

In the studies on predicting Bads Loan or Goods Loan, previous studies used Logistic model (traditional method). And this model is also really effective with model Logistic when the AUC score is up to 0.89.

In this project, not only stops at the traditional method when using Logistic Regression model, but also uses many other machine learning methods to compare, the results will be outlined in the next sections.

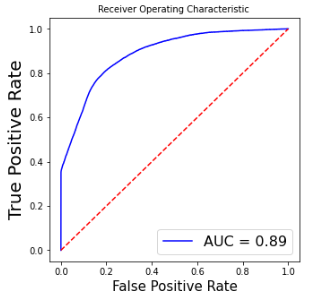


Figure 3.1. Model Logistic Regression’s AUC Curve

* 1. Random Forest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Metrics Performance** | | | |
| **Method** | **Precision** | **Recall** | **F1 - Score** | **AUC** |
| Oversampling | 0.98 | 0.98 | 0.98 | 0.95 |
| Undersampling | 0.95 | 0.93 | 0.94 | 0.96 |
| SMOTE | 0.97 | 0.97 | 0.97 | 0.95 |

**Table 6. Model Random Forest ‘s Result**

With the Random Forest model, it is very clear that all indicators are significantly improved compared to the Logistic Regression model. The dataset actually performs very well in the Random Forest model when all the metrics scores are above 0.9.

Taking a closer look at the results, we see that in the Undersampling method, the Precision, Recall and F1-Score indexes are slightly smaller than the other two methods, but in terms of AUC score, they are higher, however this is the case. that is insignificant. Although the disadvantage of the Random Forest model is that the training time of the model is quite slow, much slower than the Logistic, with this dataset the training does not take too much time, so the Random model Forest is a very good choice when training data. If using the Random Forest model, it will be better to use the Oversampling method to balance the data, which makes the metric scores best.

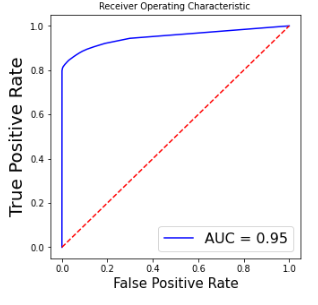


Figure 3.2. Model Random Forest’s AUC Curve

* 1. K- Nearest Neighbors (KNN)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Metrics Performance** | | | |
| **Method** | **Precision** | **Recall** | **F1 - Score** | **AUC** |
| Oversampling | 0.97 | 0.97 | 0.97 | 0.89 |
| Undersampling | 0.95 | 0.94 | 0.95 | 0.91 |
| SMOTE | 0.96 | 0.96 | 0.96 | 0.89 |

**Table 7. Model K- Nearest Neighbor ‘s Result**

With the dataset trained with the KNN model, it is clear that compared to the logistic model, the results are much better (all scores are above 0.9), which proves that the data works very well on KNN model. But when compared with the Random Forest model, the indexes are not as good, especially the AUC index (which is only approximately equal when using the logistic model) is 0.89. In addition, during the running process, the training data with model time is also very long (much longer than Logistic Regression and Random Forest). Therefore, to choose training this data between Logistic Regression and KNN models, then Logistic Regression will be the better solution. When comparing between the three models above (Logistic Regression, Random Forest and KNN), the best choice is the Random Forest model when it gives good metrics scores.

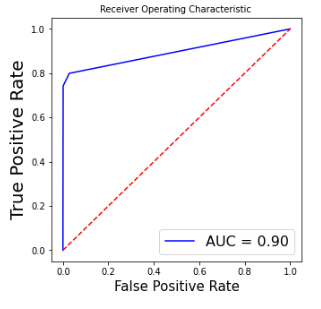


Figure 3.3. Model K- Nearest Neighbors (KNN)’s AUC Curve

* 1. Light GBM

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Metrics Performance** | | | |
| **Method** | **Precision** | **Recall** | **F1 - Score** | **AUC** |
| Oversampling | 0.96 | 0.95 | 0.96 | 0.97 |
| Undersampling | 0.96 | 0.94 | 0.95 | 0.97 |
| SMOTE | 0.97 | 0.96 | 0.96 | 0.96 |

**Table 8. Model Light GBM ‘s Result**

Looking at Table 8, like the three previous models, the Precision, Recall and F1-Score are all greater than 0.9 (namely, all greater than 0.94) – the threshold at which the training model works. works very well on the data set. First of all, about model training time, although model training takes a long time, the results received with the indicators are very good. Compared with the 3 models trained in the previous section, the results of the Light GBM model are good, the AUC index even reaches 0.97 when handling the imbalance data by Oversampling and Undersampling methods, higher than the AUC index. of the Random Forest model, but the F1-Score is slightly lower. In terms of four indicators, it is a bit better to handle imbalance data using Oversampling method. The comparison between the 4 models using the Random Forest and Light GBM models is the same, although the model training time is longer than the traditional method (using the Logistic Regression model), in terms of results, the results are better. great number of.

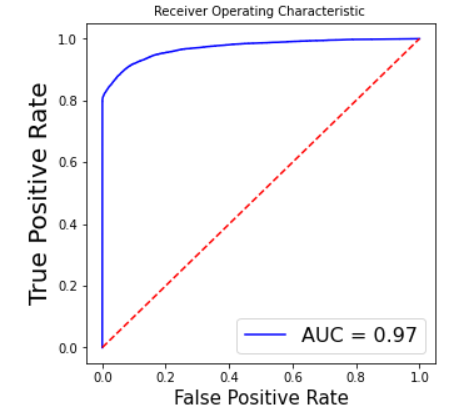


Figure 3.4. Model LightGBM’s AUC Curve

Part 4: Conclusion.

In this paper, the process of forecasting bank bad debt has been carried out. 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, sample data is used to train the dataset (about 887379 data samples) based on the stated criteria to apply the model. built. The project applied an efficient prediction algorithm of Logistic Regression, Random Forest, LightGBM and K Nearest Neighbor models to perform such tests on the training dataset. 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 predictions. The prediction performance is much better even though the training time of the model is longer than that of the Logistic Regression model. Therefore, the Light GBM algorithm and the Random Forest algorithm are stable and powerful with a small error rate, so the results are reasonable. In addition, subject to actual NPL data; Predict results close to reality. There are such logical 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 or Random Forest for data analysis. Therefore, if using this forecasting model, LightGBM (or Random Forest) or LightGBM is realistic and feasible for customers with bad debt and what is the bad debt ratio of the customer. In addition, the model correctly predicts when the right customer attributes are present, in this project, based on the dataset, nine strong and relevant characteristics have the most impact on the target variable: total ratio debt and income, annual income, year in which the transaction ended, income level, interest payable, maturity, year of loan, ratio between total payable and income.