**Risk analytics report**

**Phạm Ngọc Anh**

**1119049**

**DSEB\_61**

**Table of content:**

I. INTRODUCTION ............................................................................................................................

II.THEORETICAL BACKGROUND..............................................................................................

1. Logistic regression algorithm ............................................................................................
2. Decision tree algorithm ...............................................................................................
3. Random forest algorithm ............................................................................................

III.DATA ....................................................................................................................................

1. Dataset Description ......................................................................................................
2. Data preprocessing ......................................................................................................
3. EDA ................................................................................................................................
4. Modeling ........................................................................................................................

IV. FINDING AND DISCUSSION ..........................................................................................

**I: INTRODUCTION**

### In banking sector, Credit risk makes the bank lose the opportunity to receive interest. In addition, it also causes loss to the bank's profit and capital. If the situation becomes serious, the bank will be insolvent leading to bankruptcy

### **The bad debt status of a bank due to the failure to recover the loan, leading to the supervision of the State Bank, thereby reducing the reputation score and the affected operation range is also a consequence of the risk.**

Credit risk is the risk when the borrower fails to comply with the provisions of the credit contract, the most common of which is the failure to recover the loan amount due to the customer's inability ability to pay. Credit risk device to two types: good risk, bad risk. **Good Risk**: An investment that one believes is likely to make profit. The term most frequently refers to a loan made to a trusted individuals or businesses. Good risk is considered exceptionally likely to be repaid. **Bad Risk**: A loan that is unlikely to be repaid due to bad credit history, insufficient income, or another reason. A bad risk raises the risk to the lender and the likelihood of the borrower defaulting.

Personal customers are extremely diverse, creating a potential and "fertile" market for banks. However, credit for individual customers is riskier than corporate customers. Beside the general reasons of bank credit, personal loans also have risks stemming from the customers themselves such as financial conditions of individual or household can change very quickly due to illness, accident, unemployment or family tragedies.

Therefore, due to evaluating the customer's ability to repay, credit scoring is one of the most important issues for lending institutions. Sixty years ago, credit scoring helped credit decision making refers to application of mathematical models that transform relevant data collected from customers themselves, internal systems, credit institution into a value. In retail credit, this method not only reduces the subjective judgment of the creditor, but also maximizes the value of available information, and saves considerable labor costs.

The commonly used scoring methods based on logistic regression algorithm. Because logistic is suitable for binary classification tasks and can calculate the coefficients of each feature. But the logistic regression algorithm needs to meet some assumptions, but the actual business may not meet the corresponding assumptions and the differentiation ability of the logistic regression algorithm is difficult to improve.

At present, with the development of machine learning algorithms such as neural network, support vector machine and nearest neighbor methods.

The major objective of this study is to predict if an applicant is 'good' or 'bad' client using a robust and efficient decision tree and random forest model using a combination of specific sampling algorithms based on machine learning techniques. The results indicate that the random forest under the synthetic minority oversampling technique is able to achieve an accuracy of 97% while logistic model is 71%

**II. THEORETICAL BACKGROUND**

Description of method

a) Logistic regression algorithm

Logistic regression algorithm is a simple and more efficient method for binary and linear classification problems

Assume we have a set of observation given by (, ….(( where n is the number of observations , (i=1,..,n) are the predictor variable, is the response variable. Then the probability for classifying bad and good samples is given by :

where are the model coefficients

Denote where and we have: ln(=z

The transformation function on the left side of the equation is called logit function, and p/(1-p) is called odds ratio, the ratio of probability of occurrence of events to probability of non-occurrence. Logistic regression maps the number of real number ranges to interval [0,1], and gives the probability of event occurrence. With a critical value (e.g. 0.5) in setting, it can judge whether an event occurs or not

b) Decision tree algorithm

Decision Trees (DTs) is a non-parametric supervised learning method. This model predicts the value of a target variable by learning simple decision rules inferred from the data features. There are basic types of nodes: leaf node or terminal node (represent output- the class labels which is used to make prediction), non-leaf node (represent question ) and root node ( non-leaf top node)

How decision tree work?

The model can be "learned" by splitting the source set into subsets based on the check of the attribute value. This process is repeated on each derived subset recursively. The recursion is finished when the subset at a node has the same value of the target variable or when the split no longer adds values to the predictions. Classification and Regression Trees (CART) is one of the methods to separate nodes in decision trees. The concept of the CART method uses the binary splitting method which means that each non-leaf node splits into two new branches. This study use Gini index which similar to information gain, used to evaluate whether the division at the conditional node is good or not.

Due to calculate Gini index, first we will find the Gini at each note

where C is number of classes

, is the number of elements in the ith class.

N is the total number of elements in that node

where is gini index in parent node

K is the number of split child nodes

is the number of gini at the kth child node

M is the number of elements at node p

is the number of elements in the i-th child node

In fact, the gini index computes the parent node's gini offset by the sum of the weighted gini values of the child nodes.

c) Random forest algorithm:

Random forest model will build many trees by decision tree algorithm, however, each decision tree is different (because of random factor). on different data sets and using different attribute sets. Then result of prediction will be compiled from decision trees.

When using the Random Forest algorithm, it is need to pay attention to attributes such as: the number of decision trees to build, the number of attributes used to build the tree. In addition, there are still properties of the Decision Tree algorithm to build a tree such as the maximum depth, the minimum number of elements in a node to be split.

Since each decision tree in the Random Forest algorithm does not use all the training data, nor does it use all the attributes of the data to build the tree, each tree may make bad predictions, then each decision tree model may have high bias. Nevertheless, the final result of the Random Forest algorithm is aggregated from many decision trees, so the information from the trees will complement each other, leading to a model with low bias and low variance, or model have good predictive results.

**III. DATA**

**a) Dataset Description:**

The dataset contains two table, which are connected by ID.

One table include appliers personal information, which you could use as features for predicting i.e. applicant gender, date of birth, jobs, status marital, assets (have car, phone, income ,..)

* There are **18 variables** in **application record** data set :
* **12 categorical** variables
* **5 continuous** variables
* **1** variable to accommodate the applicant ID

Another one is about applicant’s loan payment record: record month, status of a loan. Because there should have 3% users in risk, we will determine response variable: bad risk (who have past due for more than 60 days) and good risk (remaining values)

* There are **3 variables** in **credit record** data set:
* **1 categorical** variable,
* **1 continuous** variable
* **1** variable to accommodate the applicant ID.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Feature name | Explanation | Remarks |
|  | DAYS\_BIRTH | Birthday | Count backward from current day (0),  -1 means yesterday |
|  | DAYS\_EMPLOYED | Start date of employment | Count backward from current day (0). If positive, it means the person currently unemployed. |
|  | MONTHS\_BALANCE | Record month | The month of the extracted data is the starting point, backwards, 0 is the current month, -1 is the previous month, and so on |
|  | STATUS | The amount of time the debt is past due | 0: 1-29 days past due  1: 30-59 days past due  2: 60-89 days overdue  3: 90-119 days overdue  4: 120-149 days overdue  5: Overdue or bad debts, write-offs for more than 150 days  C: paid off that month  X: No loan for the month |

**b) Data preprocessing:**

* Handling missing value, duplicated value:

Because each ID is not for unique client, data set related to consumer information has 79.5% duplicates. All duplicated information will be deleted.

For missing value in categorical variables(occupation\_type), it will be replaced by ‘Other’

For constant feature which has only one unique value (‘flag\_mobil’), it will be drop from data set

* Handling outlier
* Alter categorical variables

These feature which have 2 values yes/no will be convert to dummy variable

We do some transformations in ‘DAYS\_BIRTH’, ‘DAYS\_EMPLOYED’ to age and number of years employed of customer

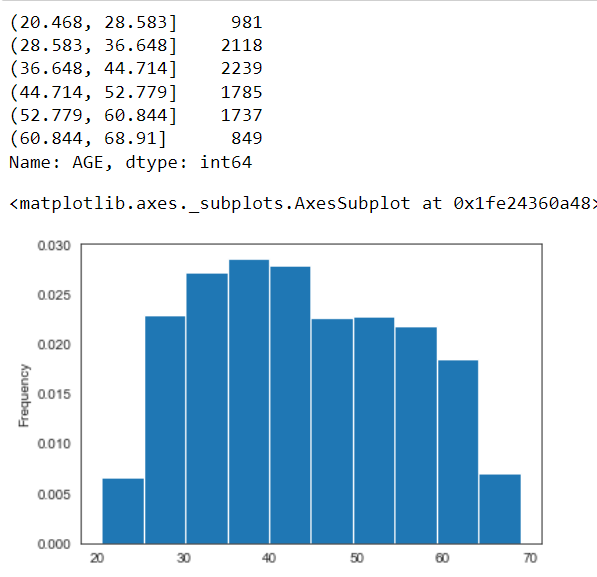
The new ‘AGE’, ‘EMP\_YEAR’ columns and total income will be divide into different interval.

Our target variable will be encoded : 0 – bad debt , 1- good debt

**c) EDA**

After preprocessing, our application data has 90058 rows and 17 columns

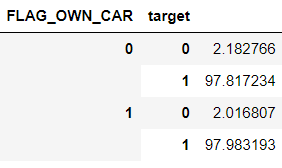
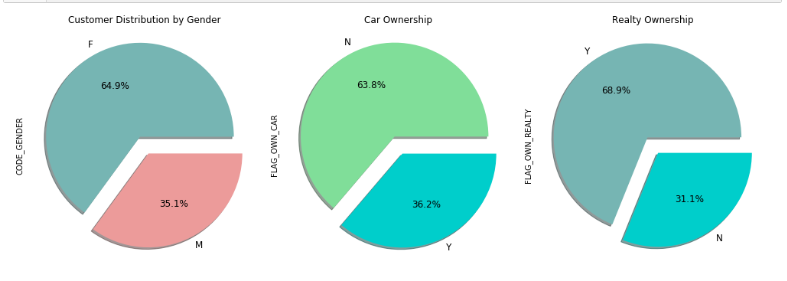
Customer age is between 20 and 69. The group from 28 to 55 years old which is working age takes up largest number of total customer, next is retirement group (after 55 years old) and the lower 28 is the group with the lowest customer participation rate.



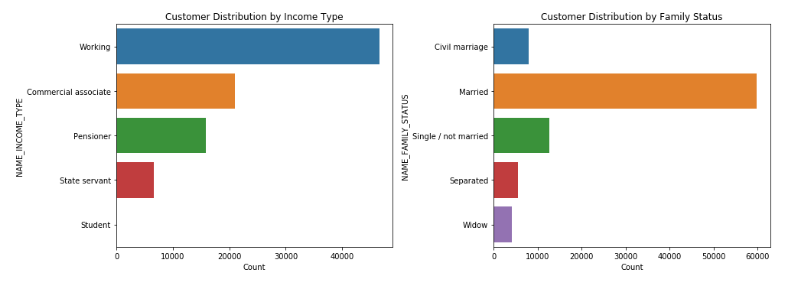
Customer Distribution by Gender: 64,9% female; 35,1% male

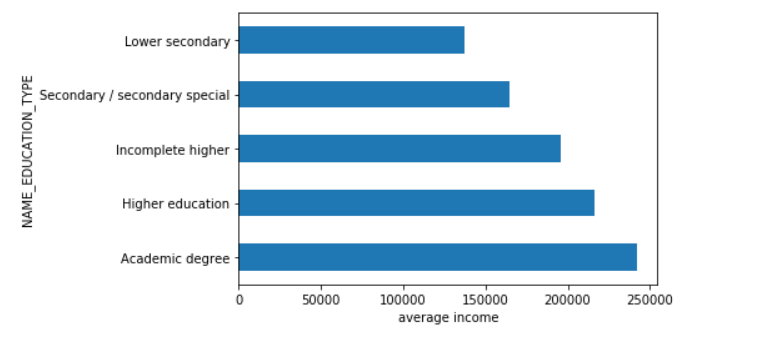
Car Ownership: 63,8% customer have car. However, proportion of bad customers who own a car is less than those who don't own a car

Realty ownership: there’s 68.9% customer own real estate

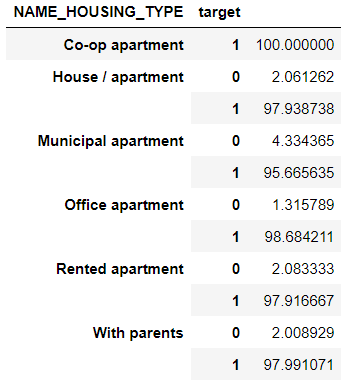


Working person and who married take up largest percentage. And obviously , the high education level the high average income. Average income in total customer is 178700$





People living rent don't have the highest proportion of bad customers. Person owning 'Office apartment' has the highest proportion of bad customers



d) Modeling

* Select features:

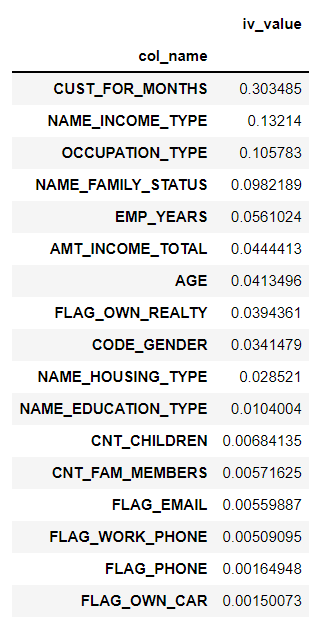
IV, Woe: concept of weight of evidence, information value and how they apply

WOE show the predictive power of an independent variable related to the dependent variable.

IV is a measure of the predictive power of a categorical variable in predicting the outcome (or label) of the classified variable. Woe is mediated to calculate IVs (cumulative IVs to be exact)

According to Siddiqi (2006), the values of the IV statistic in credit scoring can be understand as follows:

|  |  |
| --- | --- |
| IV | Ability to predict |
| <0.02 | the predictor is not useful for modeling |
| 0.02~0.1 | the predictor has a weak relationship to the Good/Bad odds ratio |
| 0.1~0.3 | the predictor has a medium strength relationship to the Good/Bad odds ratio |
| 0.3~0.5 | the predictor has a strong relationship to the Good/Bad odds ratio |
| >0.5 | suspicious relationship |



Depending IV value table, all feature have result lower than 0.02 will not be in final model.

* Splitting data: data is splinted to two set, 70% for training , 30% for testing
* Feature scaling :

Normalization and Standardization are 2 method to scale data. In this study, we use Normalization which scales data from any range of values to a range between 0 and 1.

Y=(x-min)/(max-min)

where y is the value after normalize, x is the original value

min,max is maximum and minimum value of all observation in each feature

* Our target is in imbalance status with only 1,5% bad customer in total amount. In order to improve this problem, we oversampling training data set by SMOTE technique existed in sklearn library
* Logistic regression, decision tree and random forest will be built for this data set

**IV. Finding and discussion**

1. Training data set:

|  |  |
| --- | --- |
|  | Accuracy score |
| Logistic regression | 0.72 |
| Decision tree | 0.9 |
| Random forest | 0.95 |

2. Testing data set

|  |  |
| --- | --- |
|  | Accuracy score |
| Logistic regression | 0.68 |
| Decision tree | 0.82 |
| Random forest | 0.87 |

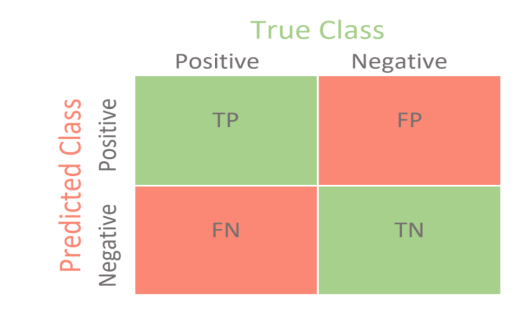
* F1 score
* the ability to classify positive samples.
* the ability to classify negative samples..

If the negative samples are important, we should focus on precision. Otherwise, we should focus on recall. In this study, the propose is predict bad debt (negative sample) so we choose precision is index to evaluate the classification. About f1-score, the model is better when it has high f1-score. Therefore, decision tree is the best model for this case

|  |  |
| --- | --- |
| Logistic regression |  |
| Decision tree |  |
| Random forest |  |

* Confusion matrix

Is a specific table layout that allows to visualize the performance of an algorithm



TP (True Positive): Index of correct predictions with reality

TN (True Negative): Indirectly accurate salary predictions.

FP (False Positive-type1 error): Number of false predictions.

**FN (False Negative - type2 error):** The number of indirectly false predictions

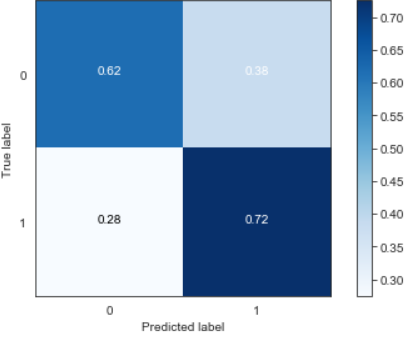
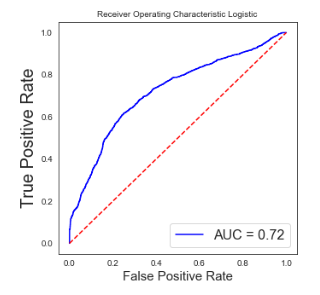
* AUCROC: a graph that shows the predictability of a binary hierarchy



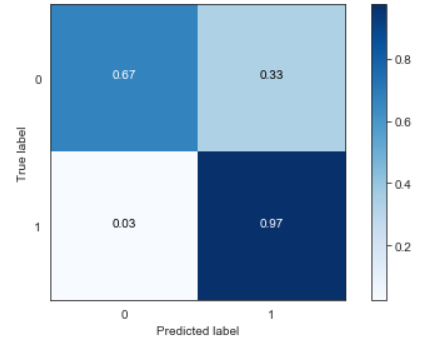
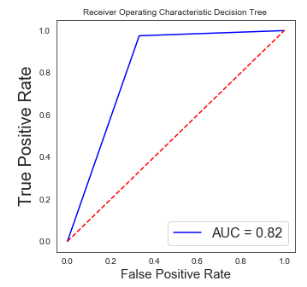
The area below the ROC curve is called the AUC

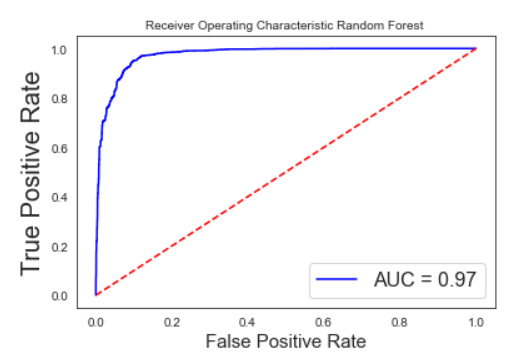
The larger they are, the more efficient the model is at classifying the data. And the more inclined the ROC curve is to the left the better

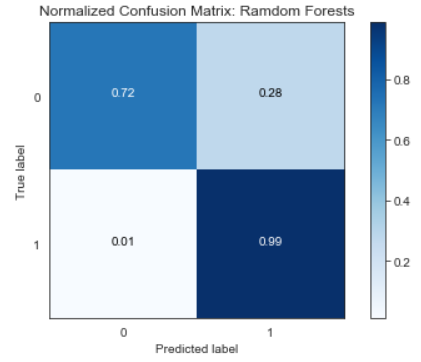
* Logistic regression



* Decision tree

* Random forest



3. Conclusion

The use of Machine Learning’s algorithms in credit risk scoring help standardize the decision-making process, the sound decisions can be made faster or automated. With feature selection, it improves traditional models based on a simple multivariate statically technique and it offers non-parametric approaches

Beside Logistic regression used in predict loan status, which is good risk or bad risk, this study has been used decision tree and random forest algorithms. The result shows that decision tree model for the data we chose is the most efficient in three model.