NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

AN8002 AI With Advanced Predictive Techniques in Finance

Business Analytics, AY 2020-21, Trimester 2

Group Project Report

Credit Default Risk Modelling

Members	Chen Dingming
	Ritika Jain
	Tan Chian Wen Melvin
	Wang Jingyi
	Yu Peichen
Group	Group 3
Instructor	Dr. Teoh Teik Toe
Submission Date	29 December 2020

Summary

Credit default risk prediction can help in balancing risk and return for the lender, finding target reliable customers, charging higher rates for higher risks, or even denying the loan when required. For Home Credit Group, it strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. This is a Kaggle Competition.

The dataset from Home Credit Group is quite complicated. We address the issue of several rows and columns having too many missing values. In addition, in particular with respect to the categorical features, there is sparse data for some of the categories that we remedy in line with contextual sense.

Using the cleaned dataset, we explore 65 continuous data features by plotting out correlation matrix and probability histogram. We find DAYS_EMPLOYED possible to be representative features. For the 52 categorical features, we explore contract type, gender, family status, education etc. The result shows that more customers ask for Cash loans than revolving loans; Laborers and sales staff are the largest occupation groups that need to get loans due to economic burden.

Recursive feature elimination with cross-validation (RFECV) was applied to select features. A special version of SMOTE which is also called SMOTE for Nominal and Continuous (SMOTENC) was performed for imbalanced data problems.

Considering the satisfied performance of decision tree models, we perform Random Forest and XGBoost model in this project. A randomized grid search is used to find the best hyperparameter combination of random forest. The result shows XGBoost has an auc of 0.73 on the test set with 65.35% recall based on 0.3 threshold, which is better than random forest model. Finally, we display the 10 top features in the XGBoost model. We interestingly find that house related factors and income stability, rather than income alone have a big impact on whether a client will default or be able to repay his or her loan.

Contents

1. Motivation.......... 1

 1.1 Introduction.......... 1

 1.2 Project Objectives.......... 1

2. Literature Review.......... 1

3. Problem Statement.......... 2

4. Data Overview.......... 3

 4.1 Data Description.......... 3

 4.2 Data Cleaning.......... 3

 4.3 Data Visualization.......... 6

  4.3.1 Continuous Features.......... 6

  4.3.2 Categorical Features.......... 8

5. Modelling.......... 9

 5.1 Data preparation.......... 9

  5.1.1 Feature Engineering.......... 9

  5.1.2 Data Sampling.......... 10

  5.1.3 Dummy Variables.......... 10

 5.2 Model Training.......... 10

  5.2.1 Random forest.......... 10

  5.2.2 XGBoost.......... 10

 5.3 Model Comparison.......... 11

6. Conclusion & Future Study.......... 11

1. Motivation

1.1 Introduction

Credit default risk is the probability that a borrower fails to make full and timely payments of principal and interest. For customers, carry higher default risk will often find it difficult to get loans. For lenders and investors, they hope to lend money to reliable customers and avoid default risk in virtually all forms of credit extensions. To mitigate the impact of default risk, lenders often impose charges that correspond to the debtor's level of default risk. A higher level of risk leads to a higher required return. An accurate prediction can help in balancing risk and return for the lender, finding target reliable customers, charging higher rates for higher risks, or even denying the loan when required.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. To make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

This is a Kaggle Competition from Real life scene: https://www.kaggle.com/c/iiitb2020-home-credit-default-risk/data. One of the earliest uses of machine learning was within credit risk modeling, whose goal is to use financial data to predict default risk. In this paper, we would apply Machine learning methods to solve this classification problem and later evaluate the result based on the real data.

1.2 Project Objectives

There are several key objectives this paper aiming at:

- Review on the current research for machine learning in Credit Default Risk, identify general process, special methods and accuracy has been achieved.
- Develop data processing methods to clean the data and generate customer features that can be used in the model.
- Perform feature selection methods to identify important features should be used. Develop machine learning model to predict if a customer able to pay the loan or not. Evaluate the result in the test set and compare model performance.
- Based on the model result and current research, describe pros and cons of our model and aspects of future improvement, as well as conduct real-life application.

2. Literature Review

The previous literature in consumer credit can be categorized into two sections , the study on consumer credit applicants, with the lender's decision to grant the loan to the applicant and studies involving whether the consumer client will be able to repay the loan or not. There are many studies on scrutinizing and improving the rejection and acceptance criteria of credit lender's decisions. For example, on investigating the borrower's and lender's behavior it was found that most of the applicants are rejected based on their credit history, their age and their income, amount of collateral whereas time spent at the current job, time spent at the current address. type of work, their family size, sex were given less importance.

According to the study with 1989 data based in the US, it examines whether the client characteristics, which predict the probability of households being credit constrained, have changed or remained the same between 1983 and 1989 in the United States. According to his study, more years of schooling of a household head would also be expected to increase future income, with consequent increases in the household's demand for credit and the supply. There is tenuous evidence that the probability of default decreases with age. In the case of family size, there is clear evidence that the probability of default increases as the number of children increases. Roszbach K. and Jacobson T. (1998) built a statistical model in order to measure the risk of a sample loan portfolio and show how the model helps to evaluate alternative lending policies. They found that income does not affect credit-granting decisions and being a male significantly decreases the chance of being granted a loan. In addition, homeowners have more chance of being granted a loan. Although researches are mostly interested in evaluating the lenders' decisions on granting loans to credit applicants, the results of previous studies are contradictory. Not many studies are done to investigate the relationship between characteristics of people that are already accepted (client) and whether they are paying back their loans on time or not. Carling K., Jacobson T., and Roszbach K (1998) examine the Swedish consumer credit clients' payment performance. According to their study, married applicants tend to pay back their loans faster. A possible reason might be the existence of two wage earners in the Swedish families, which leads to a stable flow of income. Alternatively, it could reflect the fact that married couples are simply more diligent. Surprisingly, they found a negative relationship between incomes and default risk and the size of limit having no influence on payment performance, whereas increasing the loan size delay payback. Sexton D. E. (1977) analyzes the credit risk in two types of American families: (i) low-income families; (ii) high-income families. Aim of his study is to find out whether or not the variables associated with good credit risks among high income families are similar to those for low-income families. His study does not analyze the extent of the impact of the independent variables on the dependent variable. However, its numerical results indicate that married couples and homeowners tend to pay their debt on time. On the other hand, credit default risk decreases when the income and age increase.

Loans are the major source of revenue for the bank. So the main task for the bankers is to evaluate the loan applicants and approve loans to the borrowers who are capable of the repayment. Giving loans to the people who are able to repay, increases the revenue of the bank.

Home Credit provides easier, simple and fast loans for a range of home appliances, Mobile Phones, Laptops, Two wheelers and varied personal needs.

Many problems face the issue of misclassification. Suppose the model predicts that he/she is able to pay the loan and the loan amount is sanctioned to him/her. But later due to some reason he is not able to repay the loan. This type of scenario will make a company lose its money. Hence we should come up with the model which reduces this type of risk.

3. Problem Statement

Home Credit hopes to utilize their various statistical customer data and current credit risk modeling to make default risk predictions.

The data is complicated and dirty that data processing is necessary before data exploration. there are too many potential features thus Home Credit must decide which to use.

Credit default prediction model should be accurate as expected, otherwise would cause potential loss for Home Credit. If cannot predict the probability of default for a borrower, they cannot charge higher rates for higher risks, or even denying the loan when required, and as a result Home Credit suffer loss since fail to get the money back. On the other hand, failing identify the reliable customer would cause profit loss for the company and make Home Credit less competitive.

Given the applicant data, all credits data from Credit Bureau, previous applications data from Home Credit and some more data, our objective is to build up applicable machine learning prediction model to identify the reliable customer who are very likely to repay the loan and avoid loss due to risk customers. We would apply several machine learning models and evaluate which performs the best. Based on our model, we would offer further advice for Home Credit to better success in the market.

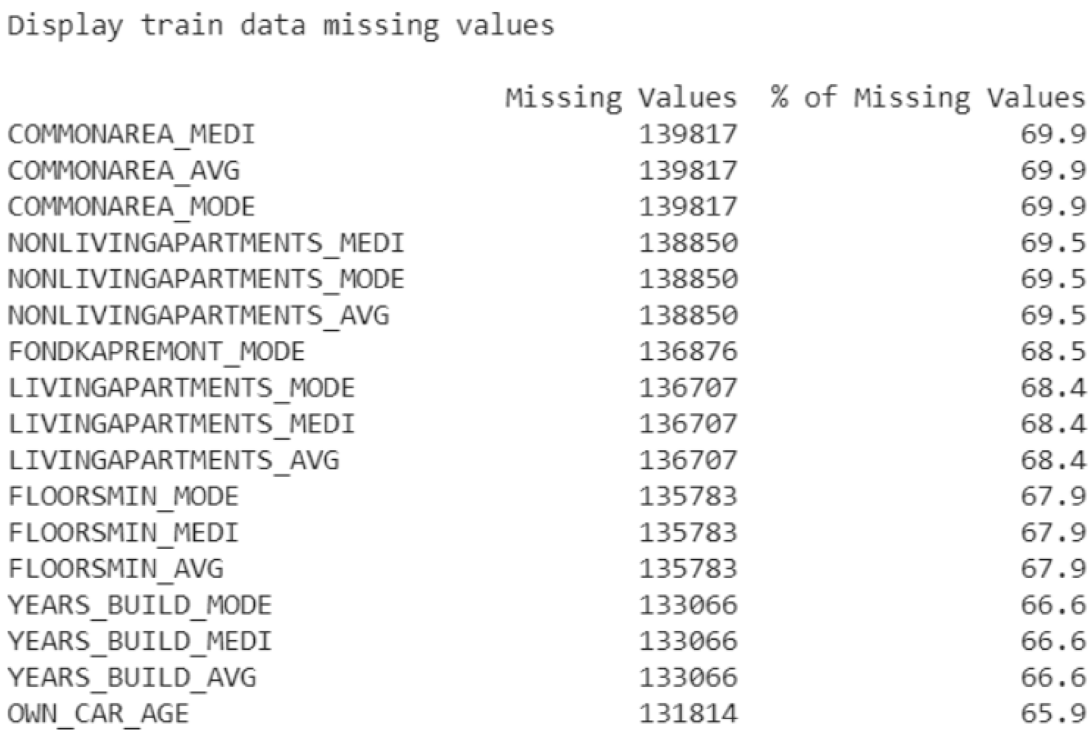
4. Data Overview

4.1 Data Description

The main data table is application_train.csv and is the dataset used for the data analytics for this project. For this project we are unable to use the corresponding application_test.csv for model testing as it does not contain the target column so that participants of this particular Kaggle competition are unable to know their model performance before submitting their work. However, the application_train.csv contains sufficient instances to facilitate both model training and model testing after data cleaning and performing a train-test split. The dataset has a total of 199882 instances and 122 columns. Excluding the primary key column and the target column, the dataset has in total 120 features.

4.2 Data Cleaning

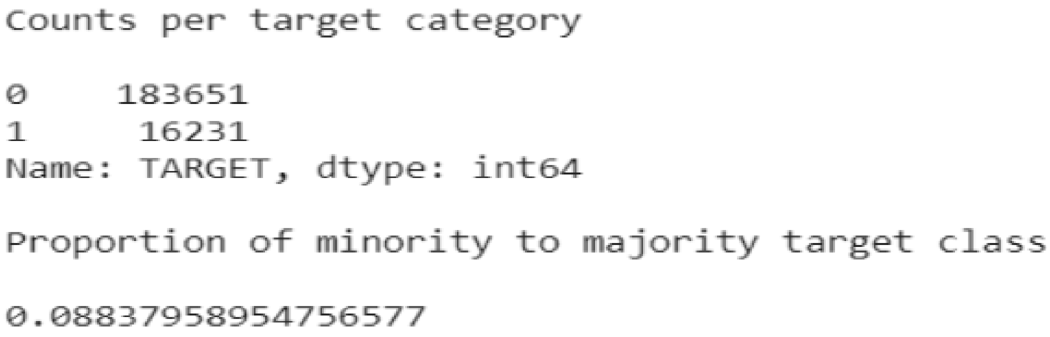
We note that several columns have many missing values. We display the number of missing values and percentage of missing values each feature has in descending order. The table below shows the columns with more than 60% missing values.



```
Display train data missing values

                            Missing Values  % of Missing Values
COMMONAREA_MEDI                     139817                 69.9
COMMONAREA_AVG                      139817                 69.9
COMMONAREA_MODE                     139817                 69.9
NONLIVINGAPARTMENTS_MEDI            138850                 69.5
NONLIVINGAPARTMENTS_MODE            138850                 69.5
NONLIVINGAPARTMENTS_AVG             138850                 69.5
FONDKAPREMONT_MODE                  136876                 68.5
LIVINGAPARTMENTS_MODE               136707                 68.4
LIVINGAPARTMENTS_MEDI               136707                 68.4
LIVINGAPARTMENTS_AVG                136707                 68.4
FLOORSMIN_MODE                      135783                 67.9
FLOORSMIN_MEDI                      135783                 67.9
FLOORSMIN_AVG                       135783                 67.9
YEARS_BUILD_MODE                    133066                 66.6
YEARS_BUILD_MEDI                    133066                 66.6
YEARS_BUILD_AVG                     133066                 66.6
OWN_CAR_AGE                         131814                 65.9
```

Figure 1 Columns with more than 60% missing values

We find the number of counts in each target category. Category 0 refers to bank loan clients who were able to repay their loan while category 1 refers to bank loan clients who were unable to repay their loan. We find that the dataset is highly imbalanced.



```
Counts per target category

0    183651
1     16231
Name: TARGET, dtype: int64

Proportion of minority to majority target class

0.08837958954756577
```

Figure 2 Displaying the proportion of the minority class to majority class

We commence the data cleaning by removing the rows that have many missing values. However, we want to do this in a fashion such that the proportion of the minority class to majority class is preserved as far as possible. The histogram on the left below shows the distribution of missing values for data rows labelled as target 1 and the histogram on the right below shows the distribution of missing values for data rows labelled as target 0.

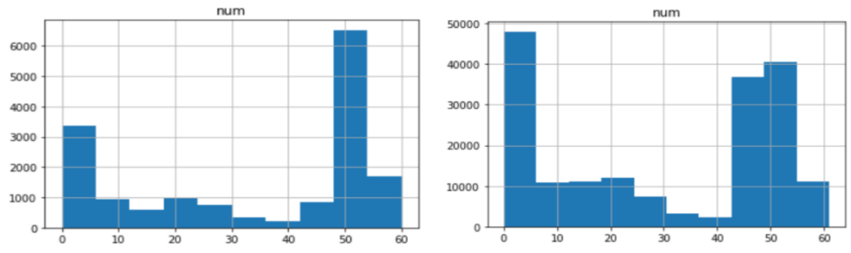
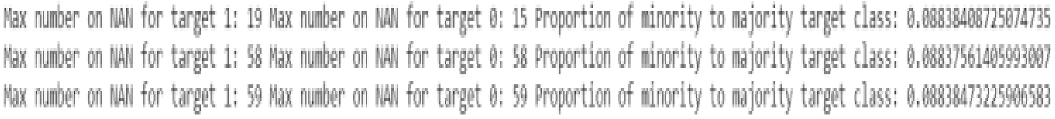


Figure 3 Histograms displaying the distribution of missing values for data rows

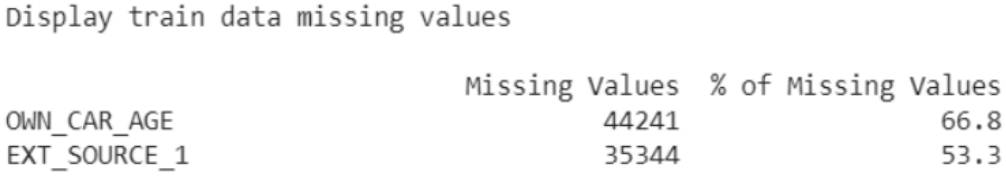
We wish to minimize the maximum possible number of missing values for data rows labelled as target 1 and data rows labelled as target 0 respectively such that the proportion of minority to majority class is preserved as far as possible. After running a double for loop, we find that 19 is the minimum maximum possible number of missing values data rows labelled as target 1 can have and 15 is the minimum maximum possible number of missing values data rows labelled as target 0 can have such that the proportion of minority to majority class is preserved.



```
Max number on NAN for target 1: 19 Max number on NAN for target 0: 15 Proportion of minority to majority target class: 0.08838408725074735
Max number on NAN for target 1: 58 Max number on NAN for target 0: 58 Proportion of minority to majority target class: 0.08837561405993007
Max number on NAN for target 1: 59 Max number on NAN for target 0: 59 Proportion of minority to majority target class: 0.08838473225906583
```

Figure 4 Preserving the proportion of minority to majority target classes after row removal

We proceed to remove all data rows labelled as target 1 that have more than 19 missing values and all data rows labelled as target 0 that have more than 15 missing values. We find that we have 60882 instances of the majority class and 5381 instances of the minority class remaining. We recount the number of missing values per column.



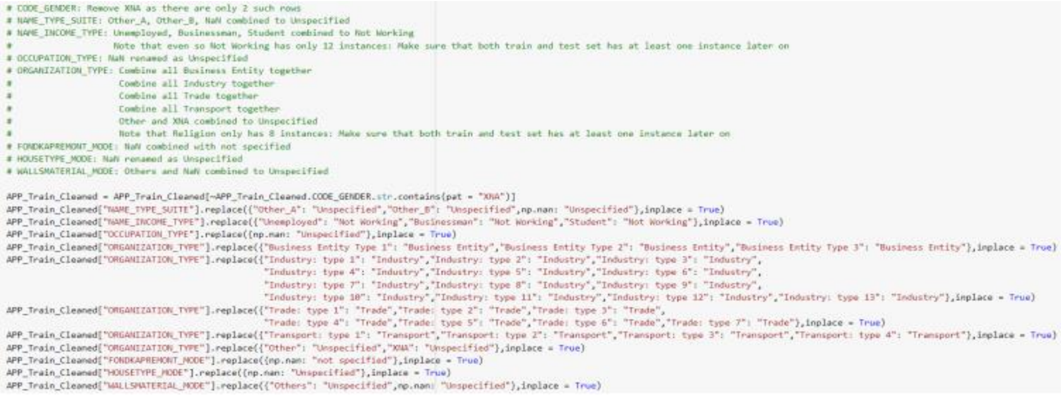
```
Display train data missing values

                       Missing Values  % of Missing Values
OWN_CAR_AGE                     44241                 66.8
EXT_SOURCE_1                    35344                 53.3
```

Figure 5 result of two columns have more than 50% missing values

This time only 2 columns have percentage of missing values of more than 50%.

The next stage of the data cleaning involves the dropping of the above two columns. Many of the columns still contain missing values but they are less than 50% of the total number of values.

Subsequently, the features are grouped among 3 main types: categorical, integer coded and continuous. We start with categorical. Issues we wish to iron out are our remaining NaN values, unknown values coded as XNA which are similar in essence to missing values and sparse data. We check that in all features that the XNA value appears, its proportion is less than 50% and there are no NaN values. This is to ensure consistency that 1) when there is a missing value for a particular feature, it is either represented by XNA or NaN but not both and 2) regardless of representation, its proportion is less than 50%. For gender, there are only two rows that contain XNA values. We remove these two rows as an unknown gender does not make contextual sense and dropping only 2 rows will not affect the proportion of minority to majority class significantly. We relabel missing values as “Unspecified” and combine sparse categories together or merge spares categories with larger categories provided that it makes contextual sense in order to mitigate several data sparseness issues as far as possible. The code with comments below explains the details on how the data processing was done for this stage of the data cleaning process.



```
# CODE_GENDER: Remove XNA as there are only 2 such rows
# NAME_TYPE_SUITE: Other_A, Other_B, NaN combined to Unspecified
# NAME_INCOME_TYPE: Unemployed, Businessman, Student combined to Not Working
#                   Note that even so Not Working has only 12 instances: Make sure that both train and test set has at least one instance later on
# OCCUPATION_TYPE: NaN renamed as Unspecified
# ORGANIZATION_TYPE: Combine all Business Entity together
#                    Combine all Industry together
#                    Combine all Trade together
#                    Combine all Transport together
#                    Other and XNA combined to Unspecified
#                    Note that Religion only has 8 instances: Make sure that both train and test set has at least one instance later on
# FONDKAPREMONT_MODE: NaN combined with not specified
# HOUSETYPE_MODE: NaN renamed as Unspecified
# WALLSMATERIAL_MODE: Others and NaN combined to Unspecified

APP_Train_Cleaned = APP_Train_Cleaned[~APP_Train_Cleaned.CODE_GENDER.str.contains(pat = "XNA")]
APP_Train_Cleaned["NAME_TYPE_SUITE"].replace({"Other_A": "Unspecified","Other_B": "Unspecified",np.nan: "Unspecified"},inplace = True)
APP_Train_Cleaned["NAME_INCOME_TYPE"].replace({"Unemployed": "Not Working","Businessman": "Not Working","Student": "Not Working"},inplace = True)
APP_Train_Cleaned["OCCUPATION_TYPE"].replace({np.nan: "Unspecified"},inplace = True)
APP_Train_Cleaned["ORGANIZATION_TYPE"].replace({"Business Entity Type 1": "Business Entity","Business Entity Type 2": "Business Entity","Business Entity Type 3": "Business Entity"},inplace = True)
APP_Train_Cleaned["ORGANIZATION_TYPE"].replace({"Industry: type 1": "Industry","Industry: type 2": "Industry","Industry: type 3": "Industry",
                                                "Industry: type 4": "Industry","Industry: type 5": "Industry","Industry: type 6": "Industry",
                                                "Industry: type 7": "Industry","Industry: type 8": "Industry","Industry: type 9": "Industry",
                                                "Industry: type 10": "Industry","Industry: type 11": "Industry","Industry: type 12": "Industry","Industry: type 13": "Industry"},inplace = True)
APP_Train_Cleaned["ORGANIZATION_TYPE"].replace({"Trade: type 1": "Trade","Trade: type 2": "Trade","Trade: type 3": "Trade",
                                                "Trade: type 4": "Trade","Trade: type 5": "Trade","Trade: type 6": "Trade","Trade: type 7": "Trade"},inplace = True)
APP_Train_Cleaned["ORGANIZATION_TYPE"].replace({"Transport: type 1": "Transport","Transport: type 2": "Transport","Transport: type 3": "Transport","Transport: type 4": "Transport"},inplace = True)
APP_Train_Cleaned["ORGANIZATION_TYPE"].replace({"Other": "Unspecified","XNA": "Unspecified"},inplace = True)
APP_Train_Cleaned["FONDKAPREMONT_MODE"].replace({np.nan: "not specified"},inplace = True)
APP_Train_Cleaned["HOUSETYPE_MODE"].replace({np.nan: "Unspecified"},inplace = True)
APP_Train_Cleaned["WALLSMATERIAL_MODE"].replace({"Others": "Unspecified",np.nan: "Unspecified"},inplace = True)
```

Figure 6 Handling of remaining missing values and sparse data for categorical features

We move on to integer coded features. There are no missing values for the integer coded features. These features are of dtype int64 and with 3 or less unique values. Those with 2 unique values are dummy encoded binary categorical features and those with 3 unique values are ordinal features. Last but not least we handle the continuous features. Features of dtype int64 and with strictly more than 3 unique values are treated to be continuous features and features of dtype float64 are naturally continuous features. The former has no missing values while the latter does. These missing values are replaced by the corresponding relevant column mean.

Finally, we are left with no missing values and any sparse data mitigated as far as possible. The resulting dataset is saved into a csv file and ready for machine learning modelling. It has 60880 instances of the majority class and 5381 instances of the minority class, 16 categorical features, 34 integer coded features and 68 continuous features. The feature columns are ordered starting with the categorical features followed by the target followed by the integer coded features and ending with the continuous features for convenient sub-setting of the data.

4.3 Data Visualization

Using the cleaned dataset, the next step before modeling is to perform data visualization to explore customer features and see whether it can give information on target credit default detection classification. This would be helpful for better understand the customers group.

4.3.1 Continuous Features

There are 65 continuous features for credit customers. The Pearson correlation matrix can explore the relation between continuous features. In this graph, we can see DAYS_EMPLOYED are correlated with several features like DAYS REGISTRATION, DAYS IN PUBLISH, CNT FAM MEMBERS, thus this might be a representative feature for customers in this dataset. What’ s more, average data features in this graph, seeing columns from BASEMENTAREA AVG to TOTALAREA MODE are slightly positive correlated with each other. The rest do not have significant relation.

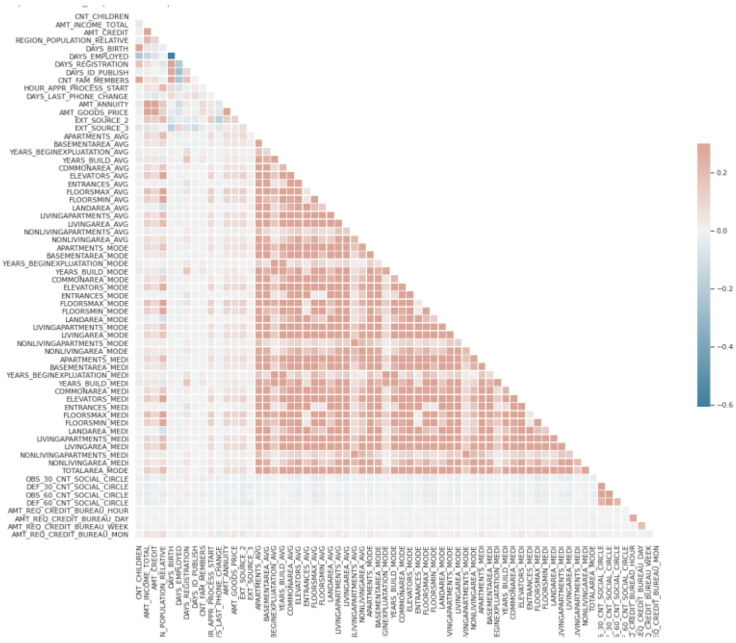


Figure 7 Correlation Matrix

Next, we want to see how the continuous data are corelated with credit default customers. We use seaborn to plot out histogram for continuous features. There are 65 subplots. Due to the imbalanced data problems, number of TARGET=1 data point is much less than TARGET=0's, but we can see many histograms have the same distribution shape for TARGET=1 and TARGET=0. Besides, AMT CREDIT may be significant classified features because the distribution shape is different.

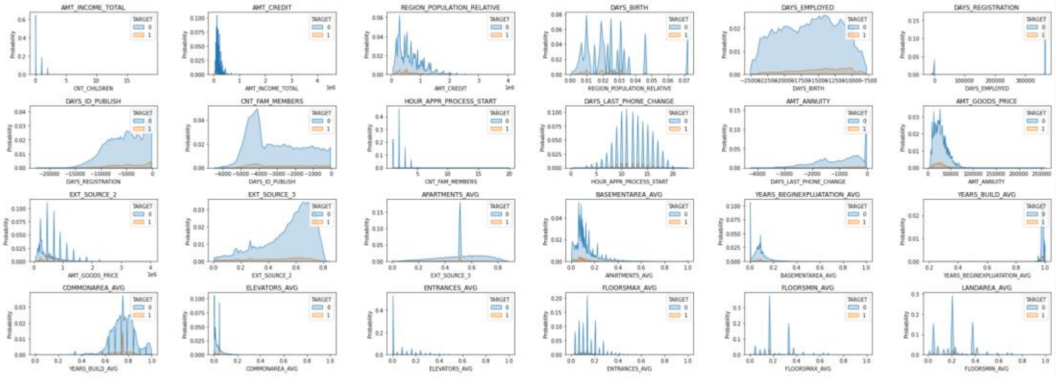


Figure 8 Continuous Subplots

DAYS EMPLOYED data point has a distributed data shape. We can use violin plot to view its shape. TARGET=1, which means risk customers, have fewer employed days than non-risk customers.

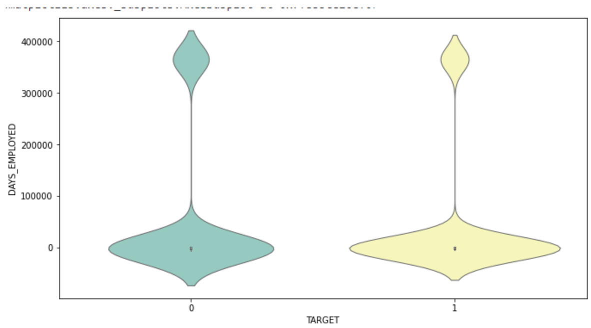


Figure 9 DAYS EMPLOYED Plot

4.3.2 Categorical Features

There are 52 categorical features in this dataset. These features have categorical meaning itself.

For "NAME_CONTRACT_TYPE" and "CODE_GENDER", more customers ask for Cash loans than revolving loans. There are more female customers than male customers in both group.

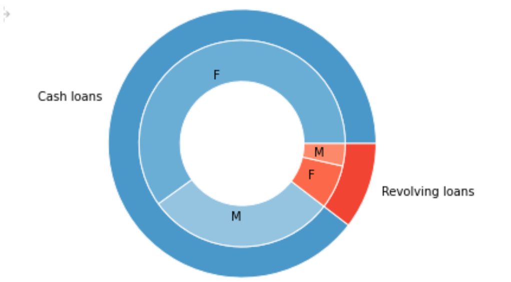


Figure 10 "NAME_CONTRACT_TYPE" "CODE_GENDER" plot

When comes to "CODE_GENDER", "NAME_FAMILY_STATUS", "NAME_EDUCATION_TYPE", married group takes up more than half of the total group, which means married group needs more money for life expenses or investment, no matter for male or female. Single/Not Married takes up more than Separated people in loans. The largest education groups are Secondary and Higher education. And Secondary groups have higher married rate than higher education group.

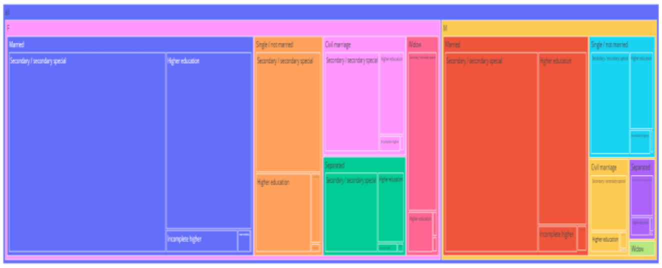


Figure 11 Tree-map by gender, family status, education

For "OCCUPATION_TYPE" and "TARGET", quite a few customers are not given specified occupation type info in the data. We can see laborers and sales staff are the largest occupation groups that need to get loans due to economic burden. At the same time, laborers have high credit default risk, thus they are more difficult to get loans and need to pay higher interest at the same time.

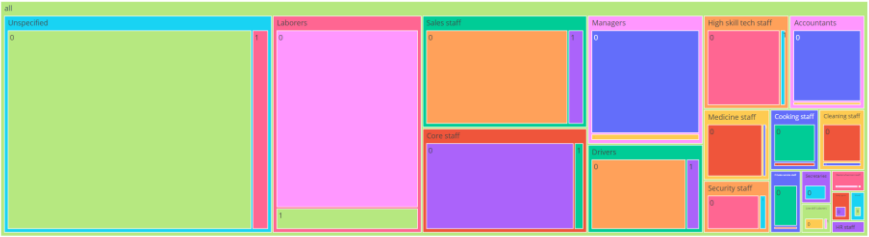


Figure 12 Tree-map by occupation, target

5. Modelling

5.1 Data preparation

5.1.1 Feature Engineering

The whole dataset is split into train and test set. In the training set, the amount of positive and negative samples is 3606 (8.12%) and 40788 (91.8%) respectively. Ordinal encoder was applied on the categorical variables to produce numerical input of our feature selection model. Recursive feature elimination with cross-validation (RFECV) was applied to select features. RFECV is to select features by recursively considering smaller sets of features based on the performance of the embedded machine learning algorithm (estimator). The estimator of RFECV was decision tree. The step of RFECV was set to be 1 which means that only one feature would be removed at each iteration. The fold of cross-validation was 5. In each cross-validation fold, the positive and negative samples were evenly shuffled to the training and testing set. Figure below shows the result of RFECV. The y coordinate is the average AUC from 5-fold cross-validation. The x coordinate is the number of features. According to the figure and text, we can note that the curve achieves its peak when there are 101 features.

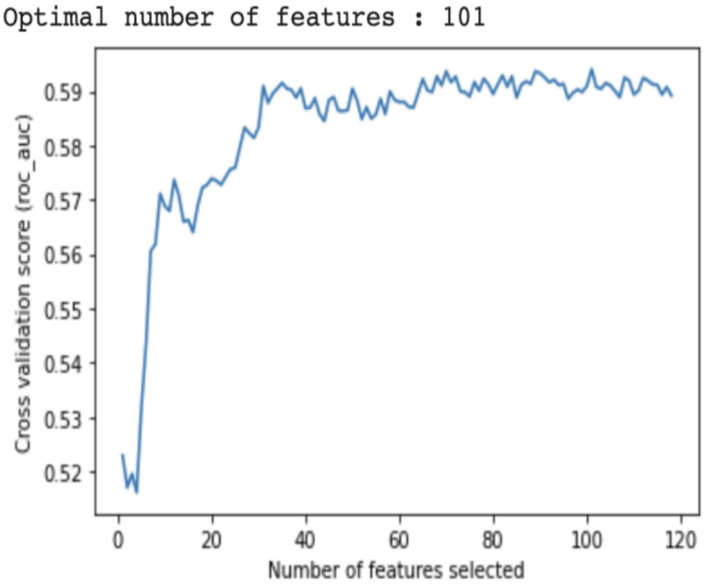


Figure 13 Determining the optimal number of features

5.1.2 Data Sampling

We have an imbalanced dataset. In such a case, we can simply get a pretty high training accuracy by considering all our samples as negatives samples. Additionally, an imbalanced dataset would easily cause the machine learning algorithm to be more biased to the majority class during the training process, which consequently leads to the low sensitivity of the minority class.

Data under-sampling and over-sampling are therefore designed to mitigate the influence of an imbalanced dataset. In this experiment, over-sampling was performed. We applied SMOTE on the training dataset which creates synthetic positive samples according to the mid-way between two near neighbors. Since our training set contains mixed categorical and continuous value, we performed a special version of SMOTE which is also called SMOTE for Nominal and Continuous (SMOTENC) so that the most common category of nearest neighbors is assigned to the newly generated sample. The Imblearn package in python was used to perform SMOTENC on the training dataset. After over-sampling, over-sampling was performed on the negative set. Two over-sampling method was performed in sequence which are Tomek links and Neighborhood Cleaning rules. After calculation, we finally obtained a training set with 20394 positive and 29184 negative samples. The new training set was then further used for the next step.

5.1.3 Dummy Variables

In this experiment, we only use tree based models which are random forest and XGBoost. Therefore, we use the “typical” (no “drop first”) one-hot encoding to create dummy variables for the categorical variables. Unlike logistic regression and neural network, tree-based models do not have an intercept term or bias term for the inputs. Every level of every categorical variable has to be involved in the selection process that selects the maximum increase in purity at every split. Regardless of encoding, dummy variables belonging to the same category should always be considered together and not in isolation.

Our final dataset for decision tree, random forest and XGBoost contains 196 attributes with 102 continuous attributes and 94 dummy attributes.

5.2 Model Training

5.2.1 Random forest

A randomized grid search is used to find the best hyperparameter combination of random forest. Grid search is run 100 times randomly out of multiple groups. The one with the best area under the roc curve (auc) is selected as the final hyperparameter for the random forest. The final hyperparameter of Random Forest is 100 estimators with 19 maximum depth. The minimum sample split is 5 and minimum sample leaf is 1. As for the result, Random forest achieves an auc of 0.71 with 40.39% recall according to 0.3 threshold.

5.2.2 XGBoost

A randomized grid search is used to find the best hyperparameter combination of random forest. Grid search is run 100 times randomly out of multiple groups. The one with the best auc is selected as the final hyperparameter for the random forest. The final hyperparameter of XGBoost is 100 estimators with 19 maximum depth. The subsample rate is 0.9 and colsample_bytree is 0.6. Additionally, gamma is 0.2 with learning rate as 0.01. As for the result, XGBoost has an auc of 0.73 on the test set with 65.35% recall based on 0.3 threshold.

Below is importance bar chart of top 10 variables in XGBoost model:

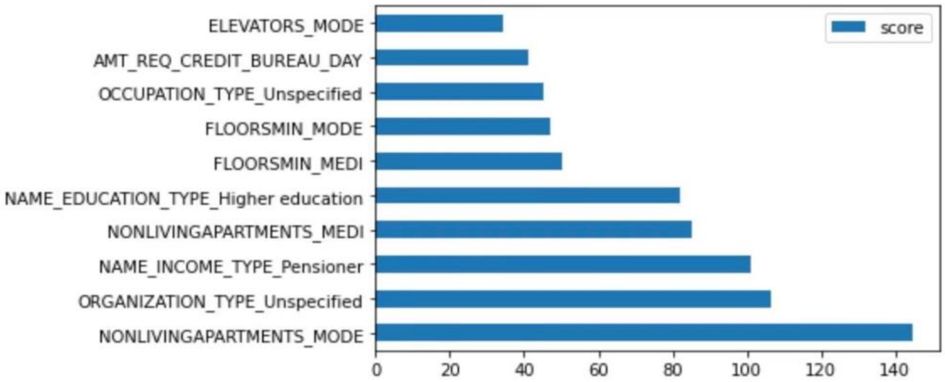


Figure 14 The importance ranking by XGBoost model

5.3 Model Comparison

Both Random Forest and XGBoost are from the decision tree family and random forest method can be regarded as the ensemble version of the decision trees. Compared to decision trees, it is more unlikely to be overfitting and robust to outliers, but still do not perfectly adapt to unbalanced data. XGBoost is a method combined boosting feature in GBDT and bagging feature in Random Forest, which make it theoretically better than both of them. In our case both accuracy and recall of XGBoost are higher than those of Random Forest, which is consistent with the theory.

6. Conclusion & Future Study

After implementing models to analysis and predict the home credit default risk, some conclusions and future improvement could be generated as below:

1) House related variables have equal level of importance than personal related variables.

Although credit default risk is more related to personal related features such as balance, income and so on, as we can see in the importance chart above, The most important variable for deciding credit risk is the square of non living apartment, which is related to the house. Besides that, floor and elevator are also variables that have an important impact on credit risk.

2) Instead of income, stability of income has more impact on credit risk.

We could say that the more income a person has, the more credit risk he or she could get from the bank. But it seems that the stability of his work, in other words, his income type, has a more important impact. We could see the variable ‘ORGANIZATION_TYPE’ in top 10 bar chart, but variables related to income amount such as all kinds of ‘BALANCE’, do not even appear in the chart. Compared to the absolute amount, it seems that the rating organization put more emphasis on the stability of income.

In this research we aim to analysis the features which are significant to decide the credit default risk of a person, and then build models to make a prediction for the risk. After data cleaning and modelling, conclusion has been found that features related to house have the same significance compared to those personal feathers, and that education, income, stability of income are decisive personal features, among which the stability of income is the most important one. However, there are some points needed to be improved and further studied:

1) Highly imbalanced data

Since the data set is highly imbalanced that we could easily get a high prediction accuracy by predicting all of them as negative, we have to implement SMOTE method to generate some synthetic samples which can rebalance the whole dataset. However, SMOTE is a method based on KNN, which means that the created sample is highly related to the original data and thus bring some unnecessary bias to the whole data set. It is also hard to select a best K value to implement the SMOTE method. Although it is definitely reasonable for us to use SMOTE in this project, we need to implement other rebalance approach to make a proper complement.

2) Correlation

Although both our models are decision trees that do not have assumptions related to correlations among variables, it is a waste to include variables which are correlated with each other. From Figure 1 we could see that many independent variables have a significant correlation, and we did not do any further cleaning because it has no impact on the quality of our model. But from the perspective of perfection, we need to put only part of those variables in our model.

Reference

[1] Singh S, Myers P, McKeown W, et al. Literature review on personal credit and debt in Australia[J]. Families at Risk Deciding on Personal Debt. RMIT University, 2005.
[2] Zhang X. Essays in credit risk management[D]. University of Glasgow, 2017.
[3] Özdemir Ö, Boran L. An empirical investigation on consumer credit default risk[R]. Discussion Paper, 2004.