Final Project for MATH 5470: Home Credit Default Risk

Hui, Chun (12284238)[*]
Department of ISOM

Liu, Shiyi (20829438) [†]
Department of BSBE

Abstract

Home Credit, a company who strives to broaden financial inclusion for the unbanked population by providing a positive and save borrowing experience, wants to make sure these underserved population has a positive loan experience. Hence, Kagglers are supposed to make use of a variety of alternative data to predict clients' repayment abilities.

In this report, we share our experiences and thinkings during the project. We mainly use Logistics Regression, LightBGM model and Decision Tree to give predictions. Correlation Screening and Principal Component Analysis (PCA) is used in the feature selection. Our final grade is 0.739. This report will be written in five parts: Introduction, Data, Feature Construction, Model, Feature Importance Analysis and Future Improvents & Discussions.

1 Introduction

As lenders, it's essential to evaluate the clients' ability of paying loans. For example, clients are usually required to fill in a questionaire for lender companies before a deal. The quesions may include incomes, loans, employment status, family members and so on. Our problem is to use these large amounts of data to give a prediction for the ability of paying. We care about the model performance and feature importance. During model evaluation, we use AUC score and confusion matrix. In feature importance analysis, we compare the selected features from different models and give our understandings.

2 Data

2.1 Data Description

In this competition, Home Credit provides 10 csv files with size 2.68GB in total. Apart from the file `HomeCredit_columns_description.csv` and `sample_submission.csv` which are explanatory documents, the remaining files include variables that related to many status of the clients. We use two forms: `application_train.csv` and `application_test.csv` involved in the analysis. That is because they have about 120 features and make up more than half of the features in file `HomeCredit_columns_description.csv`, which is adequate for attempting a machine learning model.

The raw data is as shown in figure 1. Application train data has 307 thousand samples and application test data has nearly 50 thousand samples, they both have 121 features. (training set has *TAREGT*). From the results of statistics, there are various types of data, including numerical values, integers and categories. Different types of data need different methods to deal with.

[*]Coding(Python), Report Writing and LaTax Formatting
[†]Coding(R), Slides Making and Presentation

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CRED
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000
5	100008	0	Cash loans	M	N	Y	0	99000.0	490495
6	100009	0	Cash loans	F	Y	Y	1	171000.0	1560726
7	100010	0	Cash loans	M	Y	Y	0	360000.0	1530000
8	100011	0	Cash loans	F	N	Y	0	112500.0	1019610
9	100012	0	Revolving loans	M	N	Y	0	135000.0	405000

Figure 1: The raw data.

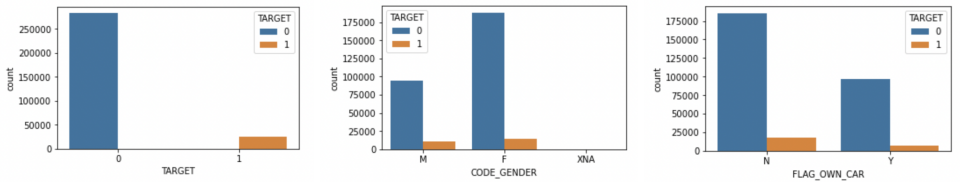

Figure 2: Distributions of the *TAREGT* label with respect to different variables.

Tentatively, we plot some distributions of the *TAREGT* label with respect to different variables. The selected examples are as shown in figure 2. The target labels are very imbalanced. Just 8% targets value 1, which is shown in the left side figure. In the middle figure, the percentage of *TAREGT=1* in males is 10.1% and in females is 7.0%, which means men have more percentage of people having difficults paying. Gender has somehow effects on the targets. Finally, in the right side figure, the percentage of *TAREGT=1* in clients owning cars is 7.2% and in clients not owning cars is 8.5%. The difference is slight, so intuitively thinking whether owning cars may not have much influence on the targets. It's consistent with the common sense that the client who owns a car is basically richer than who doesn't own cars, so the latter ones are easier to have payment difficulties.

2.2 Data Processing

2.2.1 Method 1

• **Missing Values:** There are 67 and 64 features that have missing values in application train and test set, respectively. About half of the features are incomplete. The missing percentages range from 0 to 70%.

Firstly, we delete the features with missing percentage larger than 60%. Secondly, for data of numerical type, including *float* and *int*, and data of *object* type, we use median and mode to replace missing values, respectively.

• **Dummy Variables:** For categorial data, in order to make them involved in the model, we count the number of unique categaries for each feature. For feature has two categaries, we use 0-1 label to enlabel them. For the number of catogaries larger than 2, use dummy variables to divide one feature to various dummy features.

• **Outliers:** Because there are features with extremely large values, it's common to have outliers. The idea is to drop the outliers using proper threshold, so that training and testing series are consistent in distribution. There is an examption that *'DAYS_EMPLOYED'* has erroneous values, which are replaced by median.

• **MinMaxScaler:** We use MinMaxScaler tool to normalized data, so that they could be better involved in the model.

After processing and taking intersection of two sets with respect to features, we have 218 predictors.

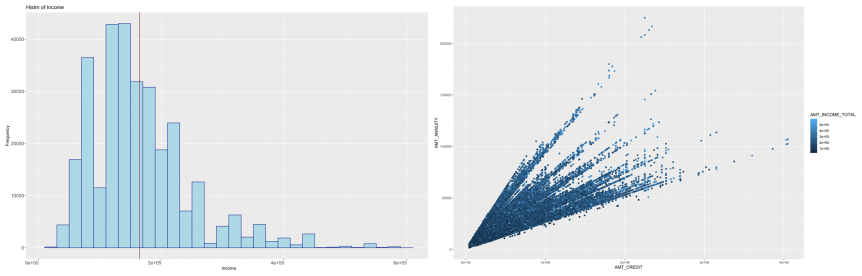

Figure 3: Data visualization

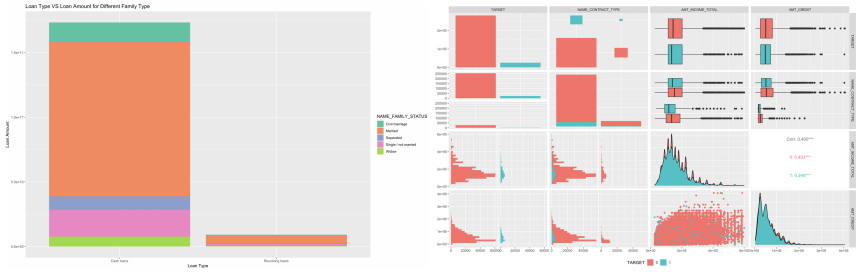

Figure 4: Data visualization

2.2.2 Method 2

The variable can be roughly divided into several parts of features: personal information, loan information, documents provided, and information from the Bureau. Firstly, to reduce time and overfitting, we delete some variables that have almost no relation to the prediction, like the house decoration and the mode of the entrance of the house. Then, to get more clear and essential observations for better training performance, some NA values, and some vague descriptions that neither fit in any valuable category of one variable, for example, the 'XNA' value in the gender column. Besides, to better quantify the data, some categorical variables are transformed into factors or numeric. Some categorical variables are recoded into binary values. Visualization of data is also used to inspect some interesting differences between the default and no default cases. The histogram in figure 3 indicates that the income of clients concentrates in the middle, a great example of the central limit theorem. The dot plot figure 3 reveals a positive relationship between the amount of loan and loan annuity In the stack bar plot in figure 4, cash loans comprise a big part and married people seem to be more likely to get loans approved. The correlation matrix plot in figure 4 provides abundant information on the relationship between different variables, the most obvious trace is that people with more income are more capable of loan payments.

3 Feature Construction

3.1 Add Ratio Features

With the help of references, we add five ratios that may contribute to the predication of labels. For example, *'AMT_CREDITL'* is the credit amount of the loan, and *'AMT_INCOME_TOTAL'* is the income of the client. The ratio of these two is similar to the debt-to-income ratio (DTI), which could help to determine if one can afford to repay a loan. Generally, clients with higher DTI ratios are riskier borrowers, because they might have trouble repaying their loan in case of financial hardship. After adding ratio features, the total number of features is 222.

3.2 Feature Selection

More features mean larger complexity of the model. We attempt to use correlation screening and Principal Component Analysis (PCA) to make feature selection.

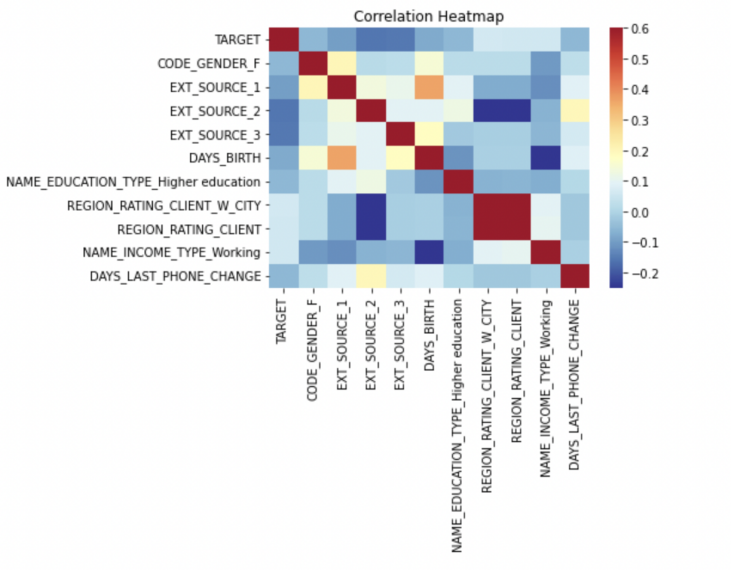

Figure 5: heatmap of the correlations between the selected top 10 features and *TARGET*.

In the simplist way, we calculate correlations between *TARGET* and every other feature. Then select one hundred features with largest absolute value of correlations. Figure 5 is the heatmap of the correlations between the selected top 10 features and *TARGET*. We finally abandon this method because it doesn't outperform the model using all the features. There are also advanced correlation-based screening methods in the literature and it will be an improving direction.

We also use PCA to selected features used in Decision Tree model, and there are 77 features to be selected.

4 Model

We split the preprocessed application train data to training set and validation set with ratio 4:1. We use three models: Logistic Regression, LightGBM and Decision Tree. For binary classification problem, the evaluation stardards we use are AUC score, specificity (True Negative Rate, TNR) and sentivity (True Positive Rate, TPR). Because the label distribution in testing data is imbalanced as well, we especially care about the value of sentivity, which is the probability of a positive test, conditioned on truly being positive.

4.1 Decision Tree

We use Decision Tree to be a baseline, and the AUC score is about 0.65. Then we use Logistic Regression, LightGBM, and LightGBM with focal loss to make improvement.

4.2 Logistic Regression

For logistic regression method, we first directly use the default prediction function, which means the threshold equals 0.5. However, the classfication performance is not good. We notice most predictions are predicting *TARGET* to be 0. This may be because the distribution for the training labels is imbalanced, so machine tends to give more 0 target to get a higher accuracy. Because of this, we reset the classifier threshold to be smaller so that we could make a tradeoff between specificity and sentivity. This idea is used in LightGBM as well. Figure 6 shows the confusion matrics of original and aftering adjusting thresholds and the AUC curve. The sentivity increases a lot from nearly 0 to 0.6, and specificity decreases from 1 to 0.77.

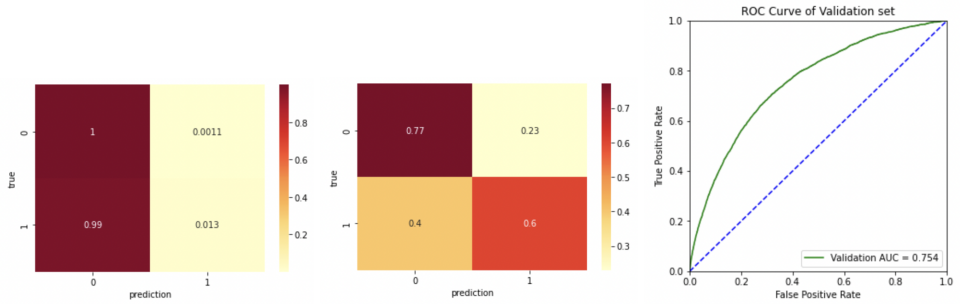

Figure 6: Confusion matrics and AUC curve of logistic regression.

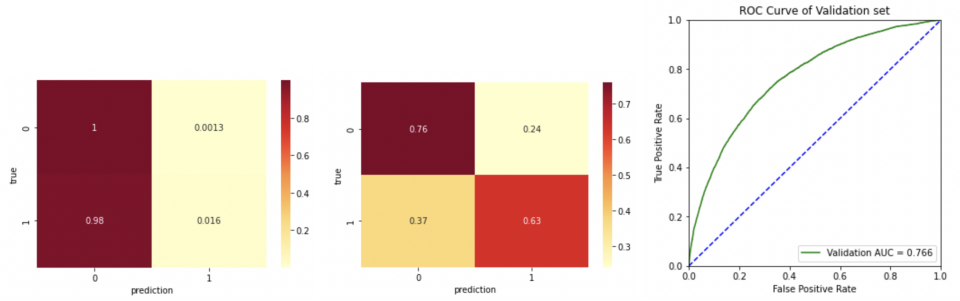

Figure 7: Confusion matrics and AUC curve of LightGBM.

4.3 LightGBM

For LightGBM, we set the number of literations is 100 and learning rate is 0.12. The original sentivity is very low as well. So we reset the threshold to 0.1 to make a tradeoff. The specificity is 0.76 and sentivity is 0.63. The AUC score is higher than logistic regression, which increases from 0.754 to 0.766, see Figure 7

We also try to use LightGBM with focal loss. The parameters α and γ are set as 0.96 and 0.38. The original sentivity sharply increases to 0.95 and specificity decreases to 0.26. So we reset threshold to 0.695. Results show that the sentivity is 0.64, specificity is 0.76 and AUC score improve to 0.77, see Figure 8

We summary the AUC score, specificity and sentivity of different models in the table 1below.

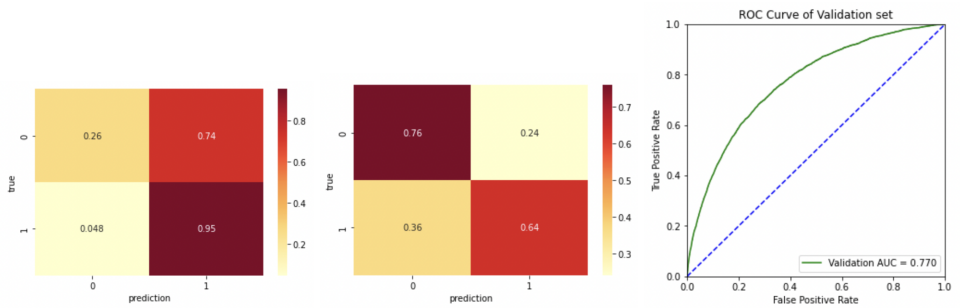

Figure 8: Confusion matrics and AUC curve of LightGBM.

Table 1: Classification performances of the four models.

Model	AUC score	Specificity	Sentivity
Logistic Regression	0.754	0.77	0.60
LightGBM	0.766	0.76	0.63
LightGBM with Focal Loss	0.770	0.76	0.64

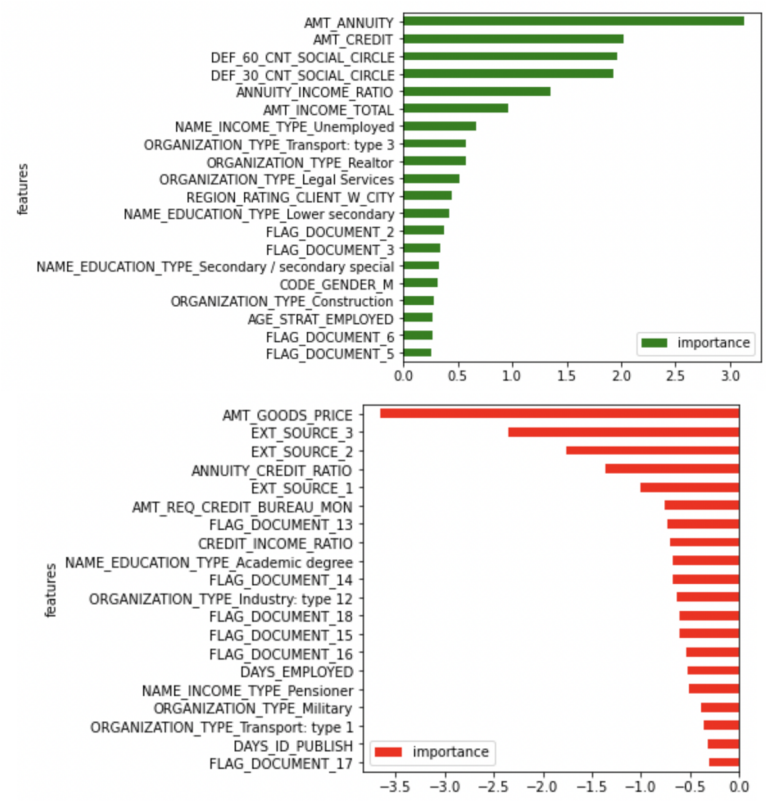

Figure 9: Top 20 features ranked by coefficients selected from Logistic Regression. The green lines represent positive and red lines represent negative.

5 Feature Importance Analysis

For logistics regression, we sort the fitted coefficients and select two sets. Figure 9 shows the top 20 features with largest positive coefficients and smallest negative coefficients. For example, *AMT_ANNUITY*, *AMT_CREDIT*, *AMT_INCOME_RATIO* and *AMT_INCOME* rank highly in positive features. That is consistent to the common sense, because the loan annuity, credit amount of the loan and income are directly related to the ability of paying loans.

However, we couldn't say with certainty that these features all contribute a lot to the prediction of labels. We need to refer to the p-value. For example, the coefficient of *AMT_GOODS_PRICE* is -3.648, and p-value is 0, which is smaller than 0.05. Therefore, we should reject the null hypothesis that $H_0 = 0$. That is, *AMT_GOODS_PRICE* is an important feature for prediction. The coefficient of *DEF_60_CNT_SOCIAL_CIRCLE* is 1.960, and p-value is 0.3084, which is larger than 0.05. Therefore, we should retain the null hypothesis. That is, *DEF_60_CNT_SOCIAL_CIRCLE* is not an importance feature, although it ranks 2nd in the positive coefficient.

For LightGBM and LightGBM with focal loss, the models automatically select the important features. We rank them by importance and plot in Figure 10. The top 5 features are the same with slightly different ranking. They are *EXT_SOURCE_1*, *EXT_SOURCE_2*, *EXT_SOURCE_3*,

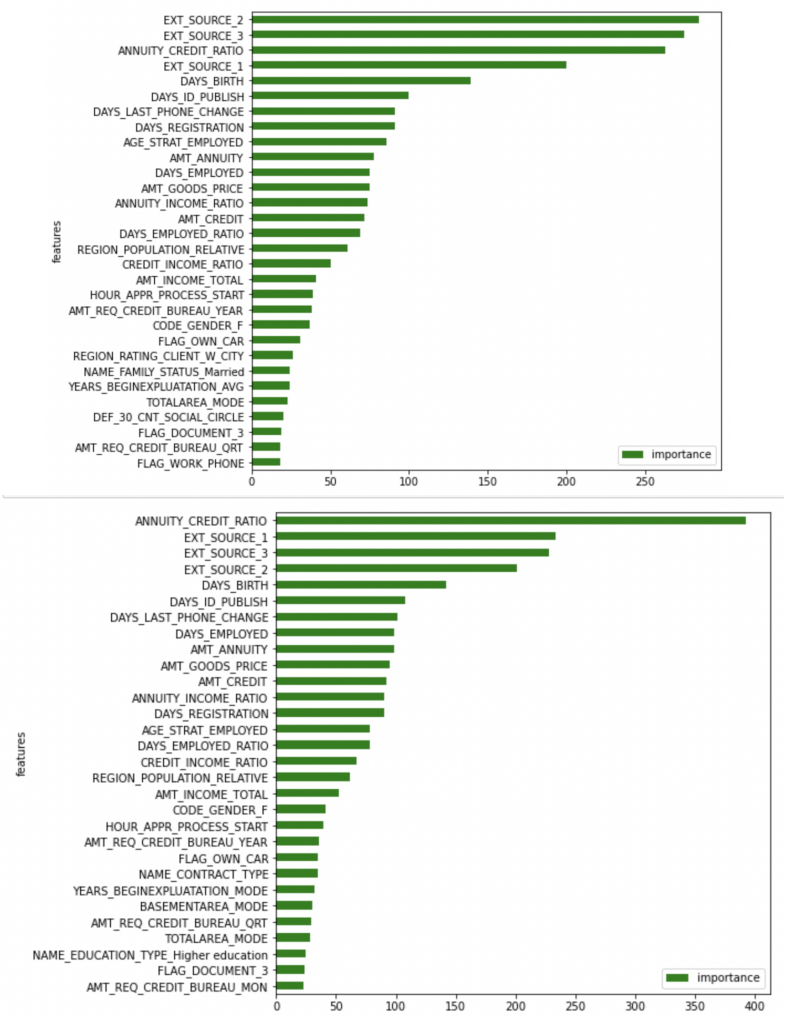

Figure 10: Important features selected from LightGBM and LightGBM with focal loss.

ANNUITY_CREDIT_RATIO and *DAYS_BIRTH*, in which 4 features are ranked high in the correlation with *TARGET*, see Figure 5.

6 Future Improvements and Discussions

In this project, we malnly use three models to give predictions. Our validation AUC score is around 0.77 but the testing score is around 0.74, which is relatively lower than 0.77. The followings are the future improvements and our discussions.

First, in the preprocessing data part, because we only use two files, there also other features could be added. More features may have a better prediction. Second, in the part of modelling, the parameters cound be tuned by cross validation or other tuning techniques, which may improve the performance. Third, for feature selection, there are many advanced statistical methods in literature. A proper feature selection method could be a bonus to the prediction. In the feature selection part, the more robust and powerful mathematical methods will save much time with possibly better results since humans are biased, as they sometimes omit some important traces or give some useless features too much attention. The universality of the model for this type of data can be limited, so how to make the model more practical and used by other countries or races, really trigger my interest, and I am willing to dig deeper about this topic.

References

[1] Focal Loss for Dense Object Detection. Lin et al. 2018.
available: https://arxiv.org/abs/1708.02002v2

[2] Kaggle Competition In 30 Minutes: Predict Home Credit Default Risk With R.
available:https://www.r-bloggers.com/2018/08/kaggle-competition-in-30-minutes-predict-home-credit-default-risk-with-r/

[3] Home Credit: Predicting the Default Risk. available: https://rpubs.com/yatingdeng/640306

[4] Wikipedia-PCA. available: https://en.wikipedia.org/wiki/Principal_component_analysis