# The model of home credit using logistic regression

**MATH6010Z Project 1**
**LAI Cong, LIU Jinghui, LU Qiaoyu, ZHEN Mengnan**
Department of Financial Mathematics
Hong Kong University of Science and Technology

## Abstract

The report introduce the model we use to predict Home Credit Default Risk, which is based on logistic regression algorithm and data preprocessing methods. The prediction results and the pros and cons of the model, as well as the improvement measures will be discussed later.

## 1.Introduction

Loans are very common in people’s lives. Everyone has different reasons for borrowing. It could be buying a dream car or house, for business capital flow, or buying some other things. Even wealthy people prefer to borrow loans to get tax breaks, it can also help them to keep cash available against unexpected and unconventional expenses in the future.

Loans are as important to lenders as they are to borrowers. Almost all banking institutions get most of their income from interest earned on loans. [1] However, lenders face the daunting task that identify the risks associated with each customer. Therefore, it is important to identify the risk of loans and make informed decisions.

Home Credit is a financial institution that creates lines of credit for people who do not have bank accounts. Unlike other credit agencies, Home Credit's main challenge in predicting loan defaults is the lack of a credit score. Thus, Home Credit's main target is the unbanked or unbanked population, who have very limited credit histories. The objective of this project is to establish an effective and efficient classification model to predict the loan repayment ability of applicants and reduce the credit default risk of Home Credit.

In the project, we have a look at the problem statement, start off with those insights gained from the Logistic regression using the dataset Application (train/test). We also try to construct new features to better explain the model of Home Credit. Finally, analyzing the results and drew conclusions to help make better predictions on the risk of loans.

# 2.Datasets & method

## 2.1 Dateset

The dataset was provided by Home Credit Group's data scientists. It contains a wide variety of personal and financial information belonging to 356,255 individuals who had previously been recipients of loans from Home Credit. These individuals are divided into training and testing sets. The training data contains 307,511 records and the test data contains 48,744 records.

The training data contains an extensive and diverse array of personal and financial information for each of its applicants. These features include common characteristics, such as marital status, age, type of housing, job type, education level.

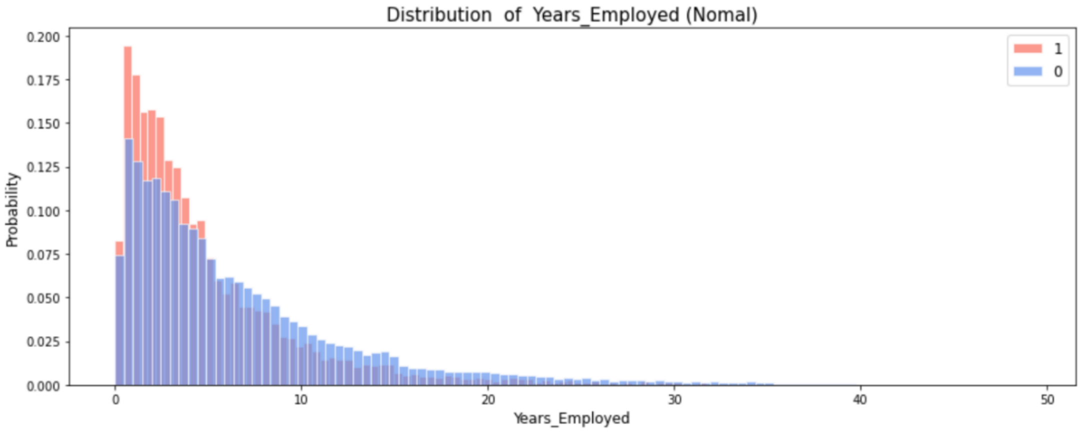
Home Credit also looks at aspects of applicants' financial backgrounds, including monthly payment on any loans or credit card balances that the applicant has previously had with Home Credit, as well as the amount and monthly repayment balances of any loans that the applicant may have received from other lenders. All of these features are spread across seven data tables.

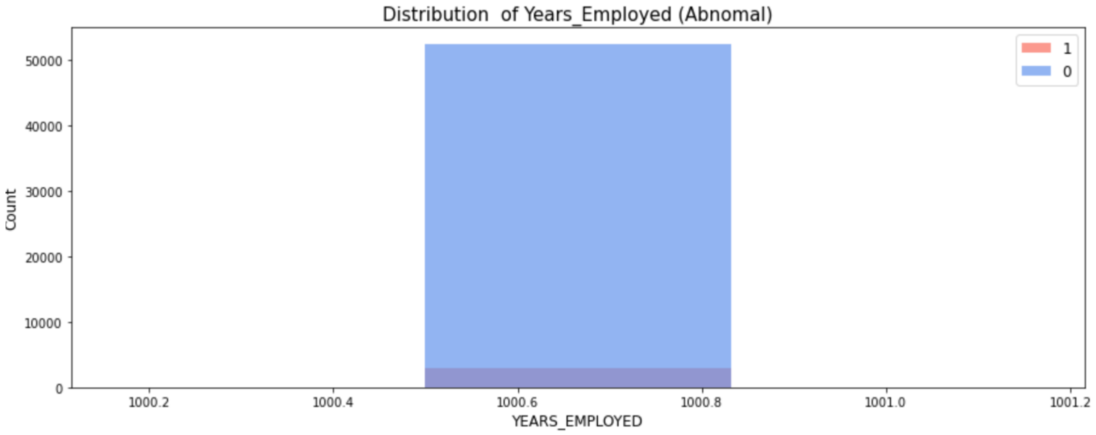
The main data table ('application_train.csv') contains 120 features that comprise applicants' personal background information. The other six data tables contain applicants' previous loan and credit card balance payment histories.

We first check the missing values for next analysis. As shown in the form, many variables have more 50% NANs. Though larger proportion of missing values should be more appention. However, considering the real situation, when a person fill the long application form, if these information was really crucial, then bank officer certainly will not let them be blank. Therefore, we tend to think less about these features.

|  | Missing Values | % of Total Values |
|---|---:|---:|
| COMMONAREA_MEDI | 214865 | 69.9 |
| COMMONAREA_AVG | 214865 | 69.9 |
| COMMONAREA_MODE | 214865 | 69.9 |
| NONLIVINGAPARTMENTS_MEDI | 213514 | 69.4 |
| NONLIVINGAPARTMENTS_MODE | 213514 | 69.4 |
| NONLIVINGAPARTMENTS_AVG | 213514 | 69.4 |
| FONDKAPREMONT_MODE | 210295 | 68.4 |
| LIVINGAPARTMENTS_MODE | 210199 | 68.4 |
| LIVINGAPARTMENTS_MEDI | 210199 | 68.4 |
| LIVINGAPARTMENTS_AVG | 210199 | 68.4 |
| FLOORSMIN_MODE | 208642 | 67.8 |
| FLOORSMIN_MEDI | 208642 | 67.8 |
| FLOORSMIN_AVG | 208642 | 67.8 |
| YEARS_BUILD_MODE | 204488 | 66.5 |
| YEARS_BUILD_MEDI | 204488 | 66.5 |
| YEARS_BUILD_AVG | 204488 | 66.5 |
| OWN_CAR_AGE | 202929 | 66 |
| LANDAREA_AVG | 182590 | 59.4 |
| LANDAREA_MEDI | 182590 | 59.4 |
| LANDAREA_MODE | 182590 | 59.4 |
| BASEMENTAREA_MEDI | 179943 | 58.5 |
| BASEMENTAREA_AVG | 179943 | 58.5 |
| BASEMENTAREA_MODE | 179943 | 58.5 |
| EXT_SOURCE_1 | 173378 | 56.4 |
| NONLIVINGAREA_MEDI | 169682 | 55.2 |
| NONLIVINGAREA_MODE | 169682 | 55.2 |
| NONLIVINGAREA_AVG | 169682 | 55.2 |
| ELEVATORS_MEDI | 163891 | 53.3 |
| ELEVATORS_MODE | 163891 | 53.3 |
| ELEVATORS_AVG | 163891 | 53.3 |
| WALLSMATERIAL_MODE | 156341 | 50.8 |
| APARTMENTS_MODE | 156061 | 50.7 |
| APARTMENTS_MEDI | 156061 | 50.7 |
| APARTMENTS_AVG | 156061 | 50.7 |
| ENTRANCES_MODE | 154828 | 50.3 |
| ENTRANCES_AVG | 154828 | 50.3 |
| ENTRANCES_MEDI | 154828 | 50.3 |
| LIVINGAREA_MEDI | 154350 | 50.2 |
| LIVINGAREA_MODE | 154350 | 50.2 |
| LIVINGAREA_AVG | 154350 | 50.2 |
| HOUSETYPE_MODE | 154297 | 50.2 |
| FLOORSMAX_MEDI | 153020 | 49.8 |
| FLOORSMAX_AVG | 153020 | 49.8 |
| FLOORSMAX_MODE | 153020 | 49.8 |
| YEARS_BEGINEXPLUATATION_AVG | 150007 | 48.8 |
| YEARS_BEGINEXPLUATATION_MEDI | 150007 | 48.8 |
| YEARS_BEGINEXPLUATATION_MODI | 150007 | 48.8 |
| TOTALAREA_MODE | 148431 | 48.3 |
| EMERGENCYSTATE_MODE | 145755 | 47.4 |
| OCCUPATION_TYPE | 96391 | 31.3 |
| EXT_SOURCE_3 | 60965 | 19.8 |
| AMT_REQ_CREDIT_BUREAU_WEEK | 41519 | 13.5 |
| AMT_REQ_CREDIT_BUREAU_DAY | 41519 | 13.5 |
| AMT_REQ_CREDIT_BUREAU_MON | 41519 | 13.5 |
| AMT_REQ_CREDIT_BUREAU_QRT | 41519 | 13.5 |
| AMT_REQ_CREDIT_BUREAU_HOUR | 41519 | 13.5 |
| AMT_REQ_CREDIT_BUREAU_YEAR | 41519 | 13.5 |
| NAME_TYPE_SUITE | 1292 | 0.4 |
| DEF_30_CNT_SOCIAL_CIRCLE | 1021 | 0.3 |
| OBS_60_CNT_SOCIAL_CIRCLE | 1021 | 0.3 |
| DEF_60_CNT_SOCIAL_CIRCLE | 1021 | 0.3 |
| OBS_30_CNT_SOCIAL_CIRCLE | 1021 | 0.3 |
| EXT_SOURCE_2 | 660 | 0.2 |
| AMT_GOODS_PRICE | 278 | 0.1 |
| AMT_ANNUITY | 12 | 0 |
| CNT_FAM_MEMBERS | 2 | 0 |
| DAYS_LAST_PHONE_CHANGE | 1 | 0 |

Second, we found extreme values in the years of employed.

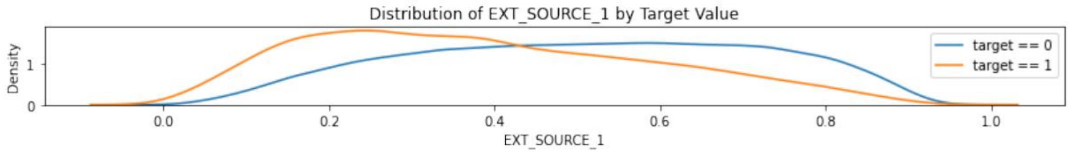


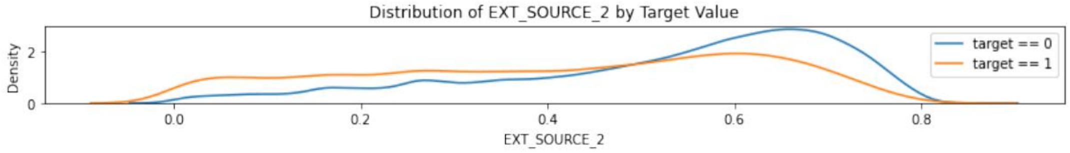


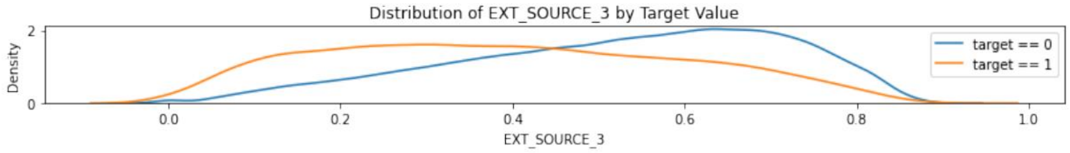
As can be seen from the figure, the shorter the years of employment, the higher the overdue rate. The shorter the working years of loan users (0-5 years), the more likely they are to be overdue. With the increase of working years, the accumulated capital tends to be stable, and the overdue rate of loan users decreases significantly.

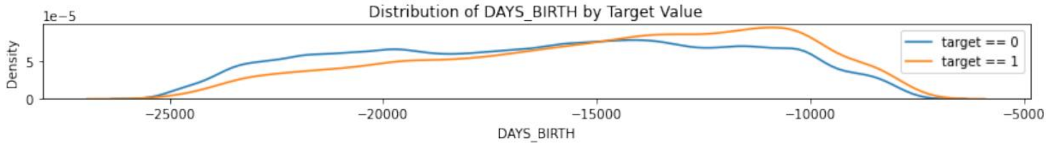
According to the separate analysis of the outlier (*DAYS_EMPLOYED* = 365243), the probability of overdue loan users in this outlier is very low. So we deleted them away.

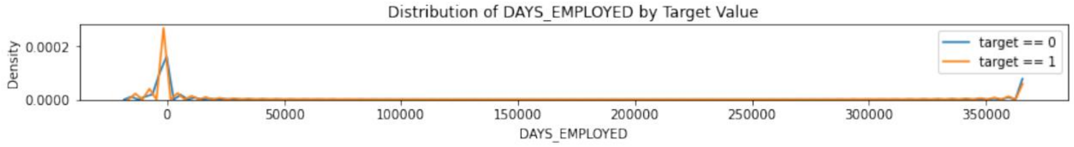
Third, we look into the main data table and found some of them show potential possibility to be significant features. Because they appear different distributions of default and nondefault records, we think they may play an important in classification. In this report, we only show the plots of considered features.

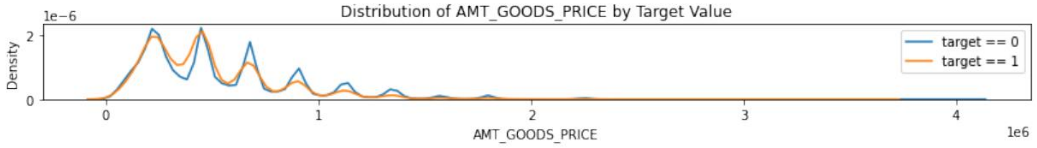

Distribution of EXT_SOURCE_1 by Target Value
Density
0
1
0.0
0.2
0.4
0.6
0.8
1.0
EXT_SOURCE_1
target == 0
target == 1

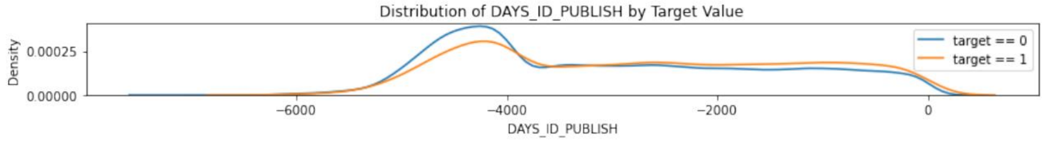

Distribution of EXT_SOURCE_2 by Target Value
Density
0
2
0.0
0.2
0.4
0.6
0.8
EXT_SOURCE_2
target == 0
target == 1

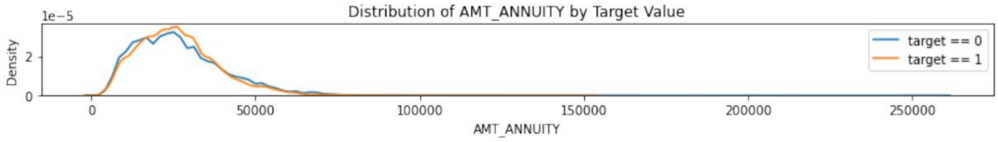

Distribution of EXT_SOURCE_3 by Target Value
Density
0
2
0.0
0.2
0.4
0.6
0.8
1.0
EXT_SOURCE_3
target == 0
target == 1

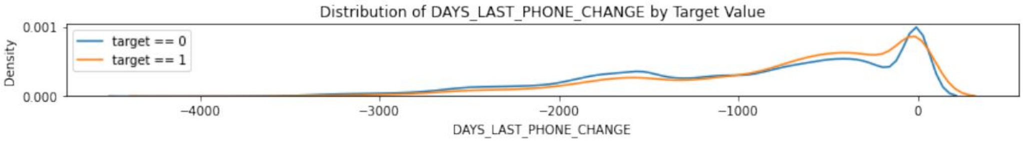

Distribution of DAYS_BIRTH by Target Value
1e−5
Density
0
5
−25000
−20000
−15000
−10000
−5000
DAYS_BIRTH
target == 0
target == 1

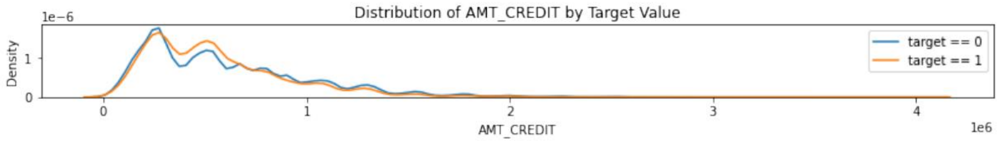

Distribution of DAYS_EMPLOYED by Target Value
Density
0.0000
0.0002
0
50000
100000
150000
200000
250000
300000
350000
DAYS_EMPLOYED
target == 0
target == 1

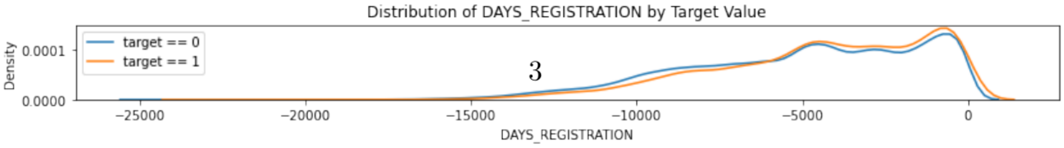

Distribution of AMT_GOODS_PRICE by Target Value
1e−6
Density
0
2
0
1
2
3
4
1e6
AMT_GOODS_PRICE
target == 0
target == 1


Distribution of DAYS_ID_PUBLISH by Target Value
Density
0.00000
0.00025
−6000
−4000
−2000
0
DAYS_ID_PUBLISH
target == 0
target == 1


Distribution of AMT_ANNUITY by Target Value
1e−5
Density
0
2
0
50000
100000
150000
200000
250000
AMT_ANNUITY
target == 0
target == 1


Distribution of DAYS_LAST_PHONE_CHANGE by Target Value
Density
0.000
0.001
−4000
−3000
−2000
−1000
0
DAYS_LAST_PHONE_CHANGE
target == 0
target == 1


Distribution of AMT_CREDIT by Target Value
1e−6
Density
0
1
0
1
2
3
4
1e6
AMT_CREDIT
target == 0
target == 1



### 2.2 method

Consider again the training data set, where the response target falls into one of two categories, 1 or 0 . Rather than modeling this response Y directly, logistic regression models the probability that Y belongs to a particular category. We consider the linear relationship between multiple explanatory variables and log odds.

The formula of logistic regression shown as below:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

and

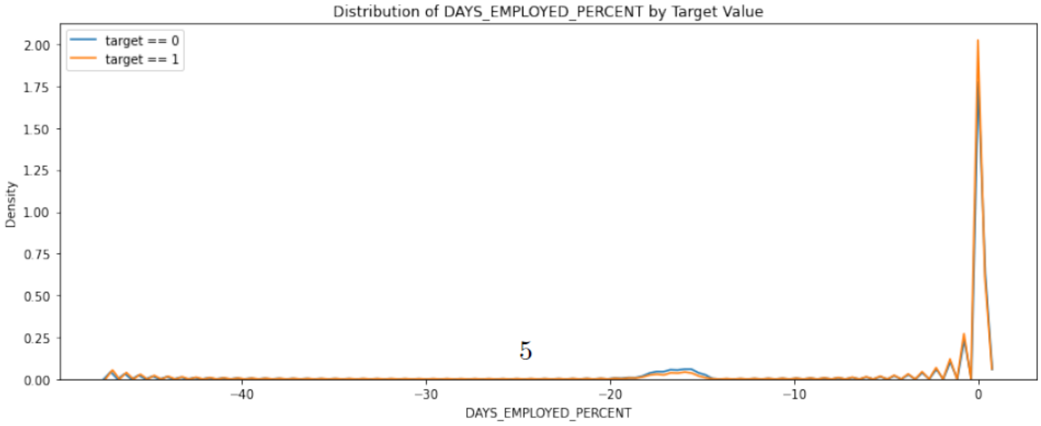
$$p = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}}$$

where usually *b*=*e*.

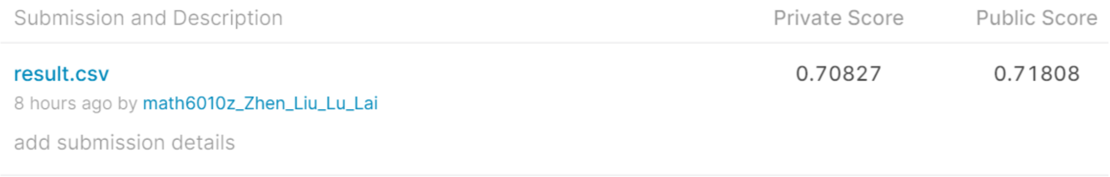
## 3. Prediction result

Initially, in order to intuitively judge whether the logistic regression algorithm is suitable for this data set, we try to do data preprocessing and analyze some feature variables with fewer missing values separately, however, the result is not ideal, which just get a prediction score of 0.55-0.66.

Therefore, we judge from the distribution of feature variables combined with common sense, as well as the pros and cons of logistic regression, and infer that some feature variables may have non-linear connections. Also, it is easily found that some feature variables still have some outliers. That may result in a serious phenomenon of multicollinearity. In order to eliminate the effects of multicollinearity between feature variables of the model as far as possible, we decide to discover relationships between feature variables in the existing model and construct some new features. For instance, the new feature *DAYS_EMPLOYED_PERCENT* explore the relevance between the default risk and someone's employed time compared to the lifetime.



Then four new variables are included in the model, after observing new feature variables' distributions. Finally, the new features perform high correlation with default rates. After repeated attempts, the model prediction gets a great improvement, which achieves a private score of 0.70827 and a public score of 0.71808.



| Submission and Description | Private Score | Public Score |
|---|---|---|
| result.csv<br>8 hours ago by math6010z_Zhen_Liu_Lu_Lai<br>add submission details | 0.70827 | 0.71808 |

The final prediction score of the model.

## 4. Analysis

There will be some discussion and analysis about the advantages and limitations of our model, as well as some possible improvements to explore in the future.

### 4.1. Advantages of the model

According to the characteristics of logistic regression algorithm, we choose suitable data cleaning method and feature construction project.  Its purpose is to minimize the impact of missing values on the logistic regression model and nonlinear variables on the logistic regression model. To further analyze the relationship between various features and default rate, our model visualizes the data and intuitively finds some rules through the graph.

The major advantages of logistic regression are concise and fast, which is easy to understand and implement. Precisely because of it, the model is quite efficient in terms of time and memory requirement, as the calculation time is just related to the number of features. Besides, with well-prepared dataset and good feature engineering, logistic regression shows a good robustness to outliers and is not particularly affected by mild multicollinearity, since logistic regression will properly adjust each weight variable for the corresponding feature variable. Hence, the model preforms an interpretable result, which well explains each feature variable’s effect towards the final prediction. This algorithm is direct enough and useful to solve financial problems.

### 4.2. Limitations of the model

There are some limitations of the model that need to be addressed. It may be easier to lead to underfit relatively, compared to Decision Tree and SVM. And the classification accuracy is not high, since it approaches to linear model and may meet some trouble when fitting the true distribution of nonlinear data. As thus, it puts forward higher and more stringent requirements for feature engineering to avoid high-dimension feature set. So we attempt to select feature variable with high correlation with the predicted results, and construct new feature variables based on this, in order to create a good prediction environment for logistic regression and improve its accuracy.

Additionally, logistic regression is extremely demanding because it is sensitive to missing values. Hence, we design a function *fillna_by_mean_of_columns()* to deal with missing values, which fill the null value by the mean of its feature. However, comparing to Light GBM model, which is able to deal with missing values automatically, our model requires another approximate method for data cleaning and balancing.

### 4.3. Future improvement measures

There are two main aspects to improve our model. The first aspect is data preprocessing, since logistic regression cannot filter feature variables by itself. Unlike logistic regression, the decision tree method is more suitable for processing nonlinear data, without the need to do data normalization. We can use Gradient Boosting Decision Tree to filter features such that logistic regression can be more efficient in this case. Besides, by Data Science, it is suggested that Sequential Forward Feature Selection could be implemented to pick up the best set of feature variables. [2]

The second aspect is classification accuracy. Some other methods can be used to improve the accuracy of the model, such as LightGBM and XGBoost. Regardless of time and memory costing, they may deal with both linear and nonlinear relationships well, such that enable the model more robust and stable.

## 5. Conclusion

In conclusion, we introduce our machine learning model used to explore Home Credit Default Risk of Kaggle, and the detailed feature engineering method and implied logistic based on logistic regression. Also, we discuss the prediction result, and further analyze the strengths and weaknesses of the model. It is known that logistic regression is fairly simple and interpretable, however, which also needs skilled and mature data preprocessing and feature engineering. Hence we propose possible improvement measures as well. In the future, we will attempt to improve the model by other machine learning algorithms to increase the prediction accuracy and minimize the default risk.

## 6. References

[1] Rao, R(2020). HOME CREDIT DEFAULT RISK — An End to End ML Case Study — PART 1: Introduction and EDA. Retrieved from https://medium.com/thecyphy/home-credit-default-risk-part-1-3bfe3c7ddd7a

[2]Data Science(2020). Credit Default Risk Prediction. Retrieved from https://www.record-evolution.de/en/credit-default-risk-prediction/

## 7. Contribution

| Coding and report: | Contribution |
|---|---|
| LAI Cong, 20747850 | 25% |
| LIU Jinghui, 20745644 | 25% |
| LU Qiaoyu, 20736916 | 25% |
| ZHEN Mengnan, 20749066 | 25% |