Midterm Project: Home Credit Report

Wenjia Xie

December 8, 2018

1. Abstract

The purpose of this project is to use data that Home Credit provided on Kaggle to find out what kinds of variables are of great importance in identifying clients who may have difficulties in repaying loan and make classification. In this project, the AIC of different models and the F1 score of classification are used to select models. As a result, the multilevel logistic regression is used to make prediction. We find out that: 1) Three different external data sources in the datasets are of vital importance in classification. 2) In terms of clients personal information, the younger the clients are and shorter they change their job before application for loan, the higher risk of failure to repay their loan. 3) Clients who are willing to update their identity document often have a lower risk of default.

2. Introduction

Home Credit is an international non-bank financial institution that serve the unbanked population by providing them a positive borrowing experience. To make sure their client can have a safe loan experience, Home Credit wants to make use of a variety of data to predict their clients' repayment abilities.

The goal of this project is to use the data that Home Credit provided on Kaggle to understand what kinds of characteristics may have strong influence on predicting clients' repayment abilities. To go further, based on these characteristics, we also want to build some models that can help Home Credit to identify clients may have payment difficulties. In this case, Home Credit may avoid some potential risk of bad loan.

3. Method

3.1 Data source

Home Credit has provided their datasets of various information on Kaggle. The total data sets include 10 files and in each file it contains some information about the client's previous credits, monthly balance, behavioral data etc. Basically, we will use the data from application {train|test}.csv to build the prediction model. It contains 307512 entries with 122 symbolic attributes. In this dataset, each entry represents a person who takes a credit by Home Credit.

3.2 Exploratory Data Analysis

3.2.1 The distribution of target

The target in the training application data indicating 0: the loan was repaid or 1: the loan was not repaid. The distribution of targets are as follows. From the plot we can see that the vast majority of the loan was repaid, which is often the case in a promising financial institution. Thus we should focus our efforts on identifying the potential unrepaid loan.

Figure 1 :The distribution of the targets

3.2.2 Personal Information in terms of loan is repaid or not

Home Credit provides a variety of data on the basic information about their clients, including their gender, age, family status, education etc. We uses these variables to group the clients and calculate the percentage of bad loans in each group. From the exploratory data analysis, we find an interesting phenomenon that compared with elder people, younger people tend to have a higher rate of failure to repay their loan:

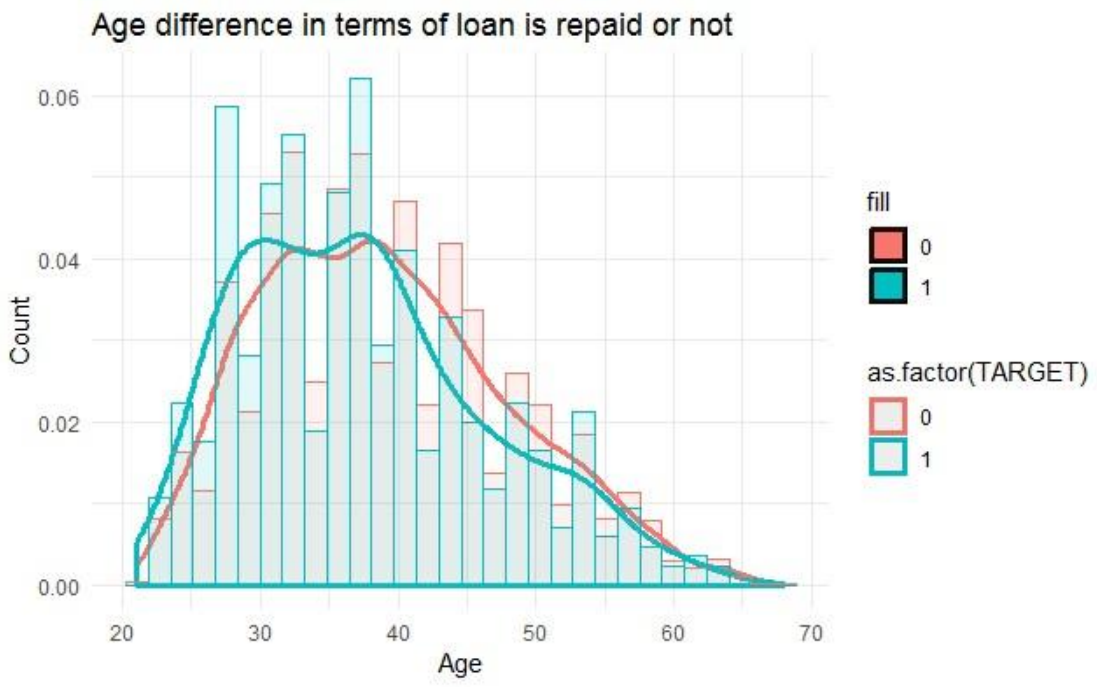

Figure 2: age difference in terms of loan is repaid or not

We can see that the youngest age group have more than 8% higher rate of failure to repay their loan than the oldest age group on average. So for the bank, maybe they can provide young people with more guidance or financial planning tips to help younger clients pay on time.

Besides the clients' age, we also look at the influence of family status and gender on repaying their loan. And we find that although male tend to have higher default rate than female(1.6%),the difference is not so significance. Also in terms of family status, the difference is too small to be noticed. (see Appendix.)

3.2.3 Employment information in terms of loan is repaid or not

In the datasets, another groups of variables may have influence on predicting the bad loan are the employment groups, including information about clients' days of the work, occupation type and income etc.

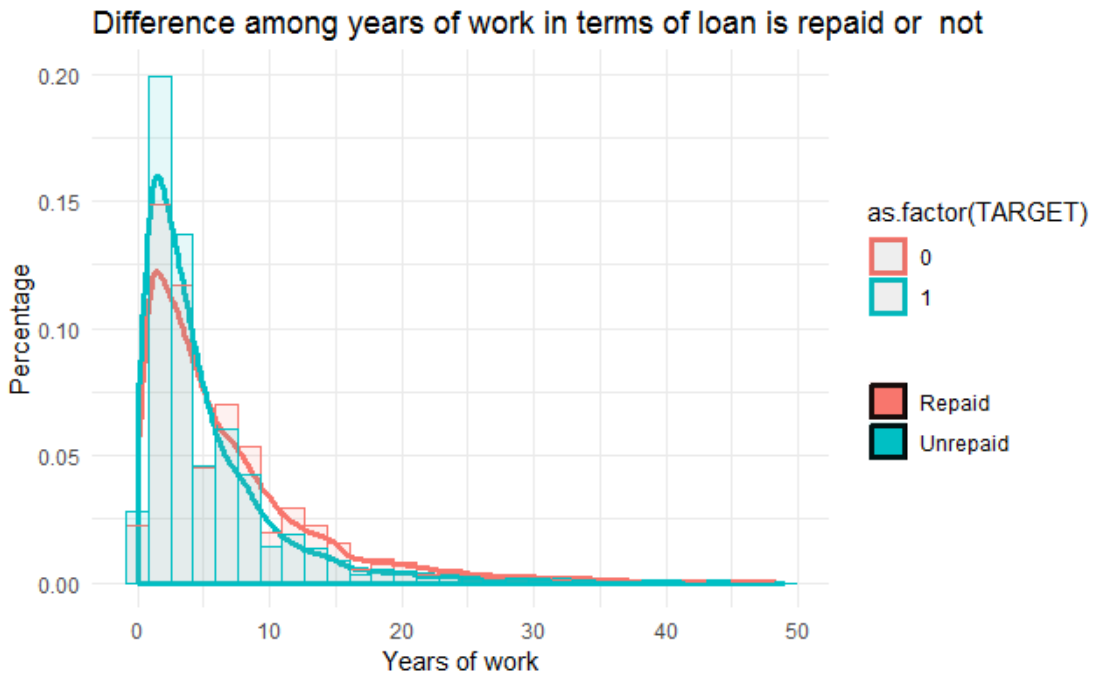

Figure 3: Difference among years of the work in terms of loan is repaid or not

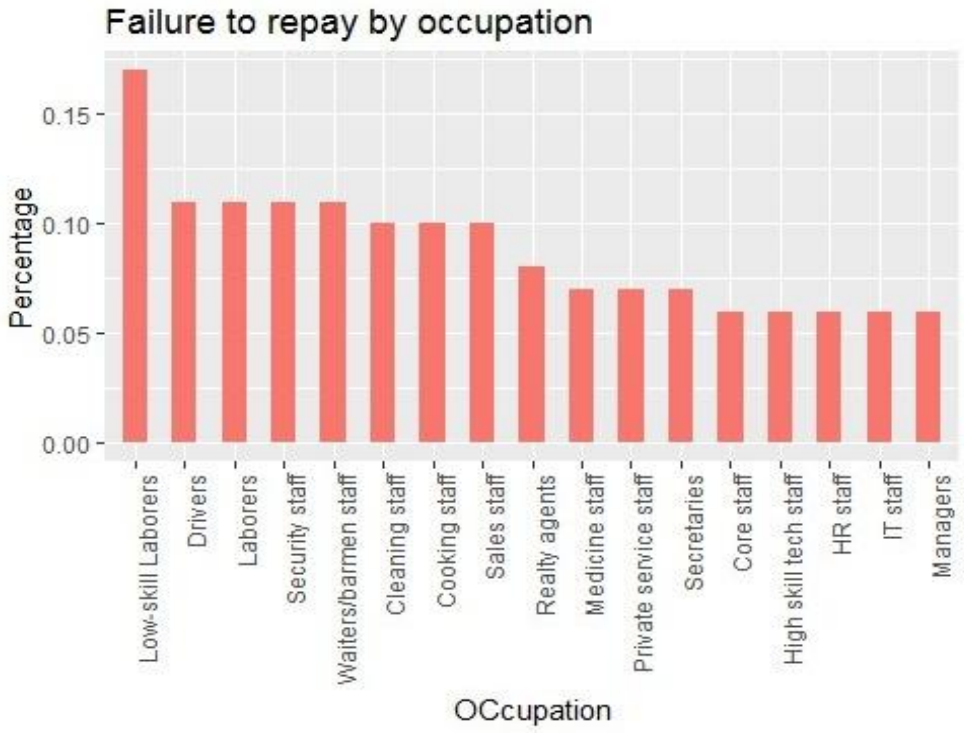

Figure 4: Failure to repay by Occupation

From the plot, we can discover that for people who have worked for a long time, they have lower probability of repaying their loan than workplace newbies. This is information that

could be directly used by the bank: The bank should take precautionary measures to people who change their work often or new to their position.

Another interesting discovery is that people of some certain occupations, like manual workers, may have significant higher default rate than others. This trend is also consistent with the trend of income: people who earn more are also more willing to repay their loan. This does not mean the bank should discriminate against manual workers, but they can pay attention to some other characteristics of these clients to make further decision.

3.3 Model Used

Before doing the modeling, I created a dataset containing the information of client's basic personal information and their previous credit score from external data source. Also, the data are preprocessed by clearing the missing values and scaling the range of the features. In the dataset, 20% of the data are randomly selected as test data, in which there are 16491 entries in total.

Names	Description
s2	Normalized score from external data source
s3	Normalized score from external data source
annulty	Annuity of the Credit Bureau credit
credit	Final credit amount on the previous application
income	Income of the client
s1	Normalized score from external data source
pop	Normalized population of region where client lives
price	Goods price of good that client asked for on the previous application
hour	Approximately at what day hour did the client apply for the previous application
gender	Gender of the client
marriage	Family Status of the Client
occupation	Occupatopn of the client
empyear	How many years before the application the person started current employment
ageyear	Client's age in days at the time of application
idyear	How many years before the application did client change the identity document
regyear	How many years before the application did client change his registration
phoneyear	How many days before application did client change phone

Table 1: the variables description

3.3.1 Logistic Model

To get a baseline model, I use the logistic model with all the features in the datasets I just created. The function "glm" with the "family = binomial" are used to train the model and then predictions are made on the testing data.

3.3.2 Multilevel Logistic Regression

From the EDA, we can find that clients from the same age group, years of work groups and occupation groups may have closer default rate; while among each groups, there exists a significance difference. To analyze the difference among the groups, we use four different multilevel logistic regressions: group by age, group by years of work and group by occupation and group by all these variables. The function "glmer" with the "family=binomial" are used to train each model and predictions are made on the same testing data.

4. Results of the model

4.1 Model Choice

To choose from above five models, I use function "anova" to see the difference. Also, they were all tested on test datasets to see the accuracy of the prediction. Under the rules of minimum of AIC and maximum of F1-score, the multilevel logistic regression model with age, occupation and years of work groups have the best performance among the five models.

The fitted model can be viewed as follows:

glmer(data = train,y~s2+s3+annulty+s1+pop+income+idyear+hour+
(1|emprange)+(1|agerange)+(1|occupation),family =binomial(link="logit"))

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
mulfit	12	31512.33	31621.49	-15744.16	31488.33	NA	NA	NA
mulfit3	16	32264.03	32409.58	-16116.01	32232.03	0.00000	4	1.000000
mulfit1	17	31571.72	31726.37	-15768.86	31537.72	694.30510	1	0.000000
mulfit2	25	31572.17	31799.59	-15761.08	31522.17	15.55712	8	0.049176

Table 2: The anova table of model comparison

4.2 Interpretation

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: y ~ s2 + s3 + annulty + s1 + pop + income + idyear + hour + (1 |
```

```
##     emprange) + (1 | agerange) + (1 | occupation)
##    Data: train
##
##      AIC      BIC   logLik deviance df.resid
##  31512.3  31621.5 -15744.2  31488.3    65950
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.3030 -0.3086 -0.2147 -0.1463 17.7061
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  occupation (Intercept) 0.02152  0.1467
##  agerange   (Intercept) 0.06857  0.2619
##  emprange   (Intercept) 0.02769  0.1664
## Number of obs: 65962, groups:  occupation, 18; agerange, 10; emprange, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.944479   0.176288  22.375  < 2e-16 ***
## s2          -1.252838   0.052660 -23.791  < 2e-16 ***
## s3          -1.868973   0.054283 -34.430  < 2e-16 ***
## annulty      0.249662   0.036324   6.873 6.28e-12 ***
## s1          -2.453932   0.098811 -24.835  < 2e-16 ***
## pop         -2.238881   1.276412  -1.754 0.079423 .
## income      -0.139732   0.037683  -3.708 0.000209 ***
## idyear      -0.016473   0.004578  -3.598 0.000320 ***
## hour        -0.016310   0.004717  -3.458 0.000545 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) s2     s3     annlty s1     pop    income idyear
## s2      -0.367
## s3      -0.431 -0.063
## annulty  0.044 -0.018 -0.066
## s1      -0.189 -0.077 -0.060 -0.050
## pop     -0.043 -0.141  0.030 -0.024 -0.039
## income   0.012 -0.077  0.106 -0.438  0.016 -0.080
## idyear  -0.156 -0.032 -0.078  0.023  0.014  0.004 -0.007
## hour    -0.256 -0.122  0.050  0.013 -0.033 -0.114 -0.016  0.006
```

From the summary of the model, we can see that:

• The coefficient for s2 is -1.25.Dividing by four yields a rough estimate that for one standard deviation increase in the normalized score from external data source s2, the probability of failure of repaying the loan decrease about 31%. The interpretation is nearly the same for s3 and s1.

• The coefficient for annuity is 0.25.Dividing by four yields a rough estimate that for one standard deviation increase in the scaled annuity, the probability of failure of repaying the loan increase about 6.25%.

• The coefficient for population is -2.23,which, when divided by 4, is 0.55,suggesting that one standard deviation increased in the normalized population of region where client lives, the probability of failure of repaying the loan decrease about 55%.

• The age groups errors have estimated standard deviation 0.28 on the logit scale. Dividing by 4 tells us that the age differed by approximately ±7% on the probability scale.

• The years of works groups errors have estimated standard deviation 0.18 on the logit scale. Dividing by 4 tells us that the age differed by approximately ±4.5% on the probability scale.

• The occupation groups errors have estimated standard deviation 0.14 on the logit scale. Dividing by 4 tells us that the age differed by approximately ±3.5% on the probability scale.

4.3 Model Checking and prediction

The binned plot of the residuals

First of all, we will look at the binned plot of the multilevel logistic regression.

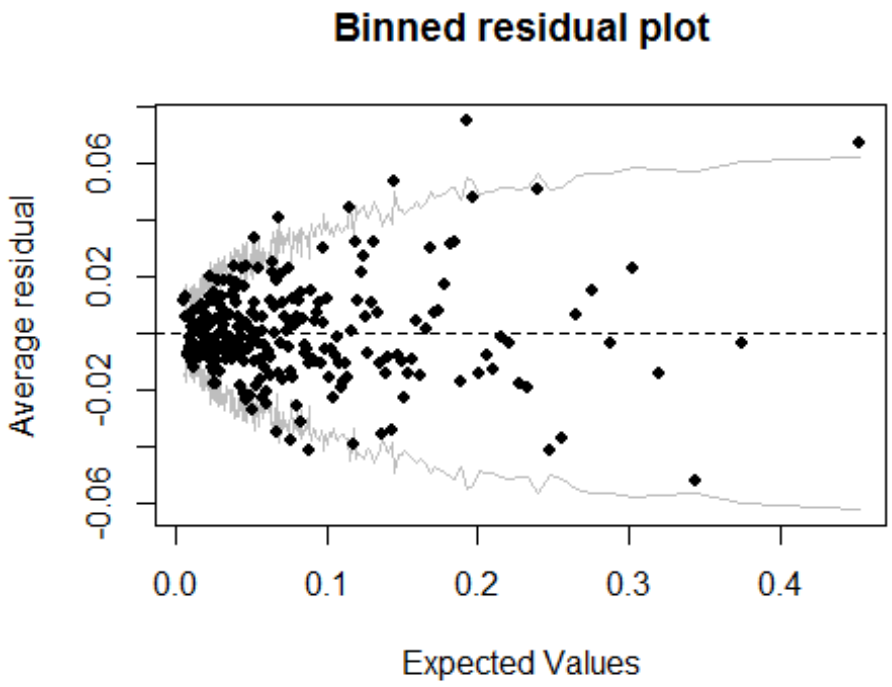

Figure 5: The binned residual plot of the models

From the plot, we can see that the majority of the points fall within the range and there is no distinctive pattern in the plot.

Testing significance of effects

To test if the coefficients are zero, we use the likelihood ratio test(LRT) in R. Although the LRT of mixed models is only approximately $\chi 2$ distributed, we can use it to roughly estimate the significance of coefficients.

```
drop1(mulfit,test="Chisq")
```

```
## Single term deletions
##
## Model:
## y ~ s2 + s3 + annulty + s1 + pop + income + idyear + hour + (1 |
##     emprange) + (1 | agerange) + (1 | occupation)
##         Df   AIC     LRT   Pr(Chi)
## <none>     31512
## s2       1 32070  559.45 < 2.2e-16 ***
## s3       1 32789 1278.35 < 2.2e-16 ***
## annulty  1 31556   45.74 1.348e-11 ***
## s1       1 32160  649.50 < 2.2e-16 ***
## pop      1 31513    2.99 0.0835554 .
## income   1 31524   13.78 0.0002054 ***
## idyear   1 31523   12.81 0.0003440 ***
## hour     1 31522   11.93 0.0005512 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the result, we can see that most coefficients in the models are significantly not equal to zero, except for the coefficient of population. Generally using MCMC methods can help us get a more reliable p-value, but due to the longer run time, we just make a simple estimation.

Prediction

```
##
## mul4.pred      0     1
##   Repaid   14441   929
##   Unrepaid   809   312

## The sensitivity is  0.2514 .

## The specificity is  0.947 .

## The F1 score is  0.2642 .
```

We use the model to predict the probability of failure to repay their debt. Since in the case of Home Credit, we are more concerned about the 1 target, I switch the cutting point from 0.5 to 0.2, in this way it can be more sensitive to predicting the default. From the result, we can see that the accuracy of predicting is about 89% and the F1- score of the test is about 26%.

5. Discussion

5.1 Implication

For a financial institution that provide loan for customers, it is important to understand what characteristics play a role in influencing clients' repayment abilities.

From our work, we can see that the scores from external data source for the clients are good reflections of their ability to repay their loans. The lower the scores are, the more cautious the bank should be of lending their money. Besides the external scores, some basic personal characteristics and employment information are also important. Generally speaking, for those who are under 35 years old and often change their job or be new to their position, the bank should be aware that they may have lower savings and wages, thus they may have a higher probability of default rate. Some other factors can also be used to identify the potential risk of lending their money, such as how many years before the application did client change the identity document. It seems that people who are willing to update their identity document frequently have a lower rate of default.

For the bank, when knowing what kinds of factors will have influence on repaying debt, it can lower the risk before lending its money. They can also take precautionary measures and provide guidance or financial planning tips to clients pay on time.

5.2 Limitation

Due to the limited time and knowledge, there are many limitations in this project.

1. The data cleaning process

Although in some variables like DAYS_EMPLOYED, I detect the abnormal values and find out there are some relationships between abnormal values and the target, in many other variables I just delete the abnormal values and the missing values. This process causes the amount of datasets decrease significantly and much valuable information are also excluded, which can also affect the training of the models.

2. The module checking process

I try to use simulations in checking the results of the models, but due to the slow efficiency I failed to run the model on my computer. If simulations can be used, they can be a good way to justify my result.

5.3 Future direction

In this project, I basically just use the data in the train data sets. There are many information of client's previous credits and monthly balance I haven't use. For further analysis, I can join each files using SQL and build a new data set. With these information, some more complicated models can be built and the accuracy of prediction may also be improved.

6. Acknowledgement

I would like to express my special thanks of gratitude to my professor, [Mr Yajima], who not only be always patient about our questions, but also give us instruction on how to perform professionally as an data scientist. Secondly I would also like to thank my friends and classmates who helped me a lot in finalizing this project within the limited time frame.