MA678 mid-term project report

Jianhaoyan

12/2/2018

1. Abstract

MA678 mid-term project is a great chance to combine what I have learned in class with my career goals. I dream to seek for a job in financial industry, so my topic is about credit card which accounts for great proportions of banks' profits. This project focuses on how to predict the probability of credit cards clients not paying back money on time. The data comes from Kaggle provided by "HOME CREDIT COMPANY."

2. Introduction

2.1 Background

Nowadays, credit cards are crucial to both banks and costumes. Credit cards give costumers good shopping experiences and bring enormous profits to banks. However, the extensive use of credit cards brings the considerable credit default risk to financial companies. So fitting the reliable models to predict the probability of credit default behavior happening in each client is essential. As we all know, mass credit card defaults can cause a financial crisis. So fitting a good model can also be crucial to the government to protect their finance from the crisis.

2.2 Data and packages preparation

Our data is messy. It has 122 variables and 307511 observations. Moreover, we can see our data has different kind of classes. What's more, missing data is common in our dataset. So it requires me to do data cleaning carefully before fitting the model. The responsive variable of my model is TARGET in which "0" means the client is paying back money on time while "1" means not.

3. EDA

In this part, we use EDA to have an overview of our data. Meanwhile, I would do some cleaning for my data. The EDA is the process of exploring data where we can find the important information.The main goal of this part is to have an initial understadings of what are the relationships between different variables

3.1 Target Distribution

Target is the most important variable in my data. So I should know the distribution of this variable.

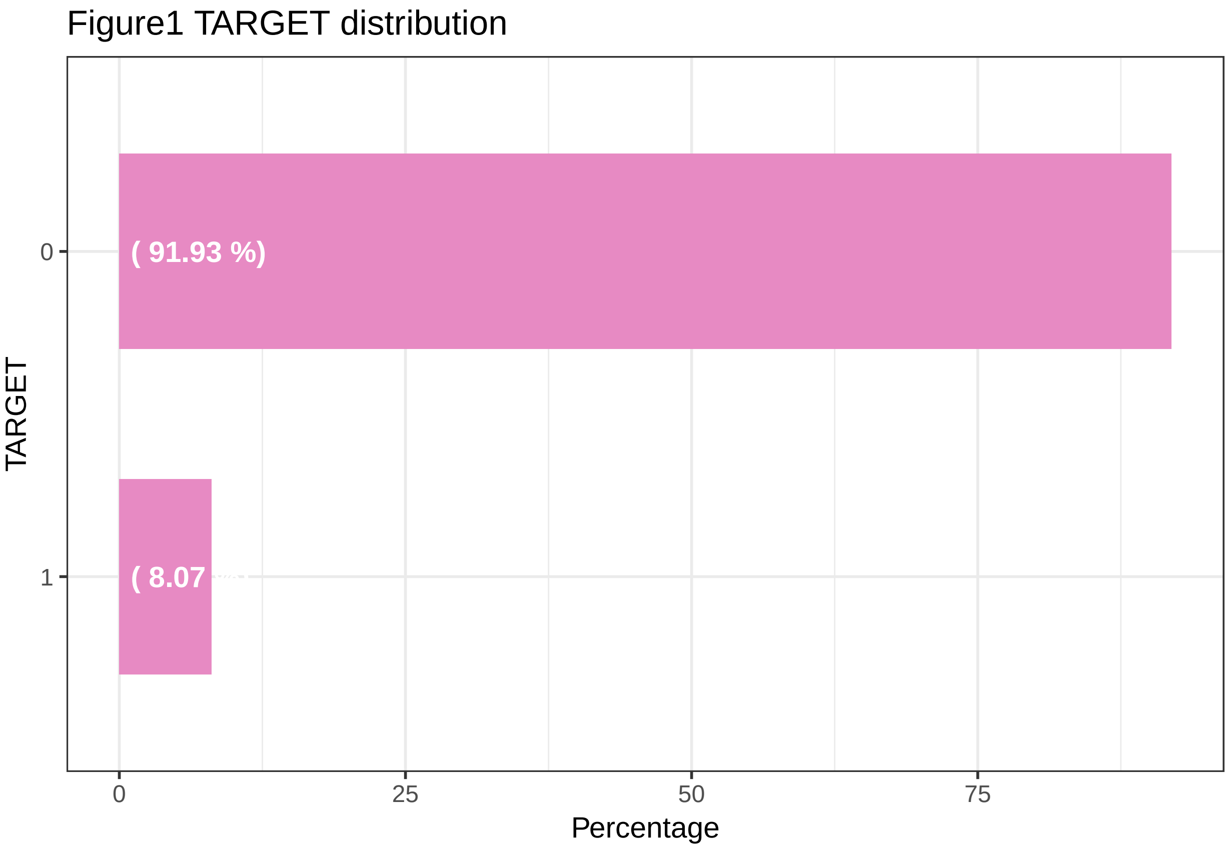


Figure1 TARGET distribution

From Figure 1, we can find that “0” takes really small proportions of the data which means ony few clients not paying back money on time. Besides, this plot tells me that my data has serious biased problems that needs me to handle.

3.2 Target VS NAME_CONTRACT_TYPE

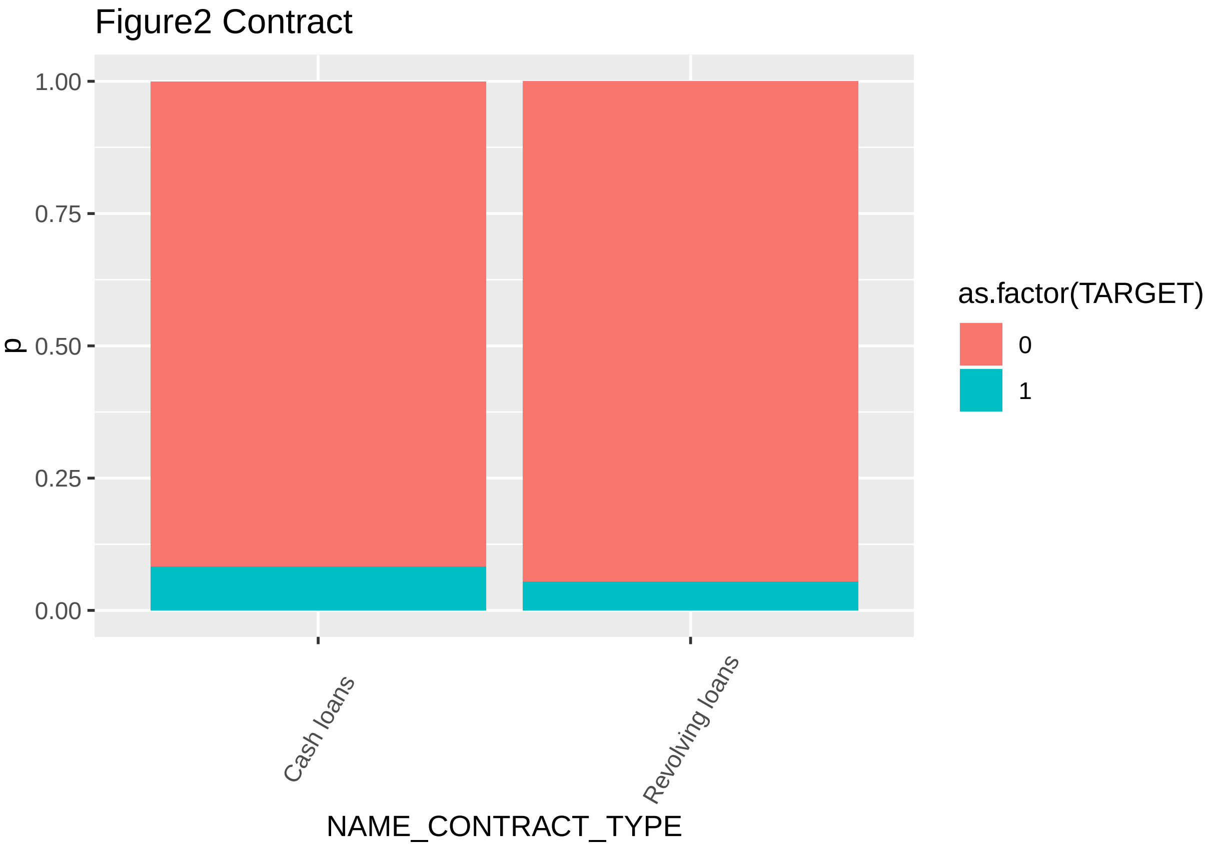


Figure 2 is about relationships between different contract types and TARGET. It is evident that TARGET 0 accounts for most of the parts of both two contract types which also implies me that my data is biased. Besides, we can find that the 0 accounts for much more probabilities in cash loans than revolving loans. So it tells us cash loans bring higher credit card default risk to banks than revolving loans.

3.3 Target VS Occupation type

Figure3 Distribution of Occupation

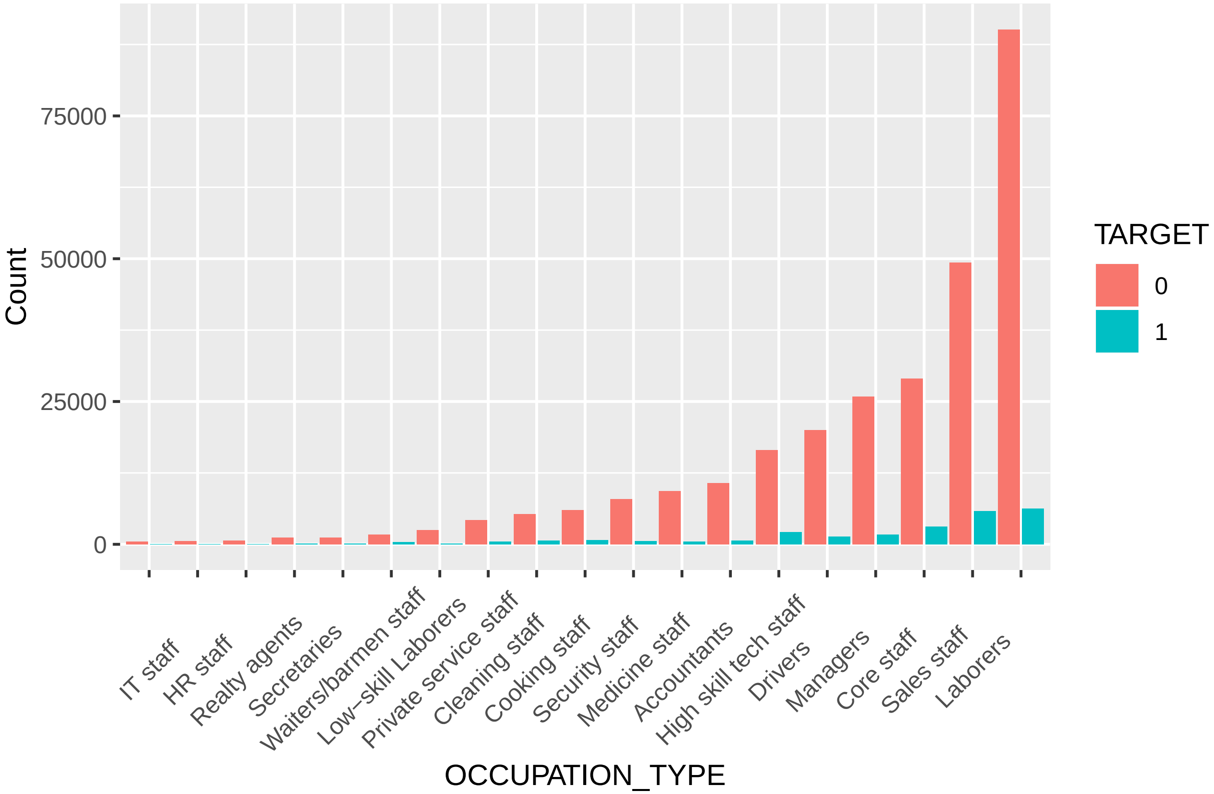


Figure4 Occupations

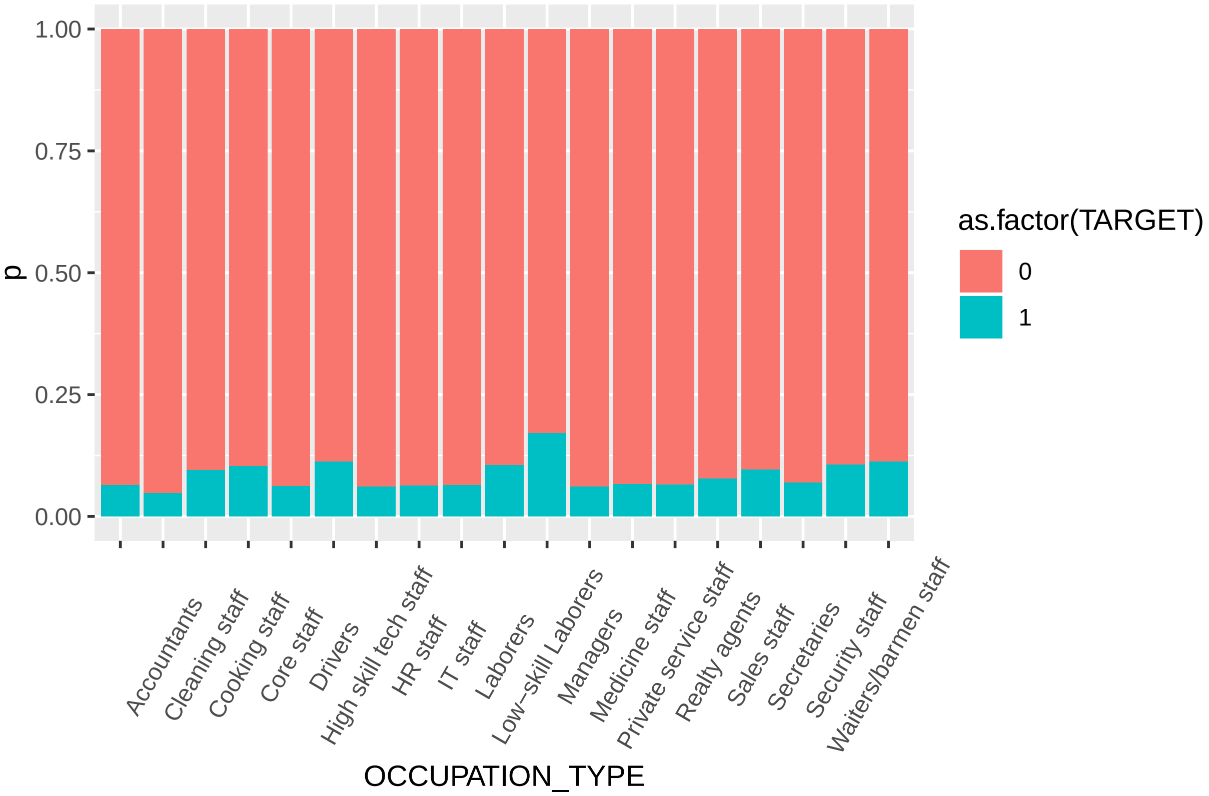
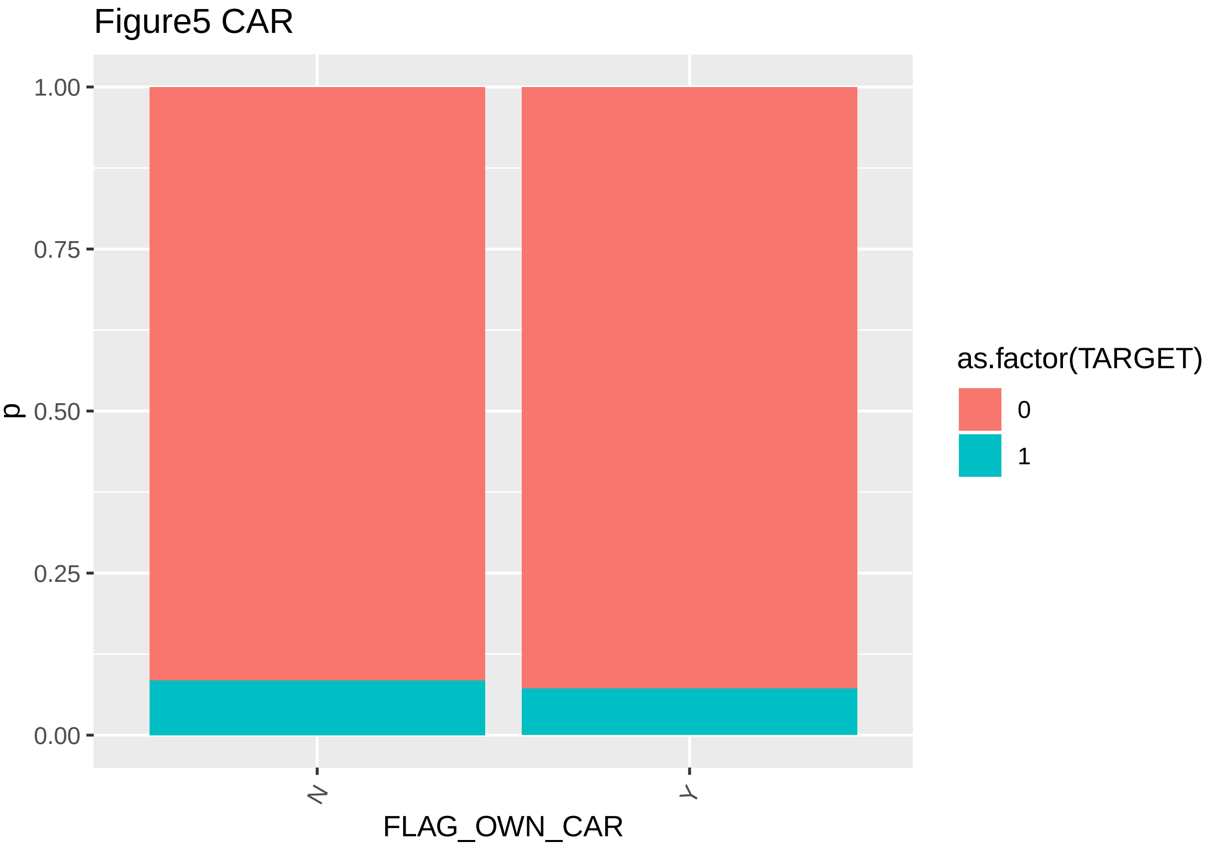


Figure3 is about Target and the occupation type. We can find the labors account for most of our clients, and labors have the most people that can not pay back money on time.

Figure4 tells us that labors may be much easier to fail to pay back money on time than other occupations. The relationship between occupations and target is obvious so that this variable can be a good predictor.

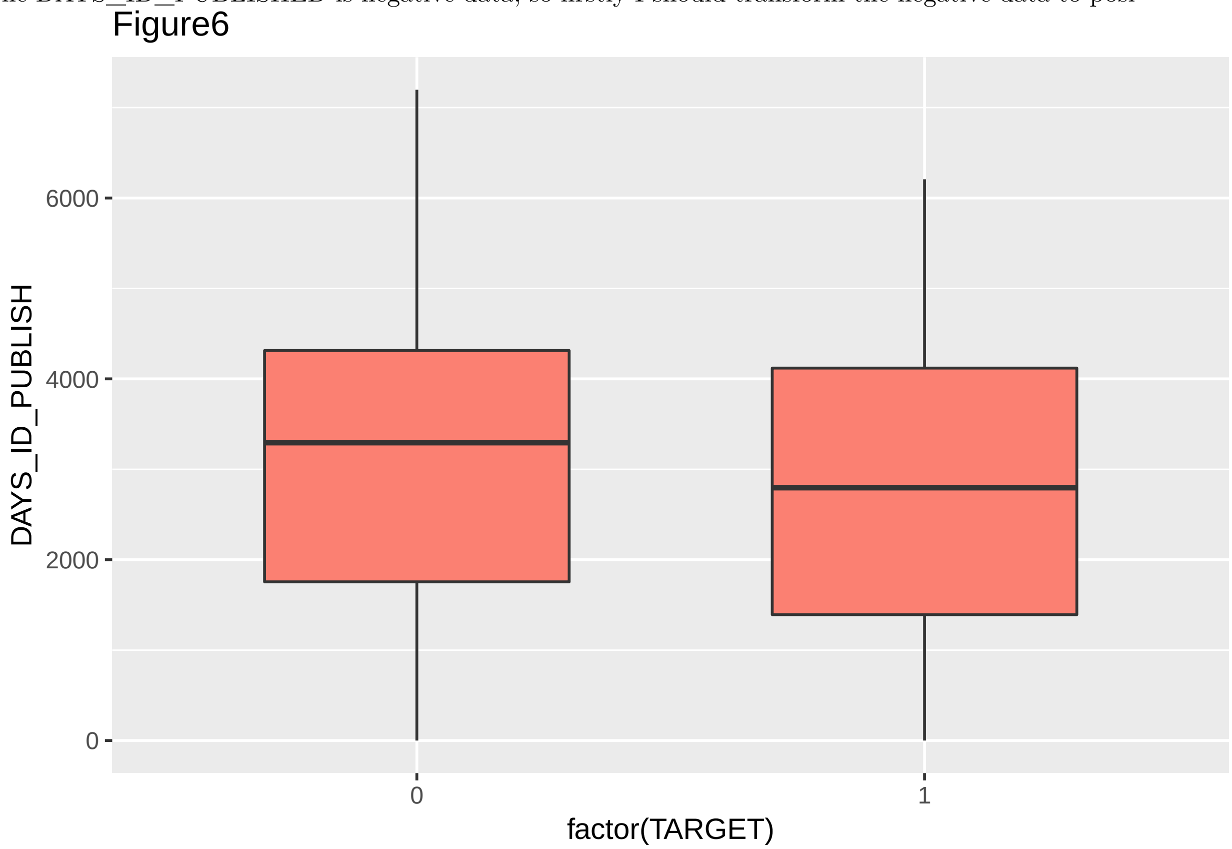
3.4 CAR VS Income



From this plot, we can find that it seems like that people without cars have larger probabilities paying back money delay than people with cars. But the difference between these two groups is slight, so the influences on TARGET brought by cars is not very obvious.

3.5 Target VS DAYS_ID_PUBLISHED

Because the DAYS_ID_PUBLISHED is negative data, so firstly I should transform the negative data to positive data.



From this plot, we can find that people paying back money on time have much longer DAYS_ID_PUBLISHED time than people cannot. And boxplot also tells me that this variable does not many outliers.

3.6 AMI_CREDIT VS TARGET

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   45000  270000  513531  599026  808650 4050000
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Figure7 Distribution of AMT_CREDIT

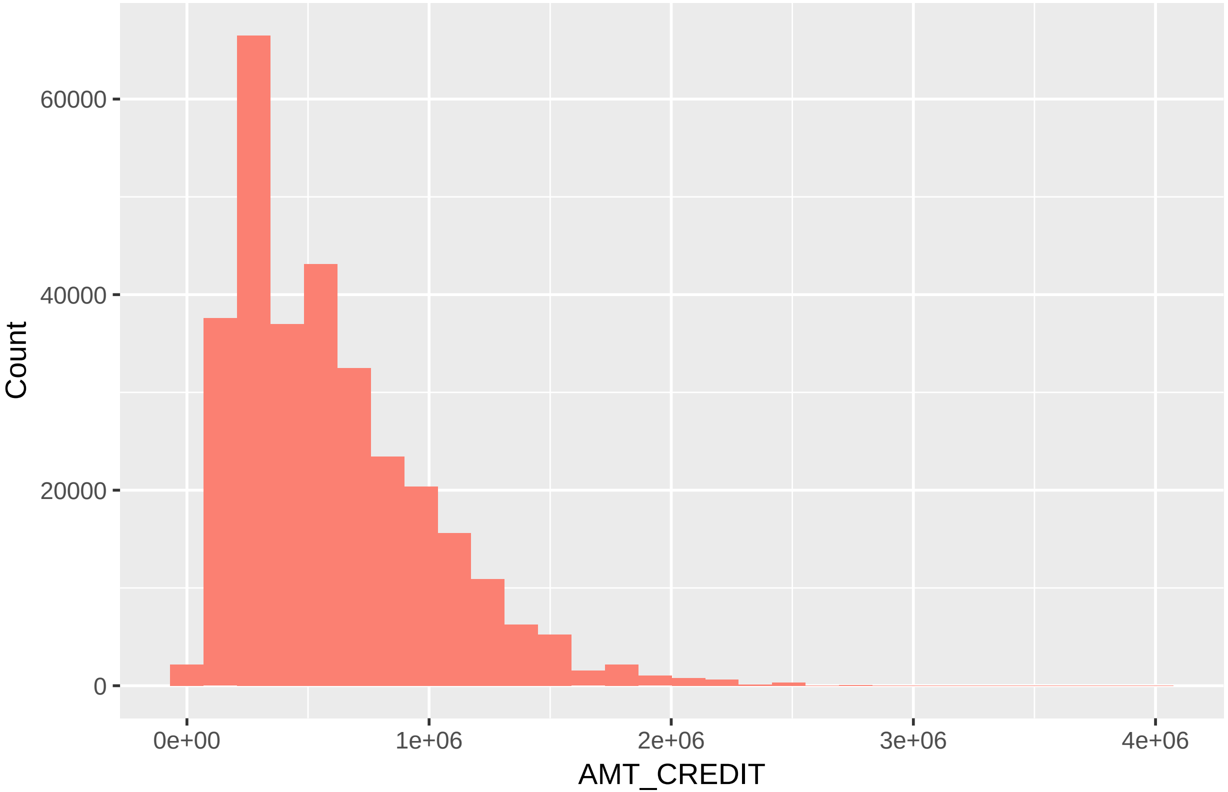


Figure8

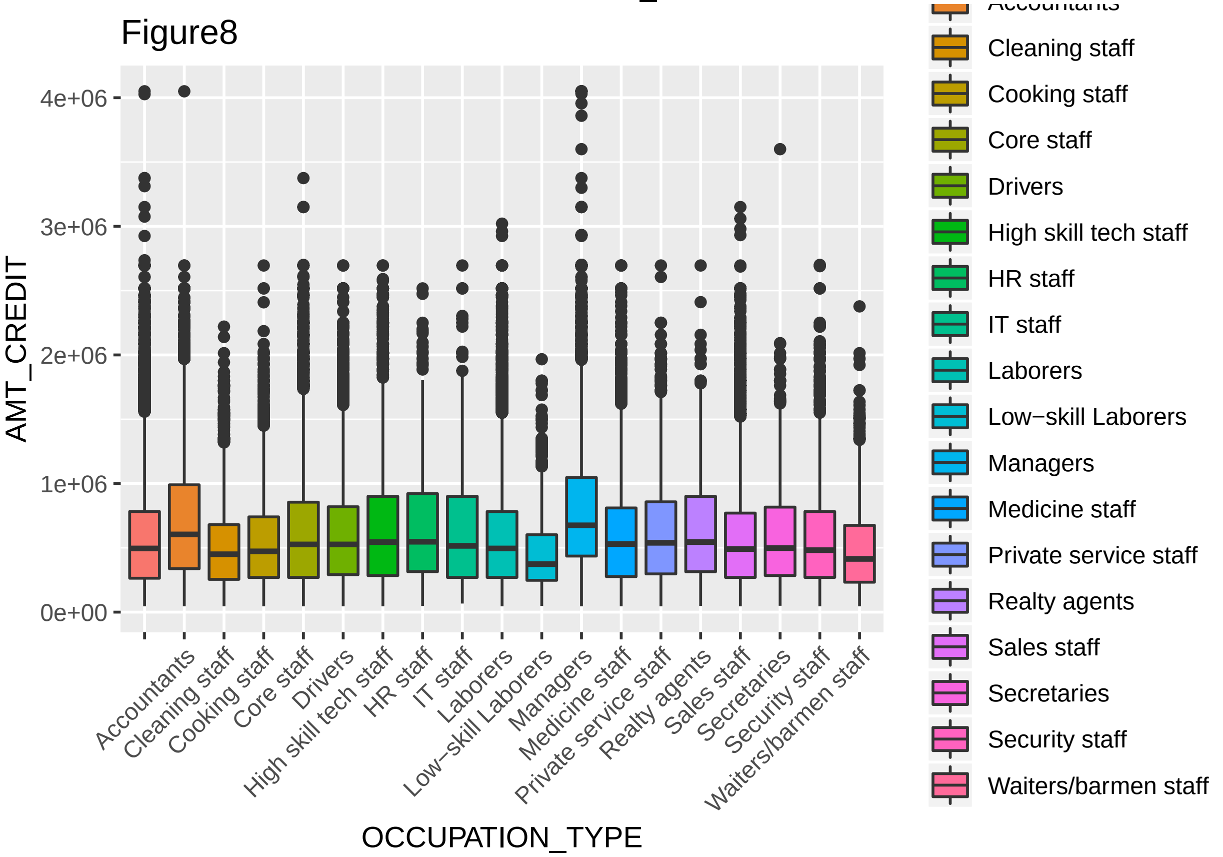
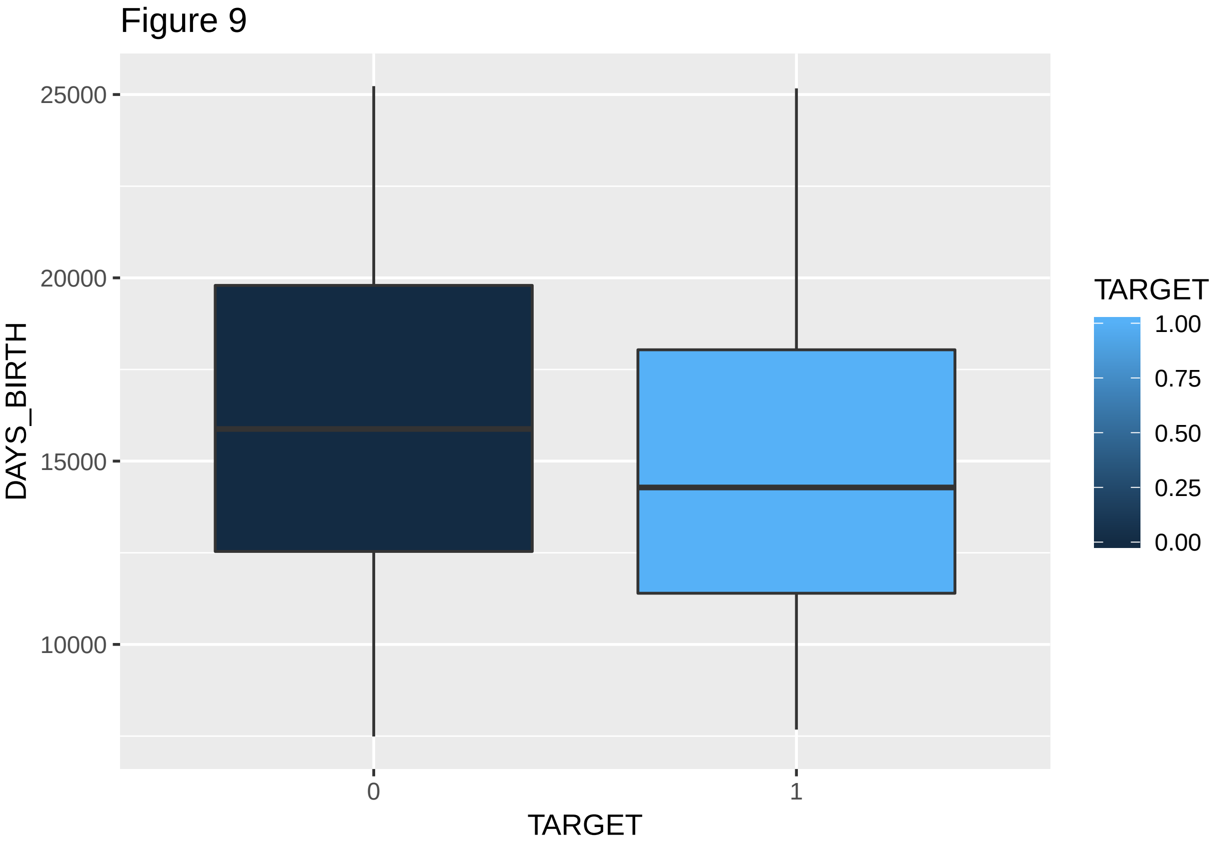


Figure7 is the distribution of AMT_CREDIT. This variable means how much money clients borrow from the bank. We can see that the AMT_CREDIT has apparent skewness which needs me to handle.

Figure8 tells me that the AMT_CREDIT has many outliers, so we should process this data to fit our model. And we can also know that the "manager" has the most higher average AMT_CREDIT than any other occupations.

3.7 Days of birth



DAYS_BIRTH is just the age of the client. From figure 8, we can find that clients are paying back money on time have more massive average days of birth than those cannot.

3.8 EXT_SOURCE VS TARGET

Figure10 EXT_SOURCE_1

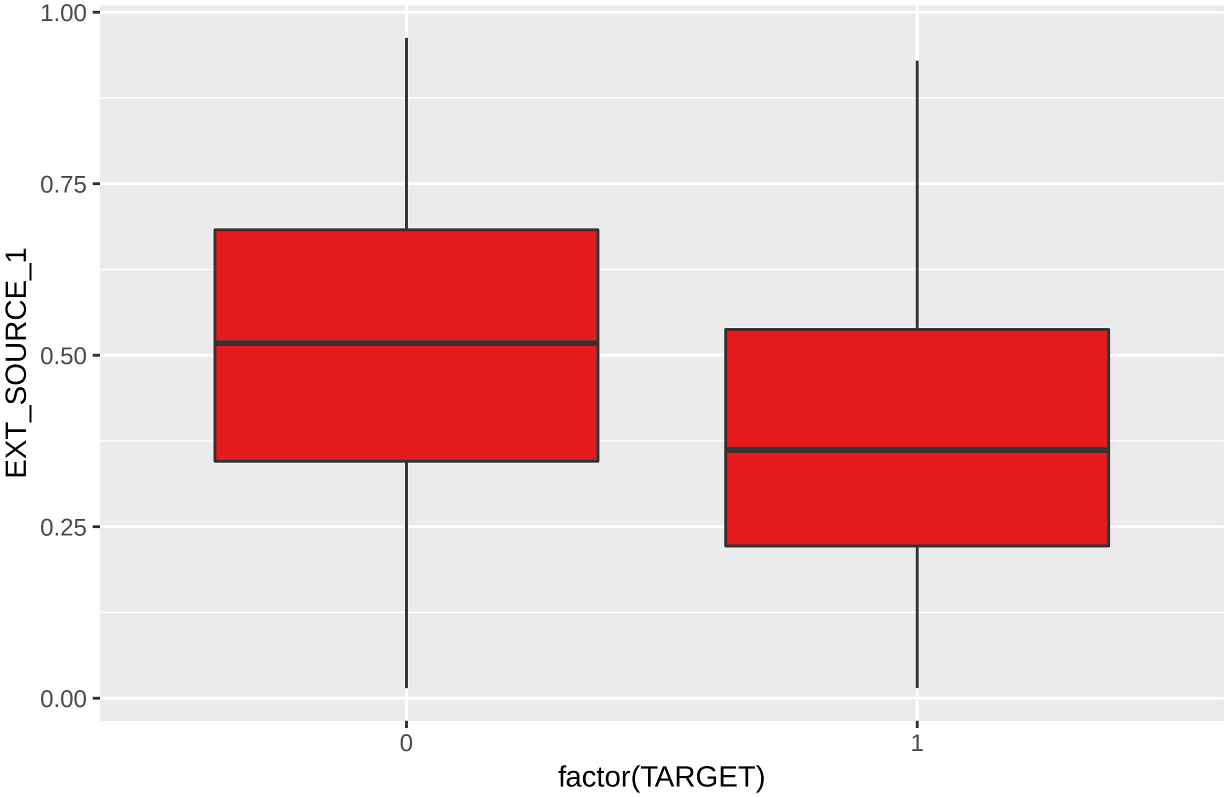
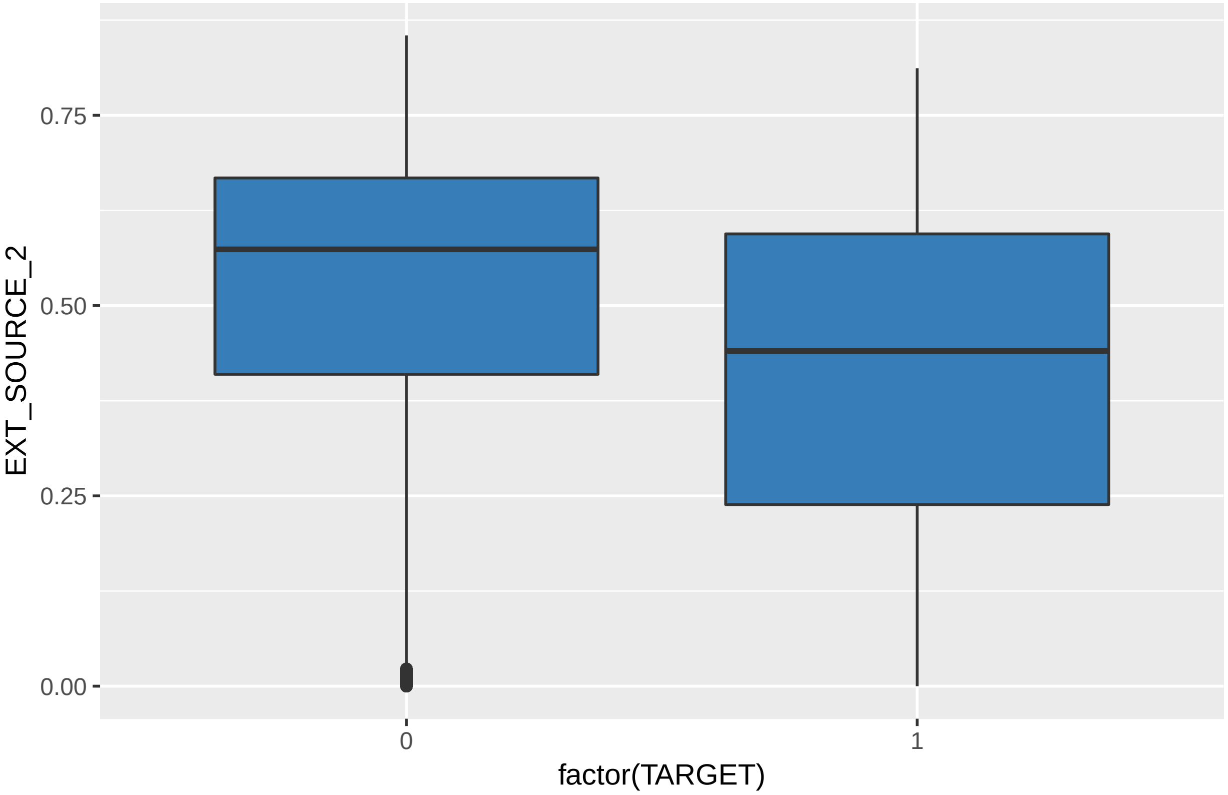
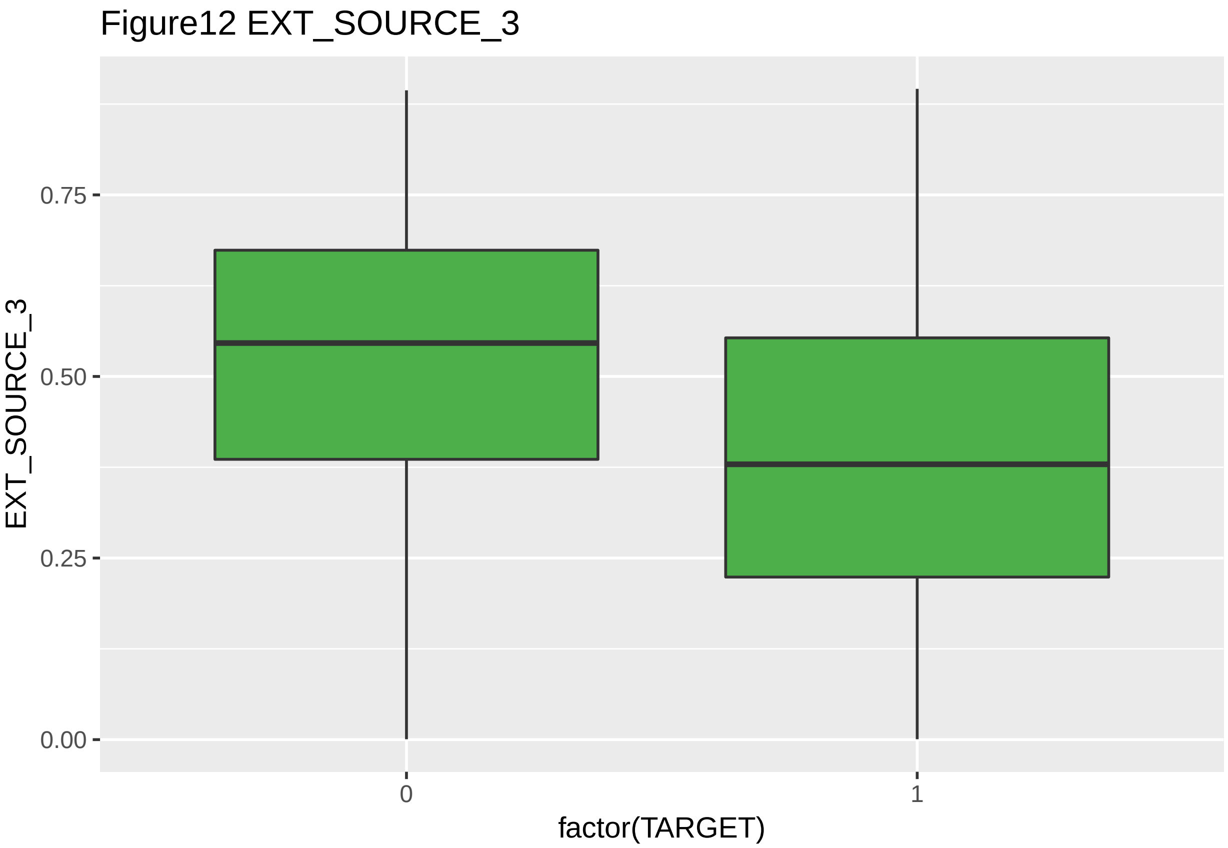


Figure11 EXT_SOURCE_2

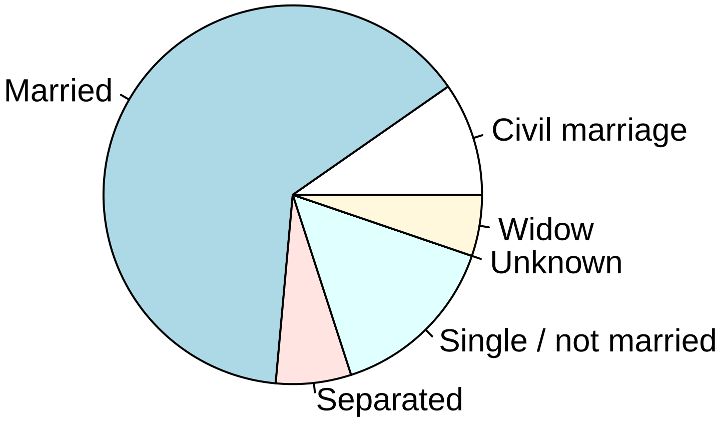




EXT_COURCE is the extra credit scores of clients.

It is evident that the extra credit scores have positive effects on the probability of clients paying back money on time. So the EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3 are important variables that can help me to fit the model.

3.9 Family Status



From this pie chart, we can find most of our clients are married.

4. Discussion

The purpose of my model is to predict the probability of failing to pay back money on time with given information about our clients. Because my responsive variable is 0 and 1, the logistic model is the first thing I think that may be suitable for my problem. However, I still need more models to compare with the logistic model to find the most suitable model. So I plan to fit three kinds of models. -Logistic model -Multilevel logistic model

4.1 Preprocessing data

From the EDA part, we can find that our data is messy. Firstly, there are lots of missing data existing. Secondly, this data is imbalanced and biased. In the responsive variable "TARGET", "0" accounts for 92%, while "1" only contains 8% of data. This characteristic of data can affect the accuracy of our model seriously. So we need to do the preprocessing containing dealing with missing data, addressing skewness problems and imbalance.

4.1.1 Extracting Features.

Because of too many variables, the extraction of features is essential. I use "baruta" to do this extraction. Because the calculation is enormous, I run this package in SCC2, so right now, I only use the outcome to set my data for the model.

4.1.2 Missing Data

For the factor and categorical variable, I choose to omit missing data. For the numeric variable, I choose to use mean to fill the missing data.

4.1.3 Skewness and outliers

From the EDA part, we find that the numeric data has skewness which brings a difficulty to fit the model. So in this part, I will do some transformations.

```
## Created from 307511 samples and 7 variables
##
## Pre-processing:
##   - Box-Cox transformation (7)
##   - ignored (0)
##
## Lambda estimates for Box-Cox transformation:
## 0.7, -0.1, 1.2, 1.5, 1.3, 0.2, 0.2
```

4.1.4 Resample

From the EDA part, we can find that my responsive variable "TARGET" has serious unbalance problem. The "1" accounts for 92 percent while "0" only accounts for 8 percent. I should acknowledge firstly I use the data without resampling to fit the model. Unfortunately, the model fails to converge. The biased data brings great difficulties to fit the good model. So this triggers me to resample the data. There are many ways to do that. We can oversample, undersample or "SMOTE". In this project, I choose to use undersample because it can also decrease the size of my data which can decrease the calculation time when I fit the model. I use "ROSE" package to resample our data in order to address the unbalanced issues of our data.

```
## Loaded ROSE 0.0-3

##
## Attaching package: 'sampling'

## The following object is masked from 'package:caret':
##
##     cluster
```

4.2 Model

In order to find the model that fits the best, I want to try different kinds of model because it can give me more space to choose.

4.2.1 logistic model:m1

Why I choose this model: - The reponsive variable in my data is TARGET which only contains 0 an 1. So this kind of data reminds me of the logistic model, and it meets my need that I want to predict the probability.

```
m1<-glm(factor(TARGET)~DAYS_ID_PUBLISH+YEARS_BEGINEXPLUATATION_MODE+
          COMMONAREA_MODE+FLOORSMAX_MODE+LIVINGAPARTMENTS_MODE+
          AMT_INCOME_TOTAL+EXT_SOURCE_1+EXT_SOURCE_2+EXT_SOURCE_3+
          AMT_CREDIT+AMT_GOODS_PRICE+factor(FLAG_OWN_CAR)+
          factor(OCCUPATION_TYPE)+REG_CITY_NOT_LIVE_CITY,
          data = db_train_1,family=binomial("logit"))
```

4.2.2 Multilevel Logistic Model

```
m2<-lme4::glmer(factor(TARGET)~
                FLAG_OWN_CAR+(1|OCCUPATION_TYPE)+
                 EXT_SOURCE_2+EXT_SOURCE_3+EXT_SOURCE_1+
                  poly(AMT_CREDIT,3)+poly(AMT_GOODS_PRICE,3)+
                  FLOORSMAX_MODE+LIVINGAPARTMENTS_MODE,
                data = db_train_1,
                family = binomial("logit"))
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : unable to evaluate scaled gradient

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge: degenerate Hessian with 2
## negative eigenvalues
```

4.3 Interpretation

4.3.1 Interpretation Logistic Model

```
summary(m1)
```

```
##
## Call:
## glm(formula = factor(TARGET) ~ DAYS_ID_PUBLISH + YEARS_BEGINEXPLUATATION_MODE +
##     COMMONAREA_MODE + FLOORSMAX_MODE + LIVINGAPARTMENTS_MODE +
##     AMT_INCOME_TOTAL + EXT_SOURCE_1 + EXT_SOURCE_2 + EXT_SOURCE_3 +
##     AMT_CREDIT + AMT_GOODS_PRICE + factor(FLAG_OWN_CAR) + factor(OCCUPATION_TYPE) +
##     REG_CITY_NOT_LIVE_CITY, family = binomial("logit"), data = db_train_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5428  -1.0040  -0.4478   1.0150   2.6804
##
## Coefficients:
##                                                Estimate Std. Error z value
## (Intercept)                                  -4.143e+00  3.975e-01 -10.423
## DAYS_ID_PUBLISH                              -4.145e-05  7.948e-06  -5.215
## YEARS_BEGINEXPLUATATION_MODE                 -1.916e-01  2.319e-01  -0.826
## COMMONAREA_MODE                              -3.515e-01  4.018e-01  -0.875
## FLOORSMAX_MODE                               -3.611e-01  1.404e-01  -2.572
## LIVINGAPARTMENTS_MODE                        -1.603e-02  2.870e-01  -0.056
## AMT_INCOME_TOTAL                             -3.303e-02  2.855e-02  -1.157
## EXT_SOURCE_1                                 -2.144e+00  1.062e-01 -20.188
## EXT_SOURCE_2                                 -3.645e+00  9.475e-02 -38.463
## EXT_SOURCE_3                                 -3.675e+00  8.827e-02 -41.637
## AMT_CREDIT                                    1.363e-01  7.908e-03  17.238
## AMT_GOODS_PRICE                              -1.327e-01  8.096e-03 -16.385
## factor(FLAG_OWN_CAR)Y                        -2.212e-01  2.639e-02  -8.383
## factor(OCCUPATION_TYPE)Accountants           -2.084e-01  7.945e-02  -2.623
## factor(OCCUPATION_TYPE)Cleaning staff         3.831e-01  9.447e-02   4.056
## factor(OCCUPATION_TYPE)Cooking staff          2.422e-01  8.208e-02   2.951
## factor(OCCUPATION_TYPE)Core staff            -1.446e-01  4.730e-02  -3.058
## factor(OCCUPATION_TYPE)Drivers                5.264e-01  5.039e-02  10.445
## factor(OCCUPATION_TYPE)High skill tech staff -7.426e-02  6.887e-02  -1.078
## factor(OCCUPATION_TYPE)HR staff               7.626e-02  2.856e-01   0.267
## factor(OCCUPATION_TYPE)IT staff              -1.727e-01  2.893e-01  -0.597
## factor(OCCUPATION_TYPE)Laborers               3.556e-01  3.458e-02  10.284
## factor(OCCUPATION_TYPE)Low-skill Laborers     7.429e-01  1.249e-01   5.948
## factor(OCCUPATION_TYPE)Managers               3.444e-02  5.401e-02   0.638
## factor(OCCUPATION_TYPE)Medicine staff        -6.763e-02  7.640e-02  -0.885
## factor(OCCUPATION_TYPE)Private service staff  2.109e-02  1.346e-01   0.157
## factor(OCCUPATION_TYPE)Realty agents          2.443e-01  2.251e-01   1.085
## factor(OCCUPATION_TYPE)Sales staff            2.558e-01  4.161e-02   6.148
## factor(OCCUPATION_TYPE)Secretaries            2.379e-04  1.835e-01   0.001
## factor(OCCUPATION_TYPE)Security staff         4.423e-01  7.740e-02   5.714
## factor(OCCUPATION_TYPE)Waiters/barmen staff   2.442e-01  1.659e-01   1.472
## REG_CITY_NOT_LIVE_CITY                        1.757e-01  4.000e-02   4.392
##                                              Pr(>|z|)
## (Intercept)                                   < 2e-16 ***
## DAYS_ID_PUBLISH                              1.84e-07 ***
## YEARS_BEGINEXPLUATATION_MODE                  0.40877
## COMMONAREA_MODE                               0.38163
## FLOORSMAX_MODE                                0.01012 *
## LIVINGAPARTMENTS_MODE                         0.95547
## AMT_INCOME_TOTAL                              0.24722
## EXT_SOURCE_1                                  < 2e-16 ***
## EXT_SOURCE_2                                  < 2e-16 ***
## EXT_SOURCE_3                                  < 2e-16 ***
## AMT_CREDIT                                    < 2e-16 ***
## AMT_GOODS_PRICE                              < 2e-16 ***
## factor(FLAG_OWN_CAR)Y                        < 2e-16 ***
## factor(OCCUPATION_TYPE)Accountants           0.00872 **
## factor(OCCUPATION_TYPE)Cleaning staff       5.00e-05 ***
## factor(OCCUPATION_TYPE)Cooking staff         0.00317 **
## factor(OCCUPATION_TYPE)Core staff            0.00223 **
## factor(OCCUPATION_TYPE)Drivers               < 2e-16 ***
## factor(OCCUPATION_TYPE)High skill tech staff  0.28092
## factor(OCCUPATION_TYPE)HR staff              0.78950
## factor(OCCUPATION_TYPE)IT staff              0.55059
## factor(OCCUPATION_TYPE)Laborers              < 2e-16 ***
## factor(OCCUPATION_TYPE)Low-skill Laborers   2.71e-09 ***
## factor(OCCUPATION_TYPE)Managers              0.52373
## factor(OCCUPATION_TYPE)Medicine staff        0.37608
## factor(OCCUPATION_TYPE)Private service staff  0.87546
## factor(OCCUPATION_TYPE)Realty agents         0.27791
## factor(OCCUPATION_TYPE)Sales staff          7.86e-10 ***
## factor(OCCUPATION_TYPE)Secretaries           0.99897
## factor(OCCUPATION_TYPE)Security staff       1.10e-08 ***
## factor(OCCUPATION_TYPE)Waiters/barmen staff   0.14106
## REG_CITY_NOT_LIVE_CITY                      1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 48342  on 34875  degrees of freedom
## Residual deviance: 42127  on 34844  degrees of freedom
## AIC: 42191
##
## Number of Fisher Scoring iterations: 4
```

The output above shows the estimate of the regression beta coefficients and their significance levels, and we can find that most of our coefficients are statistically significant.

An important concept to understand, for interpreting the logistic beta coefficients, is the odds ratio. An odds ratio measures the association between a predictor variable (x) and the outcome variable (y). It represents the ratio of the odds that an event will occur (event = 1) given the presence of the predictor x (x = 1), compared to the odds of the event occurring in the absence of that predictor (x = 0).

For a given predictor (say x1), the associated beta coefficient (b1) in the logistic regression function corresponds to the log of the odds ratio for that predictor.

If the odds ratio is 2, then the odds that the event occurs (event = 1) are two times higher when the predictor x is present (x = 1) versus x is absent (x = 0).

When coefficient estimate of the variable is positive, it means that an increase in this variable is associated with increase in the probability of default. However when the coefficient for the variable negative, it means that an increase in this variable will be associated with a decreased probability of being default.

In this model,the intercept means that, with all of variables becoming 0, the odds of default is -9.When one variable change one unit with other variables staying the same, the odd of default will increase the amount of certain coefficients.

Deviance is a measure of goodness of fit of a generalized linear model. Or rather, it's a measure of badness of fit–higher numbers indicate worse fit.

In our model, the deviance is really big which means the fittness of my model is not good. And the residual deviance decreases with the freedom decreases which means the independent variables in our model can make some sense.

ACI is the way to compare the model, and we tend to choose the model with small ACI. ###4.3.2 Multilevel logistic regression

```r
summary(m2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## factor(TARGET) ~ FLAG_OWN_CAR + (1 | OCCUPATION_TYPE) + EXT_SOURCE_2 +
##     EXT_SOURCE_3 + EXT_SOURCE_1 + poly(AMT_CREDIT, 3) + poly(AMT_GOODS_PRICE,
##     3) + FLOORSMAX_MODE + LIVINGAPARTMENTS_MODE
##    Data: db_train_1
##
##      AIC      BIC   logLik deviance df.resid
##  42175.2  42293.6 -21073.6  42147.2    34862
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.6458 -0.8091 -0.3183  0.8186  5.1434
##
## Random effects:
##  Groups          Name        Variance Std.Dev.
##  OCCUPATION_TYPE (Intercept) 0.05447  0.2334
## Number of obs: 34876, groups:  OCCUPATION_TYPE, 19
##
## Fixed effects:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -4.24071    0.10095 -42.010  < 2e-16 ***
## FLAG_OWN_CARY               -0.22327    0.02608  -8.562  < 2e-16 ***
## EXT_SOURCE_2                -3.68170    0.09397 -39.178  < 2e-16 ***
## EXT_SOURCE_3                -3.68076    0.08743 -42.098  < 2e-16 ***
## EXT_SOURCE_1                -2.17987    0.10593 -20.578  < 2e-16 ***
## poly(AMT_CREDIT, 3)1       221.17985   13.67217  16.177  < 2e-16 ***
## poly(AMT_CREDIT, 3)2        -2.10994    9.90286  -0.213  0.83128
## poly(AMT_CREDIT, 3)3        -2.99954    9.06301  -0.331  0.74067
## poly(AMT_GOODS_PRICE, 3)1 -213.00214   13.73017 -15.513  < 2e-16 ***
## poly(AMT_GOODS_PRICE, 3)2  -19.42337    9.91032  -1.960  0.05001 .
## poly(AMT_GOODS_PRICE, 3)3    5.44019    9.12009   0.597  0.55084
## FLOORSMAX_MODE              -0.38795    0.13938  -2.784  0.00538 **
## LIVINGAPARTMENTS_MODE       -0.12702    0.25220  -0.504  0.61451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Correlation matrix not shown by default, as p = 13 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```
## convergence code: 0
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 2 negative eigenvalues
```

Firstly, the display show the estimate of average intercept, coefficients and their standard errors. The interpretions of these are similar to the logistic regreesion that we discussed above. The intercept means the average odd of default with all of variables becoming 0. And the coefficient means that one unit change of one variable with other variable staying the same will cause the amount of certain coefficient change of odd of default. And we can see that coefficients are almost statistically siginificant.

4.4 Model check

The main methods I used to check my model is confusion matrix, ROC and auc. -Confusion Matrix. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.. -Accuracy. Because my reponsive variable is binarary, and my dataset is biased, there is no reason for me to use accuracy to judge my model. -ROC. ROC curve is useful to check model fittness. And it can give use the most suitable thres to classify 0 and 1. -AUC. The bigger the better

4.4.1 Logistic model

4.4.1.1 confusionMatrix

```
##  4  7 11 12 13 14
##  0  1  1  1  0  0
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 4896 2583
##          1 2422 5034
##
##                Accuracy : 0.6649
##                  95% CI : (0.6572, 0.6725)
##     No Information Rate : 0.51
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.3298
##  Mcnemar's Test P-Value : 0.02372
##
##             Sensitivity : 0.6690
##             Specificity : 0.6609
##          Pos Pred Value : 0.6546
##          Neg Pred Value : 0.6752
##              Prevalence : 0.4900
##          Detection Rate : 0.3278
##    Detection Prevalence : 0.5008
##       Balanced Accuracy : 0.6650
##
##        'Positive' Class : 0
##
```

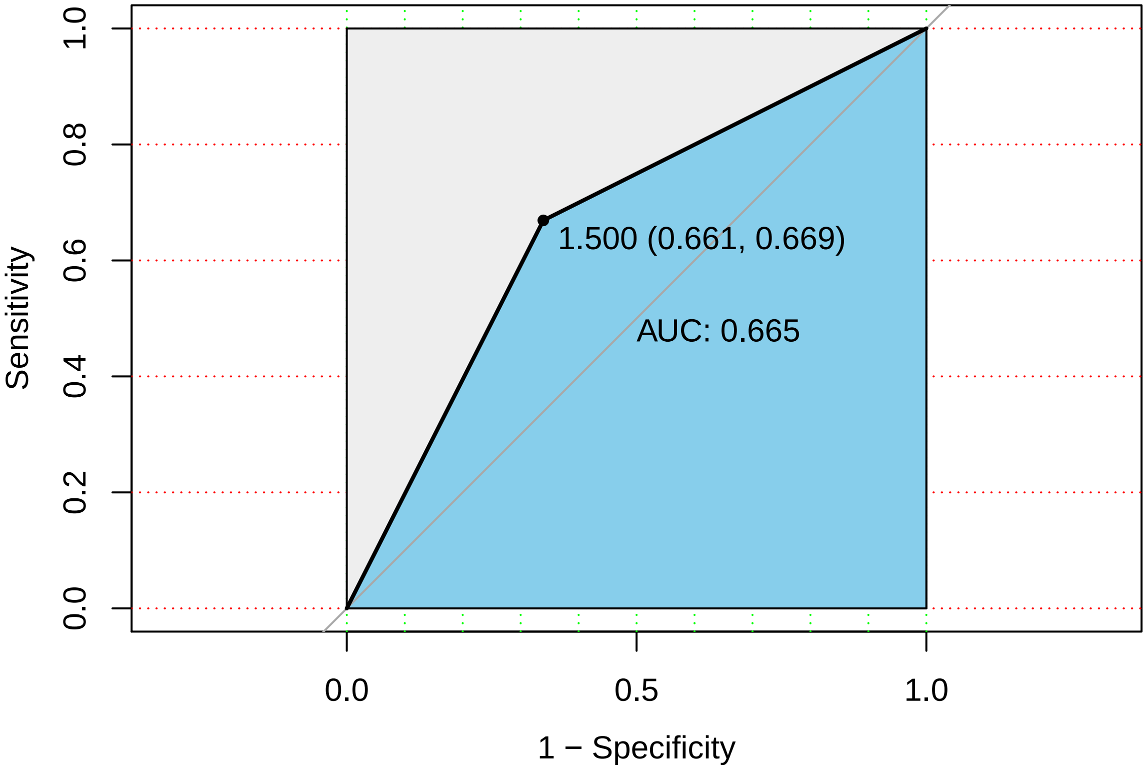
From the confusion matrix, we can find that our model does a great job. It can classify the good and bad clients fairly accuracy. The true positve rate is 69%,true negative rate is 66%.The model can classify most of the clients.

4.4.1.2 ROC and AUC

```
## Type 'citation("pROC")' for a citation.
```

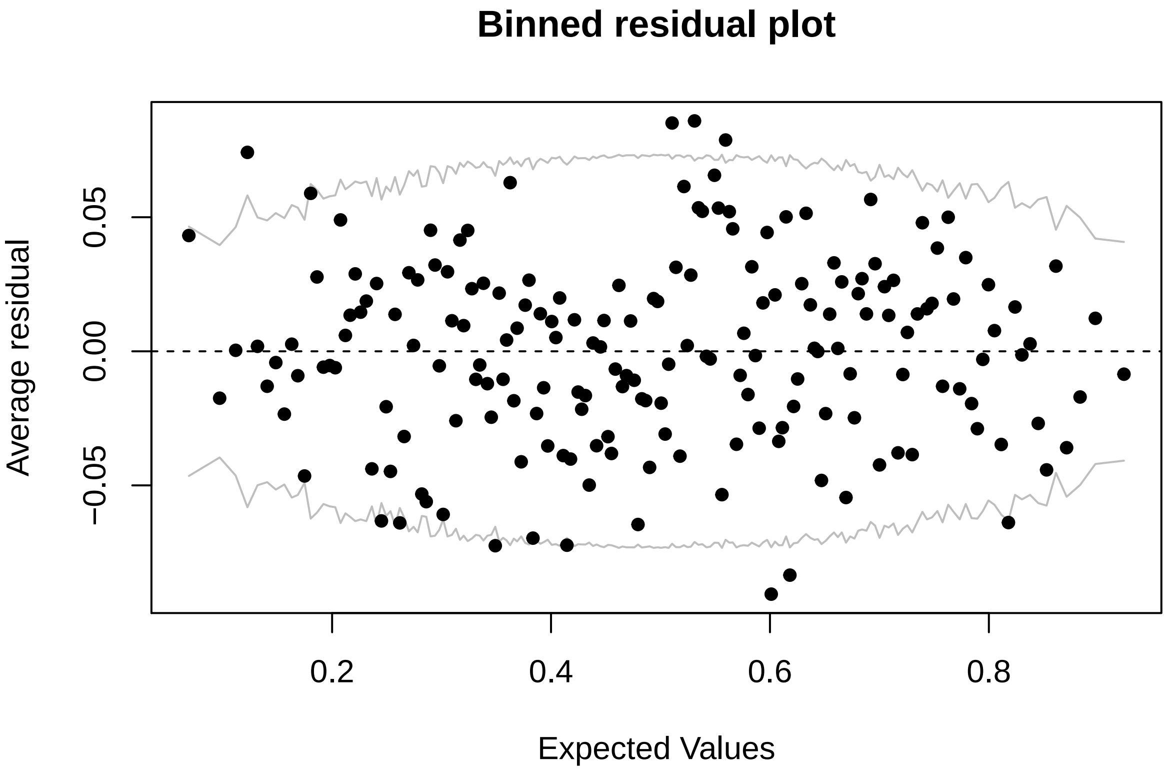
```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```



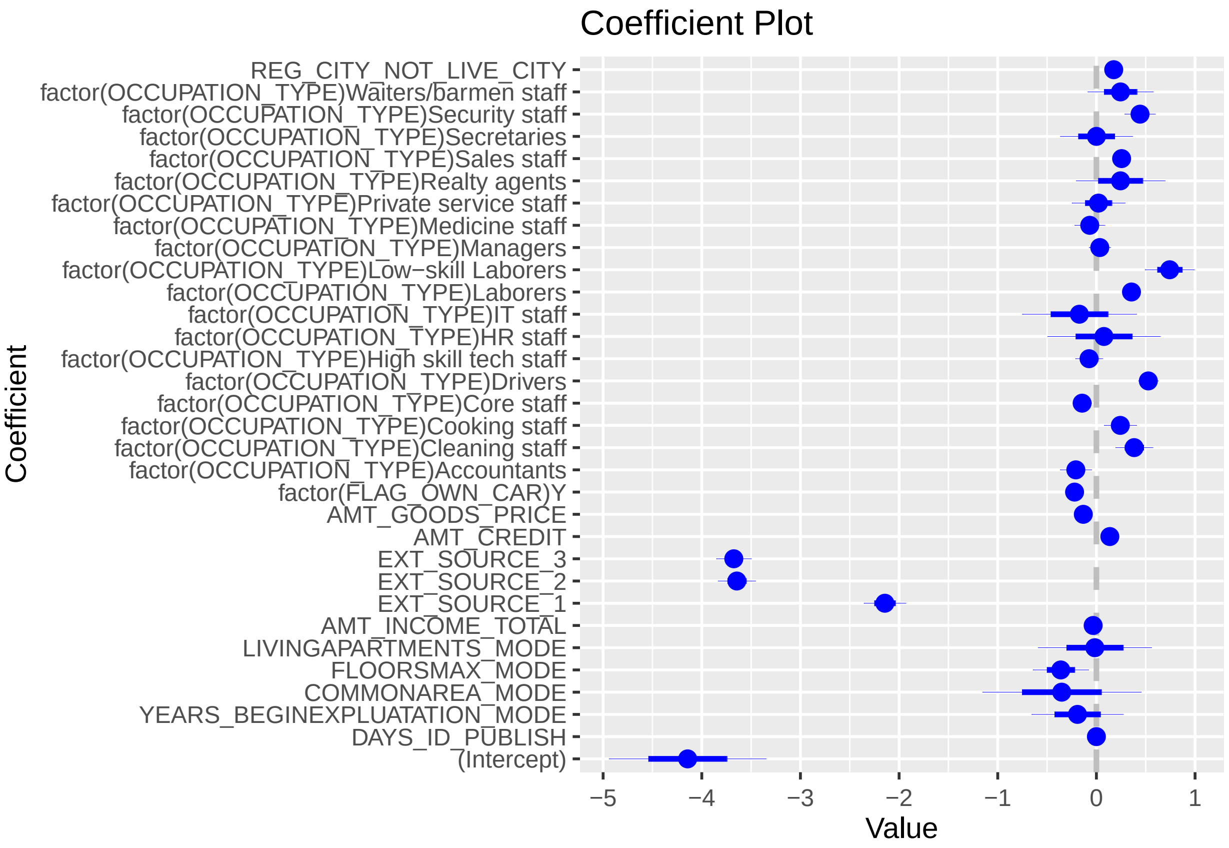
The AUC is 0.675 which is greater than 0.5, and it means our model have some classification abilities. Besides, the shape of ROC graph implies this model can do some predictions.

4.4.1.3 Binned residual plot



And we can see the binned residual plot is good because the residuals are almost inside the two lines. And these residuals are evenly distributed.

4.4.1.4 CoefPlot



4.4.2 Multilevel logistic regression model

4.4.2.1 ICC

```
## [1] 0.05165655
```

ICC of my model is 0.06 which is greater than 0, and it means the use of multilevel logistic model can make some sense.

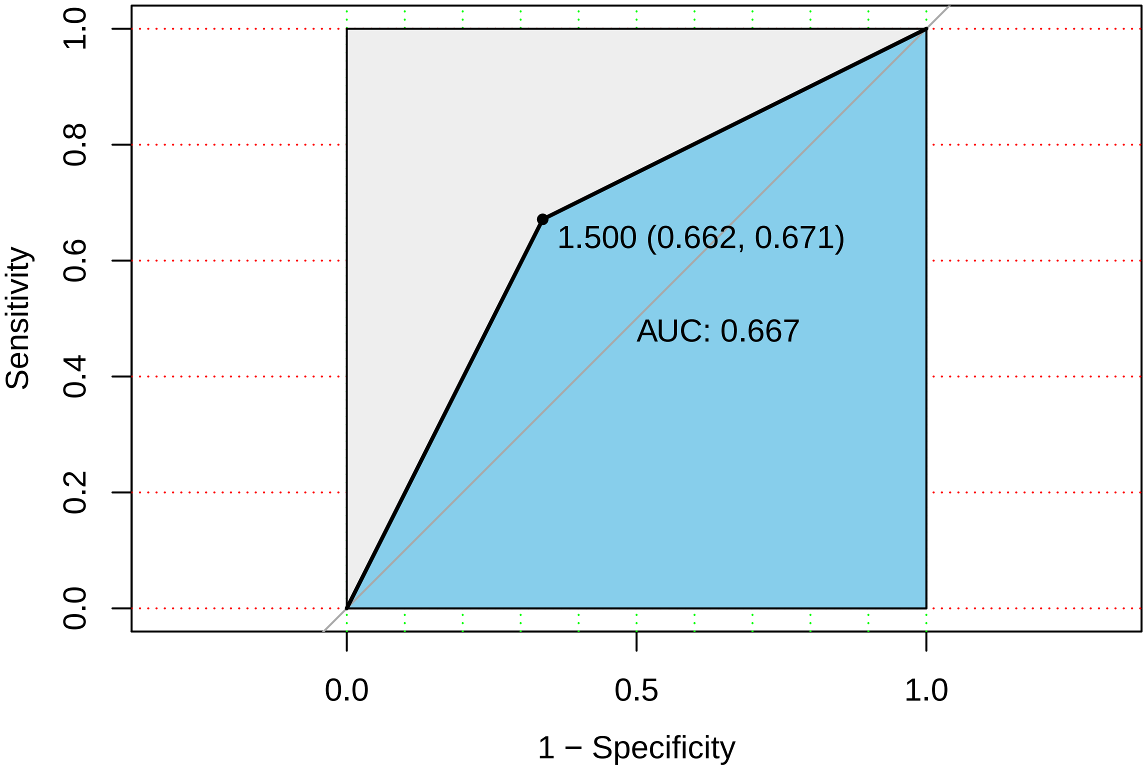
4.4.2.2 ConfusionMatrix

```
##  4  7 11 12 13 14
##  0  1  1  1  0  0
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 4912 2575
##          1 2406 5042
##
##                Accuracy : 0.6665
##                  95% CI : (0.6589, 0.674)
##     No Information Rate : 0.51
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.333
##  Mcnemar's Test P-Value : 0.01729
##
##             Sensitivity : 0.6712
##             Specificity : 0.6619
##          Pos Pred Value : 0.6561
##          Neg Pred Value : 0.6770
##              Prevalence : 0.4900
##          Detection Rate : 0.3289
##    Detection Prevalence : 0.5013
##       Balanced Accuracy : 0.6666
##
##        'Positive' Class : 0
##
```

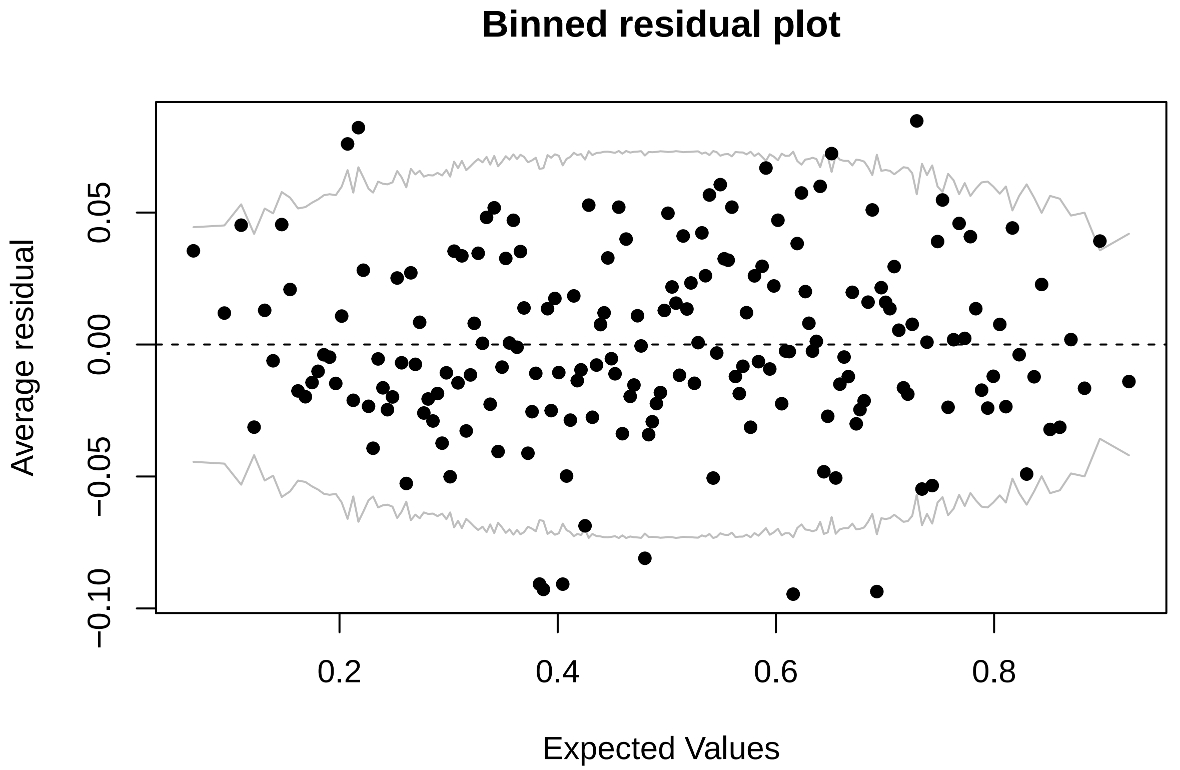
From the confusion matrix,we can find that our model can do good job on classifications. The true positve rate is 67%,true negative rate is 67%. The prediction performance of multilevel logistic regression model is worse than logistic model.

4.4.2.3 ROC and AUC



The AUC is 0.676 which is greater than 0.5, and it means our model have some classification abilities. Besides, the shape of ROC graph implies this model can do some predictions.

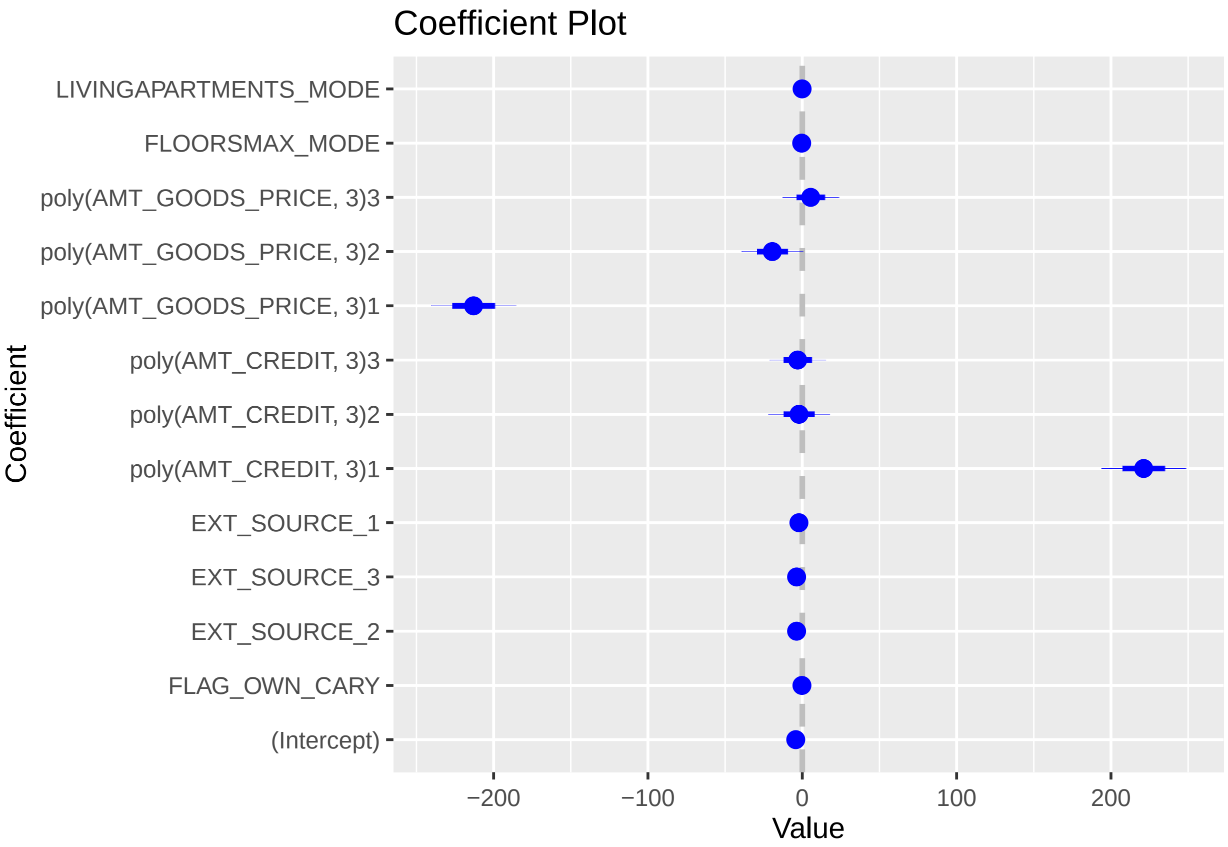
4.4.2.4 Binned residual plot



And we can see the binned residual plot is good because the residuals are almost inside the two lines. And these residuals are evenly distributed.

4.4.2.5 CoefPlot

```
coefplot.rxLogit(m2)
```



We can find that the coefficient plot informs us the model does not fit well.

4.5 Current Conclusion

When we look back these two models, it seems like that there are only little differences exsisting between logistic and multilevel logistic models. The logistic regression model has much higher prediction accuracy than multilevel logistic model. However, the ICC model tells me that the multilevel logistic model can make some sense, and the AUC of multilevel logistic model is larger than logistic model. So I think may be m2: multilevel logistic regression model can be a better model than m1: logistic regression model.

5 Discussion

5.1 Implication

This project gives me lots of inspirations. Firstly, EDA part is essential. When we do the EDA, we can have clear understandings of our data. Besides, it can help us to know how to clean our data and benefit our features extractions. Secondly, it teaches me how to deal with unbalanced data. Biased data can bring great difficulties to fit model and affect the accuracy of our model. So if we find our data is biased during the process of EDA, it is crucial to resampling. Thirdly, the preprocess of big data triggers my passion for studying the related subjects in next semster. And the big data requires us to handle carefully. Finally, the enthusiasm is critical, and we should do our best to use all study sources to improve us.

5.2 Limitations

There is no doubt that the project has limitations. Firstly the processing of missing data may be not the proper way. Secondly, there are still many variables in raw data not being used. Because the limitations of the computer, the variables and the observations are not fully utilized which can have influences on the accuracy of a model. Thirdly, I do not take non-linear regression model into considerations. I think that the linear regression model may be the best way to solve this problem.

5.3 Future Direction

In the future, I will have a try non-linear regression model such as KNN to interpret this data. And I will try another way to fill the missing data. To enhance the accuracy of my model, I plan to introduce more variables and observations into the model fitting process. Maybe I can use python next semester to remodel this data.

6. Referrence

Firstly, thank Professor Masanao for assistance with model fitting and the great lectures in this semester.Your feedback benifits me a lot.

Besides,I would also like to show my gratitude to so many outstanding data scientists for sharing their pearls of wisdom. Their kindness really

At last, I should restate that data used in this project comes from kaggle provided by HomeCredit Company. I will only use this data for nonprofit research such as final project.