HOME CREDIT

Team 5 Weifu Shi, Xiaorui Shen, Sizhe Fan, Yinghui Wei, Hang Zhang

Credit Default Analysis

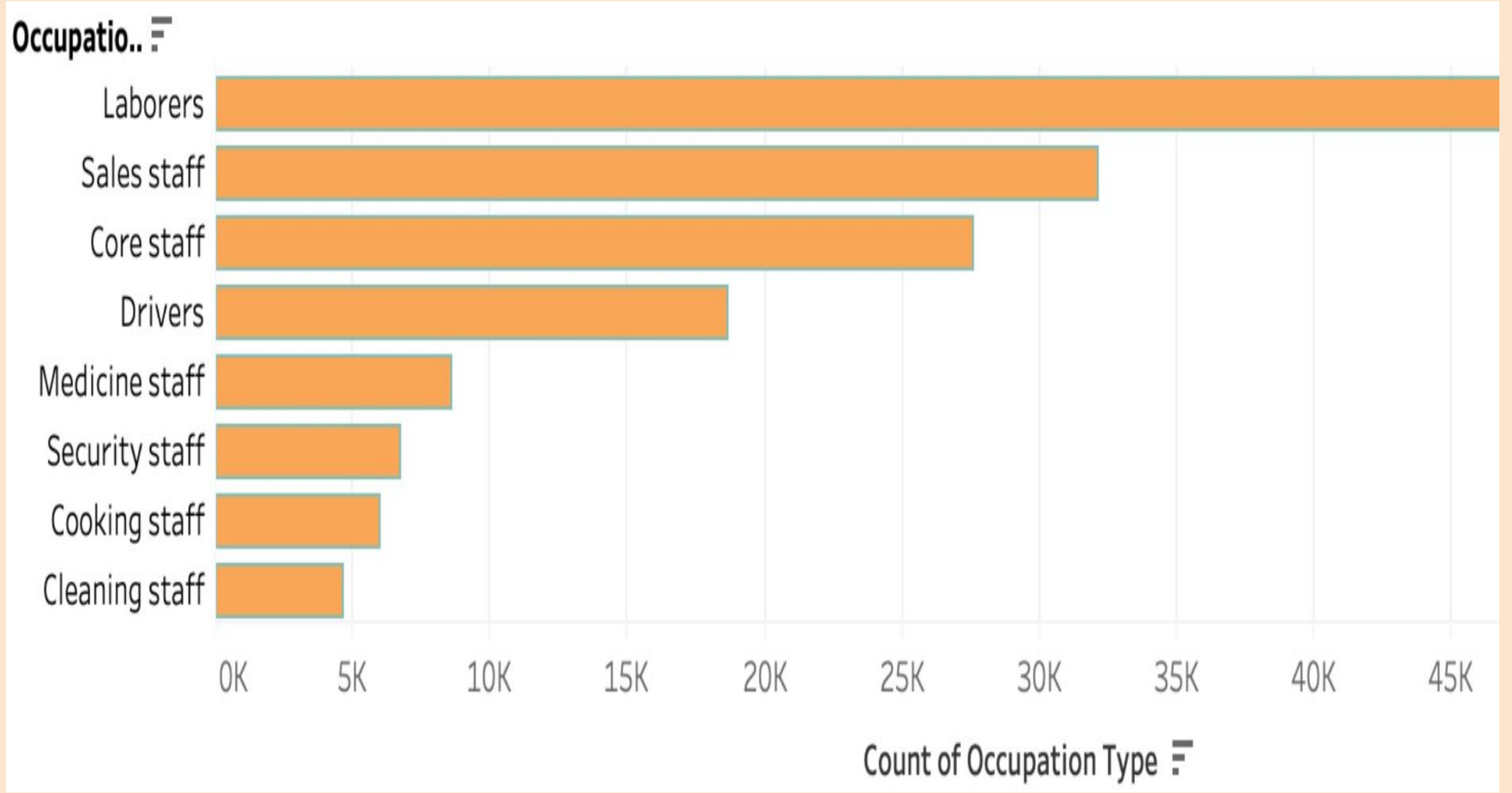
MSBA Cohort A Team 5

Weifu Shi, Sizhe Fan, Xiaorui Shen, Yinghui Wei, Hang Zhang

Abstract

The main purpose of our project is to help HomeCredit Group to better evaluate the paying capability of customers who have little or no credit history. There are two main goals: the first one is to build a supervised machine learning model to predict whether customers have difficulties repaying the loans. The second goal is to use unsupervised machine learning to cluster these customers into different groups. It is extremely important to make sure the efficiency and effectiveness of the lending model so that clients who are capable of repayment are not getting rejected. In addition, the design of our machine learning model can be a reference or case study for other financial institutions and academic institutions to have a glimpse of the credit default algorithms in the banking industry.

Introduction



The dataset we are using is the Home Credit Default Risk retrieved from the Kaggle website. The dataset contains two main types of variables. One is the demographic information of loan applicants such as birthday, gender, number of children, education, income sources, family status, and housing info. The second type is the specific attribute of the loans such as installments, loan types, and annuity etc. The target variable measures the ability of customers to repay the loan. Number 1 means that customers have difficulties in paying, and 0 means customers not. In the training dataset, most customers have no difficulties.

Data Pre-processing

Dropped Columns:

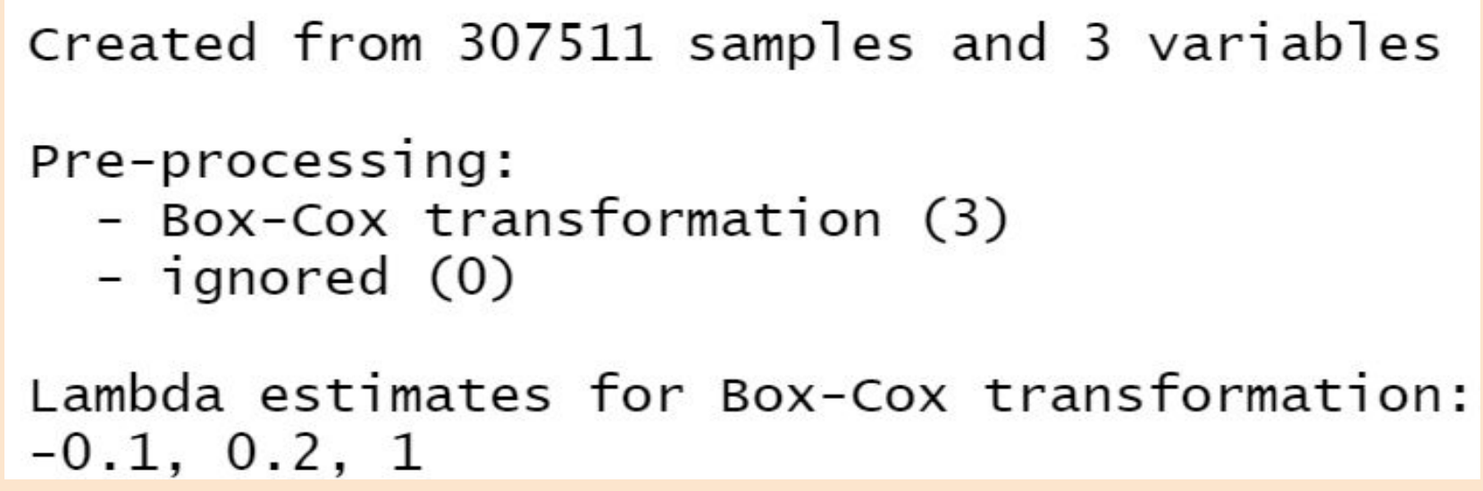
- Missing value > 50%
- 0 variance
- Highly correlated (correlation > 80%)
- Fill left numeric NAs with 0
- Fill left categorical NAs with None

Data imbalance:

- Undersampling data with 1:2 (or 1:1) ratio in train dataset
- Confusion matrix to explore specificity and sensitivity

Data Transformation:

- Box-Cox Transformation: 2 variables are normalized: AMT_INCOME_TOTAL (lambda=-0.1), AMT_CREDIT(lambda=0.2)



```
Created from 307511 samples and 3 variables

Pre-processing:
  - Box-Cox transformation (3)
  - ignored (0)

Lambda estimates for Box-Cox transformation:
-0.1, 0.2, 1
```

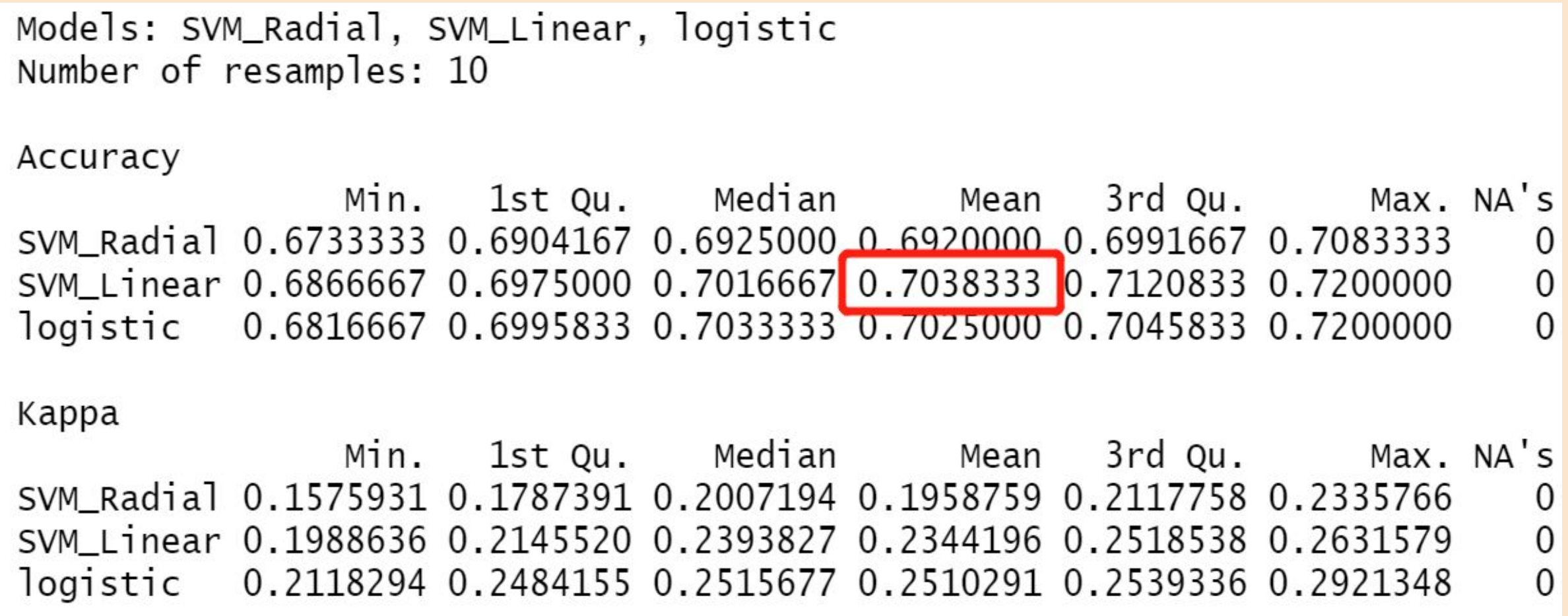
Feature Engineering:

- Combine “documents 2-21” into “Number of Document Submitted.
- Random forest

Supervised Model & Result

Support Vector Machine (accuracy: 70%) vs Logistic Regression (accuracy 68%)

Logistic Regression uses sigmoid function to classify data into 0 and 1 categories, while SVM perform non-linear classification. We compared logistic regression, radial kernel SVM model and gaussian kernel SVM model.

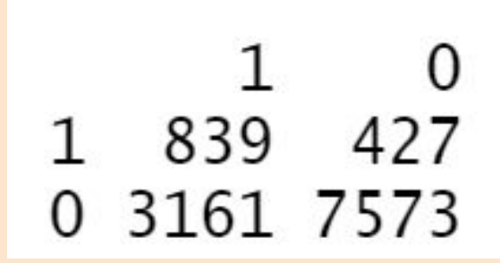


```
Models: SVM_Radial, SVM_Linear, logistic
Number of resamples: 10

Accuracy
                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
SVM_Radial 0.6733333 0.6904167 0.6925000 0.6920000 0.6991667 0.7083333    0
SVM_Linear 0.6866667 0.6975000 0.7016667 0.7038333 0.7120833 0.7200000    0
logistic   0.6816667 0.6995833 0.7033333 0.7025000 0.7045833 0.7200000    0

Kappa
                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
SVM_Radial 0.1575931 0.1787391 0.2007194 0.1958759 0.2117758 0.2335766    0
SVM_Linear 0.1988636 0.2145520 0.2393827 0.2344196 0.2518538 0.2631579    0
logistic   0.2118294 0.2484155 0.2515677 0.2510291 0.2539336 0.2921348    0
```

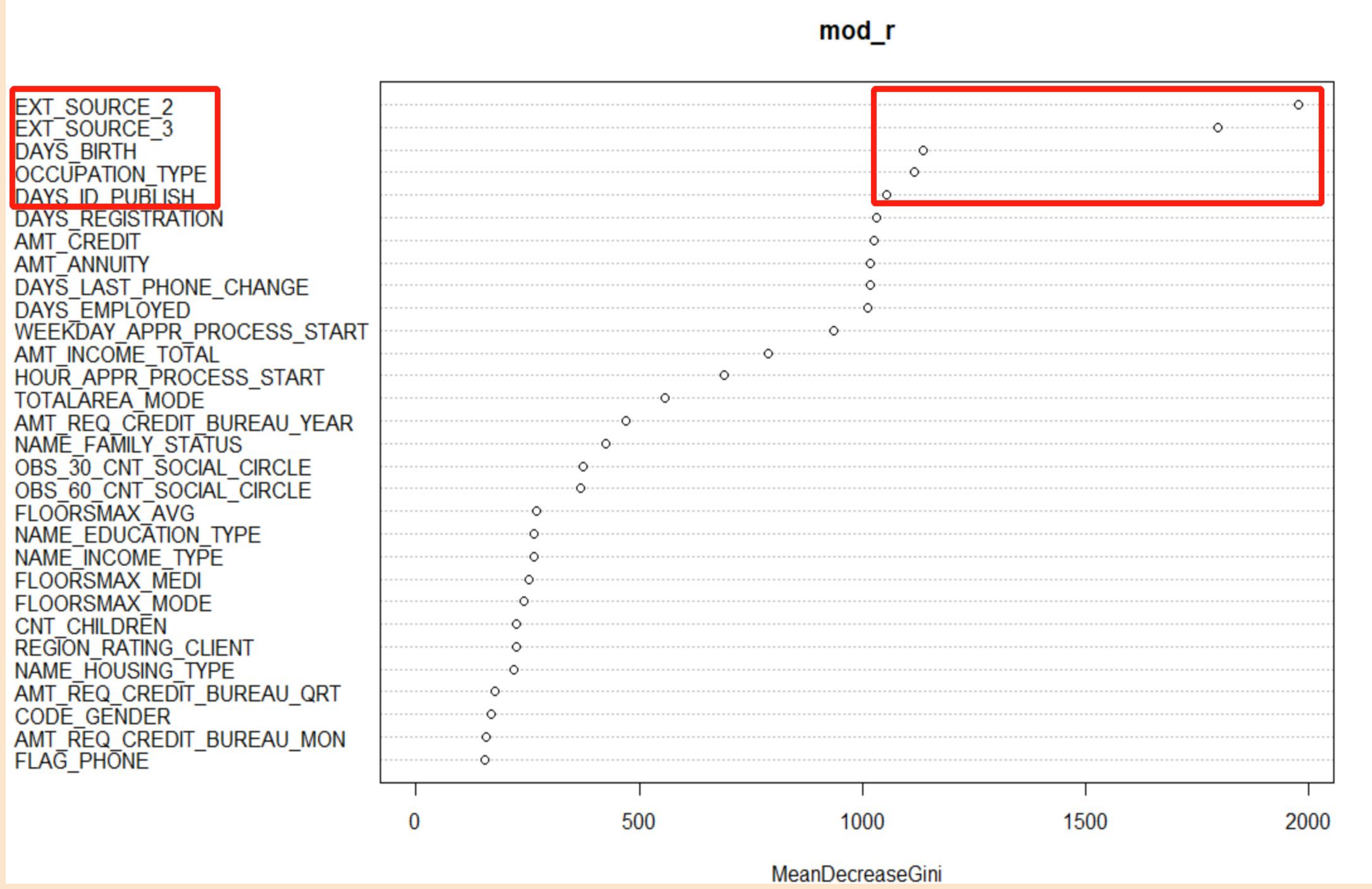
The gaussian kernel SVM model reached an accuracy 70%. with a sensitivity of 21% and specificity of 94.6%.



```
     1    0
1  839  427
0 3161 7573
```

Random Forest Classification (accuracy: 66%)

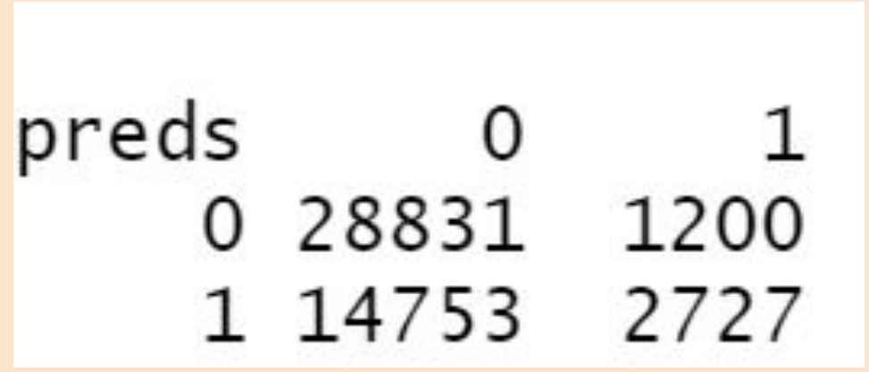
An ensemble tree-based learning algorithm, which aggregates the votes from different decision trees to decided the final class of the test object.



Prediction results for test dataset

- 30031 in class 0
- 17280 in class 1

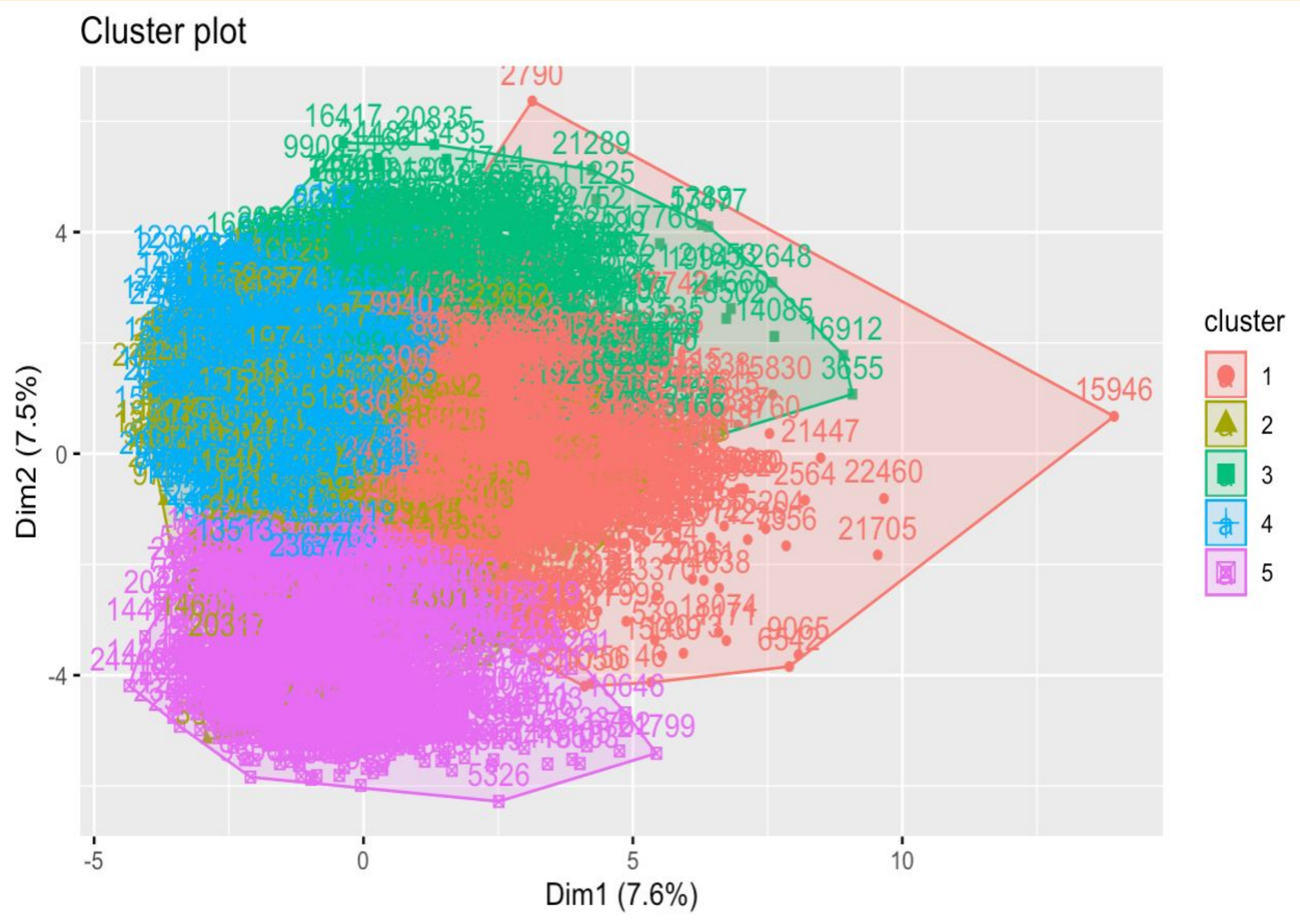
The whole train dataset used for get accuracy

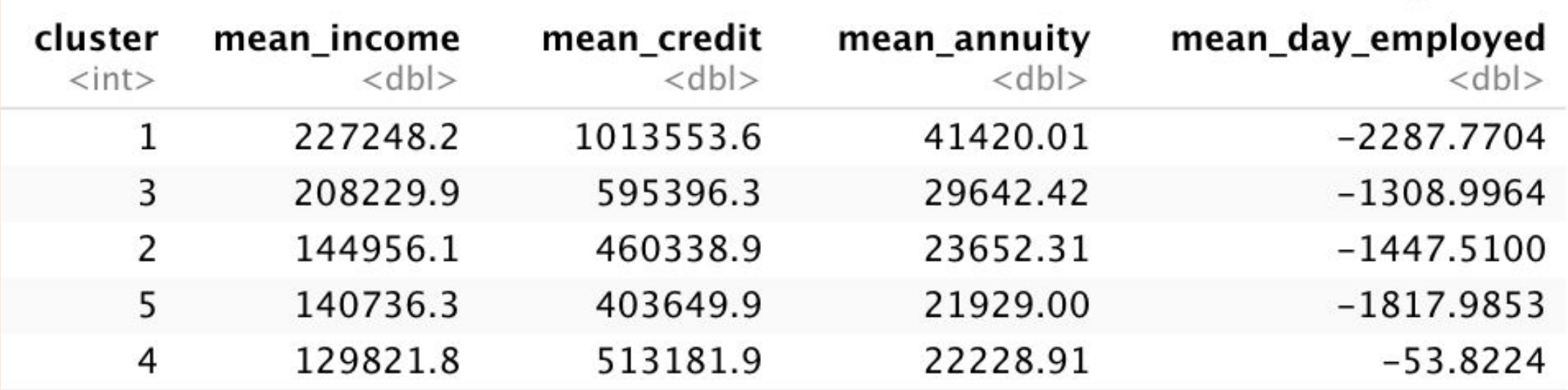


```
preds     0     1
    0 28831  1200
    1 14753  2727
```

Recall(sensitivity): 70%
Specificity: 66%

Customer Segmentation by K-Means





cluster <int>	mean_income <dbl>	mean_credit <dbl>	mean_annuity <dbl>	mean_day_employed <dbl>
1	227248.2	1013553.6	41420.01	-2287.7704
3	208229.9	595396.3	29642.42	-1308.9964
2	144956.1	460338.9	23652.31	-1447.5100
5	140736.3	403649.9	21929.00	-1817.9853
4	129821.8	513181.9	22228.91	-53.8224

We randomly sample 10,000 customers and implement K-Means unsupervised ML algorithms. In terms of Silhouette and WSS method, The optimal number of cluster would be five. As shown above, HomeCredit Group should set different levels of alarm call. For example, Cluster 3 is considered as high risk group and should be alerted immediately: Old age, low income and low employment day.

Conclusion

We finally decided to keep two supervised models and one unsupervised model. They have their own strengths in different directions. We hope to use these method to achieve the highest accuracy and best adaptability. In the future steps, we hope to build a final model that integrates all the advantages of the model to achieve better results

Acknowledgements

Professor: Mohammad Soltanieh-ha (MSBA Boston University)

Dataset Source:
https://www.kaggle.com/c/home-credit-default-risk/data