MAFS6010Z Mini-Project 1: Kaggle Home Credit Competition

Ziyi WANG[1] and Jixiang XIANG(20568852)[1] {zwangbn,yxiangae}@connect.ust.hk
[1]: Department of Mathematics, HKUST

1. Introduction

We have explored the base dataset of Home Credit Default Risk competition dataset. We did a detailed data cleaning, feature selection and built a prediction model. We noticed that the external resource score and days since when the applicant changed their application are significant to the model.

2. Data Cleaning and Feature Engineering

Data Overview

- ➢ The dataset we used is application_train.csv and application_test.csv. We combined the two dataset and separate in a later stage while training
- ➢ There are 356k rows in the dataset and 220 columns in the beginning
- ➢ After the below cleaning and enrichment, we stillhave 220 columns which resulted by:
    - ➢ Having 52 columns removed because of missing values
    - ➢ Having 52 columns added as enrichment features

Feature Engineering

- ➢ Average over multiple features are added, such as the average score of 3 external sources scores
- ➢ Sin-cos transformation are applied on the cyclic features such as *WEEKDAY_APPR_PROCESS_START*
- ➢ One hot encoder is applied on categorical features

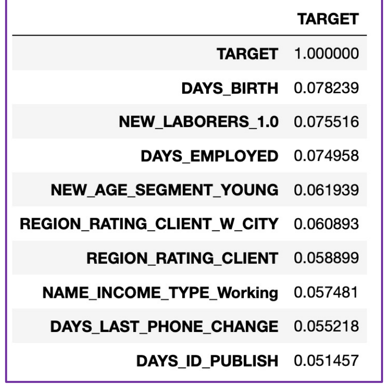
Missing Value Handling

- ➢ We dropped the columns with more than 80% missing values
- ➢ For some categorical columns, we changed some extreme value to a constant, to remove the outlier. For example, *days_unemployed* = 365243, then we change it to Null.

3. Feature Exploration

In order to understand the features before training the model, we looked at the top 10 correlated features with the target.

The features appears to be less obviously correlated with the target, which indicates that we need to use multiple features to achieve the prediction purpose.

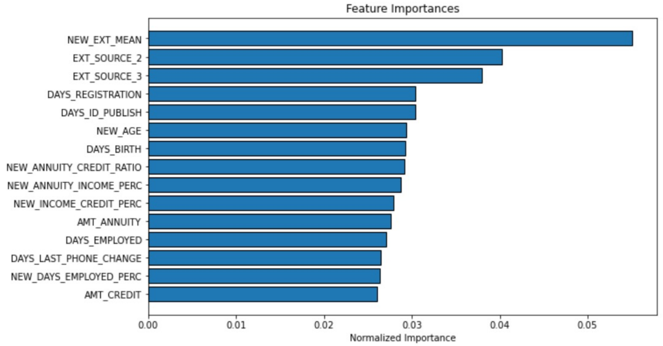


	TARGET
TARGET	1.000000
DAYS_BIRTH	0.078239
NEW_LABORERS_1.0	0.075516
DAYS_EMPLOYED	0.074958
NEW_AGE_SEGMENT_YOUNG	0.061939
REGION_RATING_CLIENT_W_CITY	0.060893
REGION_RATING_CLIENT	0.058899
NAME_INCOME_TYPE_Working	0.057481
DAYS_LAST_PHONE_CHANGE	0.055218
DAYS_ID_PUBLISH	0.051457

Top 10 features that are linearly correlated to the target

4. Feature Selection

We first train the model with all features and get a baseline model to predict the credit risk. By plotting out the importance of the feature used in this model, we can find 10 features which contribute more to the model.



5. Deal with Unbalanced Data

With two different classes of target (at-risk and no-risk), the training data is biased with a ratio of almost 10:1 no-risk v.s. at-risk. Therefore, the prediction on at-risk classes will be highly deviated if we directly use all the training data without dealing with the unbiased distribution.

With the base line accuracy being 92% and the balanced accuracy being 50.2%, we tried two methods to deal with the bias:

1. SMOTE. The returned accuracy is 80% and the balanced accuracy is 58.9%
2. Under Sampling. The returned accuracy is 66.7% and the balanced accuracy is 66.4%

6. Conclusion

Due to the biased nature of training data, although the accuracy of the classification model can reach at 92% accuracy even though we do nothing to the training data, the balanced accuracy is nearly random guess. And for a unbalanced dataset, balanced accuracy ((TPR + TNR)/2) is far more important.

Choosing only the highly correlated features will increase the balanced accuracy from 50.2% to 50.7%. Based on the selected features, using SMOTE will increase the balanced accuracy from 50.7% to 58.9% while using under sampling can increase the balanced accuracy from 58.9% to 66.4%.

Hence, using only features with high correlation and using under sampling will help build the random forest classification model.

7. Contribution

Credit Risk Dataset

- ➢ Kaggle Competition