MAFS6010Z Artificial Intelligence in Fintech Project 1

CHAN Koon Lam 20748995

LAM Chung Wai 20430732

TANG Tsz Hong 20735194

1 Kaggle Contest: Home Credit Default Risk

As a warm-up task about machine learning, we focused on the "Home Credit Default Risk" problem on Kaggle. (`https://www.kaggle.com/c/home-credit-default-risk/`) Home Credit is an international financial institution that specialized in consumer loans and personal loans. In the Kaggle contest, we are trying to predict if the applicant is the company's target with 8 data files and more than 120 predictors.

The following article includes Data Cleaning, Data Visualization, Modelling with different techniques, and evaluation with ROC, AUC. We obtained 0.739 score on Kaggle contest. The proof is in Appendix II.

2 Data Preprocessing

2.1 Data Description

In the problem, there are 8 different data files.

- application_train.csv & application_test.csv

  This is the main data set. The shape of the data frame is (307507, 122). It contains the output "TARGET", the ID of loan, and 120 predictors that are related to the personal information of the client.

  details of some important columns:
  "TARGET"- it means client with payment difficulties, the definition of late payment is he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample.

- bureau.csv & bureau_balance.csv

  This is the previous credits records in the credit bureau. In these dataset, we can know about the previous credit record of client and the current status of the Credit Bureau loan.

  details of some important columns:
  "STATUS" - C means closed, X means status unknown, 0 means no DPD (days past due), 1 means 1-30 DPD, 2 means 31-60 DPD, ... , 5 means 120+ or sold or written off DPD

- previous_application.csv
  credit_card_balance.csv & POS_CASH_balance.csv & installments_payments.csv

  This is the previous credit record in Home credit. "NAME_CONTRACT_TYPE" subdivides the record into three types, credit card, POS cash and instalments payments. More record details are stored in the corresponding csv files.

2.2 Data Cleaning and Interpretation

Dealing with NaN values

Some data contain NaN values. There are several approaches, we can either drop the data, drop the predictor, or use an appropriate approach e.g., use 0 to fill the values back and then add columns to indicate the existence of missing values (i.e. add a column “COMMANAREA_NA” to indicate whether the “COMMANAREA” has missing value).

We plot the NaN values in the application train, slot them in descending order of percentage. The predictive power of these features would be low.

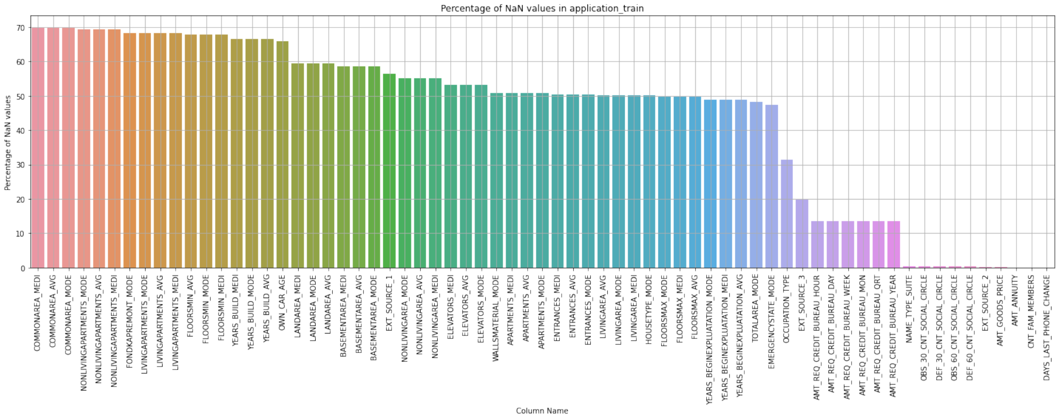


Figure 1: Percentage of NaN values in application train

Features

We also tried to visualize the data, to see if there exists some important insight. By Figure 2, we understand the outputs are mostly 0, we have to avoid the model generating all 0 as prediction. By Figure 3, it shows how different predictors affect the TARGET. We are guessing days_birth, ext_source_1, ext_source_2 , ext_source_3 are important features to the model, as they have significantly different curve on different TARGET output.

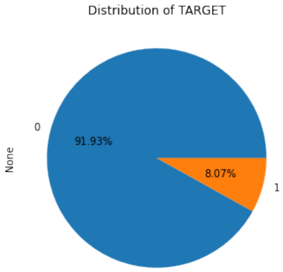


Figure 2: Distribution of TARGET

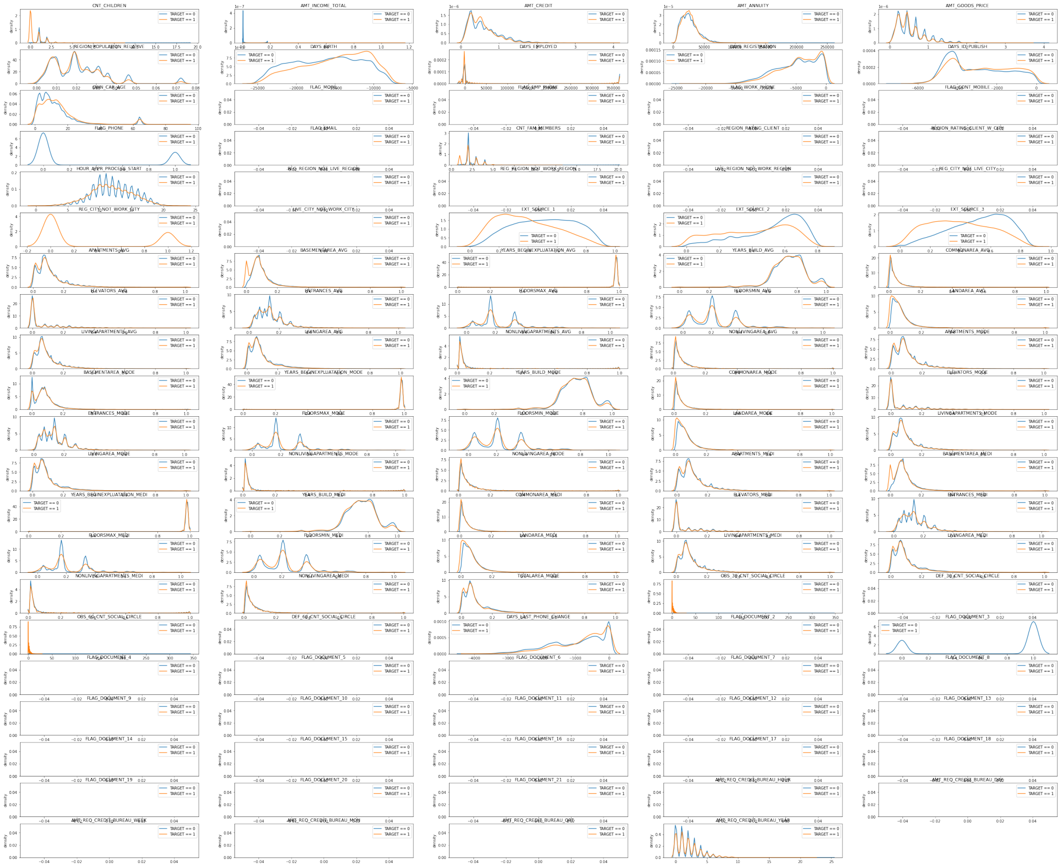


Figure 3: The plot of how different predictors affecting the TARGET. (zoom in to see details).

Oversampling

We have used oversampling on the training data to duplicate examples in the minority class. This is because the dataset is severely imbalanced which could cause the algorithm to ignore the minority class. This is effective in iterative learning models such as xgBoost and random forests.

3 Models and Evaluations

3.1 Models

As the output "TARGET" is a binary output, we have tried different classification approaches and compared their performance. We have fitted classification algorithms like Logistic Regression, Decision Tree, Random Forest, Linear Discriminant Analysis, K-nearest neighbours, Gradient Boosting; also tested 2 commonly used gradient boosting frameworks in machine learning community, XGBoost and LightGBM.

The methodology is as follows. The dataset application_train.csv is divided into training set and validation set. Then we fit the training set into different models. Finally, we will choose the best model, using the data from application_test.csv to get predictions, and upload it to Kaggle to get the Kaggle score. By the below result, XGBoost performs the best among all.

Table 1: models

Name	Computation Time	AUC (in validation set)
Logistic Regression	8s	0.63
Decision Tree	8s	0.69
Random Forest	1min 3s	0.73
Linear Discriminant Analysis	14s	0.74
K-nearest neighbours	20s	0.56
Gradient Boosting	4min 50s	0.75
XGBoost	50s	0.75
LightGBM	1min 40s	0.74

3.2 ROC AUC

To evaluate the models' performance, we often use the ROC curve. ROC curve (Receiver Operating Characteristic Curve) is the curve that plots True Positive Rate against False Positive Rate. AUC means Area under curve. AUC=0.5 means the model has no discrimination power. We often consider >0.7 AUC as a good model, >0.8 AUC as an excellent model.

The below shows the ROC curve and its AUC of the different models.

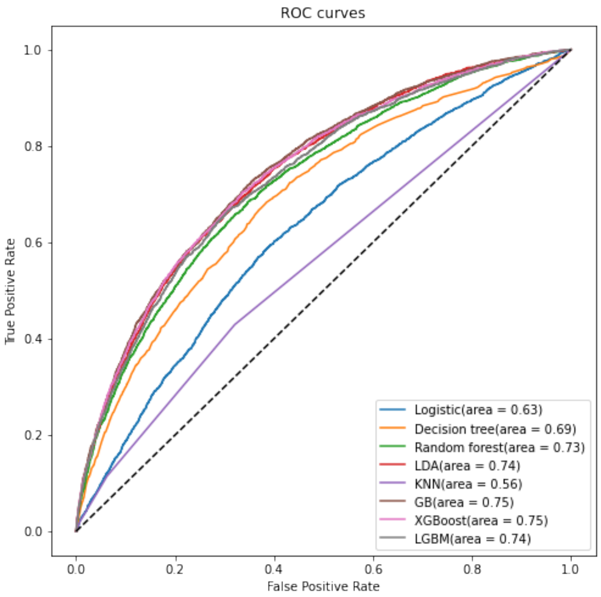


Figure 4: Plot of different ROC curve of different models

4 Conclusion

We believe the XGBoost model can best predict TARGET for this Home Credit Default Risk problem. It has AUC=0.75 in validation set and Kaggle score 0.739.

Because of computation time, we did not include all the sub dataset as the feature of the model. More in-depth investigation can be made in future analysis.

Within the XGBoost model, we found out that the following variables have the highest relative importance in affecting the output. The most important features shown in order are EXT_SOURCE_2, EXT_SOURCE_3, DAYS_BIRTH, DAYS_ID_PUBLISH and AMT_ANNUITY, which has similar insight when we visualize the features.

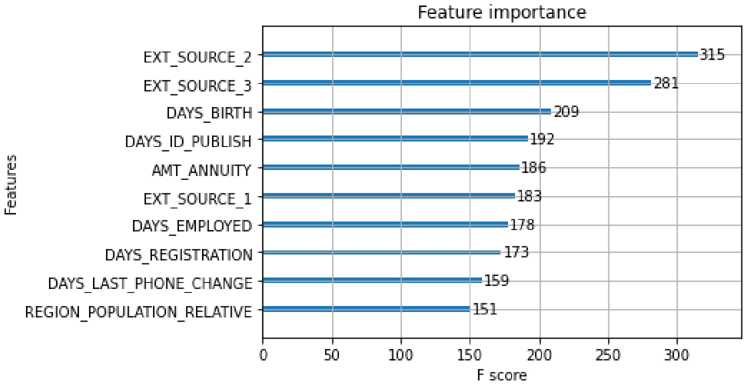


Figure 5: Feature importantce

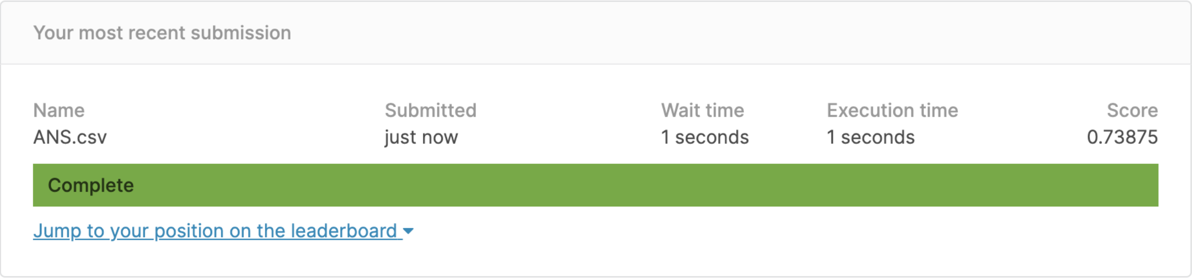
Appendix

Appendix I : Individual's responsibilities

Table 2: Contribution of groupmate

Name	Student ID	Contribution	Duty
CHAN Koon Lam	20748995	33.33%	Data preparation, Modelling, Writing report
LAM Chung Wai	20430732	33.33%	Data preparation, Modelling, Writing report
TANG Tsz Hong	20735194	33.33%	Data preparation, Modelling, Writing report

Appendix II : Kaggle score



Your most recent submission

Name	Submitted	Wait time	Execution time	Score
ANS.csv	just now	1 seconds	1 seconds	0.73875

Complete

Jump to your position on the leaderboard ▾