Predicting Credit Card Defaults

Data

Data was taken from Kaggle

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

API

https://www.fraudlabspro.com/developer

A small amount of data was also collected from an API on credit card default data, but wasn’t used as it was too difficult to merge

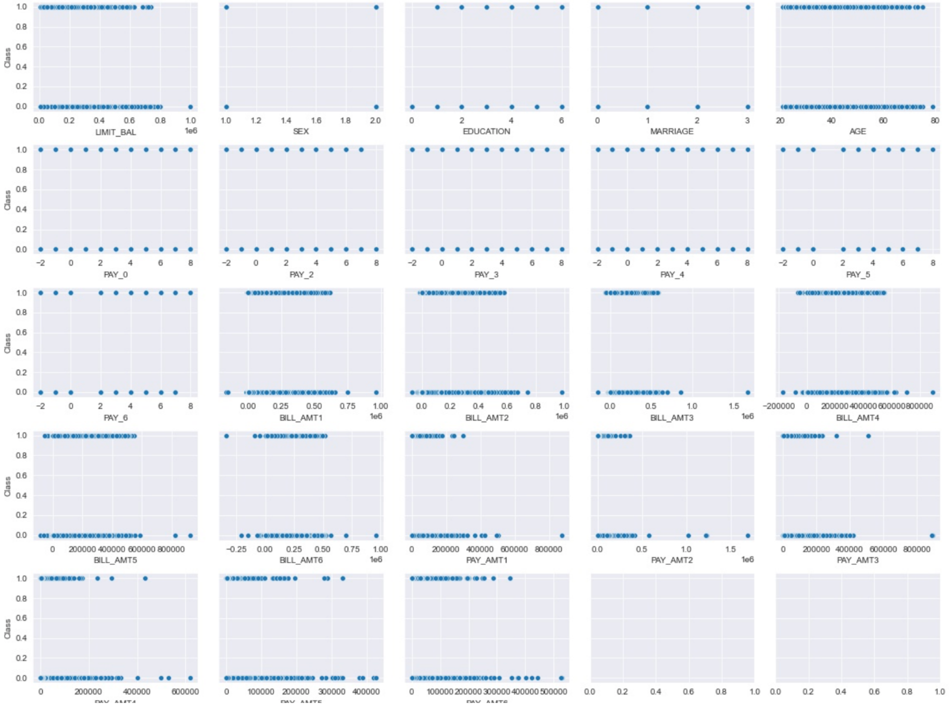
1.

EDA

Working with
our data

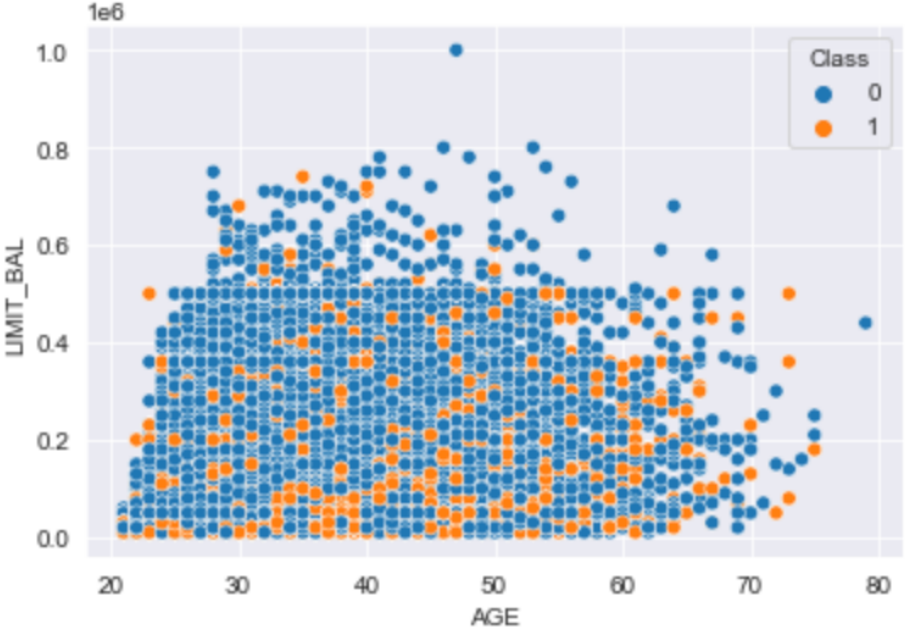
EDA

- Relationship between target and predictor variables



EDA

- Probability a customer defaults based on age and their limit balance on their account

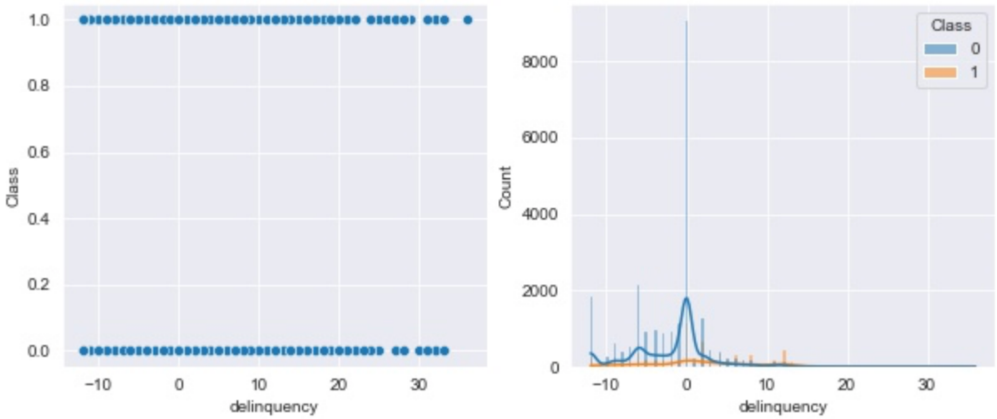


2.

Feature Engineering

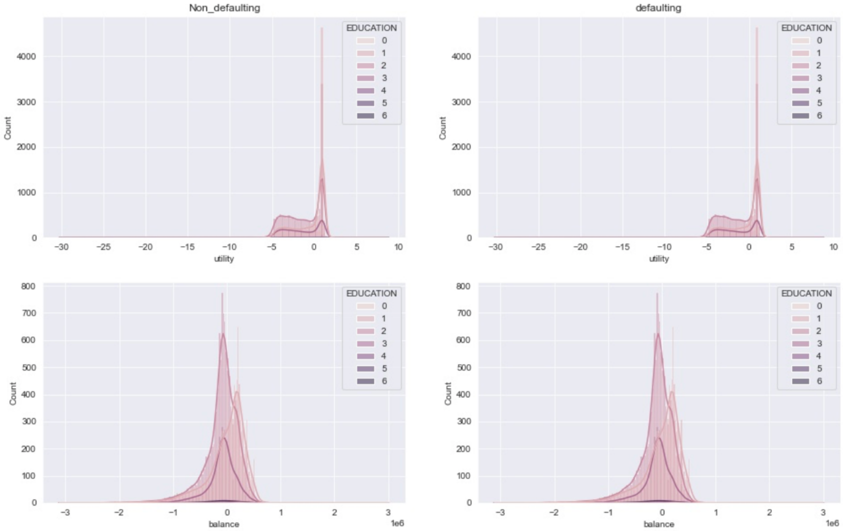
Feature Engineering

- ▣ Delinquency, probability of default, and exposure at default are studied



Feature Engineering

- Distribution of credit utility based on Education between Defaulting and Non-defaulting customers



3.

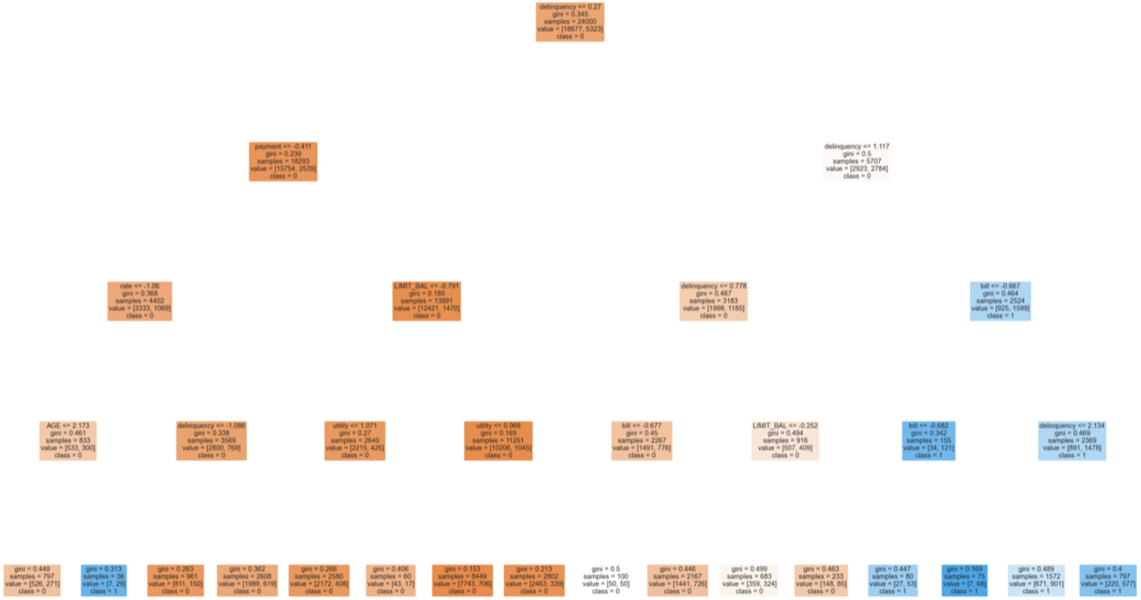
Model Fitting

Model Fitting

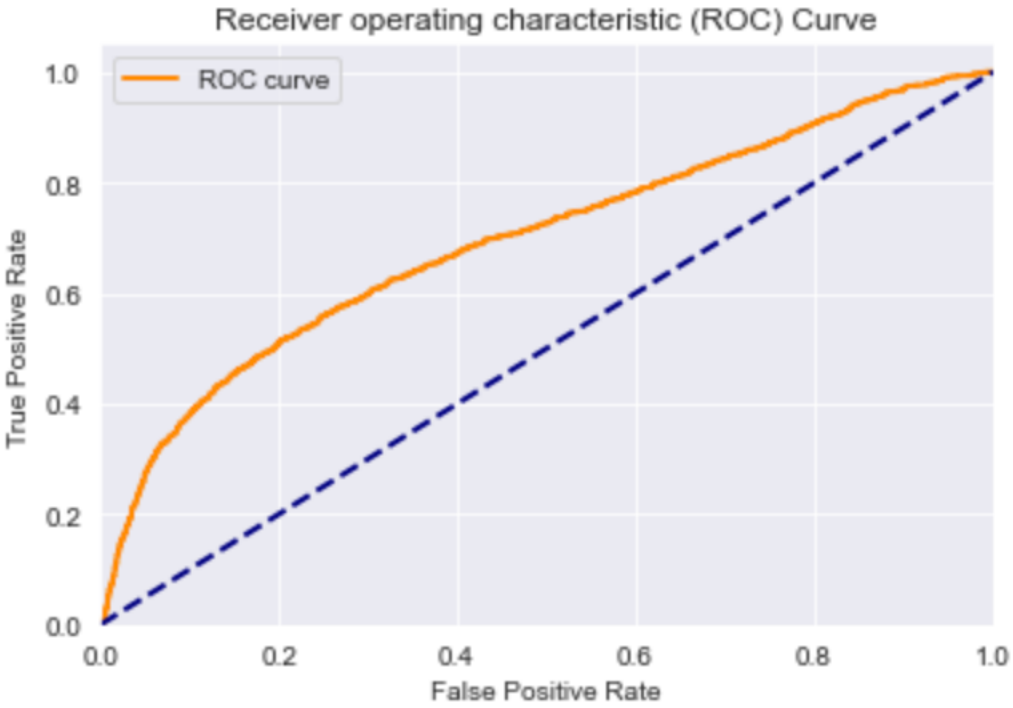
- Multiple models including logistic regression, KNN(k-nearest neighbors), decision tree, random forest, SVM(support vector machine) and gridsearch & XGboost are all tested and compared. Pipeline is used with GridSearch. SMOTE and ADASYN are implemented to deal with the imbalance problem.

Model Fitting

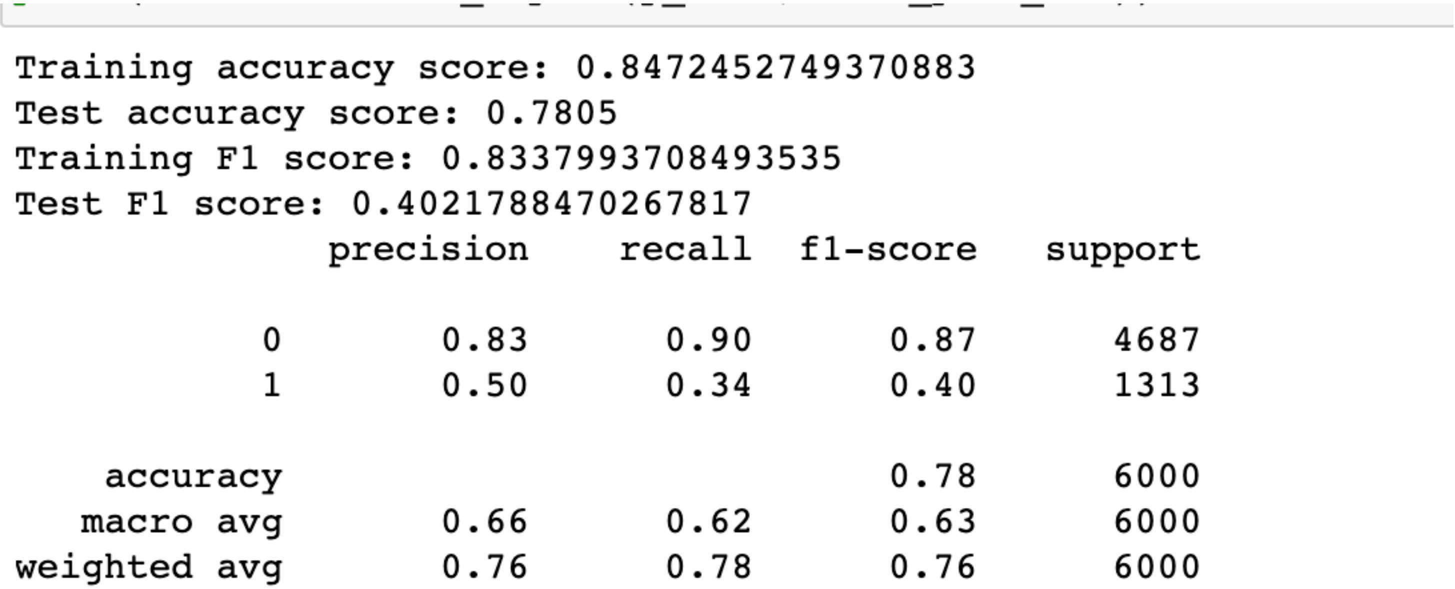
Decision Tree



ROC Curve


Receiver operating characteristic (ROC) Curve
ROC curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Results for KNN



```
Training accuracy score: 0.8472452749370883
Test accuracy score: 0.7805
Training F1 score: 0.8337993708493535
Test F1 score: 0.4021788470267817
              precision    recall  f1-score   support

           0       0.83      0.90      0.87      4687
           1       0.50      0.34      0.40      1313

    accuracy                           0.78      6000
   macro avg       0.66      0.62      0.63      6000
weighted avg       0.76      0.78      0.76      6000
```

Conclusion and next steps

- The feature SEX and EDUCATION have different probability of default payment, according to both the statistical test and model evaluation, which means male/female and different education levels have effects.
- Both continuous variable and categorical variables play important roles in the modeling. Different models mark different strong predictors.
- The credit card default payment problem has a highly imbalanced dataset. Even the data that is processed with SMOTE technique, some metrics still don't not show satisfactory results. Because the real probability of default is unknown, we may implement artificial neural network to accurately estimate the real probability of default.

Thank you!