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

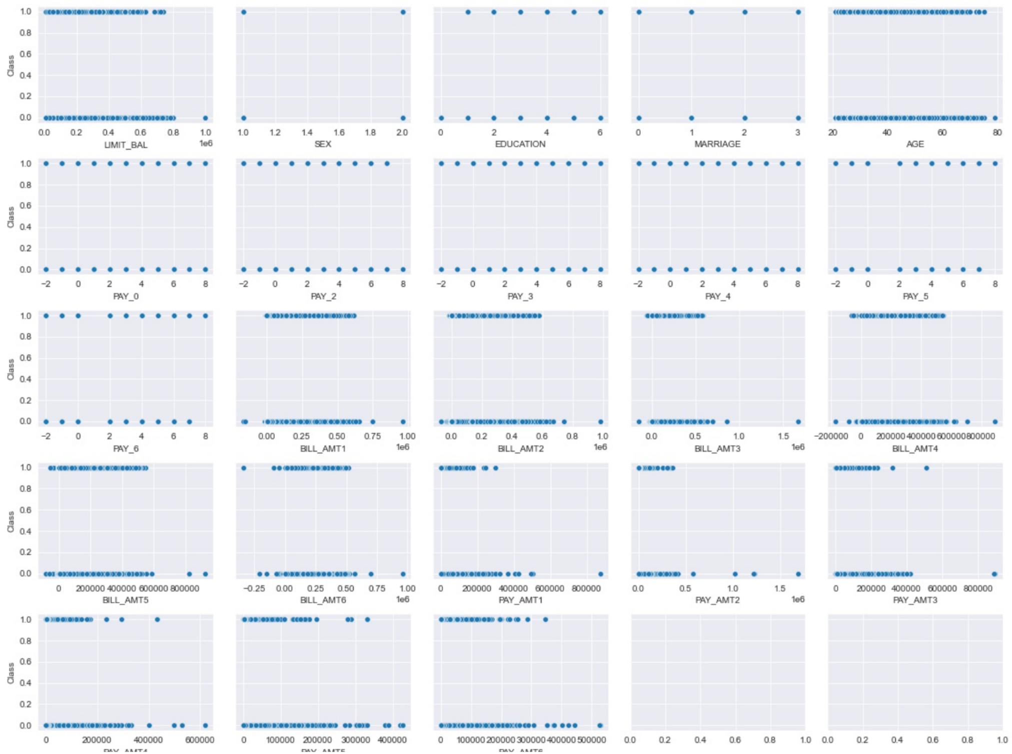
1.

EDA

Working with
our data

EDA

- Relationship between target and predictor variables



EDA

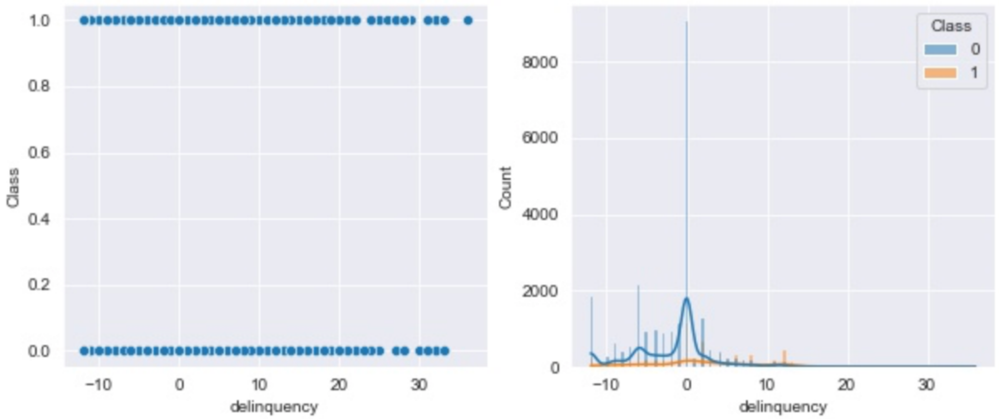
- Probability a customer defaults based on age and their limit balance on their account

2.

Feature Engineering

Feature Engineering

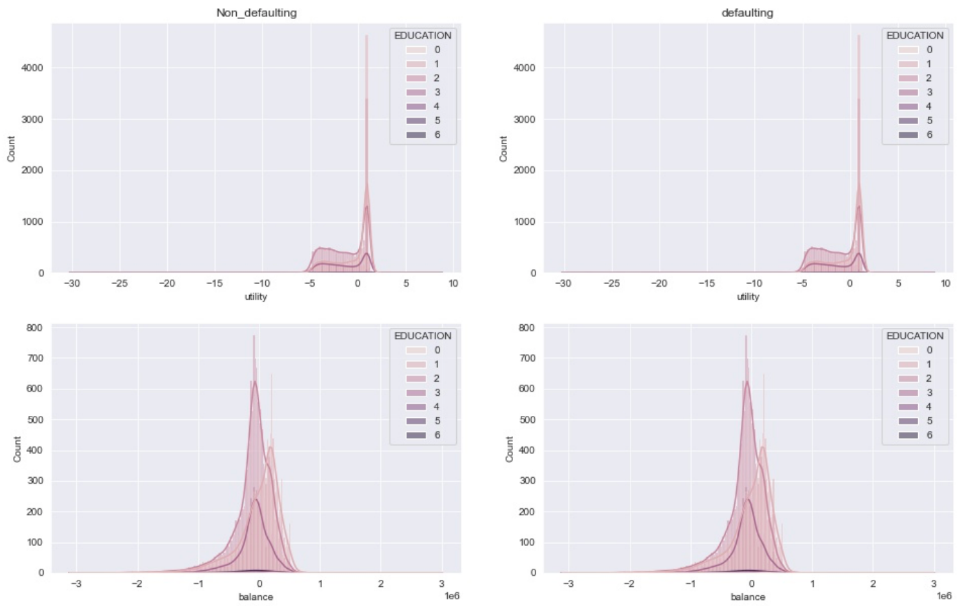
- Delinquency, probability of default, and exposure at default are studied



Feature Engineering

- credit utilization rate
- credit utility

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



3.

Model Fitting

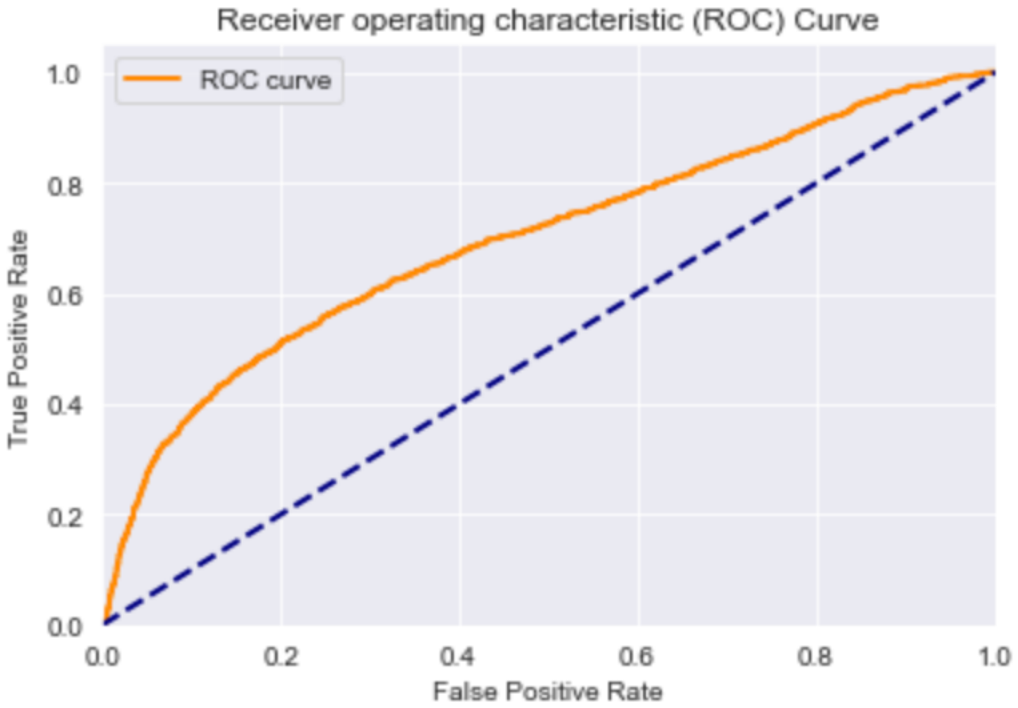
Model Fitting

- 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

Results for XGBoost

```
Training accuracy score: 0.766986132676554
Test accuracy score: 0.743
Training F1 score: 0.7637350705754615
Test F1 score: 0.5120253164556962
              precision    recall  f1-score   support

           0       0.88      0.78      0.83      4687
           1       0.44      0.62      0.51      1313

    accuracy                           0.74      6000
   macro avg       0.66      0.70      0.67      6000
weighted avg       0.78      0.74      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 significant effects.
- Both continuous variables 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 is processed with SMOTE technique, some metrics still don’t 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.
- XGboost serves as the best one in all classification models, but it also becomes expensive when we set up more hyperparameters in grid search.

Thank you!