Prediction of Default of Credit Card Clients

Yi Wang
Brown University
12/04/2019

Github: https://github.com/yahowang/Data1030Project

Introduction - Recap

- Problem to solve: target potential customers who will default
- Importance: challenging yet vital to minimize the risk of capital loss by determining whether their customers would pay off their balance in the next billing period.
- Type of problem: classification
- Data Source: UCI Machine Learning Repository

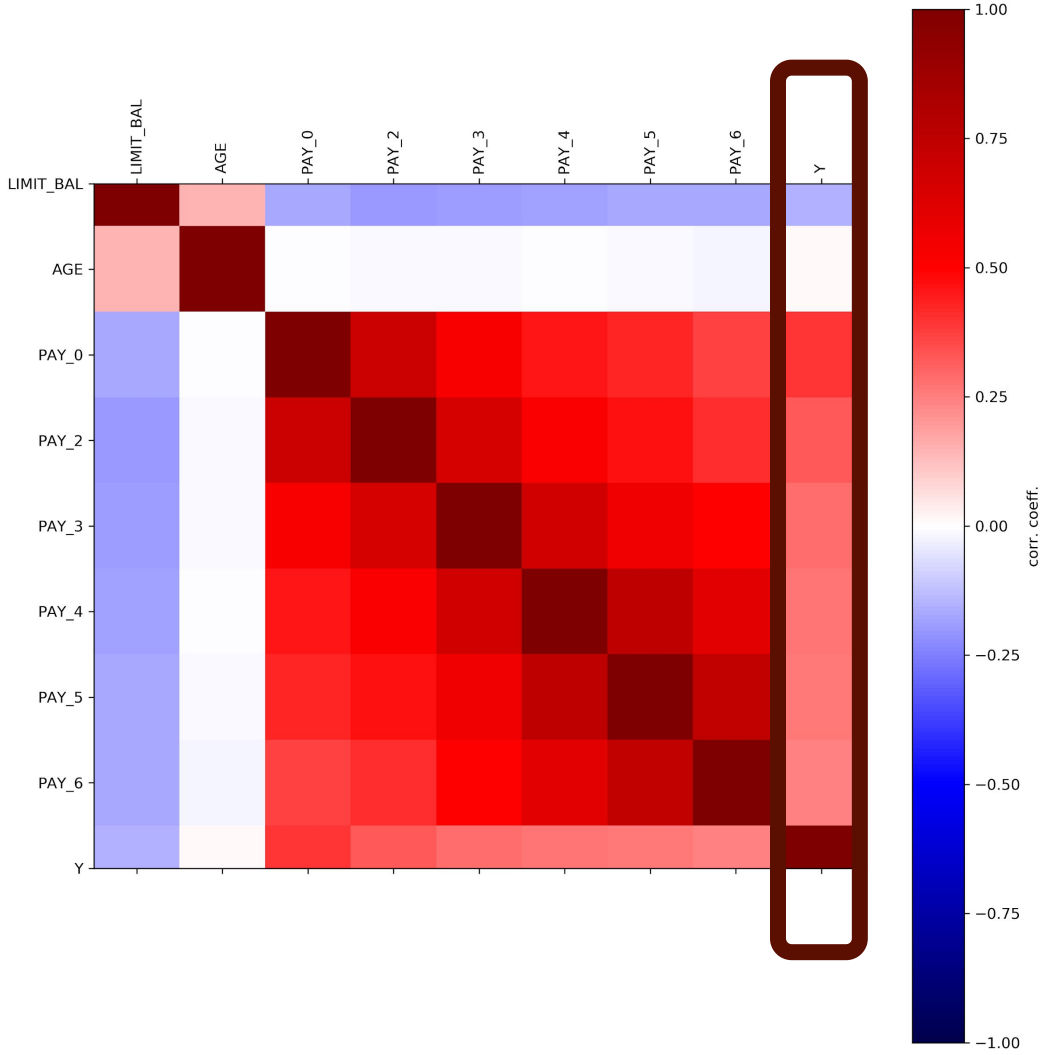
Data Preprocessing - Recap

- 30,000 instances with 24 features
- Imbalanced
- No missing values
- 4 sets of features:
  - Demographical data (One hot & Standard)
    - Gender, Education Level, Age, etc.
  - Monthly billing amount (Standard)
  - Monthly payment amount (Standard)
  - Monthly payment delay status (One hot & Minmax)
- 52 features after preprocessing

EDA - Recap: Correlation Matrix

Key points:

- More amount of given credit (limited balance) leads to less chance of default.
- The most recent payment delay status matters.



Cross Validation

- Imbalanced Data
  - Need to preserve the proportion of each class when do training/validating/testing

- Stratified K-folds cross validator
  - Based on class label
  - K = 5
  - 20% stratified test set
  - 64% stratified train set in each fold
  - 16% stratified cross validation set in each fold

Cross Validation - Models

9 Random States for each model

	Naive Bayes	Logistic Regression	Random Forest	K Nearest Neighbors
Settings		L1 Penalty 10000 Iteration max	50 inner learners	Distance as weights
Parameter 1		Alpha: logspace(-5, 5, 11)	Max_depth: 1 - 10	N_neighbors: 9 numbers from 100 to 900
Parameter 2			Min_split: 2 - 10	

Results

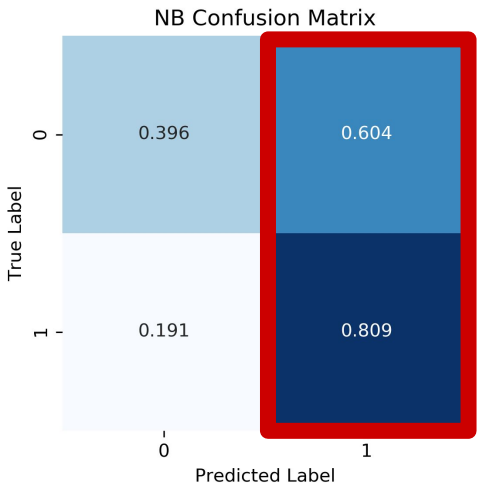
Baseline accuracy: 0.77883

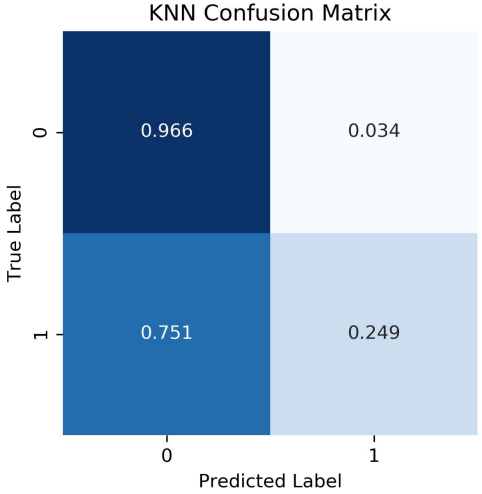
Scoring:

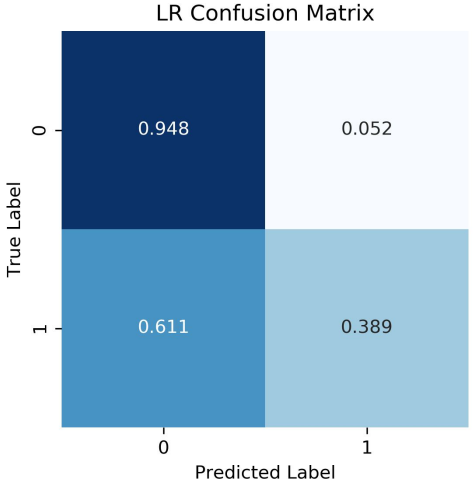
- F1 score within each fold
- Test accuracy score among the F1 chosen models

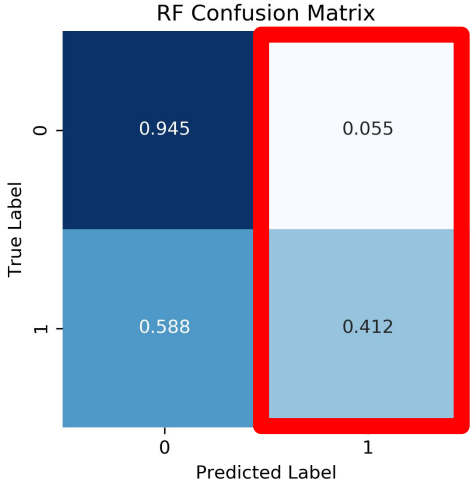
	NB	LR	RF	KNN
Mean	0.41534826991868906	0.8193703703703703	0.8214259259259259	0.8025740740740741
Standard Deviation	0.004831305070268036	0.002490448144730449	0.003100289822285545 6	0.0024483694638575 654
Best Score	0.42287361845266697	0.8241666666666667	0.8271666666666667	0.807
Best Parameter(s)		Alpha: 0.001	Max_depth:8 Min_samples_split: 4	N_neighbors: 200

Confusion Matrix

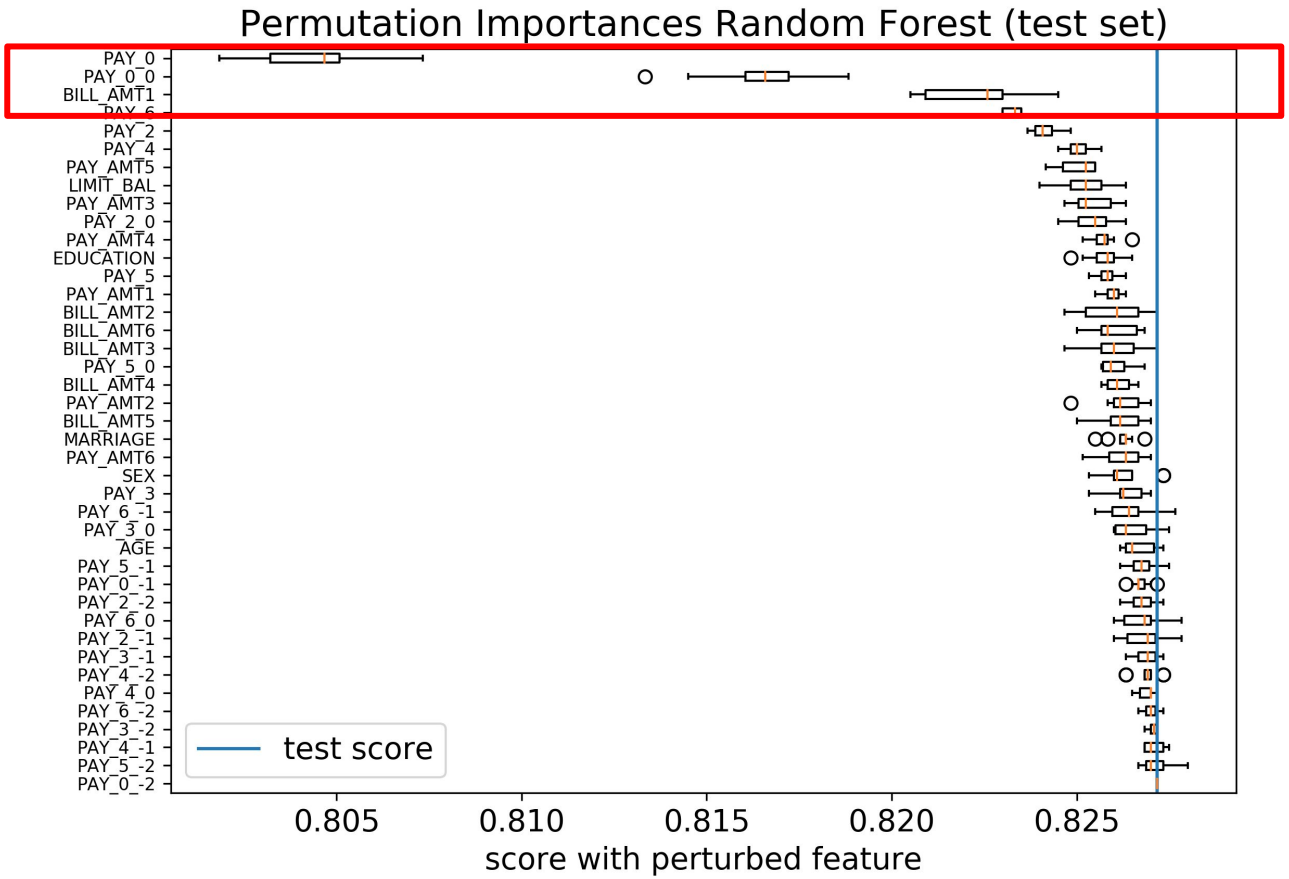








Global Feature Importance



Top 3 Important features:

- Pay_0
- Pay_0_0
- Bill_AMT1

Interpretations:

1. The most recent month delay status matters.

2. The most recent billing amount matters.

Outlook

- More inner estimators might give a slightly better performance in the random forest classifier.

- Feature reduction is another option to reduce the computing complexity while maintaining accurate models.

- Third party credit score (eg. FICO) could be an add-in feature to boost the detection rate and/or accuracy.

Thank You
Q&A