Credit Card Default Predictor

Ruoting Liang, Lichen Ni, Leiling Tao and Yuyi Zhou

Simon Fraser University

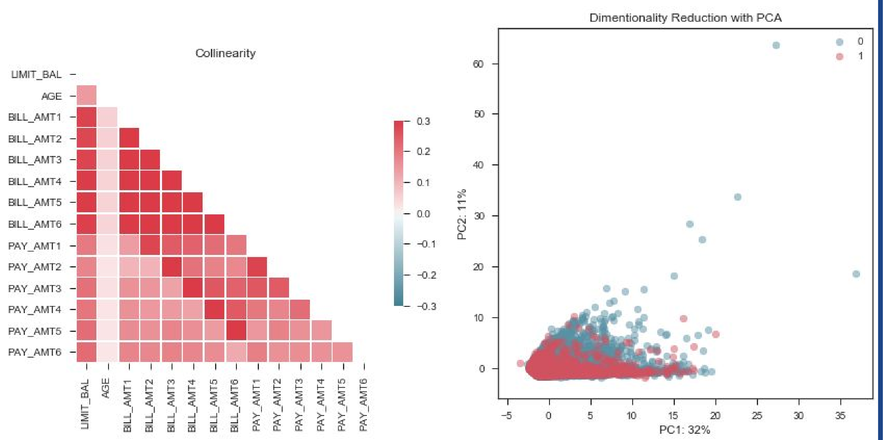
CMPT 726

Overview:

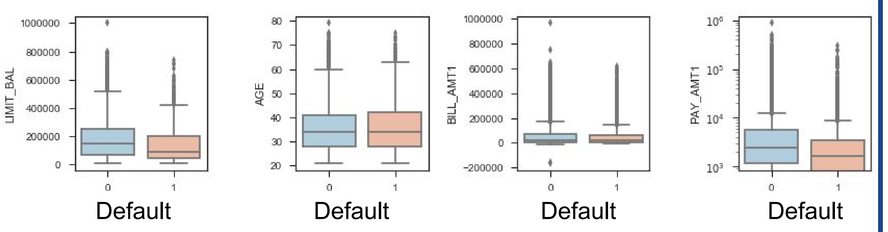
1. Each year, Canadian banks lose millions of dollars due to credit card delinquencies. Successful predictions of credit card default can help with early detection of credit delinquency and greatly decrease potential financial losses.
2. We obtained data from UCI Machine Learning Repository, in which financial information of 30000 individuals were collected. After feature engineering and extraction, we obtained 91 features.
3. To train the models, we used a total of 8 algorithms: logistic regression, decision tree, Naive Bayes classifier, random forest, extreme gradient boosting, SVM, MLP and deep learning. Additionally, we ensembled all models using majority vote.

Exploratory data analysis:

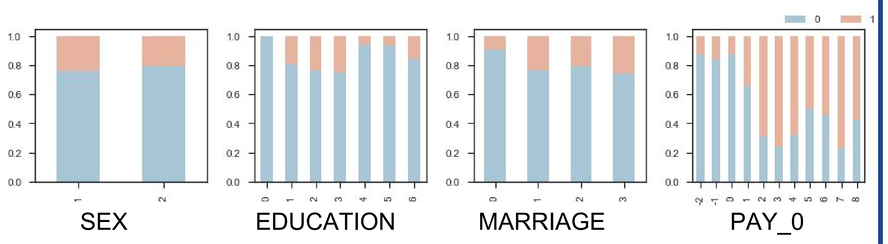
➔ Collinearity and Dimension Reduction:



➔ Continuous Variables:



➔ Categorical Variables:



➔ Strong collinearity exists among features, which may influence the performance of Naive Bayes

➔ The first 2 principal components from PCA only explained 43% variance), suggests the two classes may not be easily separable.

➔ Features that may be important are LIMIT_BAL, PAY_AMT*, EDUCATION, MARRIAGE, PAY_*

Model comparisons:

To deal with imbalanced dataset:

➔ Use F1 score as evaluation metric

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

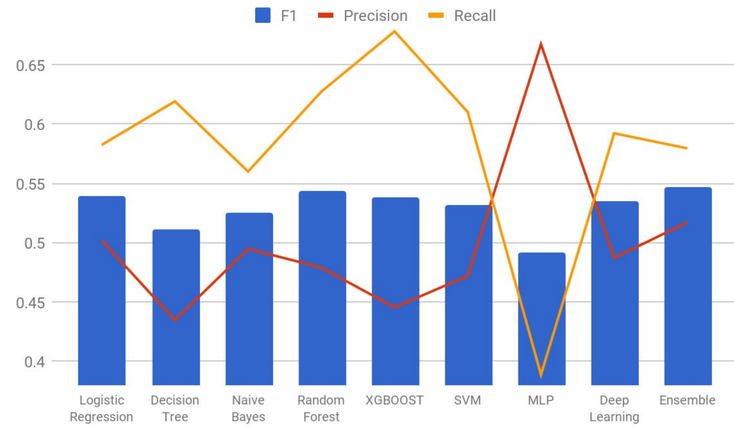
➔ Assign larger weight to minority class

➔ Up-sample minority class

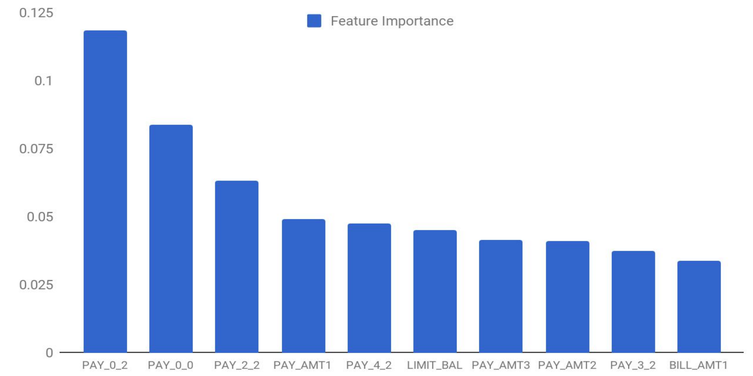
➔ Down-sample majority class

Algorithms	Precision	Recall	F1
Logistic Regression	0.5022	0.5826	0.5394
Decision Tree	0.4349	0.6193	0.5110
Naive Bayes	0.4950	0.5602	0.5256
Random Forest	**0.4792**	**0.6275**	**0.5434**
XGBOOST	0.4457	0.6783	0.5380
SVM	0.4722	0.6103	0.5324
MLP	0.6675	0.3889	0.4915
Deep Learning	0.4874	0.5924	0.5348
Ensemble	**0.5174**	**0.5797**	**0.5467**

Performance summary for all models

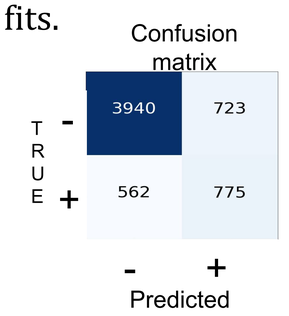


Feature importances of the best single model (Random Forest)



Conclusions:

1. Assigning different weights to samples based on their frequency or re-sampling the data has significantly improved model fits.
2. Using 3-fold cross validation for hyperparameter tuning, our best single model was random forest, under which an F1 score of 0.5434 was obtained.



3. Under the best model, the most important features are payment history and credit limit.
4. A simple ensemble of models using majority count increased model performance.