Comparison of Naive Bayes and Random Forest Classification Algorithms on Credit Card Default Dataset

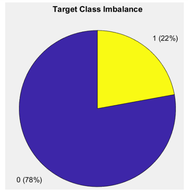
Abin Abraham and Soyeon Park

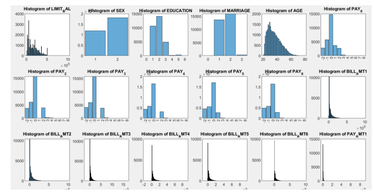
Description and Motivation

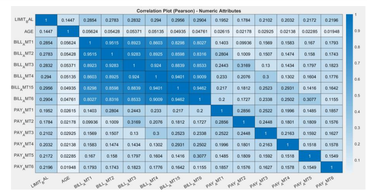
- To predict whether a credit card user will default based on demographics and history of past payment.
- Compare and contrast the result of Naive Bayes and Random Forest algorithms for a binary classification problem.
- Explore hyper-parameter tuning for machine learning algorithms.

Initial Analysis

- Dataset: Default of credit cards in Taiwan dataset from UCI Repository.
- The dataset has default payment (Yes = 1, No = 0), as the response / target variable and 23 features.
- Among those 23 predictors, 9 are categorical and 14 are numeric features, the categorical variables are transformed to categorical types
- The original dataset has 30000 instances with no missing data. The dataset is imbalanced with 22% of defaults and 78% of non-defaults.
- There is high correlation among all the bill amount feature. Limit Balance, age and all bill amounts are skewed.
- The numerical features have been normalized to mean of 0 and variance of 1.







Naive Bayes

A classification algorithm based on applying the Bayes theorem, based on an assumption of independence between all the features.

Pros

- Computationally fast and easy to implement
- Provide simple and efficient approach with clear semantics to represent and use
- It can be theoretically justified for other classifier models that can't apply Bayes theorem[1]

Cons

- Relies on independence assumption and will perform badly if this assumption is not met, meanwhile the absolute independence of features hardly met in real life

Random Forest

A classification and regression algorithm which builds an ensemble of decision trees. It uses bagging and feature randomness when building each individual tree and creates an uncorrelated forest of trees where a prediction is based on majority poll of the tree, which will be more accurate than that of any individual tree[2]

Pros

- Generally outperforms other machine learning algorithms
- Less prone to overfit and performs well with very little tuning required[3]

Cons

- Training a large number of deep trees can have high computational costs (but can be parallelized) and use a lot of memory
- Predictions are slower.

Hypothesis

Random Forest would outperform Naïve Bayes as it doesn't make assumptions of independent features on the underlying data.I.-C. Yeh, C.-h. Lien (2009)[1] has reported a validation accuracy of 0.79 and 0.83 for the NB & RFrespectively on the dataset.

Description of Methodology

Created baseline models with all the predictors and no particular tuning or pre-processing. Improved this with a model with training data after normalization, resampling to balance the target classes, ranking the features based on importance, tuning the hyper-parameters using grid search and kernel distribution for RF and NB respectively.

Choice of Parameters and Results

Naive Bayes

Hyper parameters

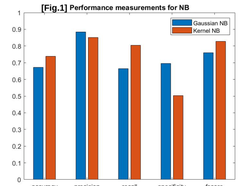
- Train/Test set: 80%/20%
- Using Gaussian Naïve Bayes, apply Kernel Naïve Bayes to increase an accuracy of the model
- Kernel density estimation: the setting of the width 90
- Validation: 10-fold cross validation to minimize overfitting for both Gaussian and Kernel Naïve Bayes

Main experimental Result

- Cross validation estimate of error decreased with Kernel Naïve Bayes (0.26) vs. Gaussian(0.34)
- Kernel method outperformed than Gaussian, the accuracy improved from 0.67 to 0.74 in Kernel, Recall is Increased (0.66 to 0.8), overall F1 Score marked from 0.76 to 0.83 in Kernel Naïve Bayes[Fig.1]

	Accuracy	Precision	Recall	Specificity	F1 score
NB_Gaussian	0.67	0.88	0.66	0.70	0.76
NB_Kernel	0.74	0.85	0.80	0.50	0.83

[Fig.1] Performance measurements for NB



Type	CV Error
NB_Kernel	0.2672
NB_Gaussian	0.3427

Random Forest

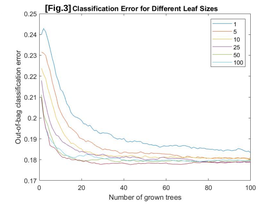
Hyper parameters

- Train/Test set: 80%/20%
- Grid search: trees:80, minleaves:50, predictors:10
- Using Tree Bagger with 80 bagging
- Validation: Out of bag error to validate the model as there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error.

Trees	MinLeaves	Predictors	Accuracy	Precision	Recall	Specificity	F1Score
80	50	10	0.8173	0.8378	0.949	0.3549	0.89
100	1	10	0.7898	0.858	0.8747	0.4917	0.8663
80	50	10	0.8173	0.8378	0.949	0.3549	0.89
1	1	10	0.6888	0.8431	0.7375	0.518	0.7868
80	50	10	0.8173	0.8378	0.949	0.3549	0.89

[Fig.2 Grid search result]

[Fig.3] Classification Error for Different Leaf Sizes



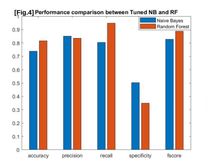
Main experimental Result

- Using grid search for hyper parameter tuning, we found that the best combination of parameter is trees: 80, minleaves:50, predictors:10 for the best accuracy and F1 score of the model. Depending on the objective the hyperparameters are selected[Fig.2]
- According to the out of bag error, the model stabilizes at about 80 trees, while at 100 trees boosting continues to improve, therefore we could choose '80' would be a sufficient number of trees similarly for min leaves thus balancing under/overfitting[Fig.3]

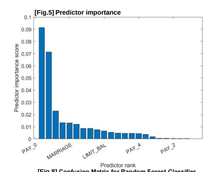
Analysis and Critical Evaluation

- Comparison of the models: Random Forest outperformed verses Kernel Naïve Bayes in terms of accuracy(RF:0.81 vs. NB: 0.74), F1 score(0.89 vs, 0.83) and AUC(ROC curve), which is consistent with the hypothesis[Fig.4,6]
- For Naïve Bayes, Gaussian is one of the most popular techniques for the continuous variables. However, we found that Kernel density estimation generates better results with our dataset in cases where we can't make a certain assumption of normality[4]. Gaussian outperformed on precision and specificity metrics. Kernel estimator would be a good alternative to explore the models and compare the results between the Gaussian Naïve Bayes model[fig.1]
- For Random Forests, we used the OOB samples to generate a different variable importance measure. The top 10 most important features determined using MSMR algorithm is - PAY_0, MARRIAGE, LIMIT_BAL, PAY_4, PAY_2, EDUCATION, PAY_5, PAY_AMT1, PAY_3 and PAY_6. The variable 'PAY_0', which is the most important predictor is a payment history in Sep, 2005 when is just before month of the respons That means just before payment status has high correlation and this has the most contribution to the card payment default[fig.5]
- Accuracy vs. Performance of the models: we should consider not only accuracy, but also focus on other measurements since out original dataset is unbalanced. For example, our dataset is a classification test ('yes' or 'no') and one of the responses is dominant(yes'1': 78%), an accuracy can be skewed as well. In this case, F1 score will be more accurate measure.
- Since detecting the credit card default is a very critical issue in terms of business perspectives. Therefore, the most important thing for the card companies is to catch the defaulters. In our results, high recall means that high prediction rate for detecting defaulters among the real credit card defaulter. In addition, higher precision means less false positive. We think that a good-model for our dataset is balanced with high recall and precision. Therefore, F1 score and ROC curve would be a good measurement besides accuracy for the models[5]

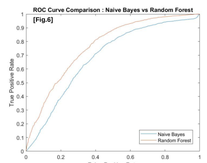
[Fig.4] Performance comparison between Tuned NB and RF



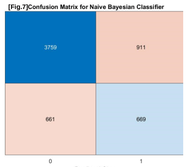
[Fig.5] Predictor importance



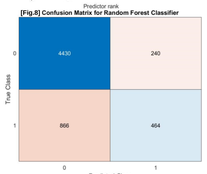
[Fig.6] ROC Curve Comparison : Naive Bayes vs Random Forest



[Fig.7]Confusion Matrix for Naive Bayesian Classifier



[Fig.8] Confusion Matrix for Random Forest Classifier



Lessons Learnt and Future Work

- Random Forest generally performs well compared to Naive Bayes but may lose its advantages sometimes eg specificity.
- Explore imbalance data set re-sampling using ADASYN/ SMOTE for imbalanced data.
- For Naïve Bayes, extend Kernel Naïve Bayes to set the kernel width adaptively and apply other optimizable Naïve Bayes.

References

1. I.-C. Yeh, C.-h. Lien, The comparison of data mining techniques for the predictive accuracy of probablity of default of credit card clients (2009)
2. Breiman, L. Machine Learning (2001) 45: 5. https://doi.org/10.1023/A:1010933404324
3. T,Hastie, R.Tibshirani, J.Friedman, The elements of statistical learning 590p,
4. G.H.John, P,Langley, Estimating continuous distribution in Bayesian classifiers (1995)
5. S.Sharma, V.Mehra, Default Payment Analysis of Credit Card Clients (2018).