Journal of Management Science and Business Intelligence, 2017, 2–2
August. 2017, pages 36-41
doi: 10.5281/zenodo.851527
http://www.ibii-us.org/Journals/JMSBI/
ISBN 2472-9264 (Online), 2472-9256 (Print)

The comparison of machine learning methods to achieve most cost-effective prediction for credit card default

Shantanu Neema[1, *], Benjamin Soibam[1]

[1]Department of Computer Science and Engineering Technology, University of Houston – Downtown

*Email: neemas1@gator.uhd.edu

Received on May 08, 2017; revised on August 23, 2017; published on August 27, 2017

Abstract

The purpose of this research is to compare seven machine learning methods to predict customer's credit card default payments in Taiwan from UCI Machine learning repository. By comparing different machine learning methods for classification; we aim to determine the best method and study the behavior of clients from each method based on a cost control perspective. Majority of customers do not default on their payments and hence a severe imbalance in classification accuracy pose a significant challenge. Objective of using various machine learning methods is to predict the best possible cost-effective outcome from the risk management perspective. Like any classification problem; the model is trained with different algorithms with re-sampling methods. A cost function is also implemented by implying a higher cost to defaulters classified not correctly. The cost function not only keeps a good balance in predictive accuracy but also a parameter well known as Mathew's Correlation Coefficient (MCC) to not compromise on losing potential customers. By varying the cost factor; we have also tried to see the behavior of each machine learning method (linear or non-linear) which will eventually help us to determine the best algorithm for the said problem. The outcome had different behavior of the results based on cost for original vs resampled data and between different methods. Depending on the trend of results (linear or non-linear) we preferred the method and type of data with non-linear trends. Non-linearity has more cofactors and hence more accuracy which was witnessed during the analysis. It was concluded that original data with Random Forest algorithm is the best in terms of a good balance on cost vs the accuracy.

Keywords: Cost factor, Predictive Accuracy, Machine Learning, Default payment

1 Introduction

In recent times, the credit card issuers regularly face credit debt crisis especially after 2008-2009 economic collapse. Many instances of over-issuing the credit cards to unqualified applicants have raised concerns. Our aim is to determine probable defaulters with reasonably good accuracy and to develop a cost-effective model where not all but the defaulters can be predicted with better accuracy while retaining good customers at the same time. It is a big challenge for any card issuing financial institution as well as for the shareholders and clients.

The use of machine learning methods has significantly increased post 2009. Butaru et al. (2016) used machine learning methods to predict delinquency across 6 major commercial banks using macroeconomic variables. Failure of commercial banks is very much related to bad credits. To evaluate the accuracy for the credit card default, many different approaches including linear discriminant analysis [Wiginton, 1980], k-nearest neighbor [Henley and Hand, 1996], classification trees [Bastos, 2007], artificial neural networks [Malhotra, 2003] etc. have been used in past. The performance of one method over the other usually depends on the problem. This paper is an attempt to address usage of the most frequent type of credit card data. This includes demographic information like age, gender, marital status etc. and the credit history showing billing and payment records to predict performance of an individual's risk when it comes to a potential defaulter. Present study tries to identify a standard method that lowers the cost along with maintaining the quality of results in terms of accuracy.

Many advanced machine learning methods can be used for classification of clients based on risky or non-risky with a predictive accuracy [Chen and Lien, 2009]. Chen and Lien used six different machine learning methods and concluded Artificial Neural Networks is the best when it comes to predictive accuracy using the same dataset. Chen and Lien, 2009 did not utilize Random forest [Leo, 2001] algorithm. These advanced methods can detect a client who might default on next payment with a high accuracy. But there is a high potential to lose many good customers as when the default detection is so specific it might categorize a lot of good customers as defaulters. But it might categorize a lot of good customers as defaulters when the default detection is so specific; it might list lot of potential good customers to fall in category of defaulters. Just to get a high predictability of defaulters, one cannot afford to lose such good customers

as it might very well prove detrimental for financial institutions issuing credit cards. A good prediction will potentially have a mix of risky and non-risky clients with a better accuracy in predicting a defaulter in a cost-effective manner.

The models developed from these machine learning methods can be modified to implement a cost factor to have a risk control [Galindo and Tamayo, 2000] by penalizing false predictions of defaulters. with a good estimation on prediction of defaulters while maintaining a good number of consumers and reduce the overall cost. Present approach of implementing a risk control is an attempt to answer questions listed below by implementation of a cost control parameter:

(1) Is there any difference in cost effectiveness of different machine learning methods?
(2) Does cost is the only factor to be considered while assessing the risk?

The seven machine learning methods used in this project are as follows:

(1) Artificial Neural Networks
(2) K-Nearest Neighbor
(3) Linear Discriminant Analysis
(4) Logistic Regression
(5) Decision Tree
(6) Naïve Bayes Classifiers
(7) Random Forest.

2 Classification Accuracy

2.1 Data Description

This research is based on a multivariate classification dataset provided in UCI Machine Learning Repository. The data contains 30,000 clients with 23 attributes with no missing information (Table 1). Attributes X1 through X23 are independent variables; and class is the dependent variable with binary classes (0,1); 0 – Not defaulted, 1 – Defaulted on credit card payment.

A preliminary insight to data shows that there is a significant imbalance since approximately 78% of the clients never default (23,364 out of 30,000). All of the 23 variables from the dataset are described below and have been utilized in this research

- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6–X11: History of past payment. We tracked the past monthly payment records (from April to September 2005); as follows: X6= the repayment status in September 2005 X7= the repayment status in August 2005 X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
- X12–X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September 2005; X13 = amount of bill statement in August 2005 X17 = amount of bill statement in April, 2005.
- X18–X23: Amount of previous payment (NT dollar). X18 = amount paid in September 2005; X19 = amount paid in August 2005. X23 = amount paid in April 2005.

A correlation heatmap of the data (Fig 1) was developed to check collinearity in the data. Very well-defined collinearity of the data is observed and therefore, one should consider use of penalized methods like Ridge or Lasso regression. A principal component analysis with LDA (Linear Discriminant Analysis) is also used to see if the results improve by reducing the dimensionality in the given dataset. For further analysis train dataset and test dataset are created with 2/3rd i.e. 20,000 clients for train data and remaining 1/3rd for the test data.

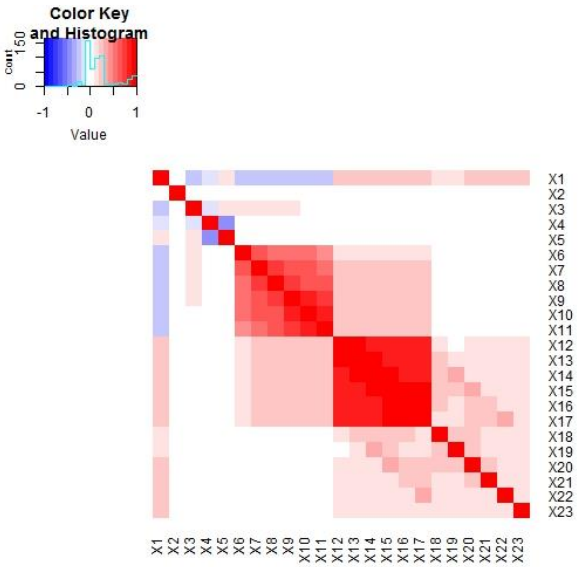


Fig. 1. Correlation Heatmap. Shows the Pearson's correlation coefficients between the different attributes.

2.2 Preliminary Accuracies

To evaluate the accuracy of default using the chosen methods in section 1, one can see 3 types of accuracies as follows

(1) Accuracy of default = No when the client is predicted as not defaulter
(2) Accuracy of default = Yes when the client is predicted as defaulter, and
(3) Overall accuracy for correct prediction of default = No and default = Yes

In imbalanced data, it is generally seen that overall accuracies might be very good but if one focuses on accuracy of default = Yes, it falls lower than 50% accurate. These preliminary accuracies are a direct result from imbalance in the dataset as explained in Section 2. A visual comparison of accuracies in all 7 methods are shown in Fig 2.

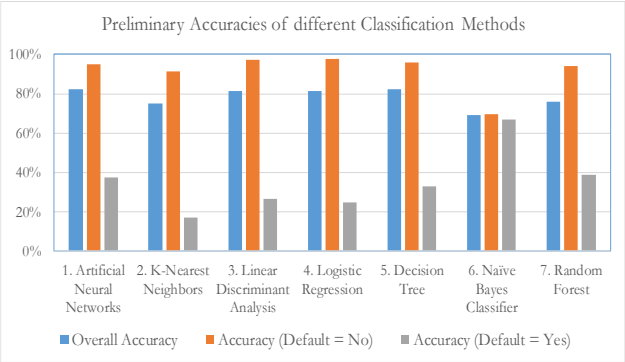


Fig. 2. Preliminary Accuracies. Shows the accuracies of default = No/Yes for all 7 machine learning methods along with overall accuracies for each method.

It can be clearly interpreted from Fig 2 that except Naïve Bayes classifiers with overall low accuracy is better than other methods in predicting both classes (i.e. default = Yes and default = No) with more than 60% accuracy. All other models provide better prediction in case when default = No but provide poor prediction (lower than 50%) in case of default = Yes. This makes them difficult to accurately predict who may potentially default on their credit card payment. One can clearly forecast a need to improve the balance in accuracies of default and no-default for all the models except Naïve Bayes which is good in predicting balanced accuracies.

3 Methodology

There are two ways to analyze data by using all seven methods to improve the balance in accuracies and keep cost effectiveness. For each method, client data is randomly divided into training data (about 2/3rd of all data) and remaining client data were used to validate the model. The dataset is used with two different approaches as shown below:

(a) Cost Function: A cost matrix shall be implemented using a cost factor (>1) for the more expensive clients presented by confusion matrix below to reclassify the classes based on cutoff probability which depends on the imbalance of the train data.

Table 1. Cost Matrix (Use of cost factor for bad clients)

		Observed	
		No	Yes
Predicted	No	0	> 1
	Yes	1	0

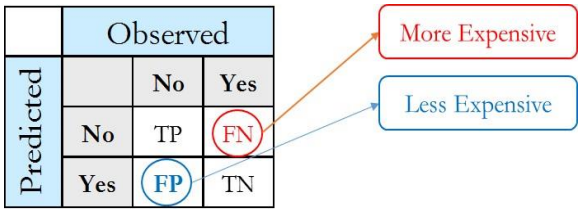
Cost Factor

Cost function also utilized another parameter widely known as Matthew's Correlation Coefficient (MCC) [Liu et al, 2015] defined below:

$$MCC = \sqrt{\frac{\chi^2}{n}} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

TP, TN, FP & FN are defined as follows:
TP – True Positive; client who did not default and predicted as not defaulter, TN – True Negative; client who default and predicted as defaulter, FP – False Positive; client who did not default but predicted as defaulter (less expensive) and FN – False Negative; client who defaulted but predicted as not defaulter (more expensive). These terms are clearly presented in Table 2.

Table 2. Confusion Matrix (Use of cost factor for bad clients)



Implementation of MCC will control the risk by minimizing the cost and have a better balance on the prediction as well. MCC shall be always above "zero" (means greater than 50% balanced accuracy; both sensitivity and specificity) and closer to +1. The selection of best model based on MCC and Cost Factor will be on getting "Less Cost & Similar/better MCC".

(b) Resample the train dataset such that the proportion of default is more balanced. This can be performed using following methods:

Subsampling methods:

- Under Sampling (choose less data with default = No)
- Over Sampling (choose more data with default = Yes), and

Synthetic data generation

- SMOTE (Minority oversampling)
- ROSE (Random oversampling)

From above 5 ways to analyze each machine learning method, following methodology is adopted to select the best model in each machine learning method (Fig 3). As discussed earlier cost is not the only parameter to choose the best model. In this paper, results with lower than 10% reduction in MCC were not selected as the best models to prevent significant reduction in customer base.

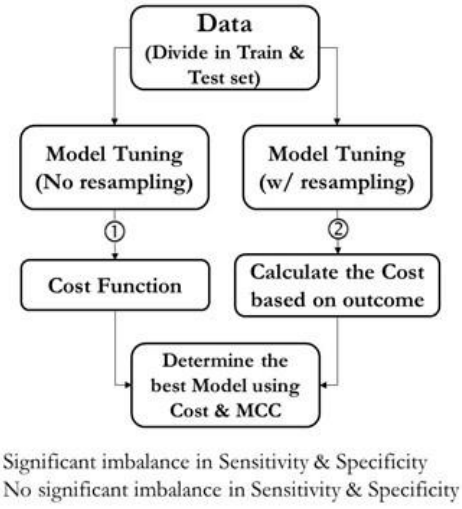


Fig. 3. Best Model Selection. Shows the flowchart on how to select the best model for each machine learning method.

4 Results

Any of the 5 ways mentioned above can be selected as the best cost-effective model for each method based on original vs resampled train data. Cost factor of 10 and 15 is chosen to understand effects of cost factor variation on performance of the model in terms of predictive accuracy and MCC. There are cases where the most cost-effective model is different for different cost factors. Furthermore, the results also justify usage of different models and, the reason behind choosing cost factor of 10 and 15.

4.1 Sample Results (ANN)

For each machine learning algorithm, the user defines range of few key parameters [Kuhn M. (2016)] and let the machine learning algorithm use the full range of these parameters. The model with the best accuracies can be called as model with best parameters. Initial part of this section summarizes results obtained from Artificial Neural Network (ANN) method (Table 3) to define the good model based on Cost and MCC (Table 2) for 30,000 clients. These sample results are presented to explain the following:

(1) Model with best parameters may not be the most cost-effective model.
(2) Models with better specificity after implementation of the cost function are the models best suited for our analysis.
(3) Models with higher drop in MCC (Old MCC vs New MCC) might be lower cost but not chosen as that is not good for our methodology.
(4) In models with resampled data; the cost function provides "Old Cost" which is the selected as cost of that model as the "New Cost" might be lower but the drop in MCC is significant drop in MCC (as much as 50% or more in many cases)
(5) A minor drop in MCC (as much as about 10%) may be acceptable if the cost reduces significantly.

Further, later part of this section summarizes (Table 4) results for all machine learning methods which includes cost of the model with (1) Original data, (2) down-sampled data, (3) up-sampled data, (4) Minority sampling (SMOTE) and (5) Random over-sampling (ROSE) All the models are chosen based on the methodology of low cost with significantly good MCC.

Table 3. Results from Artificial Neural Networks (ANN)

Cost Factor = **10** (lowest cost of **10657**)

Overall Accuracy	Old Sens	Old Spec	New Sens	New Spec	Old MCC	New MCC	Old Cost	New Cost
0.820	0.952	0.354	0.853	0.569	0.40	0.41	14613	10657
0.821	0.951	0.359	0.859	0.549	0.40	0.40	14512	11046
0.823	0.954	0.361	0.853	0.559	0.41	0.40	14451	10863
0.824	0.951	0.372	0.853	0.549	0.41	0.39	14221	11084
0.821	0.952	0.356	0.855	0.556	0.40	0.40	14566	10918
0.822	0.955	0.349	0.852	0.557	0.40	0.40	14709	10922
0.821	0.948	0.371	0.844	0.567	0.40	0.40	14285	10766
0.823	0.951	0.369	0.856	0.546	0.41	0.39	14293	11132
0.823	0.952	0.366	0.852	0.562	0.41	0.40	14351	10805

Cost Factor = **15** (lowest cost of **12204**)

Overall Accuracy	Old Sens	Old Spec	New Sens	New Spec	Old MCC	New MCC	Old Cost	New Cost
0.820	0.950	0.358	0.711	0.693	0.40	0.35	21613	12392
0.822	0.952	0.360	0.724	0.696	0.40	0.36	21551	12204
0.821	0.951	0.364	0.726	0.679	0.40	0.35	21415	12756
0.821	0.948	0.370	0.710	0.685	0.40	0.34	21238	12682
0.821	0.954	0.349	0.707	0.698	0.40	0.34	21880	12291
0.821	0.951	0.363	0.749	0.658	0.40	0.36	21443	13284
0.821	0.947	0.377	0.713	0.692	0.41	0.35	21023	12420
0.821	0.955	0.346	0.729	0.675	0.40	0.35	21965	12870
0.823	0.952	0.364	0.738	0.673	0.41	0.36	21401	12857

Model with highest accuracy
Model with lowest cost

4.2 Sample Results (ANN)

A summary of best model chosen from the cost perspective for each machine learning method are presented here. These results explain the following

(1) With different cost factors, best models may be different (i.e. with original data or with resampled data). No correlation with cost factor is observed from the results.
(2) Majority of the models with higher cost factor have shown significant reduction in MCC. This implies that the decision makers must carefully decide the cost factor to avoid risk of losing potential customers.
(3) Penalized methods have failed to show improvement in results in comparison to model with original data.
(4) SMOTE data generation method has also failed to provide good results.

4.3 Result Summary

Following conclusions can be summarized for different cost factors (10 and 15) presented in Table 4.

(1) With different cost factors; lowest cost model can be from different machine learning methods.
(2) With different cost factors; even though the best machine learning method is same, the model can be from original or resampled data.
(3) Higher cost factors have higher cost improvement but penalty on the MCC as in some models it reduces significantly.
(4) Regression and Discriminant Analysis are poorly performing models for this type of data.

At this stage; it seems difficult to establish which method is the best and based on which type of data (resampled or original). An initial look in the results are shown in table below:

Table 4. Summary of results from all Machine Learning Methods

For Cost Factor = 10

S. No	Machine Learning Method	SubSampling/ Synthetic Data	Cost	MCC
1	Random Forest	DownSampled	9478	0.37
2	Decision Tree	ROSE	9499	0.33
3	Naïve Bayes	-	9704	0.31
4	Artificial Neural Networks	DownSampled	9817	0.38
5	K-Nearest Neighbor	-	10333	0.28
6	Linear Discriminant Analysis	DownSampled	10429	0.31
7	Ridge Regression	UpSampled	10597	0.26
8	Logistic Regression	-	10676	0.26
9	Penalized Linear Discriminant	UpSampled	11235	0.37

For Cost Factor = 15

S. No	Machine Learning Method	SubSampling/ Synthetic Data	Cost	MCC
1	Random Forest	-	11435	0.30
2	Artificial Neural Networks	-	12204	0.36
3	Decision Tree	ROSE	13129	0.33
4	Linear Discriminant Analysis	-	13272	0.24
5	Naïve Bayes	-	13379	0.31
6	K-Nearest Neighbor	-	14272	0.28
7	Logistic Regression	DownSampled	14721	0.26
8	Ridge Regression	DownSampled	14721	0.26
9	Penalized Linear Discriminant	-	16205	0.37

5 Analysis and Discussion

It can be concluded from summarization of results (Section 4) that random forest is the best method for both down-sampled as well as the original data when the cost factor is 10.

Random forest models works well as for factor like cost a single model is not well suited by the fact that it has high variance due to multiple factors. On average, combined estimator using bagging based ensemble method like random forest works better as its variance is reduced.

A plot of the cost vs cost factor for random forest (Fig 4) was generated to present the difference between downs-sampled and original data.

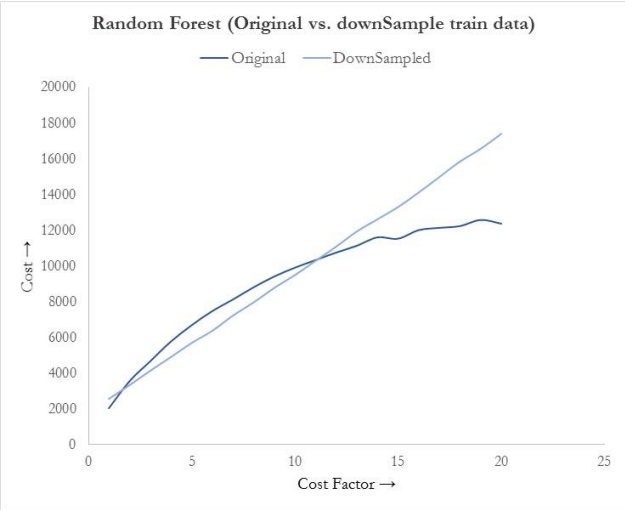


Fig. 4. Random Forest. Original vs. down-sampled train data

Similarly, in Fig 5 cost vs cost factor for ANN also represents difference between down-sampled and original data.

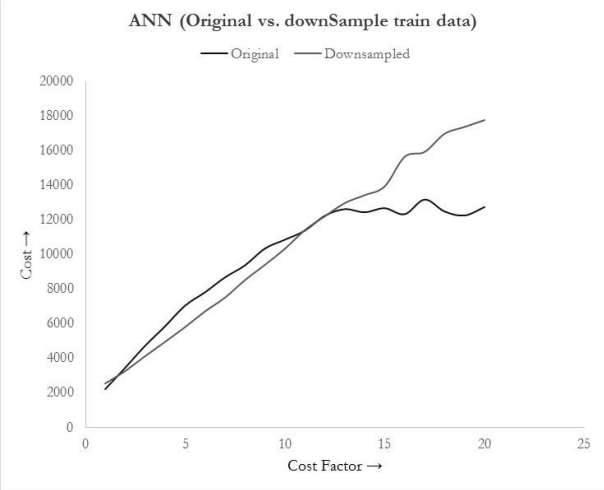


Fig. 5. Artificial Neural Networks. Original vs. down-sampled train data

From Fig. 4 and Fig. 5 it can be concluded that for low cost factor the down-sampled data presents better results than the original data. This is because; results from down-sampled data are linear in nature while the results from original data have non-linear characteristics. From risk management perspective, linear trends are in general not preferred. Non-linearity has more cofactors and hence more accurate which can be witnessed from the analysis. Therefore, it can be established that overall original data is better in terms of balance of cost vs accuracy. Also, no significant savings in cost is observed in terms of cost with down-sampled data, as compared to original data which provides greater number of savings for higher cost factors. Therefore, a non-linear model should be selected.

Similarly, random forest can be considered as best method only in the case of best cost outcome, because of the same reasons explained above. As seen in the figure below (Fig 6); random forest also has non-linear outcome with lower cost in comparison to any other method.

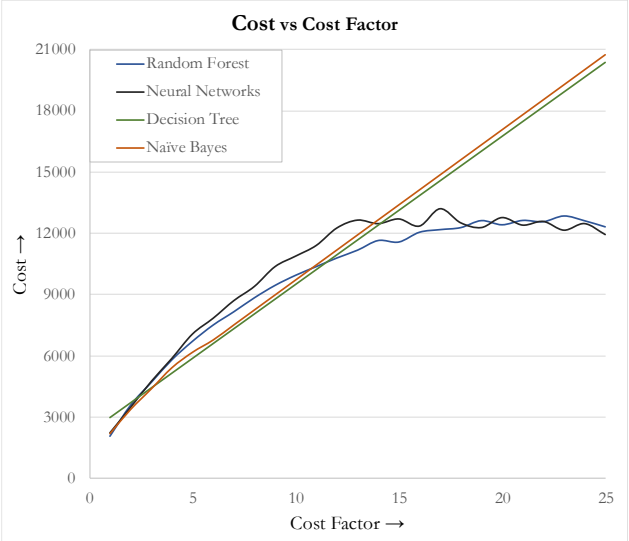


Fig. 6. Comparison of Machine Learning Methods.

For selected test data; Sample outcome of the confusion matrix is shown below:

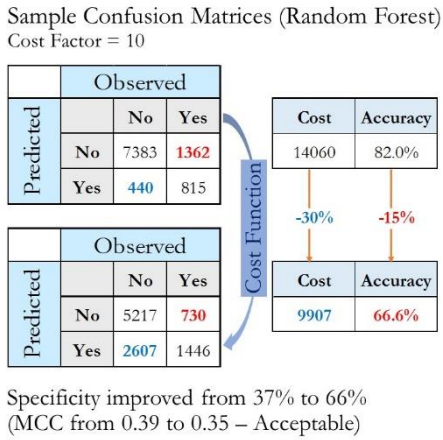


Fig. 7. Sample Results. (Confusion Matrix for 30,000 clients)

A significant reduction in the cost (30%) is observed by maintaining reasonably good MCC and Accuracy (Fig 7) is shown. Hence, random forest model should be utilized for any test set with original data.

6 Conclusion

This paper discusses 7 machine learning methods as defined in Section 1, and compares the performance of each method by considering cost-effectiveness. Each method is compared by using a cost function developed to penalized defaulters predicted as not defaulters. For a single cost factor, there are multiple results available from the confusion matrices and MCC.

However, over a range of cost factors among all 7 machine learning methods, only Random forest and artificial neural networks not only resulted in lower cost but also shows non-linearity in incurred cost per customer. Among these two methods, Random forest has the lowest cost over a larger range of cost factors.

In addition, random forest models have longer run-times and if one desire to have a better Matthew's Correlation Coefficient, Artificial Neural network is a better method. Choosing ANN model will also mean that the financial institution is more likely to take a little more risk which may be good as well.

In course of performing this analysis it was also noted that majority of machine learning methods used credit limit, billing & payment information with more importance. Random forest method was an exception and used Age as one of the top 5 variables. However, other discriminant variables like marital status, gender, education etc. are not as important as timely payments and credit limit across all 7 methods.

Overall our analysis indicates that the credit card default depends non-linearly on various factors. Therefore, ensemble method such as Random Forest and non-linear discriminators such as Neural Networks outperformed other models. We also used the Matthew's Correlation Coefficient, which has been shown to be a valid metric for evaluating model performance [Chen 2015].

Acknowledgements

I would like to thank Dr. Benjamin Soibam for letting me know about the opportunity to present our work at IIBI conference. I would also like to thank proof readers, participants, supportive family members, and Dr. Dvijesh Shastri.

Conflict of Interest: none declared.

References

Butaru, F., Chen Q., Clark B., Das S., Lo A. and Siddique A. (2016) Risk and risk management in the credit card industry, Journal of Banking and Finance, **72**, 218-239

Wiginton, J.C. (1980) A note on the comparison of logit and discriminant models of consumer credit behavior. Journal of financial quantitative analysis, **15**, 757-770

Henley, W.E. and Hand, D.J. (1996) A k-NN classifier for assessing consumer credit risk. Statistician, **45**, 77-95

Bastos, J. (2007) Credit scoring with boosted decision trees, Munich Personal RePEc Archive, 1

Makowski, P. (1985) Credit scoring branches out, Credit world , **75**, 30-37

Malhotra, R., and Malhotra, D.K. (2003) Evaluating Consumer Loans Using Neural Network, Omega, **31**, 83-96

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, **36(2)**, 2473-2480

Galindo, J. and Tamayo, P. (2000) Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications, Computational Economic, **15**, 107-143

Leo, B. (2001) Random Forests, Machine Learning, **45.1**, 5-32

Kuhn M. (2016) caret: Classification and Regression Training. R package version 6.0-73, https://CRAN.R-project.org/package=caret

Soibam, B. (2016) Performance Matrices [PowerPoint presentation]. University of Houston – Downtown

Liu, Y., Cheng, J., Yan, C., Wu X. and Chen, F (2015) Research on Matthews Correlation Coefficients metrics of personalized recommendation algorithm evaluation, International Journal of Hybrid Information Technology, **8.1**, 163-172

Liu, Y., Cheng, J., Yan, C., Wu, X., and Chen, F (2015) Research on the Matthews Correlation Coefficients Metrics of Personlized Recommendation Algorithm Evaluation, International Journal of Hybrid Information Technology, **8.1**, 163-172