Credit Card Default Prediction using Machine Learning Techniques

Yashna Sayjadah
Computing & Technology
Asia Pacific University
Technology Park Malaysia
Kuala Lumpur, Malaysia
yashna.sayjadah@gmail.com

Ibrahim Abaker Targio Hashem
School of Computing & IT
Taylor's University
Subang Jaya, Malaysia
ibrahimabaker.targiohashem@taylors.edu.my

Faiz Alotaibi
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
UPM Serdang, Malaysia
faiz.eid@hotmail.com

Khairl Azhar Kasmiran
Faculty of Computer Science & Information Technology
Universiti Putra Malaysia
UPM Serdang, Malaysia
k_azhar@upm.edu.my

***Abstract*— Credit risk plays a major role in the banking industry business. Banks' main activities involve granting loan, credit card, investment, mortgage, and others. Credit card has been one of the most booming financial services by banks over the past years. However, with the growing number of credit card users, banks have been facing an escalating credit card default rate. As such data analytics can provide solutions to tackle the current phenomenon and management credit risks. This paper provides a performance evaluation of credit card default prediction. Thus, logistic regression, rpart decision tree, and random forest are used to test the variable in predicting credit default and random forest proved to have the higher accuracy and area under the curve. This result shows that random forest best describe which factors should be considered with an accuracy of 82 % and an Area under Curve of 77 % when assessing the credit risk of credit card customers.**

Keywords—: Credit Score, Data mining, Machine Learning, Banking

I. Introduction

The Big data paradigm has revolutionized the banking industry, changing the way financial institutions operate. The aftermath of the past financial crisis has been slowly rectifying and people are nowadays better off in terms of job opportunities and financial health (Subbas & Lahiri, 2017). The needs for financial services have been increasing drastically over the past years and thus generating huge data in terms of volume, veracity, and variety (Demchenko, Laat & Membrey, 2014). Financial institutions have gained in importance and they have been pushed to provide a wide array of services namely; credit facilities, investment, mortgage and retail banking (Schubert, 2015). Therefore, to cope with the phenomenon of budding data, banks are finding ways to leverage their prevailing data sets. The banking industry is gathering massive data from every single transaction of their customers varying from their demographic details to their web history data.

With a view to addressing their myriad challenges, data analytics has been a game changer for banks to mitigate their market imperfections. Analytics has allowed banks to approach the data-driven industry in a healthier way to handle the real-time data generated customers. Nowadays, predictive and prescriptive analytics are allowing banks to explore a whole new horizon which was not possible with their previous descriptive approach. The banking sector is nowadays well-known for making efficient use of machine learning techniques involving several classification techniques to segregate customers to better predict trends (Ajay, Venkatesh & Jacob, 2016). Credit card default prediction is one of the main prediction that banks are a concern with includes credit scoring to better understand why customers are likely to default. They want to capture every small detail of their customers to keep track of payment data which is added to the credit scoring literature to better predict their default (Khandani et al. 2010; Bellotti & Crook 2013). The aim of this paper is to apply different machine learning techniques to find the correlation and predictive power of factors contributing to credit card default and recommendations to the banking industry.

A. The contribution of the paper

The objectives of this paper are:

- We identify the significant factors impacting credit card default.
- We conduct dimension reduction in the data set by feature selection techniques.
- We develop different predictive models using machine learning algorithms to identify the possibility of a customer to default.
- We evaluate the accuracy of the models developed.
- We compare the predictive models and identify the best model suitable for the data set.

B. The outline of the paper

978-1-5386-7167-2/18/$31.00 ©2018 IEEE

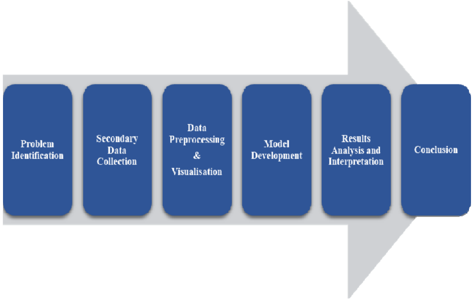
Problem Identification
Secondary Data Collection
Data Preprocessing & Visualisation
Model Development
Results Analysis and Interpretation
Conclusion

Figure 1. Research Process

II. Ease of Use

Based on past literatures, different data mining techniques such as artificial neural network, linear regression, Naïve Bayes and random forest regression have been used to evaluate the risk of customers and their likelihood to default (Bu-yun ZHANG, Shi-wei LI & Chuan-tao YIN, 2017; Ali AghaeiRad, Ning Chen & Bernardete Ribeiro, 2016; Ajay, Venkatesh & Jacob, 2016).

Among the literature, the artificial neural network was the most used technique to investigate the risk associated with credit customers. It is a technique using interconnected neuron to solve an issue just like the human brain operates. The artificial neural network was used by Bu-yun Zhang, Shi-wei LI and Chuan-tao Yin (2017) in their study of A Classification Approach of Neural Networks for Credit Card Default Detection and the findings showed their neural network has the highest processing capabilities when it involves large complex financial data. Moreover, clustering is another effective method practice by many researchers where an analysis of credit card customers' accounts is conducted by segregating them based on their behaviour thus making it better to design scorecards for those with similar behavioral pattern (Bakoben, Bellotti & Adams, 2017).

Some researchers have used the Bayesian network, which is a graphical representation model showing the probability of interconnection of variables. Xia et al. (2017) conducted a study using the Bayesian method to assess credit scoring and they developed a model that can be used as a decision support system for banks to refer to when granting credit facilities. In addition, the random forest regression model has proved to deliver meaningful insights as it is a method which uses a series of decision trees for prediction purpose. It is a robust techniques practice by researchers when studying the banking domain. A research done by Ajay, Venkatesh & Jacob (2016) have shown that the random forest method is on top of the level of accuracy in predicting credit card default. Few techniques which have a good record based on past studies will be applied to the data to evaluate the credit risk factors and to come up with a predictive model.

III. Methodology

A. *Datasets*

The dataset used is generated from credit card operations by the users. It is made up of 30000 instances, 24 attributes which are of integer and real characteristics. Since this dataset has been published and used by various researcher it is not anonymized however, the dataset incorporated a binary variable of Yes = 1 and No= 0 for example default payment outcome. The table 1 below shows the variables to be studied from the dataset.

Table 1. Description of Dataset

Attribute Name	Description
ID	User ID
Limit_Bal	Amount of the given credit (NT dollar)
Sex	Gender(1=male,2=female)
Education	Education(1=graduateschool,2=University,3=others)
Marriage	Marital status(1=married,2=unmarried)
Age	Age(year)
Pay_1 – Pay_5	The repayment status from April to September 2005
Bill_Amt1 to Bill_Amnt6	Amount of bill statement from April to September ,2005
Pay_Amt1 to Pay_Amt 6	Amount paid in September,2005
Default_Payment_Next_Month	Amount to be paid next month

B. *Data Pre-processing and Feature Selection*

For this stage, the dataset does not need to be treated for missing values as there is none. However, the id column will be removed as it has no use in this study. Also, few variables will be converted to factor from numeric. Variables such as Education, Marriage and Pay will be reviewed to merge the unknowns to make more sense of the data.

Before embarking on the data analysis, feature selection will be undertaken. This process reducing the number of variables by choosing the most important one to use during modeling.

For this assignment, the Correlation-based Feature Selection (CFS) will be utilized.

C. *Data Partitioning*

One of the common data splitting range according to past literature is the ratio of 70:30. The 70 % is allocated to the training and the remaining 30% for validation on the test data. This proportion has approved to be a good one as researchers showed that it makes classification model better and the test data makes the error estimate more accurate. For this assignment. The 70:30 ratio is applied to split the dataset.

IV. Data Visualisation

This section is devoted to data cleaning and deeper understanding of the data can be obtained by visualizing each data point. From the graphical illustrations, it can be concluded that there were more credit card clients with features like a female, single and those who just left university.

A. *Default payment next month*

The R codes below was used to plot the default payment for next month and the outcome shows that the default payment is around 22% of the total observations in the dataset and 78 % is not likely to default. The results were as follows in Figure 2:

	Overall Classification Accuracy		
Algorithms	**Accuracy**	**True Positive**	**AUC**
Logistic Regression	0.82	6659	0.75
Rpart Decision Tree	0.8206	6744	0.64
Random Forest	0.8181	6639	0.77

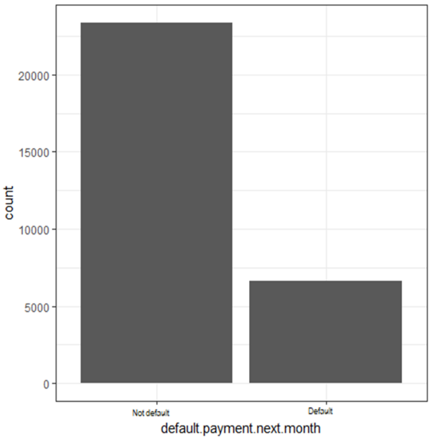

Figure 2. Default Payment

B. *Machine Learning Algorithms*

The data mining process fulfils the basic requiring of applying different techniques to look for hidden forms and patterns in the dataset to come up with the more in-depth explanation of the scenario and the possible recommendations. Among the various commonly used data mining techniques, Decision Tree, Random Forest, and logistic regression will be used to further investigate the causes of any inconsistencies in the dataset.

C. *Logistic Regression*

Logistic regression is a useful model when the researcher must test response variable of categorical nature to obtain binary outcome just like in this assignment the results are expected to either be Yes or No in relation to default payment (Harrell, 2015).

D. *Decision Tree*

The rpart package in R studio is often used for regression and decision tree classification. It is referred to as Recursive portioning designed to better understand the structure of the data before predicting categorical variable (Batal & Hauskrecht, 2010). It has two stages; splitting the data and cross-validating the data to get the full tree.

E. *Random Forest*

The random forest generates a series of bootstrapped trees based on the variables, group them using the tree in the forest and then predict the result by combining the outcome across all the trees (Gupta et al., 2016). The random forest model interprets the Mean Decrease Accuracy and the Mean Decrease Gini.

V. Performance Evaluation

Following the process of creating the predictive model, the performance evaluation of these machine learning algorithms is of equal importance. Some statistical criteria such as the F-measure, precision values, and classification accuracy will be assessed to have a better view about the quality of the outcome. Confusion matrix with true positive rate and false positive rate and Area Under the Curve (AUC) will be considered as well.

A. *Model Comparison and Discussion*

This section outlines a performance comparison between models based on the evaluation assessment provided the prior explanation. Based on the confusion matrix. Classification table and Area Under curve analysis, the outcome is illustrated in Table 2 as follows:

Table 2. Overall Classification Accuracy

According to the analysis provided, all three machine learning algorithms have similar accuracy level. Based on the confusion matrix performance evaluation, rpart decision tree has a better accuracy comparatively with a score of 82.06%. Further assessment can be made by looking at the ROC Curve and the Area Under Curve (AUC) score. The results showed that rpart has the lowest score of 64%. Therefore, it cannot be classified as the best performing algorithm in predicting credit card default. On the other hand, logistic regression and the random forest have an AUC rate of 75% and 77% respectively.

Figure 3 shows a model performance comparison of the three algorithms with x axis outlining the false positive rate and y axis the true positive rate.

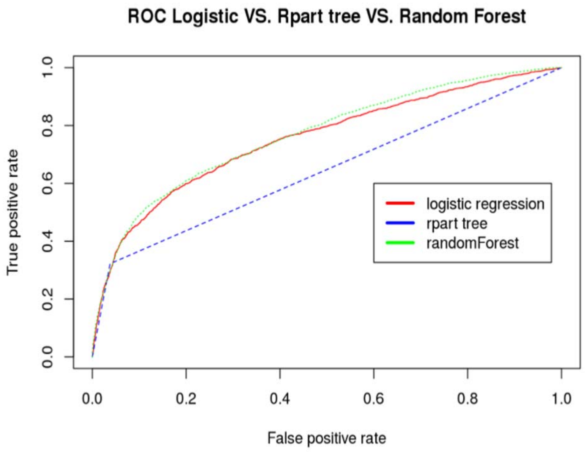

Figure 3. Model Performance Comparison

Random forest algorithm showed the best accuracy and the largest area of the curve. The random forest is built as an ensemble of Decision Trees to perform different tasks such as regression and classification. It uses subsets of both the training data and the feature space, which result in high diversity and randomness as well as low variance. It can be deduced that random forest can better predict credit card default compared to the other algorithms used in the study.

VI. CONCLUSION AND FUTURE WORKS

This study aimed at applying the machine learning techniques to predict credit card default in banks. Based on the analysis of the results, random forest has a prediction accuracy of more than 80%. Banks can use machine learning to assess credit risk of customers before granting them credit card. Banks major concern in to offer valuable products and services to their clients and in order keep up with their competitors they must stay innovative and creative. Machine learning techniques allow banks to approach their customer base in a more customized manner. By applying analytics in the business, banks can benefit in several ways. By studying the customer in terms of their risk level and applying the results from the model, it allows the bank to ingrain smart decision making into a business. It also provides greater insight into data visualization. Banks can make the most of the machine learning algorithm which can contribute in boosting their performance and image in the industry.

REFERENCES

[1]. A, A., Venkatesh, A. and Gracia, S. (2016). Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers. International Journal of Computer Applications, 145(7), p.36-41.

[2]. AghaeiRad, A., Chen, N. and Ribeiro, B. (2016). Improve credit scoring using transfer of learned knowledge from self-organizing map. Neural Computing and Applications, 28(6), p.1329-1342.

[3]. Available at: http://Rising credit card delinquencies to add to U.S. banks' worries [Accessed 13 Nov. 2017].

[4]. Azimi, A. & Hosseini, M. (2017) The hybrid approach based on genetic algorithm and neural network to predict financial fraud in banks. International Journal of Information, Security and Systems Management, 6(1). p.657-667.

[5]. Bakoben, M., Bellotti, T. and Adams, N. (2017) Identification of Credit Risk Based on Cluster Analysis of Account Behaviours. Department of Mathematics, Imperial College London, London SW7 2AZ, United Kingdom.

[6]. Batal, I. and Hauskrecht, M. (2010). Constructing classification features using minimal predictive patterns. 19th ACM international conference on Information and knowledge management, pp.869-878.

[7]. Bellotti, T. and Crook, J., 2013. Forecasting and stress testing credit card default using dynamic models. International Journal of Forecasting, 29(4), pp.563-574.

[8]. Demchenko, Y., De Laat, C. and Membrey, P. (2014) Defining architecture components of the Big Data Ecosystem. In Collaboration Technologies and Systems (CTS), 2014 International Conference on (pp. 104-112). IEEE.

[9]. Ghasemi, A., Motahari, A.S. and Khandani, A.K. (2010) Interference alignment for the K user MIMO interference channel. In Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on (p. 360-364). IEEE.

[10]. Gkoulalas-Divanis, A., Loukides, G. and Sun, J., (2014) Publishing data from electronic health records while preserving privacy: A survey of algorithms. Journal of biomedical informatics, 50, p.4-19. 18

[11]. Gupta, B., Tewari, A., Jain, A. and Agrawal, D. (2016). Fighting against phishing attacks: state of the art and future challenges. Neural Computing and Applications, 28(12), pp.3629-3654.

[12]. Hargreaves, I., Roth, D., Karim, M., Nayebi, M. & Ruhe, G. (2017) Effective Customer Relationship Management at ATB Financial: A case study on industry- academia collaboration in data analytics. Springer International Publishing.

[13]. Harrell, F. (2015). Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer International Publishing.

[14]. LaMagna, M. (2017) Americans now have the highest credit-card debt in U.S. history. [online] MarketWatch. Available at: http://Americans now have the highest credit-card debt in U.S. history [Accessed 13 Nov. 2017].

[15]. Lin, K. (2017) Taiwan's Bank SinoPac leverages advanced analytics to better understand customers' credit card usage patterns. [online] bankITasia. Available at: https://bankitasia.com/bankitasia/customer-insights--analytics/taiwans-bank-sinopac-leverages-advanced-analytics-to-better-understand-customers-credit-card-usage-patterns/ [Accessed 13 Nov. 2017].

[16]. Maiya, R. (2017) 6 Technology Trends That Will Transform Banking In 2017. [online] Huffpost. Available at: http://6 Technology Trends That Will Transform Banking In 2017 [Accessed 13 Nov. 2017].

[17]. Subba, N. and Lahiri, D. (2017) Rising credit card delinquencies to add to U.S. banks' worries. [online] Reuters. Available at: http://Rising credit card delinquencies to add to U.S. banks' worries [Accessed 13 Nov. 2017].

[18]. Tobback, E. and Martens, D. (2017).Retail credit scoring using ne-grained payment data. Department of Engineering Management, University of Antwerp, p.1-5.

[19]. Verikas, A., Kalsyte, Z., Bacauskiene, M. and Gelzinis, A. (2009) Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. Soft Computing, 14(9), p.995-1010. 19

[20]. Wu, K.Y., Zuo, G.L., Li, X.F., Ye, Q., Deng, Y.Q., Huang, X.Y., Cao, W.C., Qin, C.F. and Luo, Z.G. (2016) Vertical transmission of Zika virus targeting the radial glial cells affects cortex development of offspring mice. Cell research, 26(6), p.645-654.

[21]. Xia, Y., Liu, C., Li, Y. and Liu, N. (2017) A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, 78, p.225-241.

[22]. Yap, B., Ong, S. and Husain, N. (2011) Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Systems with Applications, 38(10), p.13274-13283.

[23]. Yeh, I. (2017) UCI Machine Learning Repository: default of credit card clients Data Set. [online] Archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients [Accessed 13 Nov. 2017]. References