Prediction of Credit Card Clients Payment Status

Xiaojun Lu
Postgraduate Student
Whitireia New Zealand
Xiaojun.lu23@gmail.com

Adrian Hargreaves
Supervisor
Whitireia New Zealand
adrian.hargreaves@whitireia.ac.nz

Dr. Sara Bilal
Supervisor
Whitireia New Zealand
Sara.Bilal@whitireia.ac.nz

ABSTRACT

Nowadays, the world of business is undergoing tremendous changes due to the current technology trends and its impact on management and organizations. Data mining is the process of discovering certain patterns and knowledge from large amounts of data. Electronic banking becomes a popular trend as financial institutions have accepted the transformation to provide electronic banking facilities to their customers in order to remain relevant and thrive in an environment that is competitive. This research aim is to provide a model to the banking industry that can manage credit card clients using data mining algorithms. The dataset used in this research is named "Default of Credit Card Clients" (DCCC) which consists of 30,000 observations and 24 attributes. Algorithms from classification and clustering categories have been applied on the dataset. The accuracy of the implemented algorithms was measured to determine which algorithm generates the best predictive model. From the classification category, C5.0 algorithm generates the best model with an accuracy of 83.60%. As well, from the clustering category, K-means gave a silhouette point of 0.55. This research work has produces a good model that can be used by the banks to identify the genuine clients based on their credit cards repaying records. Hence, banks could have a better control on risk management issues.

Keywords: Data mining algorithms, Default credit card dataset, R-software.

1. INTRODUCTION

Data mining technology first appeared in mid 1990s. It has helped many organizations to gain more profits and to be more competitive over their existing rivals through its intelligent algorithms (Koh & Gerry, 2002). It has a significant impact on the businesses performances especially in banking industry. Data mining algorithms has captured the interest of many people especially on how these algorithms work on massive data to establish data models, analyse and visualise the data (Thillainayagam, 2012). Data mining is likely to improve the banking industry in terms of detecting defraud clients and lowering costs. Due to the flexibility and easily expandable attribution of IT infrastructure such as cloud data storage analytics and data lakes, dealing with huge amounts of data becomes a significant practise (Schutte et al., 2017). Banking industry is facing various problems such as risk management and fraud detection. Sometimes the clients can't repay the money on time, and the banks have to spend huge time and work investigating, negotiating or even suing to get the money back. Often, the money won't return even tremendous efforts have been done. So with the development of data mining technology, effective and efficient work could be done based on specific requirements.

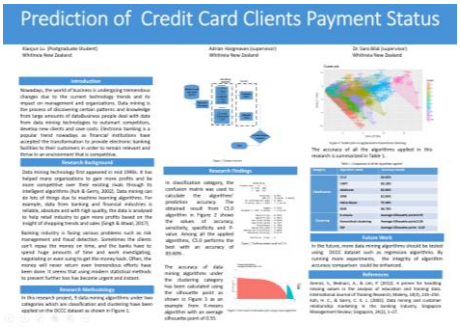

Figure 1: Prediction of Credit Card Clients Payment Status

This poster appeared at the 9th annual conference of Computing and Information Technology Research and Education New Zealand (CITRENZ2018) and the 31st Annual Conference of the National Advisory Committee on Computing Qualifications, Wellington, New Zealand, on July 11-13, 2018 as part of ITx 2018.

2. RESARCH BACKGROUND

2.1 What is data mining?

During the past decades, the development in the field of information technology has a huge impact on people's lives because it has brought a lot of changes and benefits to them. One of the new trends of IT industries is big data. Since 1990s, data mining has grown increasingly to make huge extra profits to organisations. Data mining uses sophisticated statistical processes or artificial intelligence algorithms to discover useful trends and patterns from the extracted data (Koh & Gerry, 2002). Although there are different cognitions towards data mining, all can be summarised by extracting useful hidden patterns in data via different algorithms to make reasonable predictions to benefit the organisations. (Ranjan, 2009).

2.2 What can data mining technology do?

Data mining can do lots of things due to machine learning algorithms. For example, data from banking and financial industries is analysed to help retail industry to gain more profits based on the insight of shopping trends and sales (Singh & Atwal, 2017). Machine learning applications such as advertising placement, credit scoring, fraud detection, stock trading and Web search are widely used in banking industries (Singh & Atwal, 2017).

Here are 3 types of machine learning algorithms used in data mining algorithms:

- ***Supervised learning:*** Consists of a dependent variable (to be predicted) and an independent variable. An example of some algorithms using supervised learning are regression, decision trees, random forest, KNN, logistic regression
- ***Unsupervised learning:*** Nothing to be predicted. Used for clustering population in different groups. An example of an algorithm using unsupervised learning is K-means.
- ***Reinforcement learning:*** The machine is trained to make specific decisions. It learns from the past experience and tries to capture the best possible knowledge to make accurate decisions.

3. RESEARCH METHODOLOGY

In this research project, 9 data mining algorithms under two categories which is classification and clustering have been applied on the DCCC dataset as shown in Figure 2.

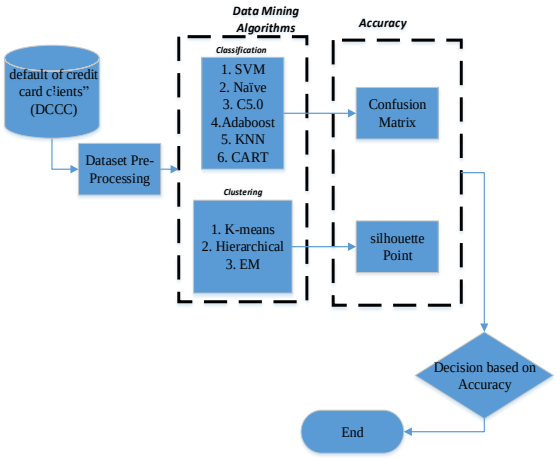

Figure 2: System Overview

4. EXPERIMENTAL RESULTS AND DISCUSSION

The accuracy of data mining algorithms under classification category has been calculated using the confusion matrix where several parameters are considered which are "accuracy", "sensitivity", "specificity", "P-value" as shown in Figure 3 from SVM.

```
          Reference
Prediction    N    Y
         N 3719  761
         Y  144  376

               Accuracy : 0.819
                 95% CI : (0.808, 0.8296)
    No Information Rate : 0.7726
    P-Value [Acc > NIR] : 5.46e-16

                  Kappa : 0.3629
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9627
            Specificity : 0.3307
         Pos Pred Value : 0.8301
         Neg Pred Value : 0.7231
             Prevalence : 0.7726
         Detection Rate : 0.7438
   Detection Prevalence : 0.8960
      Balanced Accuracy : 0.6467

       'Positive' Class : N
```

Figure 3: Confusion matrix result on SVM

The accuracy of data mining algorithms under the clustering category has been calculated using the silhouette point as shown in Figure 4 as an example from K-means algorithm.

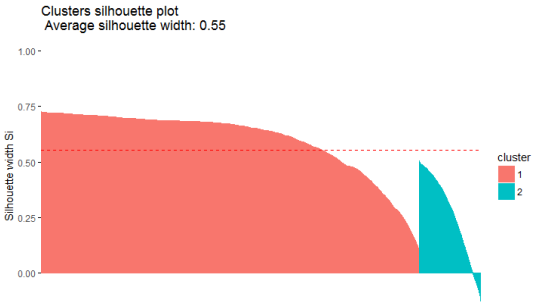

Figure 4: Plot result of silhouette point using K-mean algorithm

The accuracy of all the algorithms applied in this research is summarized in Table 1.

Table 1: Comparison of all the algorithms applied

Category	Algorithm name	Accuracy results
Classification	C5.0	83.60%
	CART	83.18%
	Adaboost	82.08%
	SVM	81.90%
	Naïve Bayes	75.38%
	KNN	69.78%
Clustering	K-means	Average Silhouette point:0.55
	hierarchical clustering	Average Silhouette point:0.36
	EM	Average Silhouette point:-0.02

5. FINDINGS

In the classification category, among all of the algorithms, C5.0 predicts the best accuracy of 83.60%. In clustering category, K-means has given a good model towards the dataset applied as the average silhouette point for 2 clusters is 0.55. Due to large size of the dataset, a subset of the dataset with 5,000 objects has been picked randomly during the implementation of the agglomerative hierarchical clustering, and the result obtained has an average silhouette point of 0.36. The EM algorithm gives poorer result with the average silhouette point of -0.02 compared with the K-means and the agglomerative hierarchical clustering algorithms. Finally, in this research work, it was found that most of the algorithms implemented from the classification category have a prediction accuracy over 75%, which means that the classification algorithms have built good predictive models based on the dataset used.

6. REFERENCES

Ahn, J. (2002). Beyond single equation regression analysis: Path analysis and multi-stage regression analysis. *American Journal of Pharmaceutical Education; Alexandria*, *66*(1), 37.

Anuradha, J., & Tripathy, B. K. (2014). Hierarchical Clustering Algorithm based on Attribute Dependency for Attention Deficit Hyperactive Disorder. *International Journal of Intelligent Systems and Applications; Hong Kong*, *6*(6), 37–45. http://dx.doi.org.whitireia.idm.oclc.org/10.5815/ijisa.2014.06.04

Gale, R. P., Hochhaus, A., & Zhang, M. –. (2016). What is the (p-) value of the P-value? *Leukemia; London*, *30*(10), 1965–1967.http://dx.doi.org.whitireia.idm.oclc.org/10.1038/leu.2016.193

García-Altés, A., Santín, D., & Barenys, M. (2007). Applying artificial neural networks to the diagnosis of organic dyspepsia. *Statistical Methods in Medical Research; London*, *16*(4), 331–346. http://dx.doi.org.whitireia.idm.oclc.org/10.1177/0962280206071839