ISSN 2278-3091
Volume 9, No.1.4, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse5891.42020.pdf
https://doi.org/10.30534/ijatcse/2020/5891.42020

Preliminary Studies on Predicting Customer Purchase Behaviour in Online Retail Business

Nur Shamsiah Abdul Rahman[1], Lim Ken Tak[1], Anis Farihan Mat Raffei[1]
[1]Faculty of Computing, Universiti Malaysia Pahang, 25150 Kuantan, Pahang, Malaysia, shamsiah@ump.edu.my, eliphas95@gmail.com, anisfarihan@ump.edu.my

ABSTRACT

Online retail business has become a popular trend in our life because it is very easy and hassle free. The increasing amount of customer in online retail business motivates the use of data mining techniques to discover the customer purchase behaviour. Before the implementation of data mining technique, the problems encountered by the online retail business are time consuming, probable human error and space consumption. These problems have degraded the rate of customer purchase in online retail business. Therefore, the objective of this research is to identify the existing methods in predicting customer purchase behaviour. The recency, frequency and monetary (RFM) based classification techniques are proposed to model the customer purchase behaviour. Next, the results obtained from the proposed models are predicted using correlation and linear regression methods to predict the customer purchase behaviour. The result from the end of the research will be further discussed. Various technique can be applied to further improve the current result in future work.

Key words: Data mining, RFM, online retail business, classification, correlation

1. INTRODUCTION

Data mining is known as wisdom breakthrough. From computer science aspect, data mining is known as disclosure of fascinating pattern and relationship between huge amount of data. To analyze this huge amount of data which is known as dataset, data mining technique needs the help of various tool such as MatLab, Orange, and artificial intelligence technique to ease the process. In simple words, data mining is used to find patterns from a large set

A shopping system, be it online such as Lazada or offline such as Tesco Supermarket is a process of buying goods or services from another person or merchant. In this research, online retail business will be more focused. The goal of online retail business is to provide a 24/7 service for customers to shop with ease, anytime and anywhere. There are several kinds of online retail business, a few examples that can be shown are Strictly Online Presence (Amazon) and Brick & Mortar Combination (Apple) of data to predict the outcome.

Customer relationship management (CRM) can be described as a strategy which handle all the company's or organization's relationship and interaction with current customers and future potential customers. With the help of CRM, it can help to raise the income of the company. CRM system act as a platform for all business unit s to interact with their clients and fulfil their need and demands effectively in order to build a long-term relationship. In CRM, there are several crucial features which can help to manage and administrate its customer and vendors in an effective manner. The features are customer needs, customer response, customer satisfaction, customer loyalty, customer retention, customer complaint and customer service [1, 2].

CRM is very important because the customers or clients are grouped according to the type of business they do or the location that they do their business. This helps the company in focusing and concentrating on each and every of the customers separately. CRM also centralized their system details which is available anytime, and can help to reduce the process time and thus increase productivity. With the help of CRM, the company can effectively deal with all their customers and provide what their customer actually needs. This will increase the customer satisfaction level and the customer will remain loyal in business with the company they deal with, which in turn can enhances the profit of the company [3].

The aim of data mining implementation in online retail business is to analyze the purchasing behaviour of customers and provide better promotion packages such as buy two products and get one free product, or services in the future. There are several data mining processes that can be used for online retail business. The processes are classification, outlier detection, clustering, association rules and regression analysis. In this research, the process of classification and clustering are chosen.

2. LITERATURE REVIEW

This section discussed about the foundation and some fundamental knowledge related to online retail business and data mining. Different types of online retail business and what kind of dataset used will be discussed in the following sub-topic, accompanied by a common explanation regarding the concept of data mining. The current implementation of data mining in online retail business also will be discussed. Apart from the implementation and explanation, this section describes two data mining techniques which is important highlighted in the online retail business, which are classification and clustering. In the classification, a decision tree induction method and rule-based method will be briefly explained. In the clustering technique, partitioning methods and hierarchical method will also be briefly explained.

The first type of online retail business is strictly online presence online shopping system such as TaoBao. As the name imply, this type of online retail business is only available online and it does not have any actual store. The second type of online retail business is brick and mortar combination online shopping system such as Adidas. This kind of online retail business is where the company has both physical and virtual storefronts. The third type of online retail business is user to user such as eBay. This system works in such a way that it acts like a platform to connect buyers and sellers from other places. The datasets used are fictional and can obtained online (eg: Kaggle, data.world, etc). According to the Privacy Act 1974, there is no online shopping system will agree to enclose their customer transaction history.

Data mining, which also known as data disclosure, is one of the process of discovering useful pattern from a large dataset and convert them into useful pattern and data statistic. Data mining is part of the process of knowledge discovery database (KDD) as shown in Figure 1. Data mining techniques can be implemented in online shopping system in many ways. The implementation of data mining in online shopping system is to model the customer online shopping behaviour. It can also analyze the buying pattern of the customer to predict their shopping behavior [4]. There are two data mining algorithm which will be discussed in the next section, which are classification and clustering algorithm.

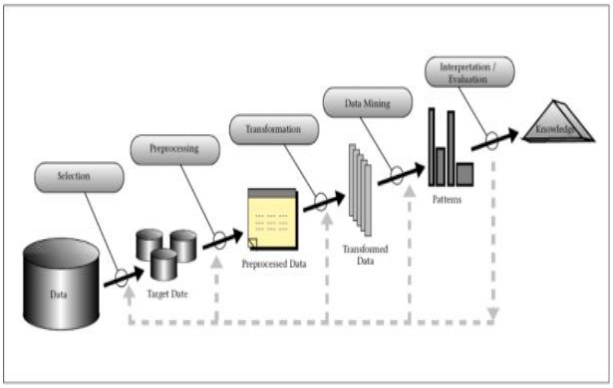


Figure 1: Overview of KKD process

2.1 Classification Algorithm

Classification is a major method which is widely used in data mining. The idea of classification is that it can predict the target class by analyze the training dataset. This can be achieved by finding the correct boundary condition which could be used to determine the each of the target class. Once the target boundary is set, the next step is to predict the target class. Figure 2 shows the concept of classification.

There are two methods discussed under classification algorithm, which are decision tree induction method and rule-based method [5, 6]. Decision tree induction method are ordered data structures for supervised learning where the object input is split into different groups in order to predict the possible result [7, 8]. The advantages are it is easy to implement and capable of handling dataset with missing value. The disadvantages are certain algorithm require distinct vale only and perform lowly if many complex interactions are present.

Rule-based method is one of the techniques that used to classify the data input using "If…Then…" rules [9, 10]. It is express in the form of IF condition THEN conclusion where condition part is the rule antecedent/precondition and conclusion part is rule consequent. The advantages are it can be easily interpreted and can trace back the rules to check out the cause and effect. The disadvantages are it is difficult to maintain large rule base and lack of open standardization rule sets.

Figure 2: Classification concept

2.2 Clustering Algorithm

Clustering is the process of separating a set of data objects into different set of clusters. Object with similar characteristic will be group together within the same cluster. The data objects within the same cluster is different form the other data objects in another cluster [11]. The purpose of clustering algorithm is to identify distinct groups from a set of unlabeled data. Figure 3 shows the concept of clustering.

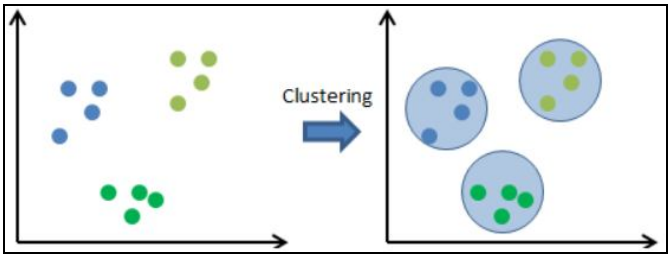


Figure 3: Clustered data

There are two methods discussed under clustering algorithm, which is partitioning method and hierarchical method. Partitioning methods works in such a way that it constructs K partitions of the data from a database of n objects. Each object must have a group and no duplication are allowed in the same group. Each group must at least have one object. The main objective of partitioning method is to divide the data points into K partitions [12, 13]. The advantages are the processing time is faster can produce more compact cluster. The disadvantages are the hard to predict K value and the number of clusters need to be specify first.

Hierarchical clustering method involves creating clusters from top-down or bottom-up approach. Hierarchical clustering method has two main type, which is agglomerative (bottom-up) and divisive (top-down). The advantages are the result in Dendrogram form is easy to understand and the number of clusters no need to be specify first. The disadvantages are it is not possible to return to previous step after splitting and this method is not suitable in large dataset. Classification is chosen to use in this research because this method tolerates missing values, easy to implement, provide flexibility and reduce the ambiguity of the user.

3. METHODOLOGY

This research follows the Cross-industry standard process for data mining (CRISP-DM) methodology. Modification has been done to remove the last step, which is the deployment part because the final results obtained can only be used as reference for future work. Figure 4 shows the research framework consists of three phases. Phase 1 starts with preliminary study. Followed by phase 2 discuss about implementation and analysis, and phase 3 explain about system evaluation.

The dataset is obtained from data.world website, namely Online Retail. Figure 5 shows the sample of the online retail dataset. During data pre-processing, all the data contain negative values, null values and time values were removed as it does not help in this research. Only United Kingdom customers is chosen because it contains approximately 90% of the whole dataset. After data pre-processing, recency, frequency, and monetary is added to further evaluate the customer.

Recency refers to the freshness of the customer activity. Meanwhile, frequency refers to the frequency of the customer successful transaction, and monetary refers to a customer's willingness to spend. To determine the range for each recency, frequency and monetary, the segmentation was done and the customers will be assigned with a new value based on the calculated value as shown in Figure 6. With the assigned new value, rules for decision tree induction method can be carry out and there will be eight rules to classify the customer. One of the rules as in Figure 7. With these eight rules, the customer is divided into four groups, which are premium, best, moderate, and churn. Figure 6 shows the process on how the classification works based on decision tree induction method. Table 1 shows the meaning behind premium, best, moderate and churn.

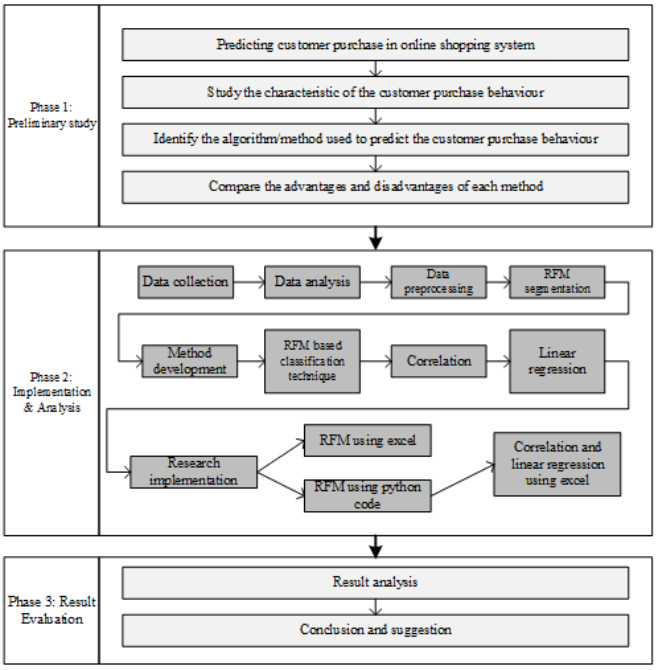
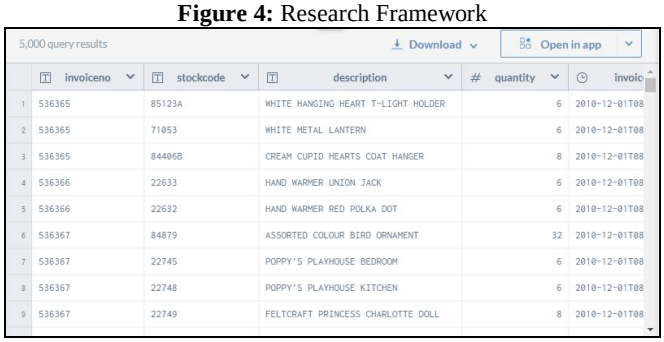


Figure 4: Research Framework


5,000 query results — Download — Open in app

	invoiceno	stockcode	description	quantity	invoic
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01T08
2	536365	71053	WHITE METAL LANTERN	6	2010-12-01T08
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01T08
4	536366	22633	HAND WARMER UNION JACK	6	2010-12-01T08
5	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01T08
6	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01T08
7	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	2010-12-01T08
8	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	2010-12-01T08
9	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	2010-12-01T08

Figure 5: Online retail dataset

Figure 6: RFM segmentation

Table 1: Description on category of Customer_value

Customer_value	Description
Premium	The customer has three good traits out of the three criteria. (Example: Low recency, high frequency, high monetary).
Best	The customer has two good traits out of the three criteria (Example: High recency, high frequency, high monetary).
Moderate	The customer has one good trait out of the three criteria (Example: High recency, low frequency, high monetary).
Churn	The customer has zero good trait out of the three criteria (Example: High recency, low frequency, low monetary).

From the obtained result, correlation and linear least regression will be carry out. Correlation is used to determine the strength of the relationship between two variables. Linear least square regression is used to predict the customer purchasing behaviour with an equation.

4. RESULT AND DISCUSSION

There are two ways in obtaining the result in this research, which are using the excel software method and python coding method. The difference between these two methods is the excel method uses regular fixed interval when dividing the customers into groups, whereas the python coding method uses quartile method when grouping the customer. Figure 8 shows the comparison between the results obtained from excel method and python coding method.

As mentioned previously, the method used by the second pie chart (RFM using python) is the even distribution of data input, so if there is an extreme value within the data input, it will not affect the distribution of the data because the distribution of data input method is based on the percentage of the data (eg: first thirty-three percent of the data will be in the first group, thirty-fourth percent of the data till sixty-sixth percent of the data will be in the second group, and the rest of the data will be in the third group). The summary that can be conclude from the RFM using Excel pie chart is that the customer is not well distributed because of the extreme value within the data. The customers in churn group (RFM Rating 4) has the most customers, which is 47% (1845 customers) from the total customer whereas premium group (RFM Rating 1) has the least customers, which is 16% (613 customers) from the total customers. The remaining two groups, the best group (RFM Rating 2) have 17% (662 customers) of the total customer and lastly the moderate group (RFM Rating 3) have 20% (800 customers) of the total customer.

Compared with the RFM result using Python method, we can conclude that the data are distributed to every group because of the group division using quartile method which is no from both of the pie charts, the result yield from RFM using Python method is more preferable as compared to the result yield from RFM using Excel method. The reason behind this is that the result yield form RFM using Excel method have zero customer in premium group and three customers in best group. This situation in general is unlikely going to happen because no grouping is made to have nearly 0% of the data input within the group. There are also too much of customer in the moderate group, which is 72% of the total customer.

By using the result from Python, Correlation and Linear Least Square Regression was carried out. The Correlation r for the result is 0.5504 with the removed of extreme values. The equation formed from the linear least square regression is:

$$y = 12.705x + 515.92,\ y = mx + C \quad (1)$$

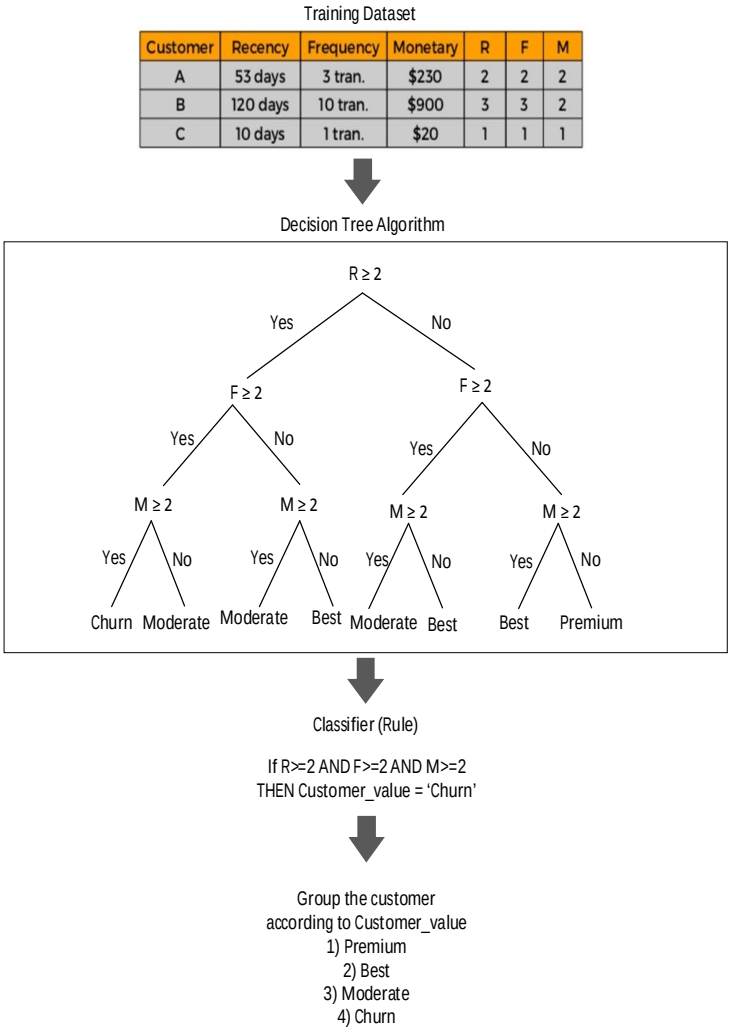


Figure 7: RFM-based classification algorithm

5. CONCLUSION

As conclusion, the customers purchase behaviour in online retail business are evaluated based on recency, frequency and monetary. Then, the customers divided into groups using decision tree induction method to determine the category of each customer, either premium, best, moderate or churn.

With the obtained result, correlation and linear least square regression was carried out to predict the customer purchasing behaviour. Since the implementation is used on a fictional dataset, the result obtained can only be used as reference purposes and the result are not suitable use to evaluate real time situation.

There are several constraints that were encountered while carry out this research. First, due to the use of online fictional dataset, the result obtained cannot be used to reflect the real customer purchasing behaviour. Second, constraint is the years span of the dataset. The model form during the research may change if the years span of the dataset increases. The third constraint is the origin of customer involved in the research. Since only UK customer is involving in this research, thus it cannot be used to predict the customer purchasing behaviour from other country. The fourth constraint is the evaluation criteria of the customer. In this research, the customer only evaluated based on recency, frequency and monetary. It is suggested to have future exploration from other sources in investigating and identifying other variables such as duration and engagement for future studies.

The method implemented in this research can be used in real dataset to predict the customer purchase behaviour. Since the method solely focus on the evaluation of the customer, which is one of the aspects for online shopping system, the products of the online shopping system can be evaluated to optimize the functionality of an online shopping system. Other linear method apart from least square regression method also can considered use to predict the customer purchasing behaviour based on the distribution of the data.

ACKNOWLEDGEMENT

This research was fully funded by Universiti Malaysia Pahang, grant number RDU1803147.

REFERENCES

1. A. F. Z. Abidin, M. F. Darmawan, M. Z. Osman, S. Anwar, S. Kasim, A. Yunianta andT. Sutikno. **Adaboost-multilayer perceptron to predict the student's performance in software engineering,** *Bulletin of Electrical Engineering and Informatics,* Vol 8, pp. 1556-1562, 2019.
2. N. S. A. Rahman, M. S. Othman and W. Al-Rahmi. **Exploring the Use of Social Media Tools Among Students for Teaching and Learning Purpose,** *Journal of Theoretical and Applied Information Technology,* Vol 91, pp. 49-60, 2016.
3. S. F. Crone and D. Soopramanien. **Predicting Customer Online Shopping Adoption-an Evaluation of Data Mining and Market Modelling Approaches,** in *Proceedings of the 2005 International Conference on Data Mining,* DMIN 2005, Las Vegas, Nevada, USA, June 20-23, 2005.
4. M. A. Remli, K. M. Daud, H. W. Nies, M. S. Mohamad, S. Deris, S. Omatu, S. Kasim and G. Sulong. **K-Means Clustering with Infinite Feature Selection for Classification Tasks in Gene Expression Data,** in *11th International Conference on Practical Applications of Computational Biology & Bioinformatics,* Porto, 2017.
5. N. S. A. Zulkifli and A. W. K. Lee. **Sentiment Analysis in Social Media Based on English Language Multilingual Processing Using Three Different Analysis Techniques,** *International Conference on Soft Computing in Data Science,* pp. 375-385, 2019.
6. B, Qin, Y. Xia, S. Prabhakar and Y. Tu. **A Rule-Based Classification Algorithm for Uncertain Data,** *IEEE International Conference on Data Engineering,* pp. 1633- 1640, 2009.
7. K. Moorthy, M. H. Ali, M. A. Ismail, W. H. Chan, M. S. Mohamad and S. Deris. **An Evaluation of Machine Learning Algorithms for Missing Values Imputation,** *International Journal of Innovative Technology and Exploring Engineering,* Vol 8, pp. 415-420, 2019.
8. B. Satish and P. Sunil. **Study and evaluation of user's behavior in e-commerce using data mining,** *Research Journal of Recent Sciences,* Vol. 1(ISC-2011), pp. 375-387, 2012.
9. J. A. Jupin, T. Sutikno, M. A. Ismail, M. S. Mohamad, S. Kasim and D. Stiawan. **Review of the machine learning methods in the classification of phishing attack,** *Bulletin of Electrical Engineering and Informatics,* Vol 8 (4), pp. 1545-1555, 2019. https://doi.org/10.11591/eei.v8i4.1344
10. J. S. Saket and S. Pandya, **An Overview of Partitioning Algorithms in Clustering Technique,** *International Journal of Advanced Research in Computer Engineering & Technology,* pp. 1943- 1946, 2016.
11. R. C. Barros, M. Basgalupp, A. de Carvalho and A. Freitas. **Automatic Design of Decision-Tree Algorithms with Evolutionary Algorithms,** *Evolutionary Computation,* Vol. 21 (4), 2013.
12. R. Heldt, C. S. Silveira and F. ernando B. Lucea. **Predicting customer value per product: From RFM to RFM/P,** *Journal of Business Research,* Article in Press, 2019. https://doi.org/10.1016/j.jbusres.2019.05.001
13. A. F. M. Raffei, N. S. Awang, N. S. A. Rahman, N. S. A. Zulkifli. **Internet of Things (IoT) Based Fire Alert Monitoring System for Car Parking,** in *7th International Conference on Electrical and Electronics Engineering,* pp. 290-293.