Implementation of Decision Tree Algorithm in Customer Recency, Frequency, Monetary, and Cost Profiling: a Case Study of Plastic Packing Industry

To cite this article: W Gata et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 662 022032

View the article online for updates and enhancements.
Implementation of Decision Tree Algorithm in Customer Recency, Frequency, Monetary, and Cost Profiling: a Case Study of Plastic Packing Industry

W Gata 1, Iskandar 2*, H Basri 3, D A Puspitawati 4, S Hidayat 5, Walim 6

1, 4, 5 Computer Science Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri), Indonesia.
2 Faculty of IT, Sekolah Tinggi Teknologi Muhammadiyah Cileungsi (STTMC), Indonesia.
3 Faculty of IT, Bina Sarana Informatika University, Indonesia.

E-mail: *Iskandar@gmail.com

Abstract—This study aims to model the form of customer profile classification on companies using the C 4.5 and Random Forest algorithms to produce the best profile classification model from customers to see a pattern of assessments of manual assessments so far. This study uses descriptive analysis method. Through classification of Customer Profiles with the Recency, Frequency, Monetary - Cost (RFM-C) model approach. After process the two models, the results obtained are the C4.5. After testing the two algorithms, the results obtained are the use of the C4.5 algorithm for companies to classify RFM-C which is expected to predict because it has higher accuracy and kappa values compared to the Random Forest algorithm. It can be concluded that the modeling of customer profile forms in companies that use the C 4.5 algorithm and random forest can produce the best profile classification model.

1. Introduction

There is a company called xyz which is engaged in plastic packing requiring classification based on sales data and customers they have. The general model used in customer grouping is the RFM (Recency, Frequency, Monetary) model, which is grouped based on the latest customer visits, the frequency of visits, and income received by the company. This concept is introduced by Bult and Wansbeek (1995)[1]. In the previous study and looking at the condition of the company and the data, it was seen that RFM was the right concept to be applied based on the needs of the company.

J. Wei and friends research conducted in the hairdressing industry in Taiwan, RFM effectively to identify its valuable customers and develop its marketing strategies[2]. In Gharavi and Tarokh's research, conducting research using decision tree algorithms to determine the development of customers in front of customers using wireless equipment can conclude that C5 is the best algorithm for implementation compared to other algorithms[3]. Research by Daoud about selling online in Marocco got result that the values of RFM variables for each cluster greater than those of the overall average are identified and show that the cluster 7 is the most important cluster because the average values of R, F and M are higher than the overall average value using SOM and K-Means Algorithm[4]. The research conducted by Hardiani and friends, the RFM method was used to cluster Savings Customer segmentation and could produce the 3 optimal clusters which included occasional customer groups, superstars, and typical customer groups[5]. Research on RFM was also carried out by Maryani and Riana, where RFM was used to cluster customers and then analyzed again using the Decision Tree algorithm[6].

In this research, The concept of RFM is added with Cost (RFM-C). Companies need the Cost attribute needed for deductions from other attributes. High costs that arise in the sale of products in a company will certainly affect the sustainability of a company because with high costs will burden the company so that the company is difficult to develop on its business so that it will affect the value of
company efficiency and competitiveness with other competitors. After the data has been adjusted to the RFM-C format, then the RFM-C data is categorized as text and labeling according to the definition given by the company. The algorithms used are C4.5 and Random Tree, both of which are categorized as Decision Tree algorithms.

2. Methodology

There are several methodologies used in processing data such as KDD, CRISP-DM, dan SEMMA. The Knowledge Discovery Databases (KDD) model is an iterative and interactive model. It has total nine steps. It refers to finding knowledge in data and emphasizes the high level of specific data mining method. Cross-Industry Standard Process for Data Mining (CRISP-DM) was launched in late 1996 by Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR. This models the refines over the years. It has six steps or phases. Sample, Explore, Modify, Model, Assess (SEMMA) model was developed by SAS institute. It has five different phases[7].

In this study, illustrated in figure 1, the methodology used was CRISP-DM which has 6 steps in its application, namely: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment in just one round.

![Figure 1. CRISP-DM For Research](image)
3. Results and Discussion

The implementation of the concept from RFM-C at this plastic packaging company will use the CRISP-DM method for the data development stage.

3.1. Business Understanding

In this research, a national private company engaged in plastic packaging manufacturing in Indonesia having its address in Indonesia, where the company has been operating since 2003 and is a domestic investment company with around 500 employees requires the concept customer profiling like RFM-C to analyze characteristics each customer concerned and group it that aims to foster trust in customers. This will see a pattern of assessments of manual assessments so far.

3.2. Data Understanding

There are a total of 172,261 customer transaction data collected from January 4, 2016 to December 4, 2017. In the data there are 278 customers, of which the total amount is IDR 228,210,723,592. Sample data is drawn in the form of Excel as in table 1.

| Sell-to Customer No. | Posting Date | Quantity | Unit of Measure Code | Amount (IDR.) | Cost |
|----------------------|--------------|----------|----------------------|---------------|------|
| XXX-000075           | 29/12/2017   | 700      | PCS                  | 364,700       | 380,324 |
| XXX-000582           | 29/12/2017   | 28,290   | PCS                  | 23,905,050    | 12,843,377 |
| XXX-000320           | 29/12/2017   | 50,000   | PCS                  | 4,000,000     | 2,094,500 |
| XXX-000306           | 29/12/2017   | 100,000  | PCS                  | 7,300,000     | 9,400,000 |
| XXX-000218           | 29/12/2017   | 1,750    | PCS                  | 1,694,000     | 1,801,783 |
| XXX-000106           | 29/12/2017   | 9,000    | PCS                  | 8,550,000     | 8,903,430 |

3.3. Data Preparation

During the data preparation stage, data cleaning, integration, and reduction were carried out beforehand. From the data determined the weight is according to the criteria in the sales data history according to the variable recency, frequency, monetary and cost. In the processing this initial data using the Microsoft Excel application program. (See table 2)

| Recency Scale | Value     |
|---------------|-----------|
| 5             | < 3 Month | Very High |
| 4             | 3 - 10 Month | High    |
| 3             | 10 - 15 Month | Medium  |
| 2             | 15 - 20 Month | Low     |
| 1             | > 20 Month | Very Low |
The process of categorizing Recency is based on the length of time the customer arrives in making a transaction as in table 2, where the closer the customer arrives the higher the qualification value. In table 3, the qualification value is seen how often the customer places an order. In table 4, obtained from the value of purchases made by customers. Likewise with table 5, the classification value is obtained from the costs incurred during production.

| Table 3 | Frequency Scale |
|---|---|
| **Score** | **Frequency (F)** | **Value** |
| 5 | >20 Times | Very High |
| 4 | 10-20 Times | High |
| 3 | 5-10 Times | Medium |
| 2 | 2-5 Times | Low |
| 1 | < 2 Times | Very Low |

| Table 4. Monetary Scale |
|---|---|
| **Score** | **Monetary (M)** | **Value** |
| 5 | > 20 Billion | Very High |
| 4 | 10 - 20 Billion | High |
| 3 | 1-10 Billion | Medium |
| 2 | 500 - 1 Billion | Low |
| 1 | < 500 Million | Very Low |

| Table 5. Cost Scale |
|---|---|
| **Score** | **Cost (C)** | **Value** |
| 5 | > 20 Billion | Very High |
| 4 | 10 - 20 Billion | High |
| 3 | 1-10 Billion | Medium |
| 2 | 500 - 1 Billion | Low |
| 1 | < 500 Million | Very Low |

While the categorization of customers is obtained from the company's assessment of the customer, which consists of: Typical Customer, New Customer, Superstar, and Dormant. Table 6, is a data sample after conversion.

| Table 6. Sales Data after Conversion |
|---|---|---|---|---|
| **Customer Code** | **R** | **F** | **M** | **C** |
| XXX-000001 | Medium | Medium | Very Low | Very Low |
| XXX-000006 | Very Low | High | Very Low | Very Low |
| XXX-000007 | Very Low | Very High | Very Low | Very Low |
| XXX-000008 | Low | Very High | Very Low | Very Low |
| XXX-000010 | High | Medium | Very Low | Very Low |
| XXX-000014 | Very High | Low | Very Low | Very Low |
| XXX-000017 | Very High | Very Low | Very Low | Very Low |
3.4. Modeling

In modeling, using software tools, namely RapidMiner, because doing modeling is very easy and has more users than other tools. RapidMiner offers an integrating environment with visually appealing and user-friendly GUI. Everything in RapidMiner is focused on processes that may contain subprocesses. Processes contain operators in the form of visual components. Operators are implementations of DM algorithms, data sources, and data sinks. The dataflow is constructed by drag-and-drop of operators and by connecting the inputs and outputs of corresponding operators. RapidMiner also offers the option of application wizards that construct the process automatically based on the required project goals[8]. (see figure 2)

![Figure 2. Design model with C4.5 and Random Forest Using RapidMiner](image)

In figure 2, excell data that has been processed is then read by RapidMiner and then validated using 10 Fold Cross Validation for each algorithm. Where the algorithm is C4.5 and Random Forest.

In each of the 10 Fold Cross Validation functions to perform training actions and testing data on the algorithm automatically as many as 10 times. In Figures 3 and 4 are designed for each algorithm.

![Figure 3. 10 Fold Cross Validation For C4.5 Algorithm](image)

![Figure 4. 10 Fold Validation For Random Forest Algorithm](image)
3.5. Evaluation

From the results of the modeling that is executed, a Decision Tree image will be formed as shown in Figure 5 for C4.5 Algorithm and Figure 5 for the Random Forest Algorithm.

![Decision Tree of C4.5 Algorithm](image1)

**Figure 5.** Decision Tree of C4.5 Algorithm

In Figure 5, the C4.5 algorithm illustrates that the root of the tree is recency compared to other attributes. Similarly, Figure 6, where the Random Forest Algorithm makes the Recency attribute as root.

![Decision Tree of Random Forest Algorithm](image2)

**Figure 6.** Decision Tree of Random Forest Algorithm

As for the results of the comparison in table 7, the accuracy of which the algorithm C 4.5 and Random Forest algorithm has the same result IE 98.75%, but Kappa from Algorithm C 4.5 has a value a little higher, namely: 0979 Random Algorithm compared to Forest.

| Table 7. Comparison Value Algorithm C 4.5 and Random Forest |
|-------------------------------------------------------------|
| **Accuracy**       | **C 4.5** | **Random Forest** |
|                   | 98.57%    | 98.57%            |
3.6. Deployment
In the Deployment phase, in this study using the C4.5 algorithm by looking at the rules of the algorithm provided by RapidMiner[9], then the rule is translated into programming. In this study, the programming language used is Visual C# and MySql database[10]. The Deployment results given to the company are depicted in Figure 7. The application is also given a module to load excel data and process data if there is new data that needs to be predicted based on the results of the C4.5 algorithm classification rule.

![Figure 7. The Results Of the Deployment Of the Algorithm C4.5](image)

4. Conclusion
After process the two models, the results obtained are the C4.5. After testing the two algorithms, the results obtained are the use of the C4.5 algorithm for companies to classify RFM-C which is expected to predict because it has higher accuracy and kappa values compared to the Random Forest algorithm. And now these plastic packaging companies can predict customer transaction data based on historical data that has been processed.

5. Acknowledgement
The author wants to show his gratitude to Universitas Komputer Indonesia that have provided insight and expertise that are very helpful in research.

6. Reference
[1] Y. L. Chen, M. H. Kuo, S. Y. Wu, and K. Tang 2009, “Discovering recency, frequency, and monetary (RFM) sequential patterns from customers’ purchasing data,” Electron. Commer. Res. Appl.
[2] J. T. Wei, M. C. Lee, H. K. Chen, and H. H. Wu, “Customer relationship management in the hairdressing industry: An application of data mining techniques,” Expert Syst. Appl., 2013.

[3] E. Gharavi and M. J. Tarokh, “Predicting customers’ future demand using data mining analysis: A case study of wireless communication customer,” in IKT 2013 - 2013 5th Conference on Information and Knowledge Technology, 2013.

[4] R. Ait Daoud, A. Amine, B. Bouikhalene, and R. Lbibb, “Combining RFM model and clustering techniques for customer value analysis of a company selling online,” in Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA, 2016.

[5] T. Hardiani, S. Sulistyo, and R. Hartanto, “Segmentasi Nasabah Tabungan Menggunakan Model RFM (Recency , Frequency , Monetary ) dan K-Means Pada Lembaga Keuangan Mikro ISBN : 979-26-0280-1 ISBN : 979-26-0280-1,” Semin. Nas. Teknol. Inf. dan Komun. Terap., no. May, pp. 463–468, 2015.

[6] I. Maryani and D. Riana, “Clustering and profiling of customers using RFM for customer relationship management recommendations,” in 2017 5th International Conference on Cyber and IT Service Management, CITSM 2017, 2017.

[7] U. Shafique and H. Qaiser, “A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA),” Int. J. Innov. Sci. Res. ISSN, 2014.

[8] M. Hofmann, RapidMiner. 2018.

[9] Hofmann, M., & Klinkenberg, R. (Eds.). (2013). RapidMiner: Data mining use cases and business analytics applications. CRC Press.

[10] Ullman, L. (2006). MySQL: Visual quickstart guide. Peachpit Press.