Customer Clustering and Classification for CRM Loyalty Program in Milk Powder Industry

Melky Simorangkir, Zulkarnain
Department of Industrial Engineering
Universitas Indonesia
Selemba, Jakarta, Indonesia
melky.simorangkir@ui.ac.id, zulkarnain17@ui.ac.id

Abstract

Changes in lifestyle during the pandemic affect people's purchasing power and habits towards the need for baby food, especially dairy products, so that increasing profits and business competitiveness becomes an important part for companies which are engaged in procuring powdered milk for toddlers. One of the strategies to increase profitability and competitiveness is to increase customer retention by predicting customer loyalty and classifying customer profiles in the Customer Relationship Management (CRM) which is often also known as the Customer Loyalty Program. This discussion tries to classify customer loyalty through customer profiles (demography, transactions, etc.) using techniques from data mining methods. Beforehand, to classify the loyal and non-loyal customers as the objective of customer loyalty program, this study used cluster analysis to map the customers based on their behavior recorded in the transactions they made and the information they provided. The objective of this study is to answer several questions such as: is the predicted performance value of loyal customers accurate? And what variables are most important in classifying loyal customers? In addition, this study also aims to understand the use of data mining methods and how to interpret the results of the techniques that have been used. The dataset contains 655,625 observations and 20 variables while for cluster analysis, there are only 7 variables are being used whereby using this method, the optimal clusters obtained are 3 clusters. With the new labels for targeted variable, labels 0 and 1 are labeled as non-loyal customer and label 3 is the loyal customer where the proportion of each label are 76%, 15.18% and 8.82% then the proportion of non-loyal and loyal customers are 91.18% and 8.82%. Using logistic regression method to classify loyal and non-loyal customers, the result showed that the model accuracy is about 96.6% where the misclassification error for the model is 3.4%. With this accuracy value, it can be recommended that this model can be used to classify customers in the customer loyalty program based on the dataset and variables that build the model.

Keywords

Data mining, K-Means, Logistic regression, Customer Loyalty, CRM.

1. Introduction

The covid 19 pandemic that occurred in 2019 until now, has resulted in changes in the lifestyle for the world community which also occurred in Indonesia, especially in the capital city of Indonesia, Jakarta. As the capital city of Indonesia, Jakarta is famous for its dense activities with high mobility and fast movement which later changed due to the pandemic that hit the city. The high mobility and fast movement lifestyle shifted into stay-at-home lifestyle is clearly seen in activities such as studying from home, working from home, and eliminating celebrations that require large crowds such as weddings, birthdays, or other social activities. Lifestyle changes during the pandemic also affected people's purchasing power and habits towards the need for baby food, especially dairy products, thus increasing profits and business competitiveness becomes an important part of the companies in this industry where one of the strategies developed is to build relationship with the customers.

When the world is experiencing a change in the purchasing power of its people, market predictions in Indonesia will always experience growth for dairy products and their derivatives. Thus, to meet ever-increasing market demands, companies need to improve quality and quantity so that customer satisfaction can be achieved. Customer satisfaction is one of the important factors in increasing customer retention, which is the goal of the customer loyalty program which will then be discussed in this study. An integrated model for classifying loyal customer using data mining techniques such as cluster and classification, is a topic that will be discussed in this study, namely K-Means and Logistic Regression models that will be used to explore the variables that build customer behavior. Moreover, this

discussion also tries to investigate the relationship between variables that build loyal customers, such as the number of transactions in the first month, the number of transactions in the second month, the number of transactions in the last six months, the amount of income in a year, the use of discounts, etc.

1.1 Objectives

The purpose of this research is to answer several questions such as (1) how many clusters are created from the customer dataset with predetermined variables, (2) how accurate the classification model on the following customer dataset is resulted, and (3) which variables are important in forming a classification model for loyal and non-loyal customers. Thus, from these questions, this discussion aims to understand the use of data mining methods and how to interpret the results of the data mining techniques used in this discussion.

2. Literature Review

In this section, the literature used in this discussion will be explained, including the relationship between Customer Relationship Management (CRM) and the Customer Loyalty Program which is the focal point of the discussion, as well as explaining the relationship between customer loyalty programs and data mining techniques as the method used in the discussion, namely K-Means and Logistic Regression.

Customer relationship management (CRM) is as a business strategy aimed at optimizing revenue and profit while promoting customer satisfaction and loyalty (Guifang & Youshi, 2010). Customer loyalty is part of the dimensions of CRM, where CRM consists of four dimensions, namely: customer identification, customer attraction, customer retention and customer development (Ngai et al., 2009). As the 80/20 rule goes, eighty percent of benefit comes from twenty percent of customers, so it is essential that companies understand customer need and develop suitable CRM strategies in order to ensure customer retention, loyalty, and satisfaction (Wang et al., 2009). (Figure 1).

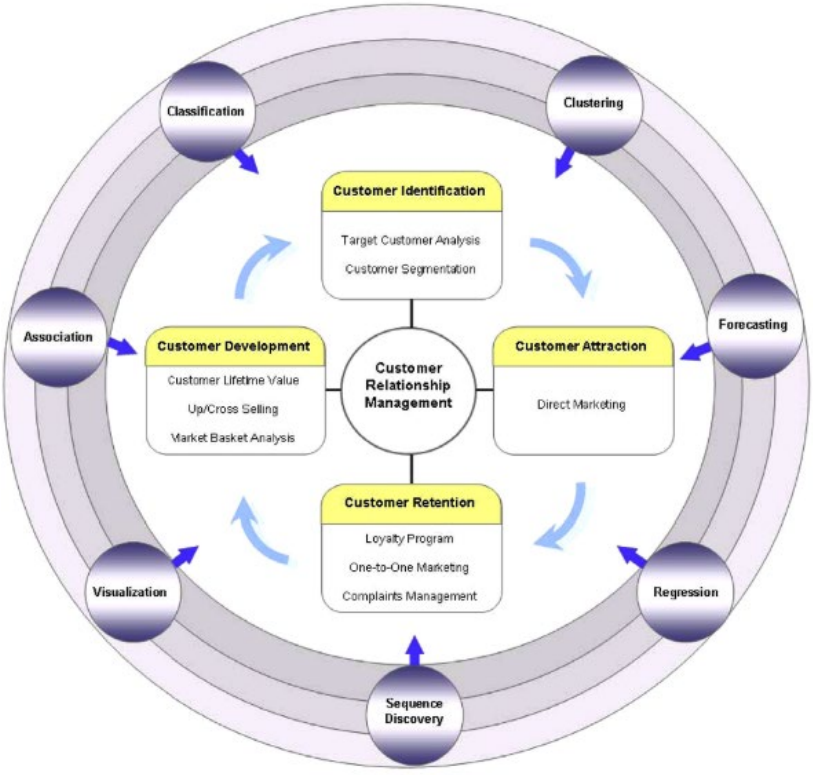

Figure 1. Classification framework for data mining techniques in CRM

In Figure 1 above, it can be explained how the framework helps to identify the main dimensions of CRM and data mining techniques for the application of data mining techniques to CRM (Ngai et al., 2009). The framework is based on research conducted by Swift (2001), Parvatiyar and Sheth (2001) and Kracklauer et al. (2004). Meanwhile, another

study uses six data mining techniques to predict an increase in the frequency of use, including Logistic Regression, Random Forest, Stochastic AdaBoost, Support Vector Machines, Kernel Factory and Neural Network (Ballings & Van Den Poel, 2015) which aims to assess the performance of the model, In this discussion, the data mining technique used focuses on K-Means clustering and Logistic Regression classification.

The k-means algorithm is a simple iterative method to partition a dataset into a specified number of clusters, K (Elbasiony et al., 2013). The algorithm is initialized by picking random K points as the initial K clusters ‘‘centroids’’, then, the algorithm iterates between two steps till convergence: (1) assignment of each point to its closest centroid and (2) relocation of each centroid to the mean of its assigned points. The k-means algorithm uses the Euclidean metric to quantify distance between points (Elbasiony et al., 2013).

Logistic Regression refers to a method for describing the relationship between a continuous response variable and a set of variables that predict it (predictor) (Larose & Larose, 2015), so that logistic regression can be written in the form y=f(x), where y is the response variable in categorical form (True/False, Win/Lose, etc.). In its application, logistic regression can also use response variables in the form of binary (1 and 0) which is also known as binomial logistic regression which aims to show how much chance an event will succeed or not, using the existing predictor variables, then Logistic regression can be formulated in the following equation:

$$Logit_i = \ln\left(\frac{prob_{event}}{1 - prob_{event}}\right) = b0 + b1X1 + \cdots + bnXn$$

In the analysis of the logistic regression model, it will produce several values as output which include Probability, Odds and Log of Odds, besides that in this discussion the AIC value is also a consideration in modeling using this logistic regression. Probability is a value that shows the probability that an event will occur or not, where the resulting value ranges from 0 to 1, while Odds is a value that shows how likely it is that an event A will occur to event A^c and the last is the Log of Odds which is a value generated ln of Odds, the value generated in the log of odds ranges from negative infinity to positive infinity. The following can be described Odds and Log of Odds in the equation below:

$$Odds\ (A) = \frac{P(A)}{P(A^C)}, and$$

$$Log\ of\ Odds\ (A) = \ln\left[Odds(A)\right]$$

Furthermore, the Akaike Information Criterion (AIC) value shows how well the resulting model is in modeling the available data so that the interpretation of the AIC value is the lower the AIC value, the better the model has accuracy in explaining the existing data.

3. Methods

This discussion follows the Cross Industry Standard Process and Data Mining (CRISP-DM) framework, which is a popular methodology for data analytics (Oztekin et al., 2016), The methodological framework used includes, (1) business understanding related to how business goals can be converted into data mining problems, which in this discussion relate to the customer loyalty program; (2) data understanding, basically is how to identify the source of the data, in this case is the CRM database, and obtain data in accordance with the research objectives; (3) the third stage is the data preparation stage, where the variable selection stage is carried out and divides the dataset into training and testing datasets; (4) Modeling, the stages where data mining techniques are used, in this discussion using K-Means Clustering and Logistic Regression; (5) Model evaluation, the stage of analyzing model using the accuracy rate value as the basis for comparison between the models used.

4. Data Collection

A structured customer profile can be extracted from the CRM database where all transaction data and customer demographics are pre-stored. The dataset that will be used is data from 2020 to 2021, where the dataset extraction process will go through the pre-processing stage, this pre-processing stage is included in the data preparation stage,

which is the stage of data transformation to produce a dataset based on research needs. Likewise with the selection of variables to be adjusted which aims to understand the effect of each variable in classifying customer loyalty.

The raw data obtained consists of 31 variables or columns with a total of 1,021,114 observations, with Cluster variable as dependent variable which is consisted with five categories (Lapsed, FANS, New Active, Regular and Loyal). Most of the data collected in Lapsed Clusters is 55.48% where this Cluster is a customer cluster with few transactions in the last 6 and 12 months with an average of 7.88 transactions overall, while the FANS category is a customer who has never made a transaction, customers who redeem vouchers are 107,169 customers or 10.5% of the total customers.

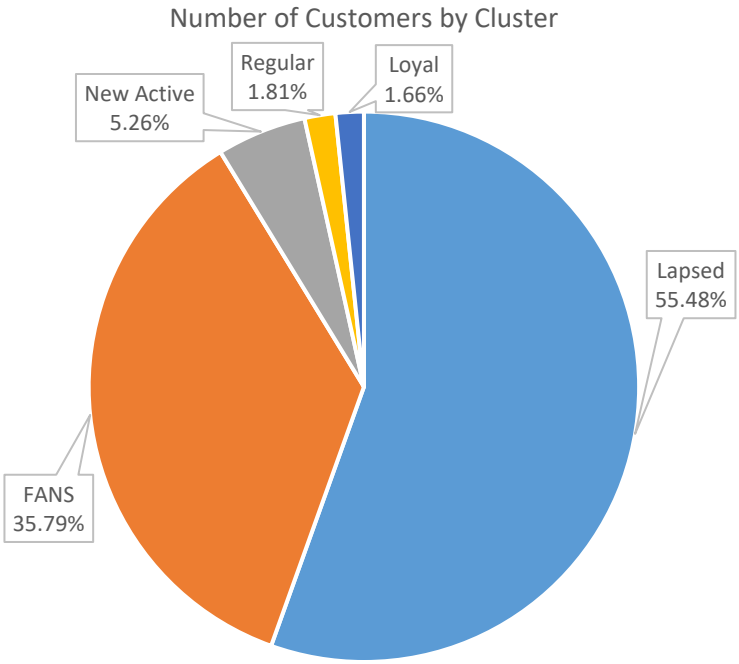

Figure 2. Number of Customers by Cluster Variable

Assuming that the FANS category does not provide sufficient information in forming the model, the next step is to remove that category from the dataset that will be used. (Figure 2).

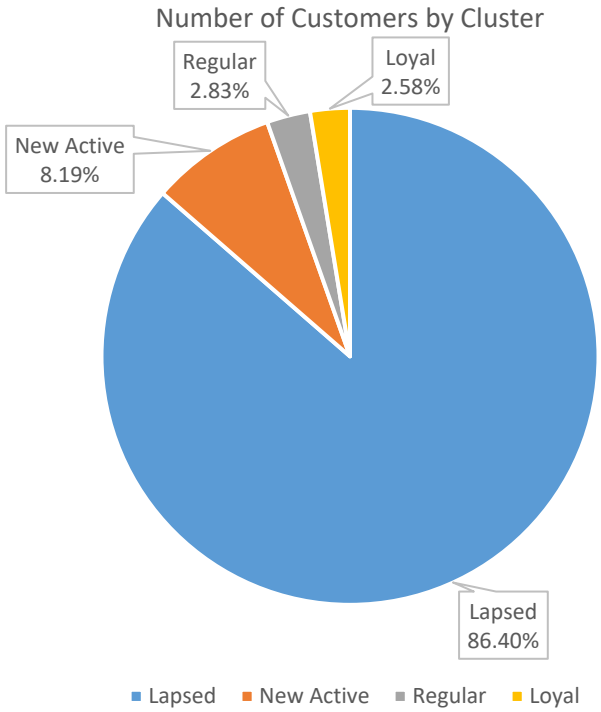

Figure 3. Number of Customers by Cluster Variable (After Removing Fans Category)

After eliminating the FANS category, the number of observations in the dataset which will then be used is 655,625 observations, in addition to the elimination of the FANS category, some variables are not used so that the number of variables to be used becomes 21 variables. This dataset will then undergo a re-clustering process using one of the data mining techniques, namely K-Means Clustering. Where in building the cluster model, the variables used are 7 variables: month_1, month_2, month_3, last_6, last_12, all and acc_age. (Figure 3).

5. Results and Discussion

With the differences in the dimensions of the data, the steps taken are to normalize the values for each variable in the model, and then the K-Means clustering model can then be carried out.

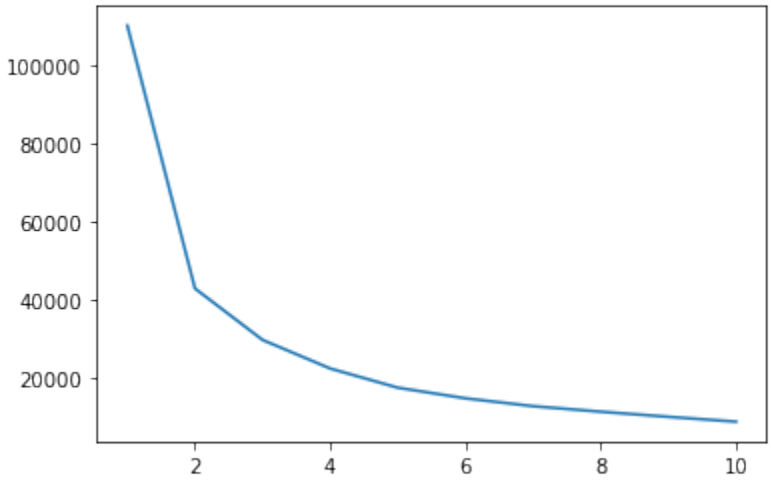

Figure 4. Elbow Method

With an inertia value of 22532.61552325621 using the Elbow Method, the optimal number of clusters is 3 clusters, by implementing a new Cluster Model, with a target variable (Loyal) of 8.82%, this dataset will be used for the Classification Model to classify loyal customers (Figure 4).

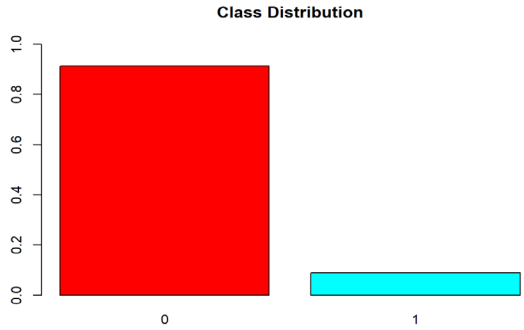

Figure 5. New Dataset Class Distribution

With the difference in the proportion between loyal and non-loyal labels of 8.82% and 91.18%, which in this case is included in the imbalance data, the next step is resampling which in this case is oversampling the following dataset, then the data will be divided into training datasets and testing datasets (Figure 5).

```
pred1 = ifelse(p1>0.5, 1, 0)
tab1 = table(Predicted = pred1, Actual = train.loyal$is_loyal)
tab1
```

```
##          Actual
## Predicted      0      1
##         0 409865  20059
##         1   8605 398411
```

```
1 - sum(diag(tab1))/sum(tab1)
```

```
## [1] 0.03424857
```

Figure 6. Misclassification Error – Train Dataset

As shown in figure 6, the misclassification error obtained in the train dataset is 3.4%, which then using the same method on the test dataset with the following results.

```
pred2 = ifelse(p2>0.5, 1, 0)
tab2 = table(Predicted = pred2, Actual = test.loyal$is_loyal)
tab2
```

```
##          Actual
## Predicted      0      1
##         0 175723   8712
##         1   3621 170632
```

```
1 - sum(diag(tab2))/sum(tab2)
```

```
## [1] 0.03438364
```

Figure 7. Misclassification Error – Test Dataset

With a consistent Misclassification error value in the training dataset and test dataset, which is around 3.4%, it can be said that the model used is consistent with an accuracy of 96.6% (figure 7).

6. Conclusion

Loyal and non-loyal customers as seen in the model will affect several variables that make up, while the model does not affect several variables such as: month_2 (number of transactions in the second month), gram2 (total grams in the second month of transactions), is_product_2 and is_product_4 (stating the last registered product is product 2 and product 4) as well as the gender variable, either female or male. As a step to increase loyal customer the interval redeem variable (the time interval between the first date of redeeming the voucher and the last date of redeeming the voucher) was added and as shown in the classification model, this variable has a tendency to affect loyal or non-loyal customers with a slope value of 3.31e- 11. The Logistic Regression Model used is a Binomial Logistic Regression Model where the label variables used are binary variables 1 and 0 (Loyal and Non-loyal), As for classifying with label variables that have more than 2 categories, we will use the Multinomial Logistic Regression Model.

References

Ballings, M., & Van Den Poel, D. , CRM in social media: Predicting increases in Facebook usage frequency. *European Journal of Operational Research*, *244*(1), 248–260. 2015

Elbasiony, R. M., Sallam, E. A., Eltobely, T. E., & Fahmy, M. M. , A hybrid network intrusion detection framework based on random forests and weighted k-means. *Ain Shams Engineering Journal*, *4*(4), 753–762. 2013

Guifang, G., & Youshi, H. *Research on the application of data mining to customer relationship management in the mobile communication industry.* https://doi.org/10.1109/ICCSIT.2010.5563538., 2010

Larose, D. T., & Larose, C. D. , Data Mining And Predictive Analytics Second Edition. In *John Wiley &Sons*. 2015

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. , Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2 PART 2), 2592–2602. 2009

Oztekin, A., Kizilaslan, R., Freund, S., & Iseri, A. , A data analytic approach to forecasting daily stock returns in an emerging market. *European Journal of Operational Research*, *253*(3), 697–710., 2016

Wang, Y. F., Chiang, D. A., Hsu, M. H., Lin, C. J., & Lin, I. L. , A recommender system to avoid customer churn: A case study. *Expert Systems with Applications*, *36*(4), 8071–8075., 2009

Biographies

Melky Simorangkir is master student in Industrial Engineering, Universitas Indonesia. His research activities are based on methodologies for customer relations management and data mining with interest in data analysis.

Dr. Zulkarnain, ST, MT. an Assistant Professor in Department of Industrial Engineering, Universitas Indonesia. His research interests include Intelligent Transport System, Data Mining, and Operations Research.