RFM-AR Model for Customer Segmentation using K-Means Algorithm

Abstract. In order to develop a marketing strategy, businesses need to implement changes in strategies and innovations in order to compete with their competitors. Business actors are not only focus on ease of transactions and services to provide customer centric strategies, but also need to carry out customer centric strategies by knowing the characteristics of customers and their needs through information on customer characteristics and transactions made by customers. RFM has been extensively used in customer segmentation research. AR model for customer segmentation using K-Means algorithm. The use of data mining techniques allows organizations to manage stored data into new knowledge. Clustering is a technique that divides data into uniform subgroups, which is then regarded as a division of a customer base. In conducting customer segmentation, previous studies mostly use several observable and objective criteria, the elbow method can produce the same number of K clusters on different types of customer clusters with different characteristics. Determination of the best number of clusters with AR variable is K=2. Clustering is divided into 3 grades are high, middle and low. The AR model for customer segmentation using K-Means algorithm is validated with customer's subscription time or since when the customer made an order, Renewal of customer transactions, Frequency, Monetary, Age, Return. The CRISP DM (Cross Industry Process for Data Mining) with a combination of DM (Data Mining) techniques is used in this research. The optimal K value for each RFM is found through the elbow method. One of the main criteria of the K value is the highest accuracy. RFM-AR generates forecasts based on the model behavior. Therefore, the model has a relevance as a predictor of potential customers based on the cluster formed based on the RFM variables. The AR variable is K=2. Clustering is divided into 3 grades are high, middle and low. The K Means algorithm is used in this research is CRISP DM (Cross Industry Process for Data Mining) with a combination of DM (Data Mining) techniques. This research will add AR (Age, Return) to the model, so the model can analyze data with similarities. In conducting customer segmentation, previous studies mostly used one type of customer clusters with different characteristics. Determination of the best number of clusters with AR variable is K=2. Clustering is divided into 3 grades are high, middle and low.

1 Introduction

Competition in the business field is getting tougher, business people are required to carry out various strategies and innovations in order to compete with their competitors. Business actors are not only focus on ease of transactions and services to provide customer centric strategies, but also need to carry out customer centric strategies by knowing the characteristics of customers and their needs through information on customer characteristics and transactions made by customers. RFM has been extensively used in customer segmentation research. AR model for customer segmentation using K-Means algorithm. The use of data mining techniques allows organizations to manage stored data into new knowledge. Clustering is a technique that divides data into uniform subgroups, which is then regarded as a division of a customer base. In conducting customer segmentation, previous studies mostly use several observable and objective criteria, the elbow method can produce the same number of K clusters on different types of customer clusters with different characteristics. Determination of the best number of clusters with AR variable is K=2. Clustering is divided into 3 grades are high, middle and low. The AR model for customer segmentation using K-Means algorithm is validated with customer's subscription time or since when the customer made an order, Renewal of customer transactions, Frequency, Monetary, Age, Return. The CRISP DM (Cross Industry Process for Data Mining) with a combination of DM (Data Mining) techniques is used in this research. The optimal K value for each RFM is found through the elbow method. One of the main criteria of the K value is the highest accuracy. RFM-AR generates forecasts based on the model behavior. Therefore, the model has a relevance as a predictor of potential customers based on the cluster formed based on the RFM variables. The AR variable is K=2. Clustering is divided into 3 grades are high, middle and low. The K Means algorithm is used in this research is CRISP DM (Cross Industry Process for Data Mining) with a combination of DM (Data Mining) techniques. This research will add AR (Age, Return) to the model, so the model can analyze data with similarities. In conducting customer segmentation, previous studies mostly used one type of customer clusters with different characteristics. Determination of the best number of clusters with AR variable is K=2. Clustering is divided into 3 grades are high, middle and low.
2 Literature Review

2.1 Data

For the need understanding the data, there are several identifying issues contained in this data. Meanwhile, the objective of data mining in this study will be carried out in this stage to detect the subset as a reference for the next stage. The data utilized in this research is sales transaction table is presented in Figure 2. The data obtained are 8,594 rows of transaction data containing 8 columns. The sales transaction table is presented in Figure 3. The data obtained from raw data contains various missing values. The selection of columns, records and data attributes is carried out to change the amount of data. The selection of columns, records and data attributes is carried out to change the amount of data. It is seen in Figure 4a. In this research, the preprocessing stage was carried out by the process of cleaning the data. The selection of columns, records and data attributes is carried out to change the amount of data. During the phase of data preparation or data preprocessing, the correlation between the missing tables was checked. The selection of columns, records and data attributes is carried out to change the amount of data. The transformation process includes the procedures to be used as input data at the modeling stage. The transformation process includes the procedures to be used as input data at the modeling stage. In this research, the preprocessing stage was carried out by the process of cleaning the data. The selection of columns, records and data attributes is carried out to change the amount of data. The transformation process includes the procedures to be used as input data at the modeling stage.

2.2 Research Method

The series of stages that can be seen in the flowchart in Figure 3. In this research, the preprocessing stage was carried out by the process of cleaning the data. The selection of columns, records and data attributes is carried out to change the amount of data. The transformation process includes the procedures to be used as input data at the modeling stage. The transformation process includes the procedures to be used as input data at the modeling stage.

Modelling

3.1 Preprocessing

The correlation between the missing tables was checked. The selection of columns, records and data attributes is carried out to change the amount of data. The transformation process includes the procedures to be used as input data at the modeling stage.

Business understanding

Data understanding

Evaluation

Deployment

Data Preparation

Modeling

Data Understanding

Start

Object move to cluster

End

Select the number of cluster center

Put object to closest cluster center

Recalculate the new cluster center

Create cluster based on smallest distance

Set initial cluster center randomly
3.2 Exploratory data analysis

EDA: Exploratory Data Analysis is used to explore the potential knowledge contained in the data. This method helps to construct a new customer chart, monthly product sales chart, monthly distribution chart, and monthly sales chart. The transaction date is a transaction time, which is the date of the transaction. The purchase transaction date is a purchase receipt date. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant. The number of sales transactions cannot be seen in Figure 4c. Random labeling is used to obtain number of sales transactions. The number of sales transactions is not significant.
ICIMECE 2023

Money

customer.

greater the intensity of that the smaller customer's value is customers with ID 63. While the nominal.

Monetary subscribers range of values 0 frequency distribution chart which transaction intensity.

majority of shopping intensity and vice intensity of.

Result

The frequency value is obtained based on the customers.

(b)

Result

value

(b)

Chart

of

number

28

transactions

Meanwhile,

monetary distribution chart which explains the results,

Figure 7b is a customer frequency data with the

of frequency data with the

(b)

Chart

of

number

28

transactions

Figure 7b is a customer
distance (median) between transactions which is quite far.

Average distance (median) between transactions which is quite far.

Return

the transaction is quite past.

value if the renewal data retur

average of 40.94 for customers with ID 987 with

Return (Days)

The age (days) value is obtained based on how

Result

the age (days) value is obtained based on how

Return

the age (days) value is obtained based on how

(b)

return value (days) will be of large

Meanwhile, if you have just subscribed, the

Age (Days)

The return value (days) will be of large

Meanwhile, if you have just subscribed, the

(b)

Return

the return data (days) with the highest

Fig.

The return data (days) with the highest

Fig.

The return data (days) with the highest
3.6 K-Means Customer Segmentation

In the graph, the value of the cluster K=1 to K=2 experienced a significant decrease which formed an elbow at K=2, so it can be concluded that the ideal K value is K=2.

Furthermore, because the graph produces a maximum number of 1000 iterations. In the graph, the value of the cluster K=1 to K=2 experienced a significant decrease which formed an elbow at K=2, so it can be concluded that the ideal K value is K=2.

The results of the cluster test on the frequency of customers are presented in graphical form in Figure 11c with the same K value, namely the range 1 to K=2.

The results of the cluster test on the age (days) and return (days) of customers are presented in graphical form in Figure 11b with the same K value, namely the range 1 to K=2.

The results of the cluster test on the monetary are presented in graphical form in Figure 11c with the same K value, namely the range 1 to K=2.

The results of the cluster test on the monetary which explains that customers tend to dominate.

The results of the cluster test on the monetary which explains that customers tend to dominate.

3.5 K-Means Scoring

The distance between clusters determines the position of a centroid.

3.4 Clustering RFM-AR with K-Means

Furthermore, because the graph produces a maximum number of 1000 iterations. In the graph, the value of the cluster K=1 to K=2 experienced a significant decrease which formed an elbow at K=2, so it can be concluded that the ideal K value is K=2.

The results of the cluster test on the frequency of customers are presented in graphical form in Figure 11c with the same K value, namely the range 1 to K=2.

The results of the cluster test on the age (days) and return (days) of customers are presented in graphical form in Figure 11b with the same K value, namely the range 1 to K=2.

The results of the cluster test on the monetary are presented in graphical form in Figure 11c with the same K value, namely the range 1 to K=2.

The results of the cluster test on the monetary which explains that customers tend to dominate.

The results of the cluster test on the monetary which explains that customers tend to dominate.

3.5 K-Means Scoring

The distance between clusters determines the position of a centroid.

3.4 Clustering RFM-AR with K-Means

Furthermore, because the graph produces a maximum number of 1000 iterations. In the graph, the value of the cluster K=1 to K=2 experienced a significant decrease which formed an elbow at K=2, so it can be concluded that the ideal K value is K=2.

The results of the cluster test on the frequency of customers are presented in graphical form in Figure 11c with the same K value, namely the range 1 to K=2.

The results of the cluster test on the age (days) and return (days) of customers are presented in graphical form in Figure 11b with the same K value, namely the range 1 to K=2.

The results of the cluster test on the monetary are presented in graphical form in Figure 11c with the same K value, namely the range 1 to K=2.

The results of the cluster test on the monetary which explains that customers tend to dominate.
4 Result And Discussion

5 Acknowledgment

REFERENCES


