Business Intelligence using the K-Nearest Neighbor Algorithm to Analyze Customer Behavior in Online Crowdfunding Systems

Chashif Syadzali[1*], *Suryono* Suryono[2], and *Jatmiko* Endro Suseno[2]

[1]Magister Program of Information System, School of Postgraduate Studies, Diponegoro University, Semarang – Indonesia.

[2]Department of Physics, Science and Mathematics Faculty, Diponegoro University, Semarang – Indonesia.

Abstract. Customer behavior classification can be useful to assist companies in conducting business intelligence analysis. Data mining techniques can classify customer behavior using the K-Nearest Neighbor algorithm based on the customer's life cycle consisting of prospect, responder, active and former. Data used to classify include age, gender, number of donations, donation retention and number of user visits. The calculation results from 2,114 data in the classification of each customer's category are namely active by 1.18%, prospect by 8.99%, responder by 4.26% and former by 85.57%. System accuracy using a range of K from K = 1 to K = 20 produces that the highest accuracy is 94.3731% at a value of K = 4. The results of the training data that produce a classification of user behavior can be used as a Business Intelligence analysis that is useful for companies in determining business strategies by knowing the target of optimal market.

Keywords: classification; data mining; k-nearest neighbor; business intelligence; user segmentation; customer life cycle; customer relationship management.

1 Introduction

Customer lifecycle refers to the stages in the relationship between customers and business people and it is important to understand customer behavior in the customer's life cycle. The customer life cycle provides a good framework by implementing data mining for customer relationship management [1].

The use of the internet is undoubtedly its contribution in online marketing. Product information becomes very accessible so that customers become more sophisticated and informative. This certainly makes customers often confused with the many offers, as a result

* Corresponding author: cacip666@gmail.com

many of the customers demand the best offer. To handle this condition, businesses must differentiate their products or services by avoiding unwanted results to become commodities. One effective way to distinguish it is with systems that can interact precisely and consistently with customers by gather customer demographics and customer behavior data and make marketing targeting. This type of targeting also designs effective promotion plans to meet stiff competition or persuade prospective customers when displaying new products [2].

Collection of user behavior data utilizing the Acquisition, Activation, Retention, Revenue, Refer (AARRR) model which is then processed using data mining techniques can improve business performance such as knowing users who are leaving (former), product development advice to developers, ways on how to increase conversions sales and others. The AARRR model is very important to find problems at each customer stage so that the company can improve the performance of Customer Relationship Management (CRM) [3].

User behavior data is informative data that can be very large and irregular, containing user data, user page data, user transaction data, and time series data from these data. Manually processing data can be time consuming and there are a lot of calculation errors, so data mining support is needed to process these data properly and a business intelligence system can be built based on customer relationship management that is useful in marketing [4]. With data mining techniques, business intelligence systems can be built that are used to analyze customer behavior patterns. A lot of customer data and irregular can be classified into several classes, using the K-Nearest Neighbor classification algorithm. The K-Nearest Neighbor algorithm can also be used to process user data, provide real time responses and recommendations to users in a faster process [5].

2 Methods

2.1.Customer Lifecycle

The term "customer life cycle" refers to the stages in the relationship between customers and business. It is important to understand the customer's life cycle because it is directly related to customer revenue and customer profitability [1]. In general, there are four main stages in the customer's life cycle: Prospects are people who are not yet customers but are in the target market; Responders are prospects who show interest in a product or service; Active Customers are people who currently use products or services; Former Customers are probably "bad" customers who don't pay their bills or who incur high costs; those who are not the right customers because they are no longer part of the target market; or those who might have transferred their purchases to competing products.

2.2.Data Mining

Data mining is used to obtain patterns and trends that exist in the data collected. These patterns and trends can be collected together and are defined as mining models that can be applied to certain business scenarios. Data mining usually involves the use of predictive modeling, forecasting, and descriptive modeling techniques as the main elements. Using these techniques, organizations can manage customer retention (retain), be used to choose the right prospects to whom to choose, profile and customer segments (by identifying good customers), set optimal pricing policies, and objectively measure and rank the suppliers which are most suitable for their needs [6].

Classification is the process of finding a model to predict the class of an unknown object class. Classification can also include checking of new features presented in a data set [7]. Classification is divided into 2 stages: the learning stage and the classification stage. The

learning phase, is a model that is used to describe a set of classes that have been predetermined by using an algorithm in a set of training [8].

2.3.K-Nearest Neighbor

The K-Nearest Neighbor algorithm is one of the supervised learning techniques that is often used for classification of pattern recognition, although it is also often used for estimation and prediction. The K-Nearest Neighbor algorithm is memory based, does not need a special model to match it and has a simple concept. No special training procedure is needed for a set of observations other than collecting vectors labeled with the class specified. All intensive calculations are performed on classifications that involve two observations, namely finding the closest k value in the training set and looking for the most votes in the iteration k and class labeling in each classification [9].

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (1)$$

Where d is distance, p is input data, q is training data, i is iteration and n is maximum iteration. The steps taken in the classification using K-Nearest Neighbor are:

1. Determine the parameter k
2. Calculate the distance between each data by the Euclidian method to calculate the distance as in equation 1
3. Sorting data based on distance from small to large
4. Take the data k specified number
5. Looking for data with the most number of k that has been determined
6. Determine the class of data from the most number obtained

2.4 Business Intelligence

Business Intelligence helps companies achieve intelligence to compete [10]. Business intelligence solutions have been prioritized by the organizations implementing these solutions [11]. Business intelligence not only supports the decision making process but also allows businesses to have a better insight into their operations by applying data analysis techniques to their information. The use of business intelligence also allows organizations to incorporate intelligent behavior into their basic functions. Business intelligence provides the support needed for businesses to make decisions using a variety of techniques and tools [12].

2.5 Information System Framework

In this classification system the initial inputs are user's age, gender, donations, count of donations, and the count of page visit. After entering the input data, the system will classify the user's behavior patterns using the K-Nearest Neighbor method. The classification process using K-Nearest Neighbor begins by giving the customer type labels into four namely prospect, responder, active, and former which are then calculated the distance between the data. After knowing the distance of each data, the classification is done on the data that has the closest distance. The Information System Framework could be seen in figure 1.

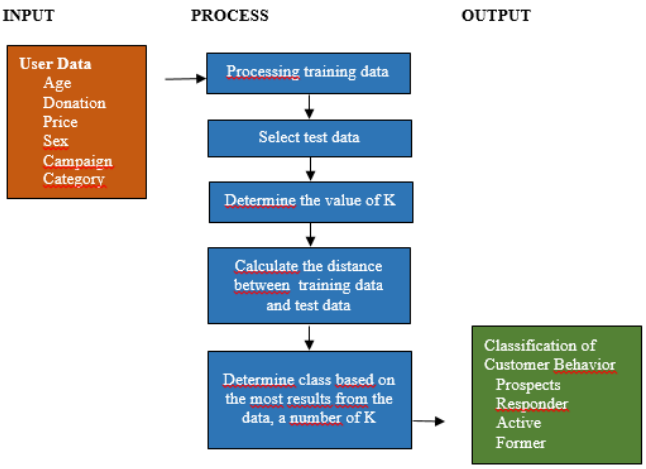

Figure 1. Information System Framework

3 Result

KNN algorithm performs the calculation of the distance between the data to be evaluated with all training data. Then KNN will sort the distance formed ascending to the order K (for example K = 10 then KNN will sort from 1 to 10). Next KNN will pair the corresponding data and look for the number of classes from the nearest neighbors and then set the class as the data class to be evaluated. If you enter the value of K = 3 then the prediction of the classification results is the top 3 data or data with numbers 1 through 3. Then from the three data will be calculated the closest class at most to get the classification results from the calculated user data.

Users who visited the website more than 50 times, made donations more than 5 times and donated more than IDR 1,000,000 were included in the active user criteria. Users who have visited the website more than 30 times, made donations more than 2 times and the amount of donations more than IDR 100,000 is included in the responder user criteria. Users who have visited the website more than 10 times, made donations more than or equal to 1 time and the amount of donations of more than IDR 50,000 is included in the prospect user criteria. Users who have never made a donation and not more than 10 times visited the website entered the criteria of the former user.

The results of this classification can be used as business intelligence analysis to assist business owners in determining target markets. In the diagram, it can be seen that a large percentage of each category of user behavior. There is also a large percentage of accuracy in the performance of the KNN algorithm along with the amount of data processed. In the diagram in Figure 2 there is information on the percentage of user classifications with a large percentage of prospects is 8.99%, responder 4.26%, active 1.18% and former 85.57%. This diagram serves to determine the comparison of user classifications. So companies can find out how many users are still active and can target program offerings to customers in the prospect category so as to increase the percentage of active users. There is also a list of KNN performance, that is, the accuracy of true data using KNN calculation with different K values. Accuracy results with a value of K = 4 have the greatest amount of accuracy which is 93.87%.

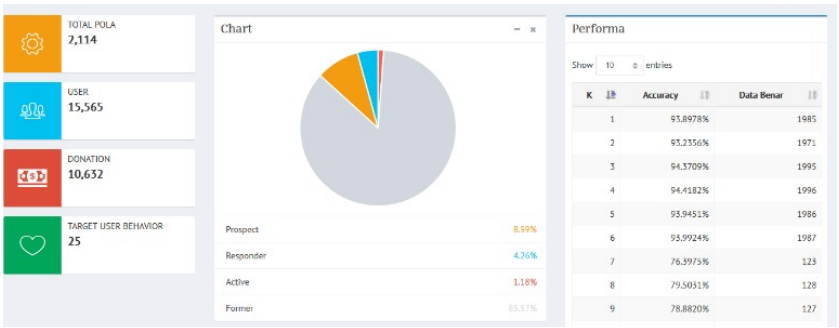

Figure 2. Percentage diagram of user classification and KNN performance

The company can also see the number of daily donations that have occurred over the past 30 days. The blue line on the chart explains the number of confirmed donations, while the red line describes unconfirmed donations. For example, on one day there were 10 donations, then only 8 donations were confirmed or had completed payments, the data in the blue line was 8 and the red line was 2. users in the active category were dominated by users aged 24 years, with 10 male and 12 women. The former category is dominated by 25-year-old users with 760 male and 860 women. The prospect category is dominated by 22-year-old users with 66 male and 94 women. Finally, the responder category is dominated by users are 38 years old, with 32 male and 40 women. Graph of donations and user summary can be seen in Figure 3.

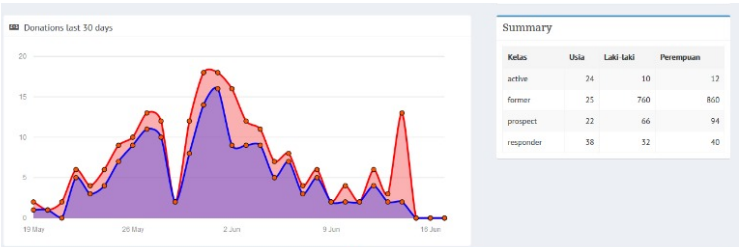

Figure 3. User donation and summary graphics

The newest member displays the last 4 users who registered. The information displayed is the username, photo and date of registration. Latest donation displays the last donation that came in. The information displayed is user id, name, supported campaign id, e-mail, donation amount and donation time date. Latest user information and final donations can be seen in Figure 4.

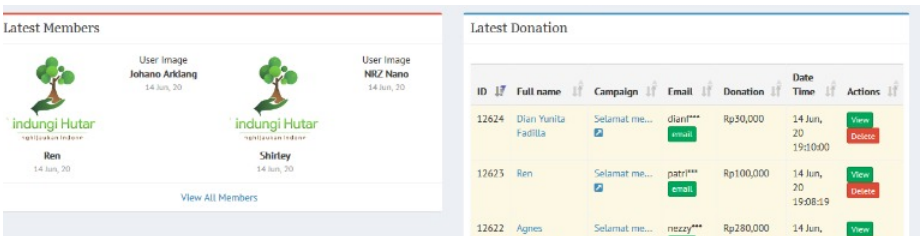

Figure Error! No text of specified style in document.**.** *Latest member* and *Latest donation*

Accuracy testing is performed to determine the level of accuracy produced by matching the results of the classification between training data and system test data using the K-Nearest Neighbor calculation results. In the system there is a "loop accuracy" button to perform repeated accuracy calculations according to the number of K parameters entered. Admin can calculate the accuracy when there is a change in the form of adding or subtracting data on training data as shown in Figure 5.

K	Accuracy	Data Benar
10	93.3222%	1677
9	93.6561%	1683
8	93.4335%	1679
7	93.4335%	1679
6	93.8230%	1686
5	93.4891%	1680
4	93.8787%	1687
3	93.3779%	1678
2	93.5448%	1681
1	93.4891%	1680

Figure 5. Repeated accuracy calculations

The dataset is obtained from the classification data of user behavior patterns that are changed in the form of arrays where the class is the target and the other data becomes the sample data. Then a number of K ranges are repeated using the foreach command. In the process of each K, a repetitive process is also conducted to conduct training on sample data. After the training is finished, the prediction process is carried out from the input data to the sample data. If the data can be predicted, then the data is correct. Then after the looping process is complete, the accuracy value of each K is obtained and stored in a database.

4 Conclusion

The use of classification systems for user behavior patterns can shorten the time in determining the classification of user behavior patterns. The system can do distance calculations to get data with the closest distance of a number of K. Learning of training data is done to calculate the distance between test data and training data using K-Nearest Neighbor. Learning user training data get the classification results of each user category, namely active by 1.18%, prospect by 8.99%, responder by 4.26% and former by 85.57%. System accuracy using a range of K from K = 1 to K = 20 produces the highest accuracy at a value of K = 4 with an accuracy of 94.3731%. The results of learning training data that produce a classification of user behavior patterns can be used as a Business Intelligence analysis. Business Intelligence Analysis can be useful for companies in determining business strategies by knowing the optimal target market.

References

[1] M. Freeman, “The 2 customer lifecycles,” *Intell. Enterp.*, vol. **2**, no. 16, p. 9, (1999).
[2] C. Rygielski, J.-C. Wang, and D. C. Yen, “Data mining techniques for customer relationship management,” *Technol. Soc.*, vol. **24**, no. 4, pp. 483–502, (2002).
[3] Q. CHEN and L. DU, “Managing Mobile Market Users Based on the AARRR Model in the Age of Big Data,” *Manag. Sci. Eng.*, vol. **10**, no. 1, pp. 58–66, (2016).

[4] L. Guoxiang and Q. Zhiheng, "Data Mining Applications in Marketing Strategy," in *2013 Third International Conference on Intelligent System Design and Engineering Applications*, (2013), pp. 518–520.
[5] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," *Appl. Comput. Informatics*, vol. **12**, no. 1, pp. 90–108, (2016).
[6] A. Maheshwari, *Business intelligence and data mining*. Business Expert Press, (2014).
[7] M. J. A. Berry and G. S. Linoff, *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, (2004).
[8] J. Han, M. Kamber, and J. Pei, "Data mining: concepts and techniques, Waltham, MA," *Morgan Kaufman Publ.*, vol. **10**, pp. 971–978, (2012).
[9] G. Suchacka, M. Skolimowska-Kulig, and A. Potempa, "A k-Nearest Neighbors method for classifying user sessions in e-commerce scenario," *J. Telecommun. Inf. Technol.*, (2015).
[10] C. S. Ishikiriyama, D. Miro, and C. F. S. Gomes, "Text Mining Business Intelligence: a small sample of what words can say," *Procedia Comput. Sci.*, vol. **55**, pp. 261–267, (2015).
[11] O. Isik, M. C. Jones, and A. Sidorova, "Business intelligence (BI) success and the role of BI capabilities," *Intell. Syst. accounting, Financ. Manag.*, vol. **18**, no. 4, pp. 161–176, (2011).
[12] A. Bologa and R. Bologa, "Business intelligence using software agents," *Database Syst. J.*, vol. **2**, no. 4, pp. 31–42, (2011).