Analysis Characteristics of Car Sales In E-Commerce Data Using Clustering Model

Puspita Kencana Sari¹, Adelia Purwadinata²

*School of Economic and Business, Telkom University
Jalan Telekomunikasi No.1 Bandung, Indonesia*

¹ puspitakencana@telkomuniversity.ac.id

² adelia@student.telkomuniversity.ac.id

Received on 23-01-2019, revised on 14-02-2019, accepted on 09-04-2019

Abstract

The number of car sales in e-commerce is currently raised along with the increasing use of the Internet in Indonesia. Purchasing of cars in Indonesia are currently getting higher, especially for used cars, caused of new traffic policies (odd/even license plate number) applied in Jakarta. This research aims to study the characteristics of clusters in e-commerce site to predict how are the car sales segmentation. Data are collected from top two e-commerce sites about car selling and buying in Indonesia. Clustering model is build using K-Means method and Davies Bouldin Index for evaluating clusters performance. The result shows there are two clusters formed for each site with similar characteristics. The first cluster is dominated by cars with lower price and older production year. The second cluster is dominated by higher price cars with latest production. The evaluation of model performance from Davies Bouldin Index shows both models are good.

Keywords : Clustering, K-Means, Car Sales, E-Commerce

I. INTRODUCTION

E-commerce is a process of buying, selling transfers or exchanging products on services or information through a computer network on the internet [1]. Business activities in e-commerce spur consumers in using the internet for various reasons in searching for, choosing products, prices, payments and shipping via the internet. This is also based on internet usage which has increased by 143.26 million people [2] [3].

This research is also based on background of new policy regulating road traffic with odd/even license plate number that currently applied in Jakarta. The new policy resulted in increment of car sales, especially for used cars, where those cars are used as complementary car with different plate number [4].

Viewed from the growth of the automotive market in Indonesia [5] in utilizing e-commerce as an online buying and selling transaction, based on TechInAsia [6], the top two in the best car buying and selling sites in Indonesia are Mobil123.com and Carmudi.co.id. Therefore, this research takes those sites (Mobil123 and Carmudi) as the objects.

Mobil123 is a portal of car sales that has largest number of car listings in Indonesia with more than 200,000 vehicles. Mobil123 becomes number one e-commerce of car buying and selling in Indonesia which contains thousands of new and used car advertisements. On this site, both sellers and buyers can easily explore cars information related to their needs and goals. Sellers can easily post cars information that are for sale, and prospective buyers can see many choices offered by the Mobil123. Carmudi Indonesia is a vehicle buying and selling site that presents thousands of vehicles sold every day. Carmudi is the one of largest online marketplace in Indonesia for used cars and new cars. Carmudi is ranked second as the best car buying and selling site in

Indonesia based on TechInAsia [6]. Carmudi works with many local dealers and showrooms to provide car listings on the Carmudi site.

Technological developments at this time have brought various kinds of benefits for the e-commerce. The usefulness of this technology can have an impact on several aspects, one of which is the aspect of digital commerce. One of them is sales in automotive products such as the sale of new cars and used cars. Sales of used cars currently in great demand because the price is lower than the new car and to purchase new and used cars at this time is not only be searched directly through the outlet provided. But with the many e-commerce it can help provide an option to finding cars by online [6]. Sale of used cars that are in great demand today is also stimulated by the needs of many activities [7].

Based on previous research by Farshid Abdi and Shaqhayegh [8] with the title "Customer behavior mining framework (CBMF) using clustering and classification techniques" is states a study in determining patterns of customer behavior and predictions of future actions by using mining techniques on telecommunication companies. And, based on research by Yan Guo and Minxi Wang with the title "Application of improved innovation algorithm in a mobile e-commerce recommendation system" [9] which was done to create the best recommendation system for increasing sales in e-commerce. So, this research will analyze the e-commerce car to find out how the characteristics of car sales on each of Mobil123 and Carmudi sites in area Jakarta. In this analysis uses a clustering model to find out the cluster optimization of each site and analyze the characteristics of car sales. From the results of this clustering, it can find out the characteristics of each area of Jakarta and help the community in determining the choice of car sales, especially in the Jakarta area.

Based on this phenomenon, the research questions in this study are:

1. How is form of cluster on car sales at Mobil123 in Jakarta ?
2. How is form of cluster on car sales at Carmudi in Jakarta ?
3. What is the comparison of characteristics in the Jakarta region from the results of clustering on the Mobil123 and Carmudi sites?

This research is aimed to make the right decision in sales on e-commerce [10]. The result of this research could be used as guide in ensuring e-commerce to enhance of car sales and as to make the good segmentation for car sales in Jakarta.

This research was organized into five section. The first section is introduction to describing the background of research, second section discuss about the literature review. The third section explains the research methods that covers implementation and simulation studies. The fourth section presents the result of simulation studied and evaluating grouping algorithms. The last section is the conclusion of research and future research.

II. RELATED WORK

A. Data Mining

Data mining is an application that uses statistics, machine learning, artificial intelligence, optimization and other analytics that are used to carry out actual research that is useful and solves a commercial problem [11]. Data mining can be utilized in knowing an event, such as analyzing and knowing a suspicious transaction, misuse of actions and used to regulate a sales position that aims to facilitate buyers of movements in it [12] [13].

Systematically, there are three main steps in data mining [14]:

1) Exploration

Is the initial processing of data consist of "cleaning" data, normalization of data, transformation of data, incorrect data handling and so on.

2) Build a model

Perform analysis of different models and choose the model with the best predictive performance. In this step methods are used such as classification, cluster analysis, associations and so on.

3) *Application*

Application means applying a model to new data to produce predictions of the problem being investigated.

B. *E-commerce*

E-commerce is a process of buying, selling transfers or exchanging products on services or information through a computer network on the internet [1]. Business activities in e-commerce spur consumers in using the internet for various reasons in searching for, choosing products, prices, payments and shipping via the internet. The e-commerce model on data mining [11] is an e-commerce based on data model that has various forms. E-commerce refers to data related on the web in mining to determine an optimal strategy for sales products and display strategic information to visitors on every site access.

E-commerce on data mining [11] is a model of data that has various forms. E-commerce refers to the related data on web mining to determine an optimal strategy for sales product and give a strategic information to visitors on every access.

C. *Clustering*

Clustering is a method used to create a series of data to form several groups based on pre-determined similarities. Clustering is data in one cluster that has a high level of similarity and data in different clusters has a low level of similarity [14].

Clustering on business is a place where a company that has a large information of data on all customers, can implemented a clustering as customer segmentation in a small group with the aim of doing analysis and strategies marketing [14].

D. *K-Means Algorithm*

K-Means is an iterative grouping algorithm that partitions the data set into a number of k clusters that have been set at the beginning. The k-means algorithm is implemented quickly, is adaptable and is commonly used in practice. K-Means is one of the most important algorithms in the field of data mining [14]. K-Means is the most classic partition-based data grouping method from one of the ten classic data algorithms. K-means classifies the objects closest to grouping point k. where the iterative centroid values of clusters are updated one by one until the best grouping results are obtained [15]. This algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^C \sum_{j=1}^{C_i} (|xi - vj|)^2$$

Where, ' $\| - \|$ ' is the Euclidean distance between and ' $'$ ' is the number of data points in cluster ' c ' is the number of cluster centers

E. *Davies Bouldin Index*

Davies-Bouldin Index (DBI) is one of method used in measuring cluster evaluation in a grouping method, this grouping is based on the value of cohesion and separation, where cohesion is defined as the sum of the proximity of the data to the cluster center point of the cluster followed. While separation is based on a distance between the cluster center points to the cluster. If the inter-cluster distance is maximal, it means that the characteristics of each cluster are small so that the differences between clusters are clearer. If the minimum intra-cluster distance means that each object in the cluster has a high level of characteristic similarity [16].

Davies Bouldin Index is a metric for evaluating grouping algorithms, where this validation is to see how well grouping has been done using the number and features attached to the dataset. The smaller the DBI value is obtained (non-negative ≥ 0), the better the cluster obtained from grouping using the clustering algorithm [17].

III. RESEARCH METHOD

The research conducted by the author uses qualitative methods. Qualitative research methods are research methods that is used to analyzing data in the form of descriptions of data that cannot be directly quantified [18]. Some steps for this research conducted are:

- 1) Data collection (crawling data from website); using the Parsehub as the tool.
- 2) Data preprocessing; doing by cleansing data that not relevant with this research.
- 3) Data processing; using k-means method to construct clustering model with Rapidminer and Orange as the tools.
- 4) Model evaluation; using Davies Bouldin Index to see how good clusters had been formed.
- 5) Data analysis; to know how the characteristics from each cluster.

A. Data Analysis Technique

In this research, data collected for analysis are from all cars listed in Mobil123 and Carmudi sites with a sample from August to December 2018 based on the start of odd/even plate number policy. Data are collected by doing web mining on Mobil123 and Carmudi sites using ParseHub. The attributes used in this study are brand, prices, location, production year, and region. The data that was successfully taken in the first phase from the two sites were 5000 data. The data were cleansed to eliminate irrelevant data. After this preprocessing by cleansing data, total data that will be processed are 2,149 from Mobil123 and 472 from Carmudi, as seen in Tabel 1.

TABEL I
 NUMBER OF DATASET FOR EACH E-COMMERCE SITE

E-Commerce	Total
Mobil123	2149 Unit
Carmudi	472 Unit

After doing the data collection and preprocessing steps, the data are processed again using the Orange application to find and determine clusters of car sales data. Orange application determine how to find the good cluster or optimal number of clusters based on Silhouette value, where the higher average value means better number of clusters.

Fig.1. Silhouette Score From Orange

After determining number of the cluster, the next step is identification of clusters. Data are processed using Rapidminer application to show visualization from each cluster and to analyze how characteristics of car sales from each e-commerce sites.

Fig.2. Visualization Using Rapidminer

B. Performance Measurement

This analysis then continued using measurement of Davies Bouldin Index. DBI is a metric for evaluating grouping algorithms, where this validation is to see how well grouping has been done using the number and features attached to the dataset. The smaller DBI value obtained (non-negative ≥ 0), the better cluster formed [17].

IV. RESULTS AND DISCUSSION

A. Characteristics of Data

From data processing for Mobil123 and Carmudi data using Orange, the cluster formed are two clusters, where the determination is based on the best optimal numbers of cluster with the highest Silhouette value. The value are 0.785 or 78% for Mobil123 and 0.707 or 70% for Carmudi. Tabel 2 are result of Silhouette score calculation.

TABEL III
 BEST OPTIMAL CLUSTER PARAMETER

Mobil123		Carmudi	
k	Silhouette Score	k	Silhouette Score
2	0.785	2	0.708
3	0.765	3	0.630
4	0.705	4	0.606
5	0.674	5	0.617
6	0.599	6	0.572
7	0.598	7	0.587

After doing pre-processing, this research get data to be processed are 2,149 from Mobil123 and 472 from Carmudi with sample data from August to December 2018. Based on these results, the number of objects for each cluster in each site can be seen in Tabel 3.

TABEL III
 RESULT NUMBER OF CLUSTER MOBIL123 AND CARMUDI

No	Label	Mobil123	Carmudi
1	Cluster 0	1.880 Units	379 Units
2	Cluster 1	268 Units	93 Units
	Total	2.149 Units	472 Units

Cluster 0 and cluster 1 are cluster results that form based of processing data using Rapidminer.

B. Results of Cluster Characteristic

Before the data are processed using k-means method, non-numeric attributes such as brand and location, were transformed into intervals scale using Method of Successive Interval (MSI) method. After that, all data can be processed using the k-Means method. Tabel 4 shows the result of clustering data process.

Attribute	Mobil123	Carmudi
Price	Cluster 0 (Left) : < 625 million Cluster 1 (Right) : 625 – 3.950 billion	Cluster 0 (Left) : < 375 million Cluster 1 (Right) : 375 – 1.775 billion
Location	Jakarta Cluster 0, 1 : Scattered in five regions of	Regions Cluster 0 : Dominant spread over five

Fig 3. Cluster Result on Data Mobil123 and Carmudi

Based on the results of Fig 3. above, it can be concluded that the cluster form are two clusters for each car sales sites, Cluster 0 and Cluster 1. The clusters at each site have different characteristics, where the first cluster is Cluster 0 has characteristic with lower sales price and older production year. The second cluster is cluster 1 has higher price with latest production. For the car sales at Mobil123, Cluster 0 has characteristic which sales with range of less than 625 million with Scattered in five regions of Jakarta, with sales of the old to the newest type of cars from 1969 to the year 2018, which sales dominant by Asia region with brands such as Toyota, Honda and Daihatsu.

TABEL IV
 REGION OF BRAND ORIGIN MOBIL123

No	Region	Brand	Total
1	America	Chevrolet	43
2		Chrysler	1
3		Dodge	2
4		Ford	35
5		Hummer	2
6		Jeep	14
7	Asia	Daihatsu	124
8		Datsun	16
9		Honda	367
10		Hyundai	66
11		Isuzu	19
12		KIA	14

13	Asia	Lexus	33
14		Mazda	76
15		Mitsubishi	112
16		Nissan	148
17		Proton	5
18		Subaru	1
19		Suzuki	107
20		Toyota	591
21		Wuling	2
22	Europe	Audi	10
23		Bentley	1
24		BMW	96
25		Jaguar	2
26		Land	11
27		Mercedes-Benz	201
28		MINI	13
29		Opel	1
30		Peugeot	2
31		Porsche	17
32		Smart	1
33		Volkswagen	13
34		Volvo	2

While in cluster 1 is dominated by cars with sales price is higher from 625 million to 3.950 billion Rupiah, distributed in five locations in Jakarta (South Jakarta, West Jakarta, East Jakarta, North Jakarta and Central Jakarta), with cars production year are quite new (from 2000 to 2018) and car manufacturers dominant are from European region, which are high-type car brands such as BMW, Mercedes and MINI. For Carmudi, it can be concluded that the first cluster form is Cluster 0 with the criteria for cars sold with price is less than 375 million, are also spread over five regions in Jakarta, with older production years (from 1981 to the latest year 2018), and car manufacturers is dominated by Asia region with brands such as Daihatsu, Honda and Toyota. Meanwhile, Cluster 1 are consisted of cars sold with higher price (375 – 1.775 million Rupiah), with latest production year, and brands mostly from Asian manufacturers.

Table V shows that the number of car units per brand. Based of this table the most car sales in Indonesia are brands of Toyota and Honda which is a production car Asia.

TABEL V
 REGION OF BRAND ORIGIN CARMUDI

No	Region	Brand	Total
1	America	Chevrolet	36
2		Jeep	1
3	Asia	Daihatsu	52
4		Honda	101
5		Hyundai	9
6		Isuzu	1
7		Lexus	1
8		Mazda	7
9		Mitsubishi	13
10		Nissan	8
11		Suzuki	53
12		Toyota	141
13		Wuling	3
14	Europe	BMW	19

15		Land	1
16		Mercedes-Benz	25

From the results of the cluster analysis above, the calculation is carried out using DBI calculation where the results for Mobil123 is 0.129 and the results for Carmudi is 0.138. The smaller the DBI value or closer to 0 indicates better cluster obtained [17]. DBI value that has been obtained for these two sites shows the clustering model in this study are quite good.

IV. CONCLUSION

From this study, it can be concluded the characteristics for both clusters have differences mainly in sale price and year production. The first cluster is dominated by cars with lowers sale price and older production year. In the second cluster mostly are cars with higher price and latest production. In Mobil123, first cluster is dominated by brands from Asian manufacture. Meanwhile, both clusters in Carmudi is dominated by Asian brands. Cluster performance evaluation is taken from Davies Bouldin Index that shows the model is quite good for both sites.

This research shows a description of the population of cars in Jakarta which dominated by Asian car brands. According to the current market share of cars where car sales are increased by 26% compared to the previous year with the highest total sales of the Honda, Mitsubishi and Toyota brands with lower prices [19]. The results of this research are aligned to the car market share where the first cluster in both sites has bigger objects than second cluster. Therefore, this result can be used as a reference for e-commerce car buying and selling sites to make car sales according to the cluster characteristics with the market segment are lower price with brands from Asian manufactures.

For the future research, this model can be enhanced by doing a comparison with other clustering method and adding more attributes.

ACKNOWLEDGEMENT

The author would like to thank to the Research Center Social Computing and Big Data which give information about this research and their support during this research. I also would like to thank to the reviewers and the editor for their valuable feedback and comments.

REFERENCES

- [1] P. Mahir . "Klasifikasi Jenis-jenis Bisnis *e-commerce* di Indonesia" . *Jurnal Neo-bis. Volume (9)*. (2015).
- [2] APIII . "Penetrasi pengguna Internet". <http://teknopreneur.com/>.(2017).
- [3] Kusuma, and Hendra, "Pendapatan Perkapita Rakyat Indonesia". <https://finance.detik.com/>. (2017).
- [4] M.N. Ghulam . "Jumlah Penjualan Mobil". <http://otomotif.kompas.com/read/2018/04/28/08420>. (2018).
- [5] Sugiharto, and Jongkie . "Data wholesales mobil kuartal 1 2018". <http://www.gaikindo.or.id/>.(2018).
- [6] R.F. Maulana." Kumpulan situs jual beli mobil terbaik di Indonesia". <https://id.techinasia.com/5-website-marketplace-otomotif-di-indonesia-ini-bisa-membantu-menemukan-mobil-impian-anda>. (2017).
- [7] W. Priyanto. "Aturan Ganjil Genap Dongkrak Penjualan Mobil Bekas ". <https://otomotif.tempo.co/read/1141936/aturan-ganjil-genap-dongkrak-penjualan-mobil-bekas/full&view=ok>. (2018)
- [8] A.Farshid and S.Abolmakare."Customer Behavior Mining Framework (CBMF) using clustering and classification techniques" . *Journal of Industrial Engineering International / SpringerLink*. (2018).
- [9] Y. Guo, M. Wang, and X. Li. "Application of an improved Apriori Algorithm in a mobile e-commerce recommendation system", *Industrial Management & Data System* , <http://doi.org/10.1108/IMDS-03-2016-0094>. (2017) : 287-303
- [10] M. Metisen, and Benri. "Analisis Clustering Menggunakan Metode K-Means dalam Pengelompokkan Penjualan Produk pada swalayan Fadhila" . *Journal Media Infotama vol.11 No.2*. (2015).
- [11] S. Kudyba. "Big Data, Mining And Analytics". *Taylor & Francis Group,LLC : by Thomas H.Davenport*. (2014).
- [12] P. Widodo, Prabowo, T. Handayanto, Rahmadya and Herlawati . "Penerapan Data Mining dengan Matlab" . *Rekayasa Sains, Bandung*. (2013).
- [13] G. Vossen. "Big Data as the new enabler in business and other intelligence" . *Journal of Computer Science/SpringerOpen*. (2014).
- [14] P. Eko . "Data Mining mengolah data menjadi Informasi menggunakan matlab".*Andi Offset,Yogyakarta*. (2014).

- [15] Z. xin, L. Qinyi et al . “Image Segmentation Based In Adaptive K-Means Algorithm” . *Journal on Image and Video Processing.* / *SpringerOpen* . (2018).
- [16] Wani, M.Arif, and R. Romana . “A new cluster validity index using maximum cluster spread based compactness measure.” *Journal of Intelligent Computing and Cybernetics./ResearchGate*. (2016).
- [17] B.J.D. Sitompul. “Peningkatan Hasil Evaluasi Clustering Davies-Bouldin Index dengan Penentuan Titik Pusat Cluster Awal Algoritma K-Means”. *Teknologi Informasi, USU*. (2018).
- [18] Indrawati . “Metode Penelitian dan Bisnis”. *Konvergensi Teknologi Komunikasi dan Infromasi. Refika Aditama*. (2015).
- [19] M.N Ghulam. “Penjualan Mobil Melonjak 26 persen Juli 2018”.<https://otomotif.kompas.com/read/2018/08/28/092200515/penjualan-mobil-melonjak-26-persen-juli-2018>. (2018).