Customer Segmentation on Online Retail using RFM Analysis: Big Data Case of Bukku.id

Mohamad Abdul Kadir[1], Adrian Achyar[2]
{adi@bukku.id[1], adrian_achyar@yahoo.com[2]}

[1,2]Faculty of Economics and Business, Universitas Indonesia, Jl. Salemba 3, RW 5, Kenari, Senen, Kota Jakarta Pusat, DKI Jakarta 10430, Indonesia

Abstract. The purpose of this research is to identify customer purchase behavior, form customer segmentation, and identify customer address of Bukku.id. this research uses customer purchase data of Bukku.co.id in the period 1 September 2017 – 17 September 2018. RFM method and clustering are used to identify customer segmentation. Then, pareto analysis results which publishers and authors need to be concerned for prioritizing effort in order to gain maximum benefit. Customer address or location has been mapped based on priority authors to determine promotion and offline marketing strategy. The results of this research show three customer cluster based on RFM and clustering analysis. Each cluster has different characteristic and it can determine which strategy suit to approach their customers. Customer profile based on authors and publisher could also benefit the company to prioritize any treatments relate to them. Better offline marketing strategy can be developed by knowing location analysis.

Keywords: Customer Segmentation, Big Data, RFM, Clustering, Location.

1. Introduction

In recent years, middle class income has been grown significantly around 53 million people by 2010 (66% growth). It is predicted to grow further than 150 million in 2014 [1]. Book can be categorized under household consumption that reach 194.4 trillion rupiahs in 2010. It is assumed that market value of book industries reaches 14.1 trillion excluding school textbook and government project book [1].

Internet become an important component in today's business life. Many organizations decide to extent their business model by exploiting online strategy to reach superior growth, profit, reputation and matching customer needs [2]. It is not only record purchase data but also potentially record location, demographic and psychographic of online customer. The abundant of online data make it possible to implement big data analytics using certain algorithm to gain customer-centric insight.

Internet is commonly used to sale and advertise products or services of multiple genre. There are benefits of using internet as media advertising such as cheaper, easily accessible, use as per convenience, price comparisons, etc. The birth of online retailing using internet started with the launch of two big online retailing websites, eBay and Amazon [3].

Customer buying behavior is a buying behavior of costumer for personal consumption, it could be individuals or household consumption [4]. This buying behavior shows how customer purchase goods and services. Comprehensive understanding of buying patterns will benefit the company for strategic marketing, segmentation, distribution, and promotion.

ICEASD 2019, April 01-02, Indonesia

DOI 10.4108/eai.1-4-2019.2287279

Customer Segmentation is considered an effective method for managing customers while developing diverse marketing strategies, it is the process of dividing customers into homogeneous and distinct groups [2]. Segmentation could be done according to customer characteristics, which are tracked online helped by certain algorithm.

Company need focusing the target customer then gaining maximize profit with win-win situation for company-customer. Customer segmentation is one of solution to optimize the result of win-win situation [5].

2. Literature review

2.1 Online retail

The increasing of online sales indicates that the way consumers purchase for and use financial services has changed [6]. There are unique characteristics of online shopping, such as: each customer ’s shopping process and activities can be tracked instantaneously and accurately, each customer’s order is associated with a delivery address and a billing address, and each customer has an online store account with essential contact and payment information. These enabled online retailers to treat customers personally with understanding of each customer and to build upon customer-centric business intelligence [7].

Regarding to customer-centric business model, Online retailers are usually concerned with the following common business concerns:

- Who are the most/least valuable customers to the business? What are the distinct characteristics of them?
- What are customers' purchase behavior patterns? Which products/items have customers purchased together often? In which sequence the products have been purchased?

2.2 Big data

Big data have been used to described data sets and analytical techniques in application that are complex and large. Then, they require advanced and unique data storage, management, analysis, and visualization technologies [7]. These data can be structured, semi-structured or unstructured and can be found in several formats, such as video, images or text from social media platforms. This explosion of data in terms of volume, structure and format calls for new approaches capable of processing and analyzing large amounts of data in real time [2].

Big data approaches derived from traditional data mining. These are able to follow information flow and analyze it in real time. Currently, organizations are using big data sources and integrate new approaches of data analysis in order achieve deeper understanding of their customers behavior and optimization of customer engagement [2].

2.3 Customer segmentation

Customer segmentation is the basic steps in improving customer's journey and achieving customer engagement [8]. It is a process of dividing the customer base into distinct and homogeneous groups in order to develop differentiated marketing strategies based on their characteristics [9]. Customer segmentation types intent to support and develop different business tasks or activities regarding marketing goals and can be analyzed by appropriate analytical techniques or tools.

Big data can play a major role in online customer segmentation, since the volume of customer data gathered online rapidly grows. Moreover, there are already available various big data tools which should be able to assist customer segmentation [2].

Traditionally, customer data can be gathered from several sources, but mostly from offline databases, purchase records and invoices. Things are more complex when it comes to the online environment, because of the amount and the variety of customer attributes that can be collected from online channels and real time. Apart from plain demographics or transactional customer data, data gathered from online channels can also reveal customer behavior and preferences [2].

2.4 RFM analysis

Recency, Frequency and Monetary (RFM) analysis is a marketing technique in analyzing customer behavior such as how recently a customer has purchased, how often the customer purchases, and how much the customer spends. It could improve customer segmentation by dividing customers into various groups for future personalization services and to identify customers who are more likely to respond to promotions [10]. The advantage is that the customers' behavior can be captured by using a relatively small number of features, which improves the transparency of the target selection models that are developed [11]. The RFM variables are appropriate for capturing the specifics of the customer's purchase behavior [11].

2.5 Cluster analysis

Cluster analysis is a traditional statistical method that was used for simple data mining. It is used for classifying things into segments whose have similar characteristics [12]. Data inputs are treated similarly in order to obtain information for deciding associations or groups. Therefore, in clustering the characteristics according to which the objects are categorized into segments or classes are initially unknown [9]. These techniques are often used for customer segmentation and can be applied to huge datasets [2].

There are several algorithms which are used in develop clustering. K-means is one of popular algorithm for cluster analysis which is a quite fast algorithm applicable in large datasets that requires predetermination of the number of clusters by the user. Another algorithm, TwoStep which processes the records in two sets and can automatically determine clusters. Then, Kohonen network/Self Organizing Map (SOM), which is a unique neural network architecture that produces a two-dimensional map of the clusters and is slower than TwoStep and K-means [9] [12].

2.6 Pareto analysis

Pareto analysis or 80/20 rule stated that low effort (input) necessary to gain high return (output). This can be used to prioritize products, services, or clients which result optimal output. 80/20 rule is counterintuitive [13].

3. Company overview - Bukku.id

Bukku.id is a website and integrated book commerce owned by PT Bukku Media Integrasi. Fundamentally, it is an ecommerce business that selling books and merchandise through web commerce and chat commerce.

For marketing activities, they use Instagram, Website, Facebook, and Line as main media promotion and campaign. To serve customer, they provide Whatsapp, Line, Messenger, Website as service channel. They create customer engagement mostly on Instagram.

4. Proposed frameworks of study

This research model (Figure 1) present an approach which uses RFM analysis in data mining tasks. The proposed model can assist managers in developing better marketing strategies to fully utilize the knowledge resulting from data mining and RFM analysis.

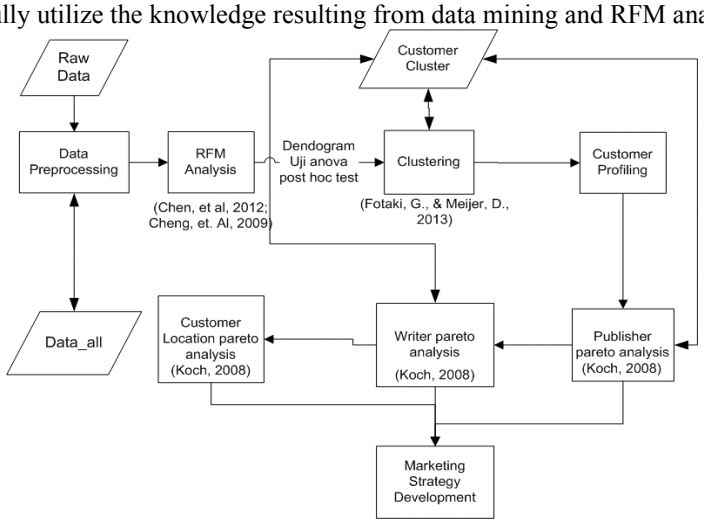

Fig. 1. Conceptual model of this research.

5. Methodology

The research flow can be explained as follow:

- Data Pre-processing: secondary data of Bukku.co.id customers are treated by researcher using data error elimination and rebuild any necessary tables (book, customer, and transaction tables). Information regarding days and months will be extracted from the data.
- Data Processing (RFM): new dataset based on category are ready to be processed. The transaction table consists of 19,813 rows, 344 book titles, 173 authors, and 37 publishers. This data processing will be using RFM analysis. RFM Analysis use recency, frequency, and monetary attributes to characterize customers.
- Clustering using K-Means: the RFM result will be clustered using K-Means clustering to obtain customer segmentation. To distinguish the RFM profile with each other, the R, F, or M value is examined whether above or below the average value by assigning high (↑) or low (↓). The high sign means that recency is below the mean value, while the low sign means that the recency is above the mean value. For recency the smaller number would be assigned high because of the customer just bought recently. In the other hand, the high means that the frequency or monetary element is above the mean value, while the low sign means the frequency or monetary element is below the mean value. The ideal customer has low recency since they just bought the product recently,

have high frequency of shopping, and have high monetary value with big amount of spending.

- Pareto Analysis: this help to identify any authors or publishers which have higher sales than others. Company may optimize profit by promoting and selling fast moving product based on pareto analysis.
- Customer location analysis: by using customer profiles, then it can be analyzed based on customer location. This will help company to develop any online-offline marketing activities and collaboration between company, publisher, author, and other relevant stakeholders.
- Generate marketing strategies based on customer profiles: marketing strategy based on customer profile and location can be formed to have more targeted marketing action.

Raw data. The raw data used in this study provided by Bukku.co.id. it is consisted of customer purchase data. It is derived from purchase transaction occurred during 1 September 2017- 17 September 2018 consist of 19,820 rows and 8 main variables as follows:

Table 1. Variable descriptions of raw dataset.

Variable name (type)	Description
Trx id (Numeric)	Transaction order per book title
Order date (Date Format)	Date of customer purchase through web or chat services
Full name (Character)	Name of customer
address (Character)	Customer address for delivery destination
Title (Character)	Book title owned by author name
Product order qty (Numeric)	Number of products requested by customer
Product price (Numeric)	Net Price of Products
Shopping price (Numeric)	Shopping value made by customer: product price times product order qty

6. Results and findings

RFM Analysis. RFM analysis is performed to all dataset. First, Recency, Frequency, and Monetary are calculated. The RFM data is clustered according to their value of Recency, Frequency, and Monetary. By using R studio, the numeric result of RFM method is shown on table 2.

Table 2. Numeric value of RFM.

No	Recency	Frequency	Monetary
1	80.70834	1	51,480

2	80.70834	1	178,000
3	80.70834	3	236,200
4	80.70834	4	277,650

Each customer has numeric value of RFM. This new dataset will form cluster by implementing clustering analysis based on RFM numeric value. Statistic summary of RFM numeric value is shown below:

Table 3. Statistic summary of numeric RFM.

Recency (days)	Frequency (times)	Monetary (Rupiahs)
1st Qu.: 68.7083	1st Qu.: 1.00	1st Qu.: 71,550
Median: 137.7083	Median: 1.00	Median: 89,000
Mean: 142.3653	**Mean: 1.26**	**Mean: 115,291**
3rd Qu.: 198.7083	3rd Qu.: 1.00	3rd Qu.: 124,800
Max: 379.7083	Max: 14.00	Max: 2,252,800

The most important value to develop cluster is mean value. Mean value of recency, frequency, and monetary respectively are 142.36 days, 1.26 times, and Rp 115,291.

Clustering Analysis. The number of clusters is determined following the dendogram cluster. K Means analysis is conducted to develop dendogram cluster. It generated the appropriate K value = 3. Now that cluster has been set, the mean value of each cluster is shown below.

Table 4. RFM cluster based on mean value of cluster.

Group	Recency (days)	Frequency (times)	Monetary (rupiahs)
2	141.12	1.09	85,604
3	143.37	1.89	218,950
1	195.42	4.03	684,362

By knowing cluster of customers based on K Means, each customer id is mapped to cluster group. So that each customer has certain characteristic like other customer in the same group. Any customer in different group has different characteristic and it can be treated differently as well.

Table 5. RFM score of customers.

No	Recency	Frequency	Monetary	Cluster	Customer ID
1	80.71	1	51480	2	4193
2	80.71	1	178000	3	4194
3	80.71	4	277650	3	4197
4	55.71	1	66000	2	4210

Customer Profiling. Knowing that cluster has been set, Mean of RFM value of each cluster are compared to mean of general RFM value. If the mean values of frequency and monetary are above the mean value of frequency and monetary of all cluster, then signed as up arrow (↑). On the contrary, it is indicated by down arrow (↓). Whereas, if the mean values of recency are above the mean value of recency of all cluster, then signed as down arrow (↓). On the contrary, the mean value of recency is indicated by up arrow (↑).

Table 6. Customer profile based on RFM.

Group	Recency	Frequency	Monetary	Total Member	Profile Name
2	↑141.12	↓1.09	↓85,604	12,992	Iron
3	↑143.37	↑1.89	↑218,950	2,288	Gold
1	↓195.42	↑4.03	↑684,362	261	Platinum

Each group of cluster has different characteristic of RFM value. Group 1 has a higher mean value of recency, frequency, and monetary so that it can be called as platinum. This group consist of 261 customers and had contributed a large income per customer to the company but tended to be longer in terms of 'last time' transaction (recency). Group 3 has a moderate mean value of recency, frequency, and monetary so that it can be called as gold. This group consist of 2,288 customers and had contributed a moderate income per customer to the company and an average 'last time' transaction not too long. Group 2 has a low mean value of recency, frequency, and monetary so that it can be called as iron. This group numbered 12,992 customers and had contributed a small income per customer and the average 'last time' transaction occurring not so long. Total customers identified are 15,541 customers.

Customer segmentation is formed by using customer pyramid as follows:

Fig. 2. The Pyramid of customer segmentation.

Discount absorption are different to each customer group. Platinum group absorbs discounts of 318 transactions out of 1038 transactions or 30.6%. Gold group uses discounts of 1555 transactions out of 4334 transactions or 35.9%. While the Iron group uses discounts of 7377 transactions out of 14197 transactions or 52.0%. Then, it can be concluded that the Iron group is most price sensitive.

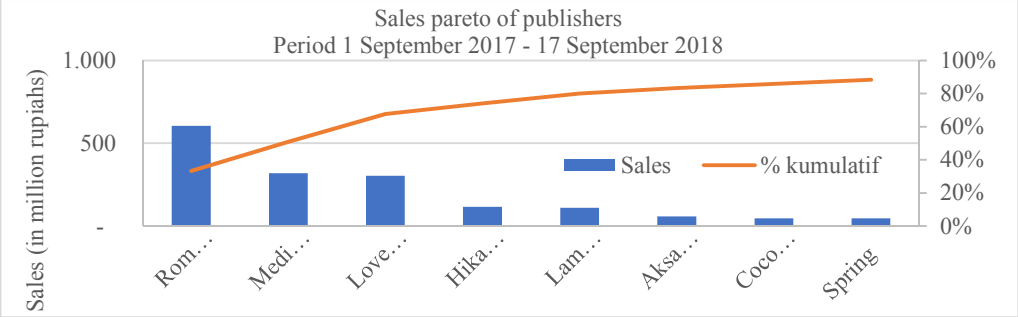

Fig. 3. Pareto analysis of publishers.

Pareto Analysis. By conducting Pareto analysis, marketing action based on publisher and author can be prioritized. Among 37 publishers and 173 authors who produced transactions at Bukku.id, there are 8 publishers who contribute up to 88% of sales and 22 authors contributing up to 88% of sales at bukku.id

Customer location analysis is needed to provide further understanding regarding marketing programs for each consumer segment. To build a handling prioritization, pareto can be formed based on the location as follows:

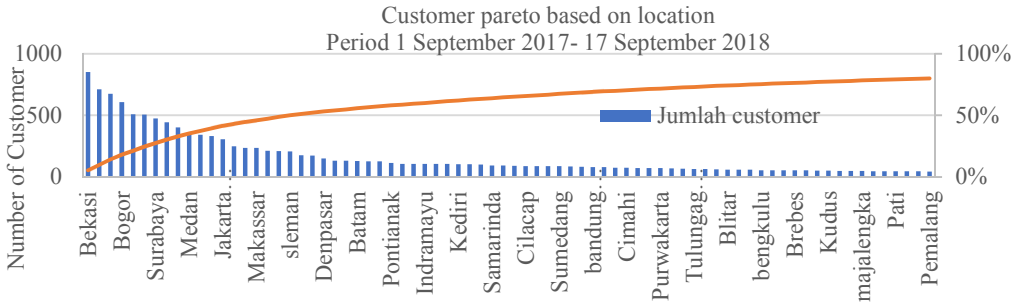

Fig. 4. Pareto analysis based on customer location.

Figure 4 shows which locations have the highest number of customers. Then, customer mapping is continued based on the author. This can allow the marketing team to target free delivery promotions for certain authors (or other attractive subsidies), book truck and book corner, seminar programs, book reviews, book festivals, logistic strategies or other offline events while making offline sales for increasing product holding ratio of customer in the target area.

7. Discussion

This study investigates customer segmentation based on customer purchase data. Unfortunately, it does not include demography and loyalty data. Customer segmentation that has been formed consist of three cluster. The three cluster represent how recent they made purchases, how often they made purchases, and how much money they spend for the company in the certain period. The clusters consist of platinum (261 customers), gold (2,288 customers), and iron (12,992 customers).

Then, marketing action based on publisher and author can be prioritized based on pareto analysis. Among 37 publishers and 173 authors who produced transactions at Bukku.id, there are 8 publishers who contribute up to 88% of sales and 22 authors contributing up to 88% of sales at bukku.id. There are 81 destination cities represent 80% of transactions and 76 destination cities represent 80% of the address of unique customers. Marketer can use this information to carry out further actions of effective online-offline marketing activities

8. Conclusion

- There are three customer cluster that has been formed: Platinum (261 customer), gold (2,288 customer) and iron (12,992 customer). Platinum has mean value of recency, frequency, and monetary respectively 195 days, 4.03 times and Rp 684,362. Gold has mean value of recency, frequency, and monetary respectively 143 days, 1.89 times and Rp 218,950. And Iron has mean value of recency, frequency, and monetary respectively 141 days, 1.09 times and Rp 85,604.
- There are 8 publishers which produce 88% of sales and 22 authors which produce 88% of sales based on pareto analysis.
- There are 81 destination cities represent 80% of transactions and 76 destination cities represent 80% of the address of unique customers. Thus, mapping the number of customers based on certain authors can be identified by pareto analysis based on the customer address.

Managerial Implication. Marketing strategies of the company can be identified based on customer segmentation. It is necessary to know the marketing strategies relevant to its objectives both short term and long term, as follows:

Table 7. Marketing objectives and strategies based on customer profile.

RFM Pattern	Marketing Strategies	
	Short Term	Long Term
R↓F↑M↑ Platinum	• Improve word-of-mouth marketing • Product review • Promote new product and bundling program	• Give Rewards and customer loyalty program • Online and offline marketing treatment activities: book review, workshop, intimate meeting • Special customer services for complaint and live chat
R↑F↑M↑ Gold	• Cross selling and bundling program • Loyalty program offering	• Improve conversion to loyalty program • Live chat • Conversion to platinum customers
R↑F↓M↓ Iron	• Improve brand activation (awareness), creative marketing campaign • Flash sale program of discount product • Cross selling	• Improve word of mouth marketing to new customer • Improve cross selling product • Customer services using chat bot • Conversion to gold customers

Other than the above strategies, the company can also do the following:

- Implement targeted promotion that can be done through several marketing channels using information technology, such as banner webpages, web notifications, direct e-mail, SMS notifications, Line and Whatsapp.
- Location analysis can help company to determine effective offline marketing programs based on pareto analysis. Company and publishers can initiate offline marketing programs such as book reviews, book festivals, and workshops. SMS and email blasts can be adjusted based on the location to increase book sales and the number of visitors to offline event. Company can also optimize free delivery programs in certain areas with a cross-subsidy pattern.

Acknowledgements. This research publication is funded by Lembaga Pengelola Dana Pendidikan (LPDP RI).

References

[1] Trim B, Gafar J & Mujib I I 2015 *Industri Penerbitan Buku Indonesia Dalam Data dan Fakta*
vol 54 (Jakarta: Ikatan Penerbit Indonesia)

[2] Fotaki G & Meijer D 2013 *Exploring big data opportunities for online customer segmentation*
(Netherland: Utrecht University)

[3] Muzumdar P 2012 Online bookstore - A new trend in textbook sales management for services
marketing *J. of Management and Marketing Research* **9** 1–14

[4] Kumar V 2010 A customer lifetime value-based approach to marketing in the multichannel,
multimedia retailing environment *J. of Interactive Marketing* **24(2)** 71–85

[5] Cheng C H & Chen Y S 2009 Classifying the segmentation of customer value via RFM model
and RS theory *Expert Systems with Applications* **36(3 PART 1)** 4176–4184

[6] Chen D, Sain S L & Guo K 2012 Data mining for the online retail industry: A case study of
RFM model-based customer segmentation using data mining *J. of Database Marketing & Customer Strategy Management* **19(3)** 197–208

[7] Chen H, Chiang R & Storey V 2012 Business intelligence and analytics: From big data to big
impact *MIS Quarterly* **36(4)** pp 1165-88

[8] Fotaki G, Gkerpini N & Triantou A I 2012 *Online customer engagement management*
(Netherland: Utrecht University)

[9] Tsiptsis K & Chorianopoulos A 2009 *Data mining techniques in CRM: Inside customer segmentation* (UK: John Wiley & Sons)

[10] Birant D 2011 Data mining using RFM analysis *Knowledge-Oriented Applications in Data*
Mining InTech

[11] Kaymak U 2001 Fuzzy target selection using RFM variables *Proc. Joint 9th IFSA World*
Congress and 20th NAFIPS Int. Conf. (Cat. No. 01TH8569) vol 2(C) pp 1038–43

[12] Turban E, Delen D & Sharda R 2018 *Business intelligence: A managerial approach (4th edition)*

(UK: Pearson Education)
[13] Koch R 2008 *The 80/20 Principle: The Secret to Achieving More with Less* (New York:
Doubleday)