Customer segmentation analysis and customer lifetime value prediction using Pareto/NBD and Gamma-Gamma model

Van-Ho Nguyen
University of Economics and Law

Thanh Ho
University of Economics and Law

Customer segmentation analysis and customer lifetime value prediction using Pareto/NBD and Gamma-Gamma model

Kim-Giao Tran[1,2], Van-Ho Nguyen[1,2], Thanh Ho[1,2,*]

[1]University of Economics and Law, Ho Chi Minh City, Vietnam

[2]Vietnam National University, Ho Chi Minh City, Vietnam

*Corresponding author: thanhht@uel.edu.vn

ABSTRACT

Customer segmentation helps the organization to manage customer relationships expertly and gain a deep understanding of customers. With the proliferation of machine learning methods, this study performs data science algorithms into traditional marketing such as Recency, Frequency and Monetary (RFM) model, or RFM model with K-Means clustering for customer segmentation, and Pareto/Negative binomial distribution (NBD), Gamma-Gamma model for predicting customer lifetime value (CLV). This study experiments on a dataset of 121,317 historical transactions of the bicycle retail industry, including individual customers and retailers. By the retention analysis, it shows that customers mostly prefer coming back after at least three months as bicycles are long-termed usable. The combination of segmentations from RFM models and the CLV results of each customer can show that whether the customers, who are in well-evaluated segments, have the proportionally high value to the business or not. This can also be seen as a basis for cross-checking the completeness of these techniques. The evaluation metrics with high accuracy in training and evaluating the Pareto/NBD and Gamma-gamma models show that the CLV are well-trained before being compared to the RFM segmentation. This study confirms the connection between the traditional RFM, the RFM model with K-Means, and CLV techniques as the results show that customers in good segments, have high predicted CLV as well. Based on the empirical results, the proposed research models can be applied in other businesses that will help them get the effective business strategies for each customer group depending on their financial and human potential.

Keywords: RFM model, customer segmentation, clustering, CLV, Pareto/NBD, customer retention

1. Introduction

In the intense competition and complexity of the business environment, customer segmentation helps the marketing departments easily define the pivotal solution to attract each group of customers. Based on the data segmentation, customers are classified into different groups according to distinguishing similarities such as gender, age, income, products of interest, and purchasing behaviors (Anitha & Patil, 2019). These characteristics are analyzed and categorized based on the historical purchasing data of the business. Recency, Frequency, and Monetary (RFM) has been very famous in marketing as a tool to identify a company’s best customers by calculating and analyzing their spending habits. RFM analysis weights customers' importance by scoring them in three measurements such as how recently they have made a purchase (Recency), how often they have bought (Frequency), and how much they have spent (Monetary) (Thanh & Son, 2021).

Besides using RFM for customer segmentation, customer lifetime value (CLV), retention rate and churn rate are a combination of robust metrics to measure customer satisfaction. While CLV is the discounted value of future profits that the customer spends on the company (Glady et al., 2009), the retention rate shows the ability of a company to keep its existing customers (Ismail et al., 2015). In contrast, the churn rate is the percentage of customers moving out of a cohort over a particular period.

As Kotler and Keller described customers' churn as a phenomenon that results in a waste of money and efforts (Kotler & Keller, 2006), choosing to focus on retaining old customers and turn them from potential customers into loyal customers will help businesses reduce more costs than building advertising campaigns to attract new customers. However, the problem is that when a business has a lot of customers and all of them have made many transactions with the business, it is tough to know if they are still attached to the business or not. Besides, businesses cannot calculate precisely when a customer will leave but can only predict based on probabilities, so this is even more difficult.

While the studies above mainly focused on the RFM model or the Pareto/NBD and Gamma-Gamma models only. This study identifies the goal of combining the techniques to find the relationship between them, and desires to provide the managers a multi-dimensional view of their customers. For more particular, this study wants to detect the bonds between the traditional RFM segmentation and the K-Means clustering with RFM. Then applying the CLV results from Pareto/NBD and Gamma-Gamma models to have a more precise insight of the customers' value of each group. The relationship between RFM and CLV is also included. That makes it easier for managers to decide whether to implement appropriate marketing strategies for each customer group as well as to assess whether existing customer care policies are still appropriate for retaining customers or not. Besides, this study also first briefly uses the retention analysis to find out the customers' purchasing behaviors and patterns before going straight to the further analysis.

The following content of the article is Section 2, including the theoretical basis and related studies, to identify models and algorithms suitable for the set goals. Section 3 is the methodology that describes relevant issues and experimental processes. After the experimental process, the results and discussion of the identified customer segments are mentioned in Section 4. The last Section is the conclusion and implications of the study.

2. Theoretical background and related work

This section provides the literature overall and some related researches based on the purpose of this study.

The RFM model is usually used to classify customers and define their behaviors. RFM records the customers' transactions under three factors:

(1) Recency is the distance between the last purchasing date of that customer and the date of implementing the model;

(2) Frequency is the total transactions of that customer;

(3) Monetary is the actual money that the customer had spent on businesses' products or services.

The most well-known clustering methods by RFM are customer quintiles (Miglautsch, 2000) and clustering by K-Means.

Clustering using the K-Means algorithm is a method of unsupervised learning used for data analysis. It generates k points as initial centroids randomly, with k is chosen by users. Each point is assigned to the cluster with the closest centroid. Then the centroids are updated by taking the mean of the points of each cluster (Anitha & Patil, 2019; Ismail & Dauda, 2013; Madhu et al., 2010). The data points may move to different clusters after each iterative approach. The chosen centroids are defined when there are no point changes clusters or the centroids remain. The algorithm uses mainly Euclidean distance to measure the distance between data points and centroids (Dwivedi et al., 2014). The formula to calculate Euclidean distance between two multi-dimensional data points $X = (x_1, x_2, x_3, \dots, x_m)$ and $Y = (y_1, y_2, y_3, \dots, y_m)$ is described as equation (1):

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_m - y_m)^2} \qquad (1)$$

Although K-Means is the most common algorithm to classify clusters, it still has some drawbacks. Because the centroids are first chosen randomly, the results can turn out different for different runs. Besides, defining the right number of clusters is also a tremendous problem to deal with. Thanh and Son (2021) used the Elbow method to find the optimal number of clusters then using the Silhouette method to re-evaluate the results above, while Anitha and Patil (2019) only used the Silhouette score to find the optimal k. These studies pointed out the efficiency of the clustering method in Data Science and also performed the clustering results in RFM analysis and provided customers' different behaviors in specific clusters.

The Elbow method is used to determine the number of clusters of a dataset by using the visual technique. The graphic obtained the results from the Sum Squared Error (SSE) calculation, which measures the difference between points in clusters. The more the number of clusters k, the smaller the SSE value will be. If the value of the former cluster and the value of the later cluster draw an angle between them, the cluster at the elbow flexion point will be the chosen cluster or the cluster with the biggest reducing value compared with its former will be chosen (Thanh & Son, 2021; Humaira & Rasyidah, 2020; Nainggolanet et al., 2015). The formula of SSE calculation is described as equation (2):

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_j} d(x_{ij}, m_i)^2 \qquad (2)$$

Where m is the centroid of the data point x and k is the number of clusters. The graphic which obtained the values of the SSE calculation for the different number of clusters will perform the visual looks as an elbow arm. The Elbow method is easy to implement and adequately fitted with perplexing, huge data, but its weakness is the user must choose the number of clusters based on experience (Humaira & Rasyidah, 2020).

Along with the Elbow method, the Silhouette score is also an effective way to see how well each cluster is separated from the others. In the two studies (Anitha & Patil, 2019; Humaira & Rasyidah, 2020), the authors give two different theories about the range values of the Silhouette score. After researching more deeply, the Silhouette score is informed to be in the range $[-1, +1]$, if it is scored near +1, the clustering quality performed well, if it is valued at 0, we can say there is no distinction between the clusters, and if it is near -1, the clusters were not distributed well (Ogbuabor & Ugwoke, 2018). The formula to calculate Silhouette score is written as equation (3):

$$Silhouette\ score_i = \frac{b_i - a_i}{\max{(a_i, b_i)}} \qquad (3)$$

With a is the average intra-cluster distance (the mean distance between i and the data points in the same cluster), and b is the average inter-cluster distance (the mean distance between the data point i to all the data points outside its cluster).

Pareto/negative binomial distribution (NBD) model is one of the most classic used RFM models to calculate CLV. The model mostly used the recency, frequency, and length of the customer's observation period to predict the customer's future purchases (Hellerslia & Talal, 2020). The Pareto/NBD model is developed by Fader and Hardie. They also described the model that is based on five assumptions (Fader et al., 2004):

(1) The transactions made by a customer in a period of length t follow a Poisson distribution with transaction rate λ. It means that they can purchase randomly whenever they want in their active period, but the rate (in a unit time) is constant.

(2) Heterogeneity in transaction rates across customers follows a gamma distribution with shape parameter r and scale parameter α.

(3) Each customer has an unobserved lifetime t. In other words, the point at which the customer becomes inactive or churned is distributed exponentially with the dropout rate μ.

(4) Heterogeneity in dropout rates across customers follows a gamma distribution with shape parameter s and scale parameter β.

(5) Each customer has a varied transaction rate λ and the dropout rate μ.

The Gamma-Gamma model is the extension of the Pareto/NBD model. While the Pareto/NBD model only focuses on the recency and frequency factors, the Gamma-Gamma model uses the monetary component to predict the average future purchasing value (Avinash et al., 2019; Hellerslia & Talal, 2020). The Pareto/NBD and Gamma-Gamma models are a powerful combination to calculate CLV. While Pareto/NBD predicts future purchases, the Gamma-Gamma model allows us to assign a monetary value to each of those future purchases. To ensure to have the best estimated CLV, these models can be evaluated in the holdout period before making forecasts.

3. Methodology and proposed research model

Figure 5 describes methodology and proposed research model with three main stages:

(1) Stage 1 is customer segmentation analysis according to RFM. From the input data is a dataset extracted from the sales department of Microsoft's Adventure Works sample data, perform data preprocessing, and calculate recency, frequency, and monetary values for use in the RFM model. After preprocessing the data and realizing the difference between the data points, the study will implement the standardization for input data, then using some methods related to K-Means to find out the optimal number of clusters for segmentation;

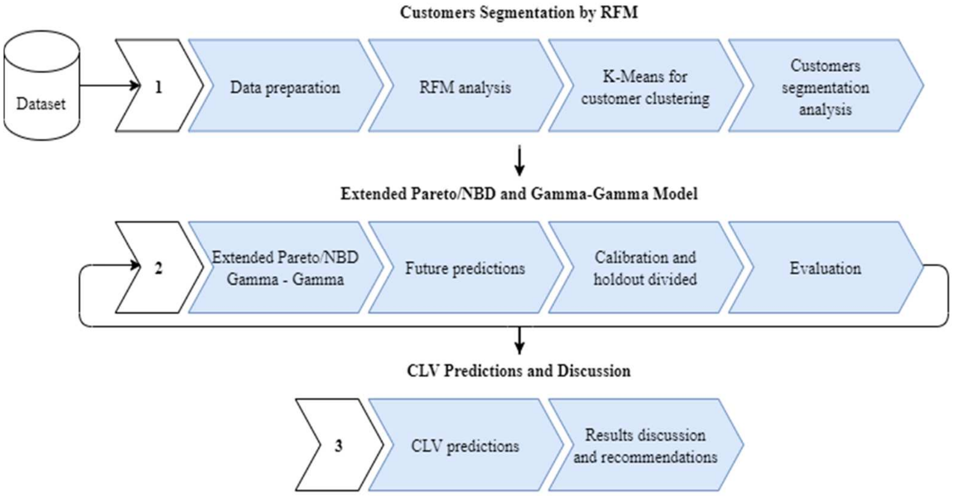
Customers Segmentation by RFM
Dataset
1
Data preparation
RFM analysis
K-Means for customer clustering
Customers segmentation analysis
Extended Pareto/NBD and Gamma-Gamma Model
2
Extended Pareto/NBD Gamma - Gamma
Future predictions
Calibration and holdout divided
Evaluation
CLV Predictions and Discussion
3
CLV predictions
Results discussion and recommendations

Figure 5: Overview of the proposed research model

(2) Stage 2 is to use Pareto/NBD and Gamma-Gamma model to predict the number of purchases and revenue that customers yield in the future, the root of these two models is to be exploited and developed from the RFM model. Build the above two models on

the training set and re-evaluate the predictions on the test set to see the accuracy of the model. Repeat this loop by changing the indexes in the model until the model gives the most optimal results;

(3) Stage 3 is from the two optimal models above, performing customer lifetime value (CLV) prediction.

4. Experimental result and discussion

4.1. Customer segmentation using RFM

The first stage of the experimental process including preprocessing data standardization, RFM data construction, and K-Means customer segmentation (Figure 5).

4.1.1. Dataset and data preprocessing

The study uses a dataset of customer transactions extracted from the dataset of the company Adventure Works Cycles. This is a multinational company that manufactures and sells bicycles to the North American, European, and Asian markets. The extracted dataset records 121,317 transactions of the company from 06/2011 to 07/2014. This includes both individual customers and retailers. To analyze the optimal customer segment for each different market, the study filtered out the transactions made in the US (United States) market for use in further analysis.

4.1.2. Customer retention analysis

Before jumping into clustering customers by the RFM model, the study will briefly analyze the company's customer retention situation to find out the insight of its business status. Adventure Works is a business that manufactures and sells bicycles for both individuals and resellers, but bicycles are non-essential and can be used in the long term, the number of customers who have one transaction only over 3 years is very high, at 74.31% (Figure 6). Meanwhile, the number of customers who used to repeat transactions with the company only accounted for 25.69% but brought even higher revenue than the others over time. In particular, there would be a sudden increase in the revenue that this group of customers brought to the business every 1 month.

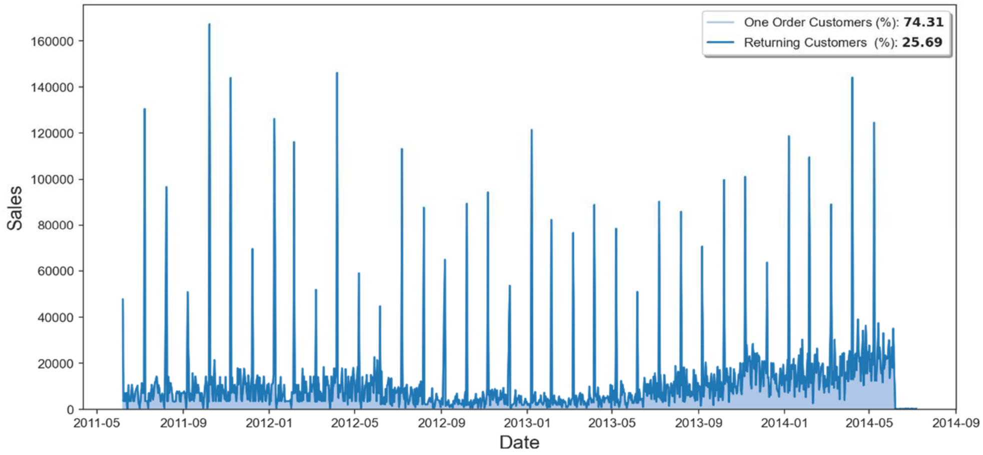

Figure 6: Sales by customers over time

It can be seen in Figure 7 that, in the period from 05/2011 to 06/2013, approximately two years, the number of regular customers was higher, almost all customers returned to make transactions again with the company. However, starting from 07/2013, when the business had a sudden growth in attracting more customers, the number of customers leaving when they only transacted once with the business was very high.

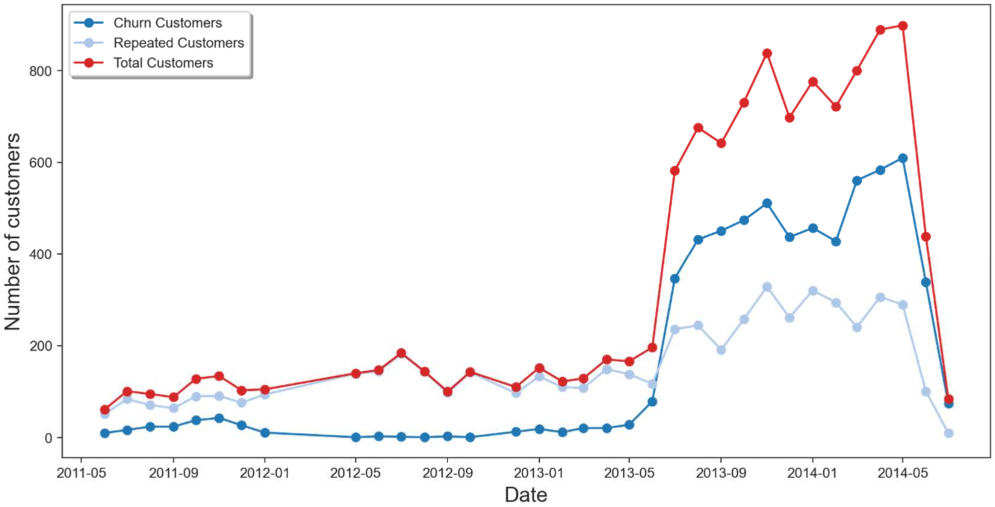

Figure 7: Number of churn and repeated customers over time

Grouping customers according to cohort, also known as grouping customers according to the timeline from the customer's first transaction (Croll & Yoskovitz, 2013). The formula to calculate the retention rate is described as equation (4):

$$Retention\ rate = \frac{Sum\ number\ of\ customers\ alive\ in\ each\ month}{Sum\ number\ of\ initial\ customers} \quad (4)$$

The retention rate of each cohort is shown on the horizontal axis of Figure 8. With the analysis of customer retention rate using the heat chart, it can be seen:

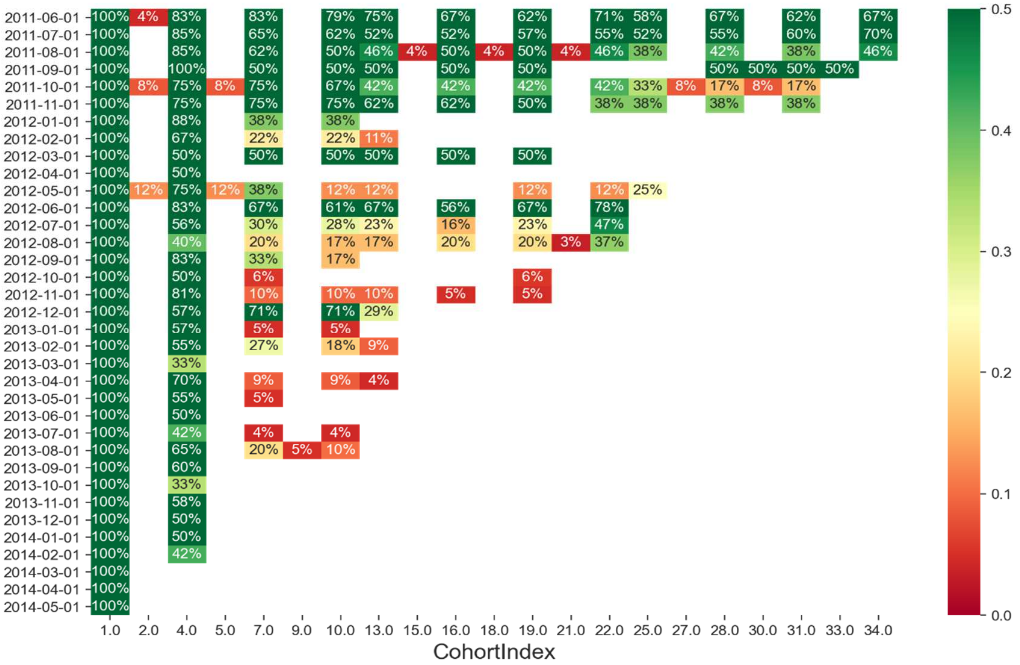

Figure 8: Retention rates in cohort analysis

Customers of the business did not transact regularly once a month, but on average, customers came back every 2-3 months. With the group of customers having transactions from the beginning of the observed period, from June 2011, only 4% of customers returned to transact in the next month. However, with a cycle every 2-3 months, the customer retention rate of the business at this time was very high, in the 34th month, it still maintained 67% of the total number of original customers.

In contrast, the retention rate for new customer groups decreased considerably. Generally, the company's customer retention policy was appropriate for the period before 2012 and was able to retain this group of loyal customers until the end of the period. However, it seemed to be no longer suitable for new customer groups, especially when the business in later period promoted marketing and attracted more customers but cannot keep them. Businesses should focus more on customer care policies as well as targeted marketing campaigns to attract returning customers.

4.1.3. Customer segmentation based on RFM scores

This is the traditional and the simplest way that can explain how the RFM model works. The RFM model is famous for transforming transactional data, which basically includes CustomerID – unique customer code, SalesOrderID – unique invoice code, ProductID – unique product code, InvoiceDate – date of the transaction, Quantity – quantity of purchased items, Unit Price – the price of one item, Country – country of the transaction, into profitability scores (Zaki et al., 2016). After calculating recency, frequency, and monetary for RFM analysis, the characteristics of the statistical distribution of these factors such as average, minimum value, maximum value, as well as quartiles are described in Table 8. The average last purchased date is 206 days ago with nearly 1.5 purchases and 1473.8 revenue in total.

Table 8
Quartiles description in RFM table

	Recency	**Frequency**	**Monetary**
Mean	206.377101	1.466626	1473.809070
Min	0.000000	1.000000	1.374000
Max	1122.000000	12.000000	58662.190608
1st quartile	91.000000	1.000000	21.490000
2nd quartile	177.000000	1.000000	69.990000
3rd quartile	277.000000	2.000000	2294.990000

While the authors in (Zaki et al., 2016) ranked customers in quintiles. This study chose to rank them based on the quartiles. Following the related works, customers with the most recency value will have a 1 R score. In contrast, the ones with the lowest recency received a 4 R score because the customers with more recent transactions are considered more valuable to the business. This step was implemented repeatedly for the frequency and monetary but in a reverse way, which means the highest frequency and monetary received 4 scores and the ones with the lowest had 1 score. Noted that the customers' value is proportional to RFM scores. By mapping the RFM scores, we had the worst valuable customers who had an overall RFM score of 111. On the contrary, the ones with an RFM score of 444 were considered as top customers to the company.

This study divided customers into segments based on the exemplary segmentation of the RFM scores, which Jasmin described as a graphic in her blog (Jasmin, 2020). This study used the exemplary in the reverse RFM scores.

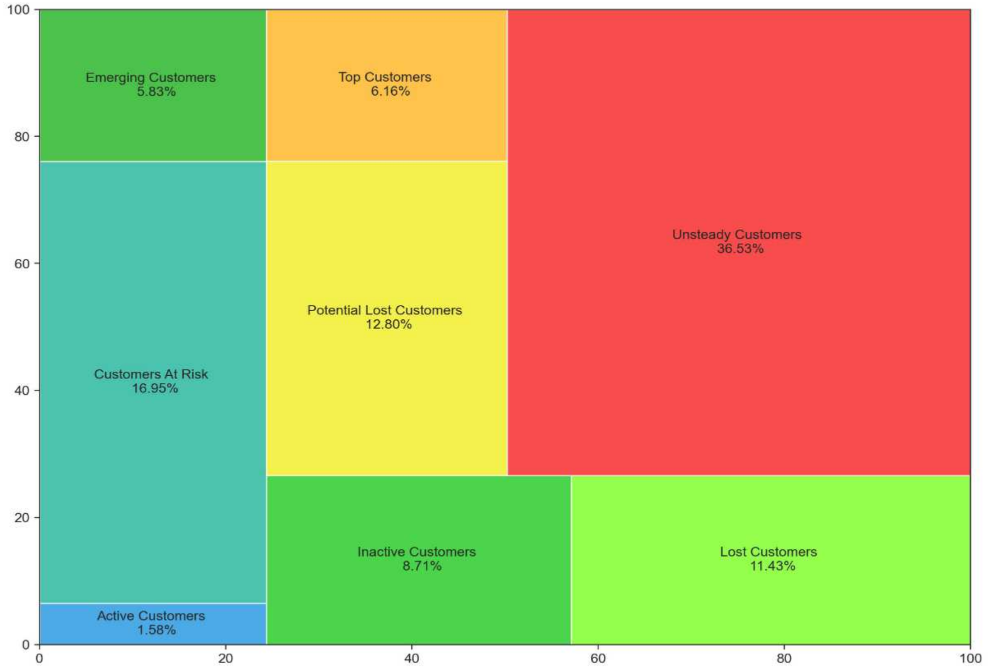

Figure 9: Customer segmentation distribution

Figure 9 shows the RFM segmentation labeling results in a treemap. Based on different segments, the company needs a specific strategy to develop its business status. The company had a huge number of new customers as Unsteady Customers (36.53%) with high monetary value. It is advised to build a long-term relationship with these customers by cross-selling strategy or specific promotions. Besides, the Customers At Risk segment, which accounted for 16.95%, is also a potential group from which the business can exploit. With a very high monetary value but having stopped trading for a long time, finding a way to contact and pull these customers back will bring a great benefit to the business. Top Customers and Active Customers accounted for a small percentage but the profits that they brought were considerable, the company cannot lose them. Finally, the company had quite many Inactive and Lost customers. The study mentioned that the company with main products such as long-term usable bicycles could have many churned customers, but the managers could research more insight into these customer groups to find the exact churned reason to re-engage these customers as much as possible.

4.1.4. Data standardization

After preprocessing the data and preparing the input data for the RFM model with the corresponding recency, frequency, and monetary values, it was found that there is a huge difference between these three values, which can affect the model run time and the accuracy of the algorithms. The study conducted data standardization according to the standard distribution method (Standard Score), also known as Z-Score (Ismail & Dauda, 2013) to bring the data to a distribution range that where the mean value of the observations is 0 and the standard deviation is 1. The formula for standardizing the data is described as equation (5):

$$x' = \frac{x - \mu}{\sigma} \tag{5}$$

With x is the initial value before standardization, μ is the mean value of the observations, and σ is the standard deviation of the observations. After standardizing the data, it can be concluded that the three current recency, frequency, and monetary values weight equally when included in the analysis in the K-Means clustering model.

4.1.5. The optimal number of clusters for K-Means

As having described these models' literature in the theoretical basis part, this study uses the Elbow method and Silhouette score to find the most optimal number of clusters of the dataset. The result for the number of clusters from 2 to 9 is described in Figure 10. The graphic has the visualization as an elbow and the SSE line shows that the elbow flexion point is around $k = 3$ or $k = 4$. Silhouette score will be implemented to re-evaluate the quality of finding optimal k in the Elbow method to get the final result.

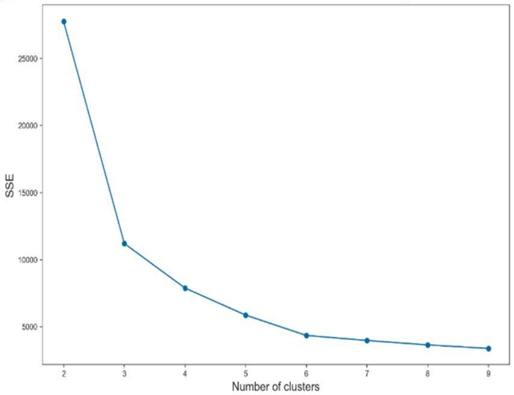

Figure 10. Elbow method result

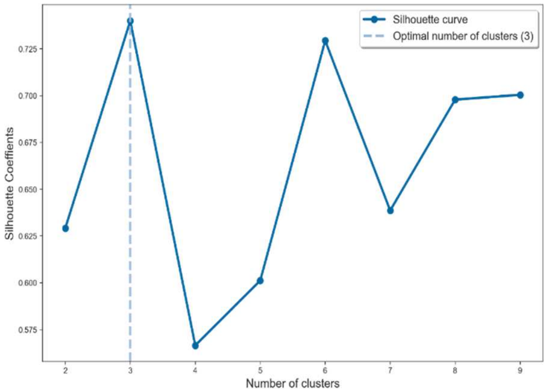

Figure 11. Silhouette score result

Figure 11 illustrates the results that Silhouette scored each cluster. It can be seen that $k = 3$ with 0.74 score is the highest among other clusters. The indicator means that with $k = 3$, the distance from data points to their centroid in each cluster is optimized and the cluster eccentricity barely occurs. Therefore, the study will use the number of $k = 3$ to cluster customers into different levels based on three factors of the RFM model (Recency, Frequency, and Monetary).

The number of customers and the average value of recency, frequency, and monetary of each cluster after being divided into 3 clusters by K-Means are all described in Table 9. It can be seen that the Gold cluster includes the least number of customers who have the most transactions, relatively recent purchases and bring the highest revenue for the business. In the other two groups, they have quite similar frequency and monetary mean value, but the average recency value of one group is nearly twice more than the other one. The group with the most recency value is labeled as the Bronze group because of without recent transactions with the business.

Table 9.

Each cluster description

Cluster	Customers	Mean recency	Mean frequency	Mean monetary
Gold	216	149.527778	8.606481	11299.873848
Silver	4665	148.998928	1.231726	1162.314786
Bronze	3329	290.471012	1.332532	1272.754954

Figure 12 describes the clustering result with $k = 3$ in three-dimensional space. The Silver level is the group with the highest convergence, but there is still confusion between the data points in the Silver and Bronze levels. The Silver group contains customers who have more stable recency and frequency indexes while the ones in the Bronze level have even higher monetary value but stopped trading for a long time. Besides the Gold group with its distinction from the others, Silver and Bronze were labeled mainly based on the average recency value.

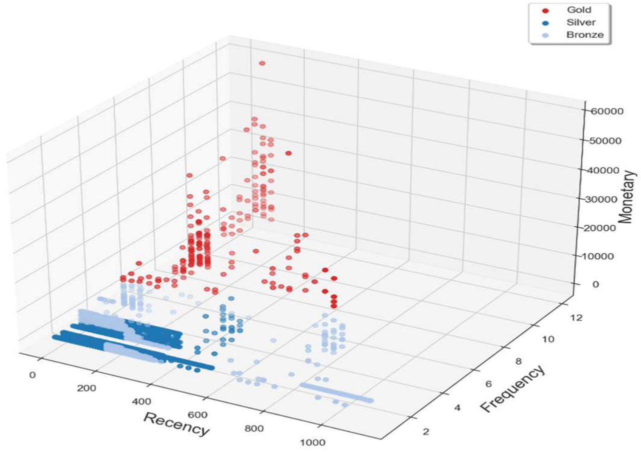

Figure 12: Clustering result by RFM level

4.2. Predicting CLV using Pareto/NBD and Gamma-Gamma model

The second stage of the experimental process including Pareto/NBD and Gamma-Gamma model data construction (Figure 5), calibration and holdout dataset divided for evaluation, then using the models to predict CLV.

4.2.1. Constructing input data for Pareto/NBD and Gamma-Gamma model

Because the Pareto/NBD and Gamma-Gamma models use the RFM basis, the data used for these models are quite the same as the previous data construction. The difference is Pareto/NBD only considers customers with repeated transactions, which means the customers with only one transaction have $frequency = 0$ and $monetary = 0$ in this model. The Pareto/NBD model also uses another factor, which is the customer lifetime (T) calculated as the distance from the customer's first purchase date to the model implementation date.

The data for the Gamma-Gamma model is the same as the data of the Pareto/NBD model but only the rows have frequency and monetary bigger than 0.

4.2.2. Calibration and holdout dataset

The calibration dataset starts at the beginning of the observed period from 07/06/2011 to 07/07/2013, and the holdout period spans from 08/07/2013 to 07/07/2014, exactly 365 days. The percentage is approximately 70% in calibration and 30% in the holdout dataset.

4.2.3. Predicting future purchases using Pareto/NBD model

Figure 13 illustrates the number of purchases in the calibration dataset on the x-axis and the corresponding average number of purchases in the holdout dataset on the y-axis. As can be seen, the model predicted that the customers with the higher purchases in the calibration would also have higher purchases in the holdout, except for a slight reduction in customers with 5 purchases in the calibration. In contrast, in actual data, the holdout set shows more unpredictable volatility.

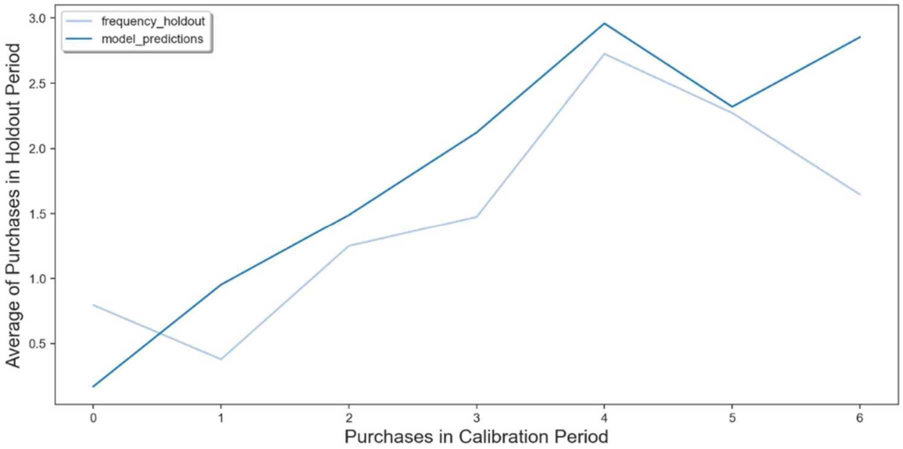

Figure 13: Actual and predicted purchases of Pareto/NBD model in holdout dataset

Evaluating models can be the most important step in Data Science. This study used some indicators to evaluate the quality of the prediction model. The formulas of these indicators are described as equations (6), (7), and (8) (Chicco et al., 2021):

$$MAE = \frac{1}{m}\sum_{i=1}^{m} |X_{\mathrm{i}} - Y_{\mathrm{i}}| \tag{6}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m} (X_{\mathrm{i}} - Y_{\mathrm{i}})^2 \tag{7}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m} (X_{\mathrm{i}} - Y_{\mathrm{i}})^2} \tag{8}$$

With X is the actual values and Y is the predicted values. MAE is a measurement of errors between actual and predicted observations. MSE measures the average of squares of the errors or can be understood as the average squared difference between the actual and estimated values. RMSE is also a measurement to evaluate the differences between two observations. The more these indicators move near to 0, the fewer errors the predictions have. Table 10 shows that the Pareto/NBD predicted the future purchases fairly well while the evaluation values are all small, nearly 0.

Table 10

Pareto/NBD purchases prediction evaluation

Types	Results
Mean Absolute Error (MAE)	0.7904071127295603
Mean Squared Error (MSE)	0.8850164704192321
Root Mean Squared Error (RMSE)	0.9407531399996665

4.2.4. Predicting the future average order value using the Gamma-Gamma model

The estimated results of the Gamma-Gamma model are shown in Figure 14. The histogram plots the monetary value distribution of the actual and estimated observations. It shows that predicted results tend to be smaller than reality and both predicted and actual monetary values are concentrated near zero.

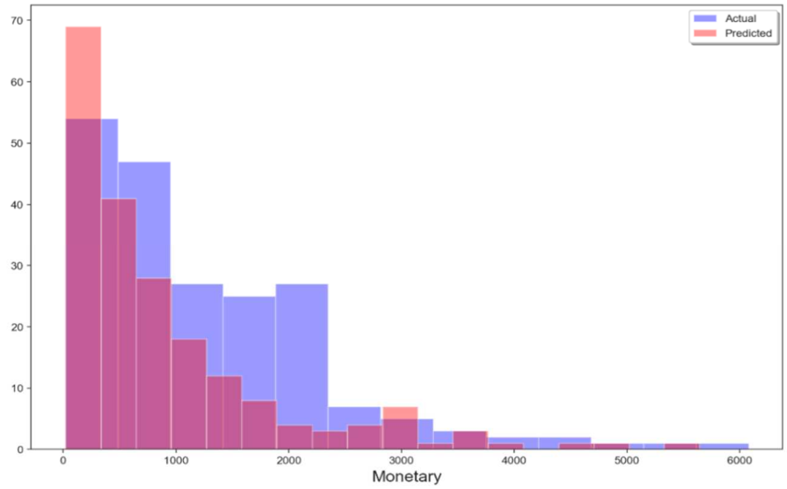

Figure 14: Actual and predicted of the Gamma-Gamma model in the holdout set

Because the difference in monetary values is much larger than the factor used in the previous model. Instead of using metrics such as MAE, MSE, and RMSE, which are often used for normalized, standardized datasets or have values close to zero, here the chosen option is dividing monetary values into 5 bins according to ordinal variable and K-Means. Then use the confusion matrix and the F1-score to evaluate the accuracy of the model.

The confusion matrix Figure 15 shows that the Gamma-Gamma model worked well in the holdout set while the predictions were mostly divided into the right bins. The F1 score is 0.9 which means the estimation had high accuracy.

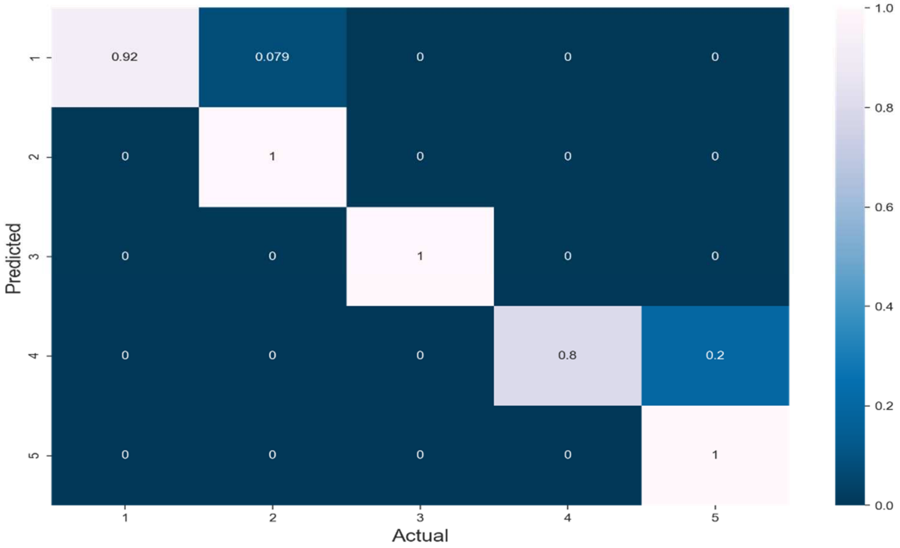

Figure 15: Confusion matrix of actual and predicted monetary value

After training and evaluating the models to check their quality, the Gamma-Gamma model was implemented again in the initial dataset to check if it had the overfitting problem or not. Fortunately, the model also performed well in the original data set, proved when Figure 16 shows that the actual and predicted monetary values have a linear correlation, and the histogram gives the results of the true and predicted values almost overlapping (Figure 17). After training and fixing the Pareto/NBD and Gamma-Gamma model, finding that the two models worked quite well and the evaluation was very high, the study would apply these two models to predict the CLV for the company's customers.

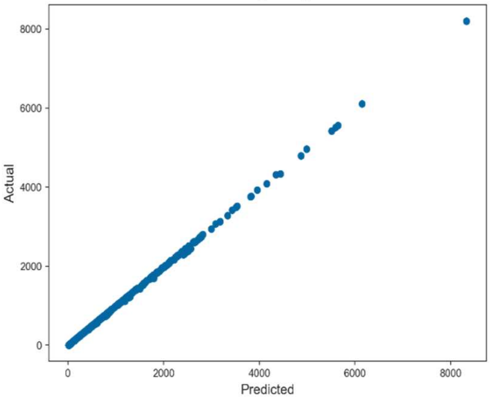

Figure 16: Scatter plot of actual and predicted monetary value in initial dataset

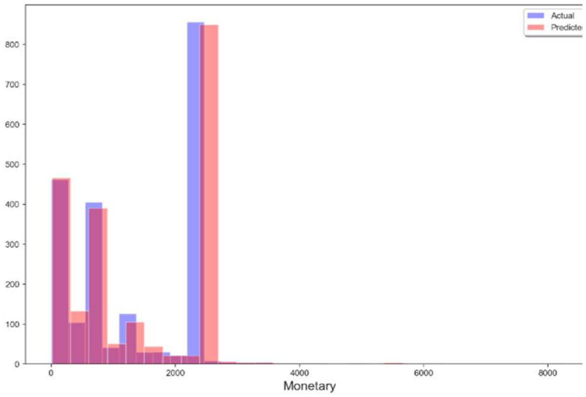

Figure 17: Histogram of actual and predicted monetary value in initial dataset

4.2.5. Predicting CLV values

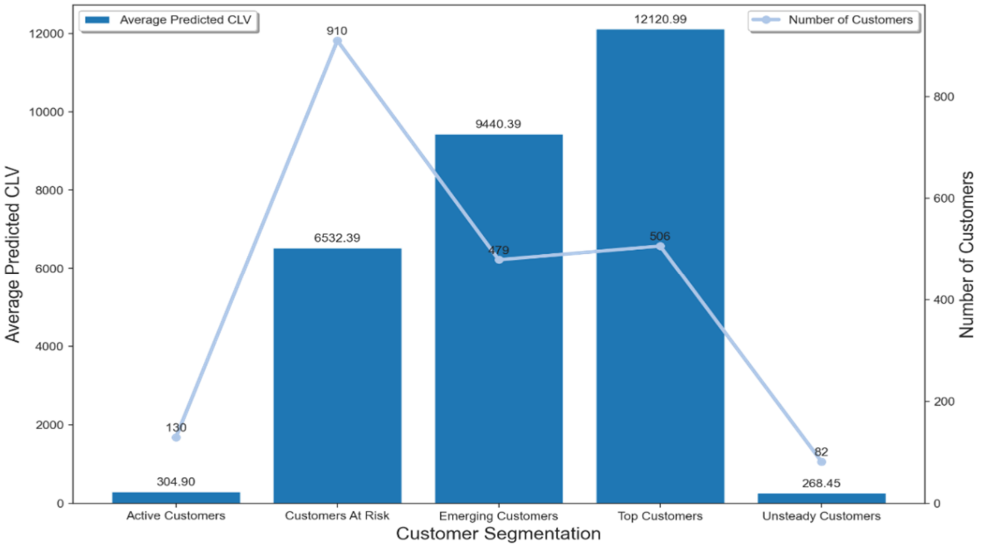

Figure 18: Average predicted CLV by customer segmentation

Among the 5 groups containing repeated customers of the business in Figure 14, the model predicted that the Top Customers group has the largest customer lifetime value. The Active Customers and Unsteady groups have a fairly low value because they did not make many transactions with the business during the observed period. Although the Customers At Risk group had not traded with the business for a long time, the number of orders and the amount of revenue that this customer group could bring is very large for the business, its estimated CLV is fairly high.

4.3. Discussion

The study had found out the relationship between the original RFM customer segmentation and the RFM customer clustering by K-Means. Figure 19 describes the total number of customers in 8 segments divided by RFM score and 3 clusters classified by K-Means. As can be seen that the customers in Gold level were mostly divided into the Top Customers, Emerging Customers, and Customers At Risk, these were also the three which had the most predicted CLV in the previous analysis. This shows that the models used in the study are closely related to each other.

Besides, as mentioned above, the Bronze and Silver groups had almost the same Frequency and Monetary indexes, only a big difference in Recency. Therefore, the K-Means model for clustering has not been fully effective. Since each run of K-Means gave different clustering results, and the user had to base on those results and label the clusters for each customer group, this requires a lot of expertise in the field to be able to effectively cluster and label the group thoroughly. However, the K-Means clustering and the predictions of the Pareto/NBD and Gamma-Gamma model were matching when the 3 top customer groups in CLV including customers in Gold level, and the two next most CLV groups contained mostly customers from the Silver group.

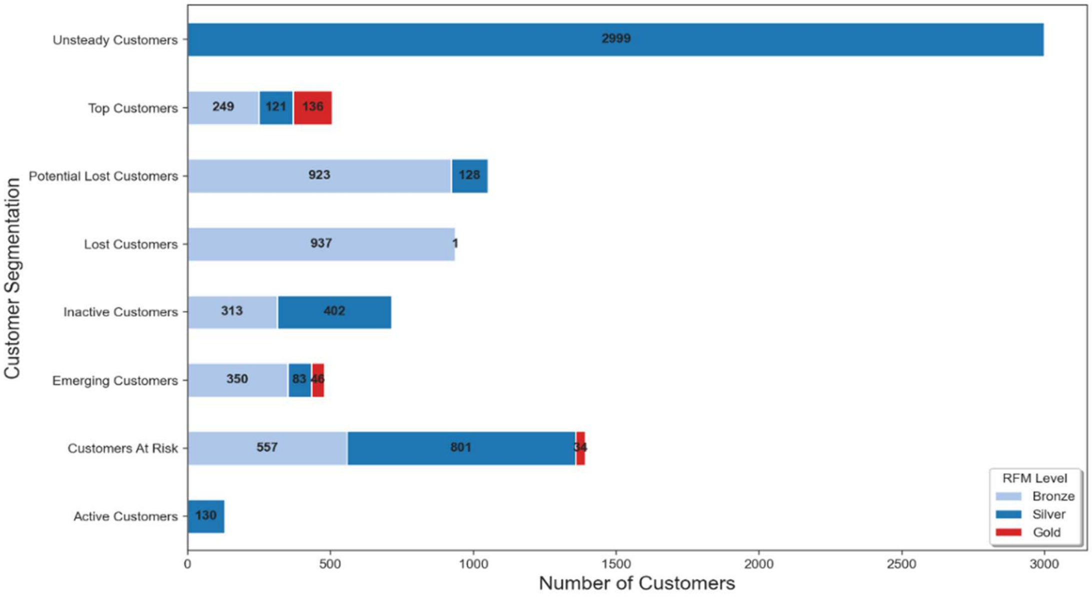

Figure 19: The number of customers by segmentation and RFM level

5. Conclusions and implications

By combining marketing and business knowledge with information technology, a clearer view of the Adventure Works company was realized. RFM is easy to apply and flexible method for implementing customer segmentation. As Mark Patron commented that RFM did not provide the company the profitability and the potential of a customer (Patron, 2004), using the combination of RFM and CLV results to find hidden potential customers in the business will be very profitable. Managers can base on that to implement customer care policies such as discounts and customer gratitude programs for Gold and Silver customers or use cross-selling strategy to maximize profits from existing customers as well as attract new customers. The model in this study was defined to use most effectively for this dataset. They can be developed to use particularly based on the company's needs.

References

Anitha, P. & Patil, M. M. (2019). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University – Computer and Information Sciences*, 1-8. doi:10.1016/j.jksuci.2019.12.011

Avinash, A., Sahu, P. & Pahari, A. (2019). Big Data Analytics for Customer Lifetime Value Prediction. *Telecom Business Review, 12*(1), 46-49. Retrieved from http://publishingindia.com/tbr/

Chicco, D., Warrens, M. J. & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*(3), e623. doi:10.7717/peerj-cs.623

Croll, A. & Yoskovitz, B. (2013). *Use the Data to Build a Better Startup Faster* (1 ed.). Cambridge: O'Reilly Media.

Dwivedi, S., Pandey, P., Tiwari, M. S. & Kalam, A. (2014). Comparative Study of Clustering Algorithms Used in Counter Terrorism. *IOSR Journal of Computer Engineering (IOSR-JCE), 16*(6), 13-17. doi:10.7763/IJKE.2015.V1.18

Fader, P. S., Hardie, B. G. S & Lee, K. K. (2004). “Counting Your Customers” the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science, 24*(2), 275-284. doi:10.1287/mksc.1040.0098

Glady, N., Baesens, B. & Croux, C. (2009). A modified Pareto/NBD approach for predicting customer lifetime value. *An International Journal, 36*(2), 2062-2071. doi:10.1016/j.eswa.2007.12.049

Hellerslia, G & Talal, D. (2020). *Comparison of classical RFM models and Machine learning models in CLV prediction.* Norway: Master of Science.

Humaira, H. & Rasyidah, R. (2020). Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm. *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018* (pp. 24-25). Padang: EAI. doi:10.4108/eai.24-1-2018.2292388

Ismail, M. & Dauda, U. (2013). Standardization and Its Effects on K-Means Clustering Algorithm. *Research Journal of Applied Sciences, Engineering and Technology, 6*(17), 3299-3303. doi:10.19026/rjaset.6.3638

Ismail, M., Bawa, M. & Safrana, M. J. (2015). Impact Of Marketing Strategy On Customer Retention In Handloom Industry. *5th International Conference, SEUSL*, (pp. 16-25). Sri Lanka.

Jasmin. (2020, November 12). *Machine Learning In Customer Segmentation With RFM-Analysis.* Retrieved from Nextlytics: https://www.nextlytics.com/blog/machine-learning-in-customer-segmentation-with-rfm-analysis

Kotler, P. & Keller, K. L. (2006). *Marketing Management* (12th ed.). New Jersey: Pearson Prentice Hall.

Madhu, Y., Pathakota, S. R. & Srinivasa, T. M. (2010). Enhancing K-means Clustering Algorithm with Improved Initial Center. *International Journal of Computer Science and Information Technologies, 1*(2), 121-125.

Miglautsch, J. R. (2000). Thoughts on RFM scoring. *Journal of Database Marketing & Customer Strategy Management, 8*(1), 67-72. doi:10.1057/palgrave.jdm.3240019

Nainggolan, R., Perangin-angin, R., Simarmata, E. & Tarigan, F. A. (2015). Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method. *Journal of Physics: Conference Series*, 1-6. doi:10.1088/1742-6596/1361/1/012015

Ogbuabor, G. & Ugwoke, F. N. (2018). Clustering Algorithm For A Healthcare Dataset Using Silhouette Score Value. *International Journal of Computer Science & Information Technology (IJCSIT), 10*(2), 27-37. doi:10.5121/ijcsit.2018.10203

Patron, M. (2004). Applying RFM segmentation to the SilverMinds catalogue. *Journal of Direct Data and Digital Marketing Practice, 5*(3), 269-275. doi:10.1057/palgrave.im.4340243

Thanh, H. T., Son, N. T. (2021). An interdisciplinary research between analyzing customer segmentation in marketing and machine learning method. *Sci. Tech. Dev. J. - Eco. Law Manag*, *6*(1), 2005-2015. doi:10.32508/stdjelm.v6i1.850

Zaki, M., Kandeil, D., Neely, A. & McColl-Kennedy, J. R. (2016). *The Fallacy of the Net Promoter Score: Customer Loyalty Predictive Model*. UK: Cambridge Service Alliance.

From service recovery to post-recovery customer satisfaction: A review the role of customer control and transparency

Luu Quang Minh[1], Nguyen Ngoc Duy Phuong[2,*]

[1,2,*]International University, Vietnam National University Ho Chi Minh City

[*]Corresponding author: nndphuong@hcmiu.edu.vn

ABSTRACT

Service business has trouble retaining their customers because of poor complaints handling and underestimation of service recovery process. This research paper aims to describe the relationship between service recovery and customer satisfaction and suggest potential service recovery implications to increase customer satisfaction. This research is conducted by comparing different service recovery definition from different researchers' point of view and identify common trends and patterns on customers perception of justice, hence, giving support ideas that can help increase customer satisfaction. The outcome of this study will help service managers to understand service recovery and service recovery procedure to recover satisfaction and retain customers. This research paper lacks analyzing the difference in post-recovery satisfaction difference in age groups, genders, and ethnicities, as well as the difference between group service failure and individual service failure.

Keywords: Customer satisfaction, service recovery, service failure.

1. Introduction

Service recovery is a term that appears following service failure, as service failure commonly exists when delivering service to customers. Service failure appears suddenly in the service delivery process because the process contains multiple steps that involve not only the human resource of the firm, but also with the help of their own customers. Therefore, service failure occurs when firms are not able to maintain a consistent quality level, business does not meet customers' expectations, and is unavoidable (Gustafsson, 2009; Johnson et al., 2002; Krishna et al., 2011; Pranić & Roehl, 2013). The incident raises a concern about customers' dissatisfaction, negative words-of-mouth (WOM), and customers preservation for service firms. Service firms usually does not notice the long-term effect of losing their customers as they think that they can fill in the gap of previous customers with new ones. A study in 2006 by Ang and Buttle shows that keeping loyal customers provide a significant financial benefit such as reducing cost of acquiring new customers, rather than trying to acquire new customers. However, service failure will not cause customers to go away to their competitors, but dissatisfaction will (Cheng et al., 2019). Service recovery is a tool to communicate and build stronger relation with customers in the form of interaction, empathy, and action (Krishna, Dangayach, & Sharma, 2014). In the same research, service recovery provides customers with confidence and confirmation, aim at achieving an immediate and long-term satisfaction and relationships. Other researchers (Baron, Harris, Elliott, Reynolds, & Harris, 2005; Mansori et al., 2014, as cited in Cheng et al., 2017) emphasizes the importance of service recovery because of its usefulness in increasing customer satisfaction.

Handling service failure with service recovery is not an easy thing to do, because firms could never be able to know how dissatisfied their customers are. Moreover, many service firms and service business researchers (Albrecht et al., 2018; Cheng et al., 2017; Xu et al., 2018) believe that compensation such as money, coupons, rooms, service, or seats upgrade can offset customer dissatisfaction. However, Krishna (2014) found out that although monetary and