Customer Segmentation using Machine Learning

Muhamad Harith Arash (2291111)

January 2023

Abstract

In the world of business, maximising profit and avoiding losses are the primary goals of any company. Therefore, the phenomenon of customer churn is one of the primary concerns that companies or organisations try to minimize; they prefer to sustain existing customers and gain new ones. Thus, the study of different types of customers and their relationships with a company's products can provide strong insight for business marketing strategy. Customer churn happens when consumers are no longer interested in the company's services or products and have made their transactions a long time ago. Losing clients would not only cause heavy losses but also pose a threat to the company. Due to the presence of several rivals in the same industry, it is more important to re-engage dissatisfied customers than to find new ones. It has been noted that obtaining new customers is more expensive than keeping an existing one. One of the best ways to counter such phenomenon is by doing churn prediction analysis, it analyses customer behaviour by predicting consumers who are likely no longer interested based on historical data and suggesting several improvement methods. If consumers churn, the company will suffers a loss that includes not just lost income but also the costs of further marketing to acquire new customers. Hence, many organisations set their business goal to reduce the rate of client turnover by taking all necessary actions. In this project, the dataset from the UCI Machine Learning repository will be used for this study. Transaction records from December 1, 2010, to December 1, 2011, make up this collection. The dataset is about a web-based retail gift retailer in the United Kingdom. The data will undergo several stages, such as cleaning, preprocessing, modeling, and evaluation. In the end, all of the customers will be segmented according to their characteristics. Here, segmentation of customers would be done by using the RFM technique and several machine learning algorithms.

1 Introduction

In the world of technology and the rapid expansion of online retail commerce, there is an increasing number of large-scale online business platforms, which has increased competition among e-commerce firms. Because of the huge number of people who shop online, a large amount of historical data is generated and it can be analysed by e-commerce companies. They can learn from the information they have and use it to attract more customers and improve customer satisfaction and loyalty by providing more personalised services and impulsive marketing campaigns [1].

There are different types of customers; some prefer to shop for a certain period of time, some shop immediately, etc. The task of identifying these patterns is difficult for any organization. One of the biggest threats to a company's development is the customer who breaks ties with the company, known as "customer churn".

Therefore, companies should better understand how their services can better meet the needs of their consumers. Understanding customer behaviour is at the heart of any successful marketing campaign. Hence, targeting a certain group of potential customers with the exact services that they require may increase enterprise profits and establish better customer relationships. Also, companies can plan a concrete strategy to encounter any customer churn in the future.

The purpose of this study is to identify people's purchasing behaviors, categorise them according to similar characteristics by using clustering techniques, and provide insight for companies on how to tackle their consumers so they can maximise their profits. One of the best methods to describe customer behaviours is the customer segmentation process. An unsupervised learning algorithm that identifies the subgroups in the data such that data points in the same subgroup (cluster) are extremely similar while data points in different clusters are very different. Examples of such algorithms include KMeans, HAC, DBSCAN, and many others.

2 Background Information

2.1 Customer Lifetime Value

Customer lifetime value is one of the fundamental matrices that are crucial in business development; it measures how much a company can maximise its revenue from the average customer over the course of a relationship. Each provides important insights into how customers interact with businesses and if the product in the overall marketing plan is working as expected. The most common elements used to predict how customers will spend their money on a company's products are recent customer transactions, purchasing frequencies, and total product amount [2].

2.2 Customer Churn Rate

Customer retention measures how many customers stick with a business for a given period of time. Almost all B2B and B2C businesses use this crucial statistic, also known to as the churn rate. In general, the lower the churn rate of the customers, the more likelihood that customers satisfy with services of the firms, and companies can learn to understand what are the reasons that customers no longer interested their products [3].

2.3 Customers type

In every business's structure, customers are the most important assets. In order to be successful in marketing and generate the highest profit, it is necessary to identify and categorise different types of customers.

In the retail industry, customers can be segmented into five main types:

Loyal customers: Loyal customers are the highest purchasers of a company's products compared to other competitors; they are easy to appease, and they contribute the majority of the company's sales revenue. They are loyal and place a high value on a product, as the term implies.

Additionally, they are more likely to suggest the company's products to others. So, it is crucial to ask for and include their views and opinions in a company's decision-making process. Loyal customers should be prioritised by businesses if they wish to expand. [4].

Browsing customers: This type of customer draws a huge amount of traffic to the company's platform, but they have a small likelihood of buying products. The revenue gained by this group is small as a percentage of sales. They have no specific need and simply browse products and buy something if they are interested. Sometimes, these customers save a lot of items in their online cart but never check out, so spending too much time on this group is unnecessary as it generates the least amount of sales revenue. However, if businesses can identify this behavior, they can use advertising to pique the interest of buyers, leading to a purchase. [4].

Discount customers: Discount customers are the least loyal segment of customers; they rarely purchase products at full price and buy when the price is considered lower. However, they have a role in turning over a company's stocks because they are one of the main contributors to a company's cash flow. They are resilient to upselling and generally move on to other platforms when better prices are available elsewhere [4].

Impulse customers: This group is the second-most attractive segment for companies trying to upsell their new products. Customers in this group do not have a specific shopping list and purchase products spontaneously. Also, they are typically attracted to any product recommendations, but their buying behaviour is unpredictable. They are also one of the major contributors to company sales revenue. Promoting a lot of new products to these customers may improve company profitability [4].

Need-based customers: The goal of this group is to buy products based on their needs and not browse other products. In other words, they shop regularly and need exactly what they want. They purchase for a particular purpose or event, which makes upselling difficult. It's important to know that need-based clients are susceptible to being attracted to competing firms. Hence, companies should start a good relationship with this clientele in order to keep them from leaving e-commerce platforms. Need-based clients may become loyal ones with the correct type of supportive interpersonal relationships [4].

3 Related Work

The data set used in this paper [5] was taken from a UK-based online retail gift store and is stored in the UCI Machine Learning repository. Pre-processing of this data set includes removing NAs, confirming numerical values, and removing incorrect data points. The data was then aggregated in order to provide data sets based on invoices and customers. Each data point has a variable turnover associated with it. All customer aggregated data is put through three algorithms and writers found that the accuracy of Random Forest, Support Vector Machines, and Extreme Gradient Boosting is increasing after running all three models. However, the amount of time needed to complete the computation likewise gets longer in the same order.

4 Hypothesis

It may be reasonable to predict a strong correlation between customer revenue in different customer segments. Also, it could be predicted that customers in the best segments have a higher lifetime value and retention rate compared to other segments.

5 The Data

This Online Retail data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/01/2011 and 30/11/2011 from the UCI Machine Learning repository [6]. The business mostly provides one-of-a-kind all-occasion giftware. Many of the company's clients are wholesalers.

InvoiceNo: It is known as an invoice number with a nominal data type. All entries consist of a 6-digit integral number uniquely assigned to each transaction; if the code starts with the letter ”c,” it indicates a cancellation.

StockCode: It is a product code with a nominal data type. All data comprises a 5-digit integral number uniquely assigned to each distinct product.

Description: it is a product name with nominal data type.

Quantity: The total quantities of each product per transaction in numeric type.

InvoiceDate: This is a numeric data type for invoice date and time. The datetime was generated when a transaction was made.

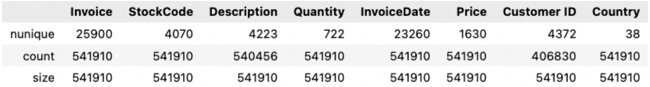
UnitPrice: A unit price for each product in sterling (£).

CustomerID: A 5-digit integral, unique customer number in nominal data type.

Country: The name of the country in nominal data type that a customer resides.

6 Data Processing

Initially, the uncleaned data had 54,1910 data entries. The total number of unique transactions (invoice) is 25900, 4070 different stocks from 4372 different customers, and 38 different customers. The table 1 below shows a brief summary of the uncleaned dataset:



	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
nunique	25900	4070	4223	722	23260	1630	4372	38
count	541910	541910	540456	541910	541910	541910	406830	541910
size	541910	541910	541910	541910	541910	541910	541910	541910

Table 1: *Original Dataset.*

All 5268 duplicate entries (roughly 1% of the total) have been removed. There are no missing values (Null) in the dataset, but due to the large number of revenue, approximately 20 data entries are considered outliers. This is to avoid overfitting in the modelling process. Finally, approximately 10K data entries were removed because they were classified as cancelled orders or did not pertain to customers.

7 Analysis

7.0.1 Descriptive Analysis

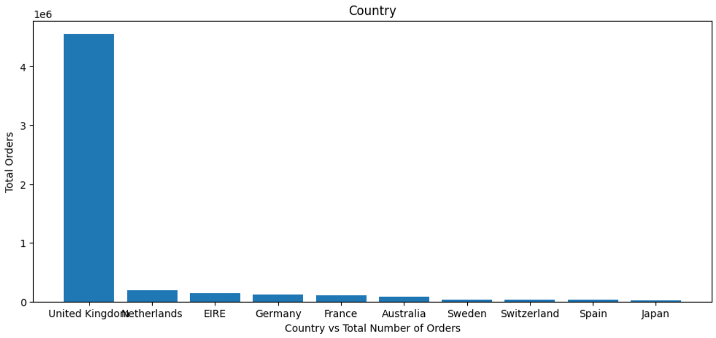


Figure 1: *Total number of orders for top 10 countries.*

The bar chart (figure 1) shows the total number of orders from the first 10 countries in decreasing order. More than 90% of orders are coming from the United Kingdom, and all other countries' numbers of orders are small in the data.

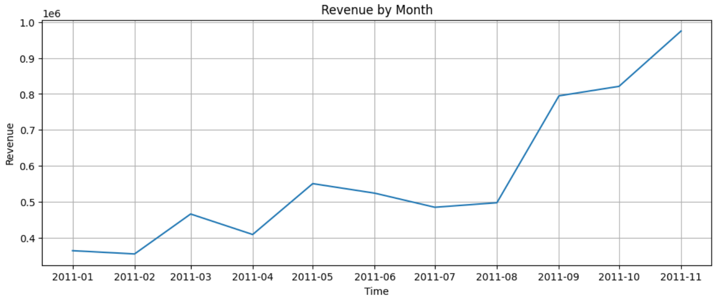


Figure 2: *Monthly Revenue from January 2011 until November 2011.*

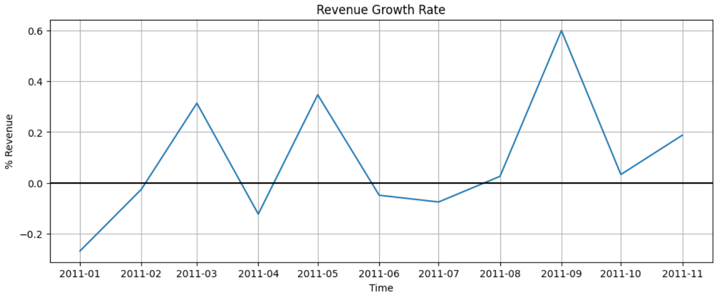


Figure 3: *Monthly Growth from January 2011 until November 2011.*

Figure 2 shows a line graph. The revenue increased throughout the 11 months; however, there are several months where sales dropped significantly; based on Figure 3, the months of April, June, and July sales are below zero. Companies need to investigate why it happened in these months. Whether it was due to fewer active customers, fewer orders, or customers preferring to buy cheaper products, further deep-dive analysis is required by companies to know the reasons why this happened.

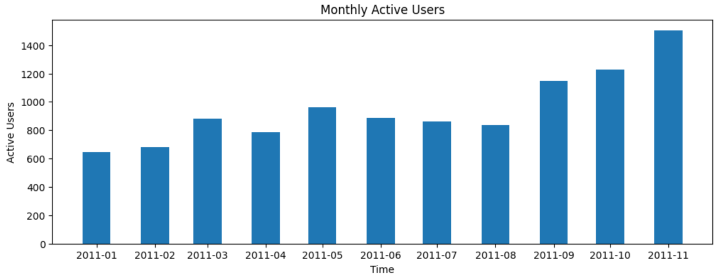


Figure 4: *Monthly Active Users.*

According to the bar chart (figure 4), the total number of active customers increases in a non-linear fashion, with November having the highest number. The drop in sales over in April (784), June (889), July (859), and August (834) are most likely due to a lower number of total active customers than usual.

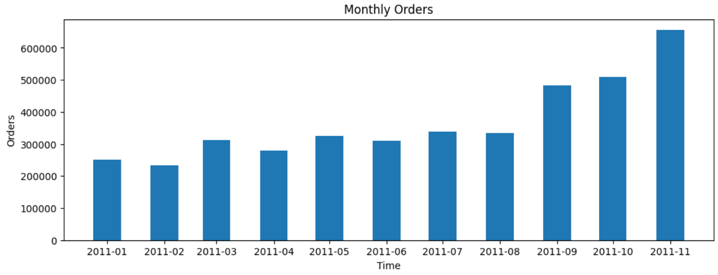


Figure 5: *Monthly Orders throughout the year.*

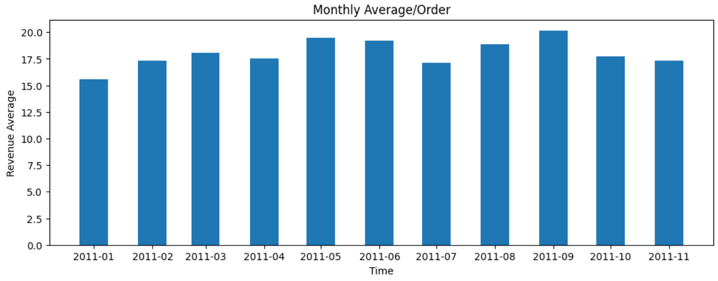


Figure 6: *Average Revenue every order within one month.*

According to figure 5 of the bar chart, total order count decreased in April and June (311k to 278k, -10% each) and (324k to 309k, -4% each). Figure 6 shows that the total number of active customers directly affected the number of orders. Therefore, the value of the monthly average revenue per order also dropped. Especially in April (18.0 to 17.0).

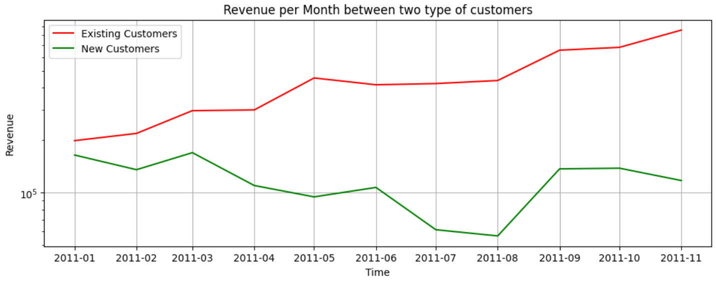


Figure 7: *New customer comparison for every month.*

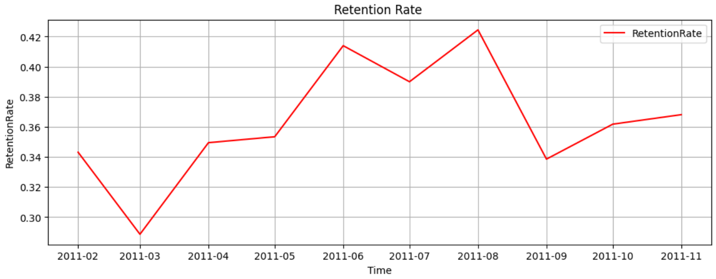


Figure 8: *Monthly Retention Rate.*

In order to get more insight about customer profiles, further important metrics such as new customers every month and monthly retention are required. The new customer ratio is a good indicator of how companies lose customers every month. While the monthly retention rate can tell businesses how many customers continue to purchase from the previous month within a given time frame.

From the monthly new customer line chart (figure 7), the existing customer trend shows a positive trend. But the number of new customers have a slight negative trend throughout the time given. From March 2011 (169860) to August 2011 (56534.360), the monetary value gain from new customers began to decline. In Figure 8, the monthly retention rate dropped until March but increased until June, then fluctuated and jumped back to previous levels from June to August.

7.0.2 Predictive Analysis

(A) Customer Segmentation Analysis

(a) RFM Technique:

RFM analysis is the method used to quantitatively rank customers based on the recency, frequency, and monetary value of their historical data transactions. The technique basically scores each customer based on the last transaction they made, the total number of transactions, and the total revenue of their transaction.

Ecommerce companies may provide services or products that are specifically tailored to certain market segments by understanding the demands of various consumer groups [7].

(b) Process Flow: The model, as shown in the figure 9 below, consists of multiple stages for categorising into different groups, which is nothing more than customer segmentation. First, the entire dataset was downloaded, and then the data was preprocessed or cleaned by removing duplicate and cancel order transactions. Each customer's recency, frequency, and revenue scores are determined using the RFM model, and their cluster identification is determined using the Kmeans or EM algorithms.

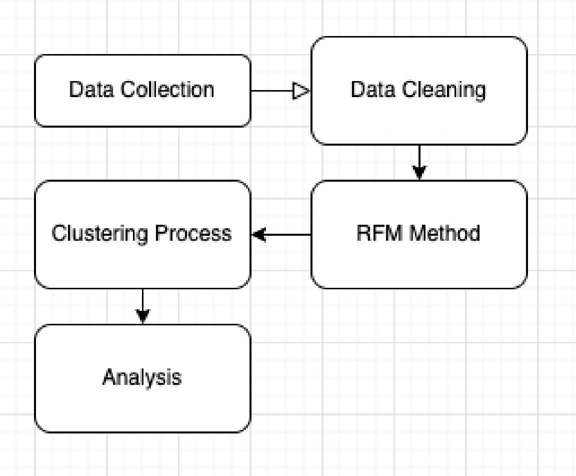


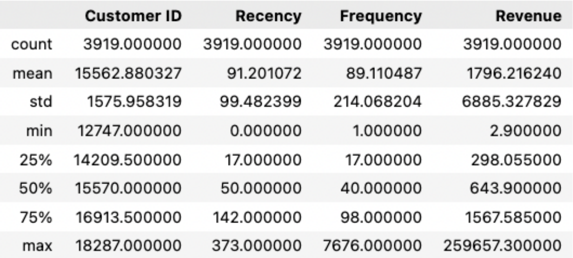
Figure 9: *Process Flow.*

(c) RFM Implementation:

(i) Recency Calculation: The value of recency for each customer can be calculated by subtracting the most recent date in the dataset to the most recent purchase date of each customer. The most current date discovered for the whole dataset is 2011-12-09. For the purpose of calculating the recency value for each client, this acts as a benchmark date.

(ii) Frequency Calculation: The total frequency for each customer can be calculated by counting the total number of times they have made purchases.

(iii) Revenue Calculation: Revenue or monetary value generated from each customer can be calculated by multiplying a unit price with the total quantity in each transaction. The total amount spent by each client is calculated by adding together the sums of these totals for each transaction.

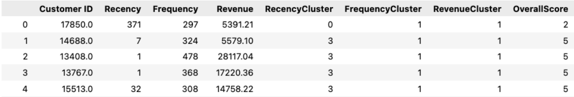


	Customer ID	Recency	Frequency	Revenue
count	3919.000000	3919.000000	3919.000000	3919.000000
mean	15562.880327	91.201072	89.110487	1796.216240
std	1575.958319	99.482399	214.068204	6885.327829
min	12747.000000	0.000000	1.000000	2.900000
25%	14209.500000	17.000000	17.000000	298.055000
50%	15570.000000	50.000000	40.000000	643.900000
75%	16913.500000	142.000000	98.000000	1567.585000
max	18287.000000	373.000000	7676.000000	259657.300000

Table 2: *Description of RFM values.*

From the table 2 above, the average recency and frequency of the customers' purchases are around 91 days and 90 times, respectively. Each customer spends an average of $1796.22 in total. It is clear that compared to recency and frequency values, the revenue value range is considerably wider. Therefore, before supplying data as input to the EM algorithm, log transformation must be used to scale numbers appropriately.

All recency, frequency, and revenue are scored (0, 1, 2, and 3) and entered into the new columns RecencyScore, FrequencyScore, and RevenueScore. All of these scores are calculated and grouped together to get the OverallScore, as shown in the table 3 below. After calculating all overall scores, the score can be used to segment customers and define different characteristics for each group.



	Customer ID	Recency	Frequency	Revenue	RecencyCluster	FrequencyCluster	RevenueCluster	OverallScore
0	17850.0	371	297	5391.21	0	1	1	2
1	14688.0	7	324	5579.10	3	1	1	5
2	13408.0	1	478	28117.04	3	1	1	5
3	13767.0	1	368	17220.36	3	1	1	5
4	15513.0	32	308	14758.22	3	1	1	5

Table 3: *RFM score table.*

(d) Expectation-Maximization (EM Algorithm) : The technique of clustering is described as one that separates the entire set of data into clusters that have similar behaviours seen in the data.

Expectation Maximization (EM) looks for parameters of the distribution that agree best with the data [8]. It determines the probability that each data point will be in each cluster. After it is computed, the new parameters for distribution will be updated, and the process will be repeated until stabilised or sufficient iteration is achieved.

(e) EM implementation : Before the clustering process, choosing the appropriate number of clusters is essential. Therefore, the elbow method will suggest the optimal values needed to properly classify customers based on all the metrics calculated previously.

Elbow Method: This approach uses k-means clustering on the dataset for a range of k values, say from 1 to 10, and computes an average score for each k value for each cluster. By default, the distortion score is calculated by adding the squares of the distances between each point and its designated centre and using the value located at the elbow curve to determine how many clusters to use [9].

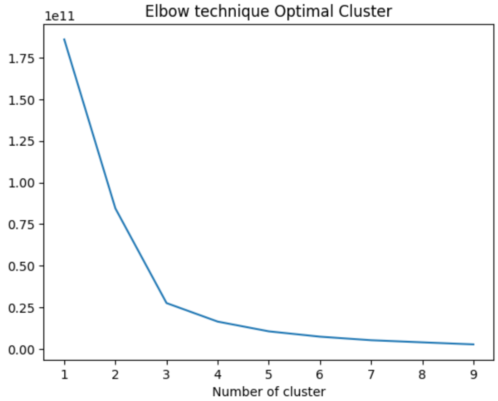
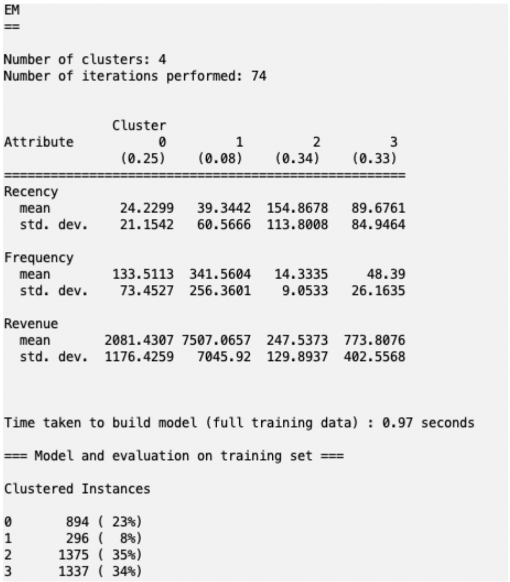


Figure 10: *Elbow line chart.*

As figure 10 above illustrates, the total of square distances decreases as the number of clusters increases. Therefore, the optimal number of clusters is 3 or 4. For this project, the value of 4 was picked because it had more distributions than the value 3.

(f) Result : Figure 11 and table 4 below show the summary report of the EM model been used to classify all customers based on their RFM scores.



```
EM
==

Number of clusters: 4
Number of iterations performed: 74


               Cluster
Attribute            0          1          2          3
                (0.25)     (0.08)     (0.34)     (0.33)
=======================================================
Recency
  mean         24.2299    39.3442   154.8678    89.6761
  std. dev.    21.1542    60.5666   113.8008    84.9464

Frequency
  mean        133.5113   341.5604    14.3335      48.39
  std. dev.    73.4527   256.3601     9.0533    26.1635

Revenue
  mean       2081.4307  7507.0657   247.5373   773.8076
  std. dev.  1176.4259    7045.92   129.8937   402.5568



Time taken to build model (full training data) : 0.97 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       894 ( 23%)
1       296 (  8%)
2      1375 ( 35%)
3      1337 ( 34%)
```

Figure 11: *Cluster summary report.*

Cluster	Recency	Frequency	Revenue	Category
0	24.23	133.51	2081.43	New
1	39.34	341.56	7507.07	Best
2	154.88	14.33	247.54	Lost
3	89.68	48.39	773.81	Risk

Table 4: *Cluster category.*

All customer groups' descriptions can be seen below, along with some suggestions that business owners should make when creating marketing plans for each group.

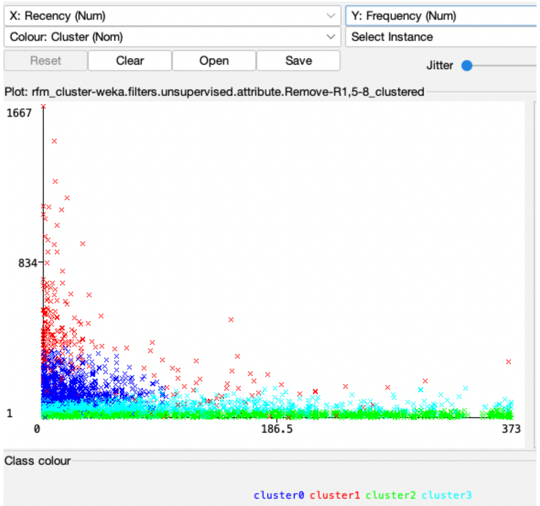
New Customer (Navy Blue): Customers who transacted recently and have lower purchase frequency, with low amount of revenue gain. This category need to handled with care by improving relationships with them. Company should try enhance their purchasing experience by providing good quality products and services, and customer care services.

Best Customer (Red): Most frequent spenders with the highest revenue gainand had transacted recently. Company can use this group to test new products and can increase company revenue by advertisement or small discounts.

Lost Customer (Green): Customers with the least monetary spending and the least number of transactions. They made their last purchase long ago. These customers are no longer interested with the company products and have exited from customer base. The best ways is to learn why this group of people no longer in the system so the precaution can be made.

At risk Customer (Cyan): Customers who made their last transaction a while ago and made less frequent and low revenue gain. Customers in this group have high probability of churning. Many actions should be taken such as advertising, discounts or gifts. Company should investigate why they are leaving in order to avoid same cases happen in the future. The best

Below are some recency, frequency, and revenue visualisations in 4 different clusters.



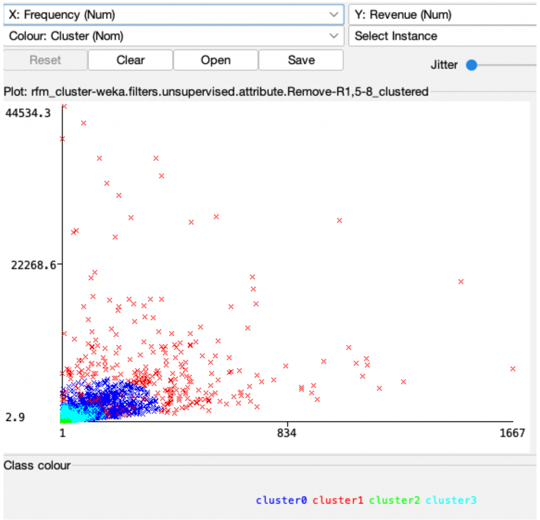
X: Recency (Num)

Y: Frequency (Num)

Colour: Cluster (Nom)

Plot: rfm_cluster-weka.filters.unsupervised.attribute.Remove-R1,5-8_clustered

Figure 12: *Recency vs Frequency.*



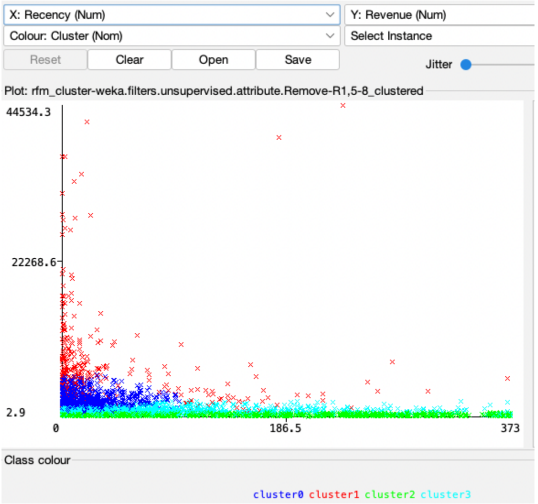
X: Frequency (Num)

Y: Revenue (Num)

Colour: Cluster (Nom)

Plot: rfm_cluster-weka.filters.unsupervised.attribute.Remove-R1,5-8_clustered

Figure 13: *Frequency vs Revenue.*



X: Recency (Num)

Y: Revenue (Num)

Colour: Cluster (Nom)

Plot: rfm_cluster-weka.filters.unsupervised.attribute.Remove-R1,5-8_clustered

Figure 14: *Recency vs Revenue.*

Based on figures 12, 13, and 14, the total number of customers for all four clusters can be seen in the recency, frequency, and revenue charts.

The red-colored cluster (cluster 1) represents the best or loyal customers, with the highest revenue, highest frequency, and lowest recency. Around 296 customers belong to this group.

The next cluster, navy blue (cluster 0), represents the new customer with a total number of 894, with frequency and recency a bit lower than cluster 1, but still contributing a huge amount of revenue to the company.

Next, the total number of lost customers is in the third cluster around 1375, the green-colored one (cluster 2). This group has the lowest frequency and highest recency value compared to the rest of the clusters.

Finally, the cyan-colored customer at-risk cluster (cluster 3) has approximately 13,000 customers. There are the potential customers that will leave at any time if no action is taken, as their group is close to the lost customers.

The overall score indicates whether customers belong to different groups based on similar characteristics. Best customers and new customers are two distinct groups that should be retained to ensure the company's security and maximise revenue.

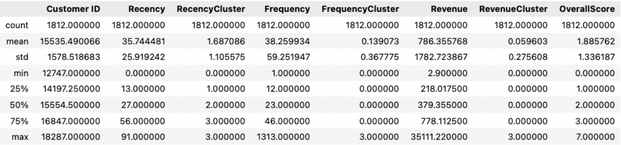
(B) Customer Lifetime Value Analysis

(a) Definition: Customer lifetime value (CLVT) is a measure that shows how much money a company can expect to make overall from a single customer account over the course of the customer relationship [10]. The formula for CLVT is as follows:

$$\text{Customer Value} = \text{Average Order Value x Purchase Frequency}$$

(b) Data Preprocess: The data must be divided into two parts in order to predict and calculate CLTV. For this project, the time window of 3 months of data will be used to predict customer CLVT for the next 6 months. Therefore, the prediction will be made only for existing customers. The expected number of groups should be 3.

Based on the formula above, the customer lifetime value is equivalent to the monetary value received from customers. Extra matrices can be calculated using the RFM technique to assign customers to different groups. The overall score for the past three months of data is shown in the table 5 below:



	Customer ID	Recency	RecencyCluster	Frequency	FrequencyCluster	Revenue	RevenueCluster	OverallScore
count	1812.000000	1812.000000	1812.000000	1812.000000	1812.000000	1812.000000	1812.000000	1812.000000
mean	15535.490066	35.744481	1.687086	38.259934	0.139073	786.355768	0.059603	1.885762
std	1578.518683	25.919242	1.105575	59.251947	0.367775	1782.723867	0.275608	1.336187
min	12747.000000	0.000000	0.000000	1.000000	0.000000	2.900000	0.000000	0.000000
25%	14197.250000	13.000000	1.000000	12.000000	0.000000	218.017500	0.000000	1.000000
50%	15554.500000	27.000000	2.000000	23.000000	0.000000	379.355000	0.000000	2.000000
75%	16847.000000	56.000000	3.000000	46.000000	0.000000	778.112500	0.000000	3.000000
max	18287.000000	91.000000	3.000000	1313.000000	3.000000	35111.220000	3.000000	7.000000

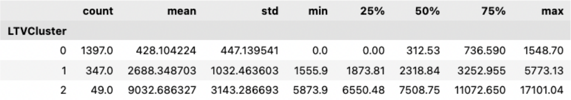
Table 5: *3 months data to predict 6 months .*

(c) K-Means algorithm : K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on [11]. The algorithm can be understand with the following steps:

- Initialize some k points, calculate means with random selected points.
- Update the centroid coordinates after categorising each item to its closest mean.
- Repeat the method for a specified number of iterations or until the clusters forming are stable and the centroids do not change, at which point the ultimate cluster is formed.

There is no cost specified in the dataset. The customer's lifetime value for the next six months is equivalent to their revenue. To see correlations between LTV and the features calculated early, all calculated CLVT is merged with 3 months of data.

By applying K-means clustering, all customers will be classified into 3 different groups of LTV clusters (high, mid, and low).



	count	mean	std	min	25%	50%	75%	max
LTVCluster								
0	1397.0	428.104224	447.139541	0.0	0.00	312.53	736.590	1548.70
1	347.0	2688.348703	1032.463603	1555.9	1873.81	2318.84	3252.955	5773.13
2	49.0	9032.686327	3143.286693	5873.9	6550.48	7508.75	11072.650	17101.04

Table 6: *Customer lifetime value.*

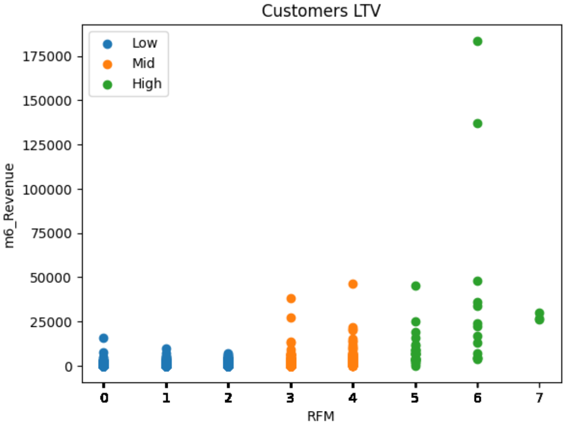


Figure 15: *Correlation of LTV with RFM.*

From table 6 and figure 15 above, the positive correlation is quite obvious, as a higher RFM score means a higher LTV, and vice versa. From the table, it indicates Cluster 2 is the best segment with an average 9.03k LTV, whereas Cluster 0 is the worst with 428.

(d) Linear Regression Model :

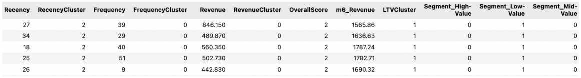
Linear regression is the most basic and widely used model in predictive analysis. It is able to predict any scalar regression problem by finding the best coefficients for each parameter and reducing the mean squared error (MSE) between predicted and actual values.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2.$$

The MSE is calculated by adding the squared differences between the observed and predicted values and dividing the result by the sample size. The lower the MSE value, the better the model [12].

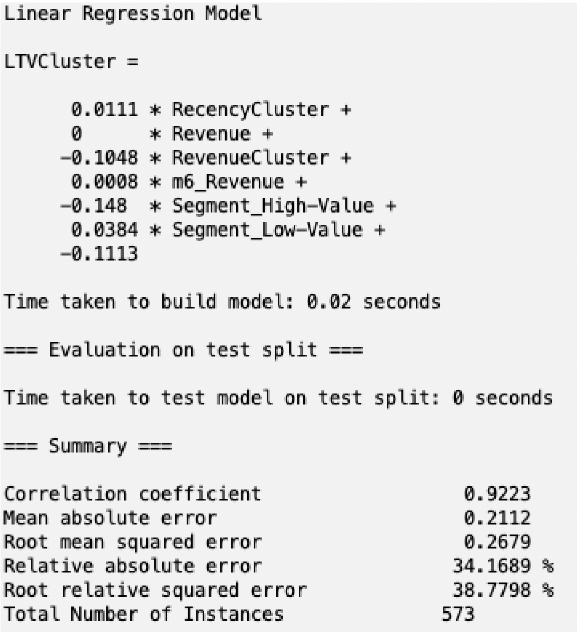
R-squared, also known as the coefficient of determination, measures the accuracy of the fit. It provides the relative measure of the percentage of the dependent variable variance that is fitted to a regression model. The measurement ranges from 0 to 1 [13].

Table 7 shows the final data built with RFM scores and each customer segment based on the LTV cluster.



Recency	RecencyCluster	Frequency	FrequencyCluster	Revenue	RevenueCluster	OverallScore	m6_Revenue	LTVCluster	Segment_High-Value	Segment_Low-Value	Segment_Mid-Value
27	2	39	0	846.150	0	2	1565.86	1	0	1	0
34	2	29	0	489.870	0	2	1636.63	1	0	1	0
18	2	40	0	560.350	0	2	1787.24	1	0	1	0
25	2	51	0	502.730	0	2	1782.71	1	0	1	0
26	2	9	0	442.830	0	2	1690.32	1	0	1	0

Table 7: *Data Before Building.*



```
Linear Regression Model

LTVCluster =

      0.0111 * RecencyCluster +
      0      * Revenue +
     -0.1048 * RevenueCluster +
      0.0008 * m6_Revenue +
     -0.148  * Segment_High-Value +
      0.0384 * Segment_Low-Value +
     -0.1113

Time taken to build model: 0.02 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correlation coefficient                  0.9223
Mean absolute error                      0.2112
Root mean squared error                  0.2679
Relative absolute error                 34.1689 %
Root relative squared error             38.7798 %
Total Number of Instances              573
```

Figure 16: *Linear Regression model.*

Based on the summary report (Figure 16), after applying the linear regression model, the R-squared value for the test set is 0.9223, which is good, and the model can predict the future LTV segments of all customers. Also, the value of RMSE is 0.2679, which is low but still acceptable for future prediction.

By knowing CLTV, companies can develop positive return of investment (ROI) strategies and make decisions about how much money to invest in acquiring new customers and retaining existing ones.

8 Conclusion

In today's expanding e-commerce climate, the best approaches to survive and win market competition are to analyse consumer behaviour, acquire new potential customers, and retain existing customers by boosting their fulfilment.The groundwork for a company's future marketing analyses and initiatives may be laid with the help of customer turnover insights. By retaining consumers and their loyalty over time, the development of goods can be improved to meet the needs of the customer.

In this project, four different customer groups have been identified with a simple understanding of customer behaviour through the RFM method and EM algorithm. Each customer has been assigned a group according to their scores, and business owners can target these groups to take any actions to minimise the churn rate. Some suggestions on how to retain customers have been discussed above.

There is also a strong correlation between a customer's overall score and monetary value; the higher the value of the RFM score, the more revenue can be gained. According to customer segmentation analysis, new and best customers contribute the majority of the company's revenue, whereas risk customers have the potential to abandon the company's products. Also, companies should learn the reasons why some customers are no longer interested with their products so the can improve in the future.

Based on customer life-time value analysis, customers in the best segments have a higher CLVT and retention rate compared to other segments. The assumption can be considered accurate as the measurement of R-Squared of the linear regression model evaluation is greater than 90% and RMSE is 0.2679, which considered to be small. Companies can take necessary action for these customers, such as promoting new products or giving some promotions, so the sales can be maintained in the future.

Finally, customer segmentation and lifetime value can be analysed with many machine learning algorithms. Companies can utilise these resources to better understand their customers and improve their efficiency.

So far, we have learned how to segment customers based on their characteristics; in future work, we can improve customer segmentation analysis by implementing other methods such as predicting each customer's churn probability using logistic regression or a decision tree and including other metrics for the best modelling of customer behaviors.

References

[1] N. Bagul, P. Berad, C. Khachane, and P. Surana, "Retail customer churn analysis using rfm model and k-means clustering," *International Journal of Engineering Research and Technology (IJERT)*, vol. 10, Mar. 2021.

[2] S. Staff. "What is customer lifetime value? the complete guide to clv." (2022), [Online]. Available: `https://www.shopify.com/blog/what-is-customer-lifetime-value`.

[3] J. FRANKENFIELD. "Churn rate: What it means, examples, and calculations." (2022), [Online]. Available: `https://www.investopedia.com/terms/c/churnrate.asp`.

[4] C. Team. "Types of customers." (2022), [Online]. Available: `https://corporatefinanceinstitute.com/resources/accounting/types-of-customers/`.

[5] T. Shoaib, "Customers churn prediction in retail store," Sep. 2018. DOI: `10.13140/RG.2.2.30545.38242`.

[6] D. Chen, *Online Retail II*, UCI Machine Learning Repository, 2019.

[7] Hokay. "Rfm and cltv to know your customers better." (2022), [Online]. Available: https://www.analyticsvidhya.com/blog/2022/01/rfm-and-cltv-to-know-your-customers-better/#:~:text=RFM%20and%20CLTV%20are%20two,marketing%20strategies%20for%20different%20segments.

[8] C. B. Do and S. Batzoglou, "What is the expectation maximization algorithm?" *Nature biotechnology*, vol. 26, no. 8, pp. 897–899, 2008.

[9] Wikipedia contributors, *Elbow method (clustering) — Wikipedia, the free encyclopedia*, [Online; accessed 29-December-2022], 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Elbow_method_(clustering)&oldid=1128459330.

[10] A. Caldwell. "What is customer lifetime value (clv) and how to calculate?" (2022), [Online]. Available: https://www.netsuite.com/portal/resource/articles/ecommerce/customer-lifetime-value-clv.shtml#:~:text=Customer%20lifetime%20value%20(CLV)%20is%20a%20measure%20of%20the%20total,and%20the%20total%20average%20profit..

[11] JavaTPoint. "K-means clustering algorithm." (2022), [Online]. Available: https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning.

[12] D. F. Benoit and D. Van den Poel, "Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 475–10 484, 2009, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2009.01.031. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417409000712.

[13] J. Frost. "How to interpret r-squared in regression analysis." (2022), [Online]. Available: https://statisticsbyjim.com/regression/interpret-r-squared-regression/.