FINAL PROJECT

CUSTOMER SEGMENTATION USING SPARK

GROUP 12

2023

SUPERVISED BY

Prof. Anwar Hossain

PREPARED BY

Mohamed Adel Mohamed

20398555

Mohamed Malek Elsamouly

20399120

Mustafa Atif Shalata

20399125

Manar Salah AlShabrawy

20398571

Table of Contents

Abstract

It is not wise to serve all customers with the same product model, email, text message campaign, or ad. Customers have different needs. A one-size-for-all approach to business will generally result in less engagement, lower-click through rates, and ultimately fewer sales. Customer segmentation is the cure for this problem.

Finding an optimal number of unique customer groups will help us understand how our customers differ and help us give them exactly what they want. Customer segmentation improves customer experience and boosts company revenue. That's why segmentation is a must if we want to surpass our competitors and get more customers. Doing it with machine learning is definitely the right way to go.

This report represents how we used Spark's API, MLlib, in customer segmentation for a dataset of online purchase history for a number of customers that were collected based on some features. The data got preprocessed and prepared for modeling using the K-means algorithm and then the segments were grouped based on the results found.

As a result, it ended up with three segments that we labeled as Champions, Loyal, and Potential based on the feature of each group. The results of the K-means model were compared using the RFM model in the traditional way.

1. Introduction

Nowadays, we can personalize everything. There is no one-size-fits-all approach. For business, this is actually a great thing. It creates a lot of spaces for healthy competition and opportunities for companies to get creative about how they acquire and retain customers. One of the fundamental steps toward better personalization is customer segmentation. Customer segmentation simply means grouping your customers according to various characteristics and it is a way for organizations to understand their customers. Knowing the differences between customer groups make it easier to take strategic decisions regarding product growth and marketing. But, doing segmentation manually can be exhausting, so using machine learning will be a great choice.

1.1 Project Description

The purpose of this project is to use spark's API, MLlib, to analyze customer data and segment customers of a company that sells some products in order to find out how well the selling performance of these products is. The segment of the customers will be based on their buying behavior in the market. The segment characteristics are recency, frequency, and mentor value.

1.2 Data Description

The dataset that has been used in the project contains 541909 samples of the online purchase history of 2400 customers. As Fig.1 shows, the data has been collected based on eight features:

- InvoiceNo: Reference number of an invoice
- StockCode
- Description: Name of the product
- Quantity: Quantity sold
- InvoiceDate: Date of the issued invoice
- UnitPrice: Unit price of the product
- CustomerID: Unique number of the customer
- Country

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Fig1. First five records of the dataset

2. Data Preprocessing

2.1 Feature Engineering

We extracted three features out of the eight features to use them as an RFM model. RFM stands for **recency**, **frequency**, and **monetary value** which are the new features that got extracted from Quantity, InvoiceDate, and UnitPrice. Those features got aggregated using customer ID. Fig2. shows how the dataset looks after feature extraction.

CustomerID	Recency	Frequency	MonetaryValue
12346.0	326	2	0.00
12347.0	3	182	4310.00
12348.0	76	31	1797.24
12349.0	19	73	1757.55
12350.0	311	17	334.40

Fig2. Dataset after feature extraction

2.2 Data Shapes and Distribution

2.2.1 Recency Distribution

Fig3. shows the distribution of the recency feature

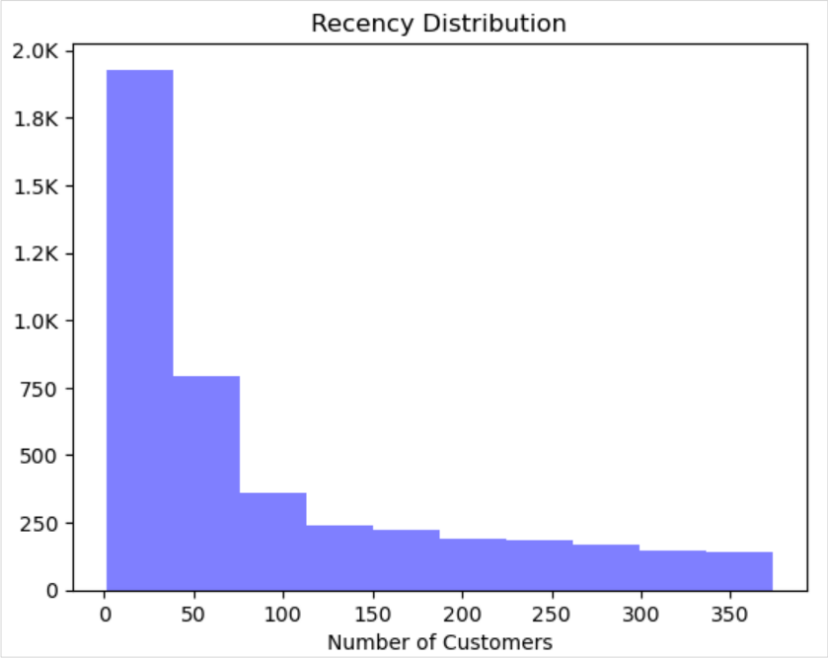

Fig3 Recency Distribution

2.2.2 Frequency Distribution

Fig4. shows the distribution of the frequency feature

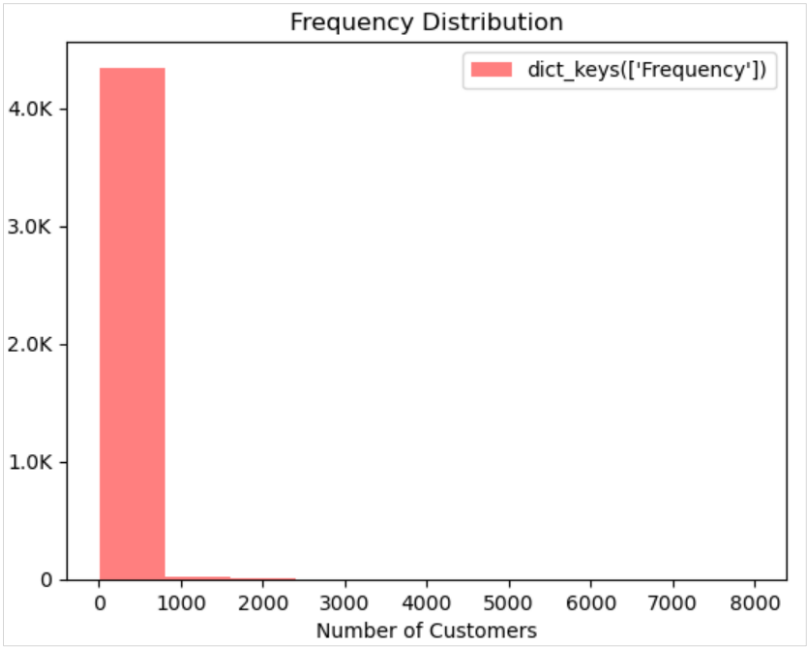

Fig4. Frequency Distribution

2.2.3 Monetary Distribution

Fig5**.** shows the distribution of the frequency feature

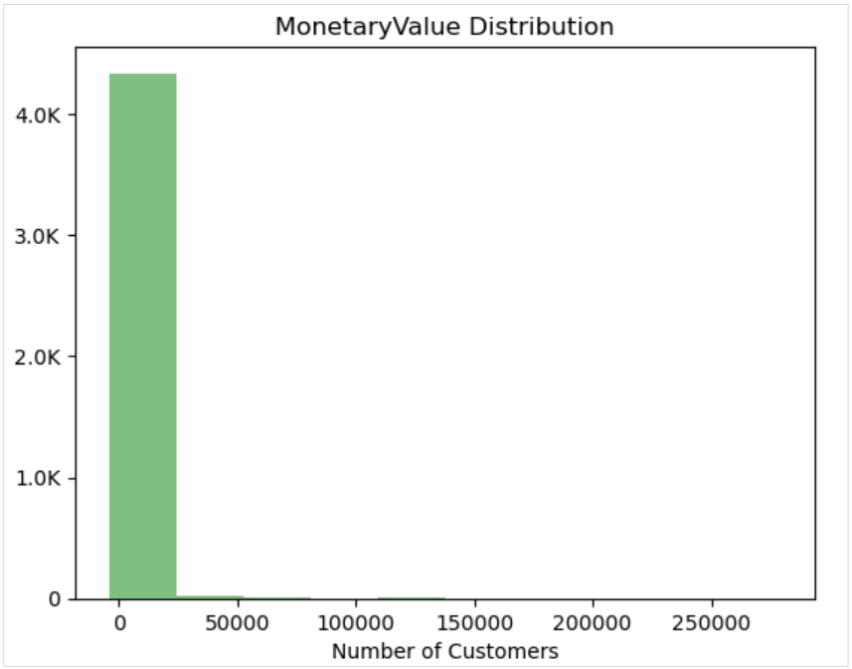

Fig5. Monetary Distribution

2.2.4 Box Plot

Fig6**.** shows the feature box plot.

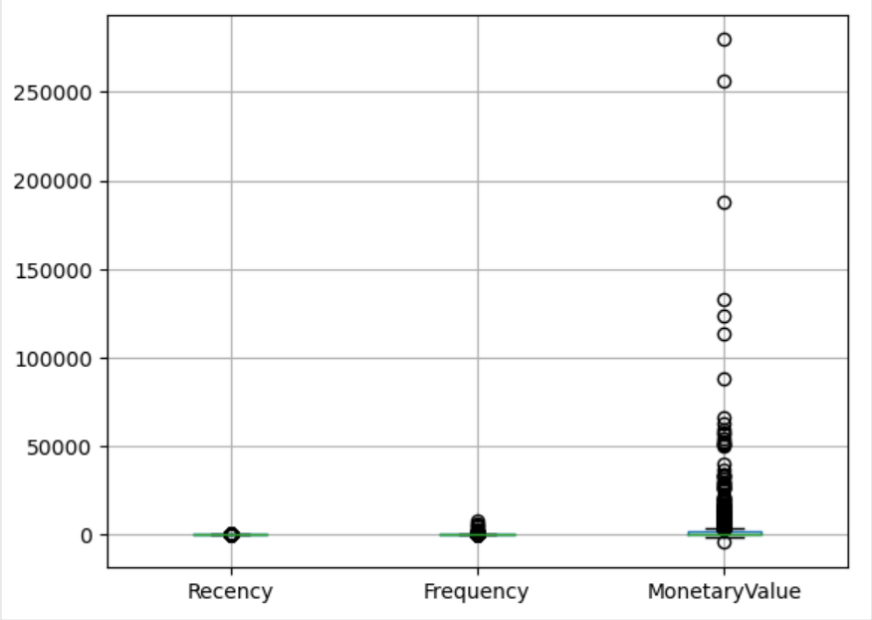

Fig6. Features Box Plot

2.3 Data Manipulation

- **Null Values:** we found 135080 records with no cluster IDs and this will lead to wrong aggregations so we dropped them.
- **Negative Values:** we found customers with a negative monetary value which may indicate a data entry error as there are 50 entries so we dropped them.

2.4 Scaling

The range of the features is different as was shown in the distribution plots above and this may cause bias in the mode so we used z-score normalization.

3. Modeling

3.1 Traditional RFM Model

3.1.1 Building RFM Model

To build an RFM model, we assigned a recency score, frequency score, and monetary score to each unique customer. The raw data, which can be collected from a customer database from previous transactions, is then compiled in a spreadsheet or database.

3.1.2 Dividing Customer Segment

We divided the RFM database into tiered groups for each of the three values of the RFM score. Tier designation is based on the greatest to the least. So, for our dataset, we grouped them into four groups. Tier four for monetary value is assigned to the high spenders and tier one is assigned to the lowest spenders as shown in Fig7. Fig8. shows the sorted datasets after converting RFM score values to integers. Also, we can see a description for it in Fig9.

	Recency	Frequency	MonetaryValue	features	ScaledFeatures	recency_score	Frequency_score	MonetaryValue_score	RFM_Score
0	326	2	0.00	[326.0, 2.0, 0.0]	[3.2886410581769927, 0.008567711958295188, 0.0]	1	1	1	111
1	3	182	4310.00	[3.0, 182.0, 4310.0]	[0.030263568020033674, 0.7796617882048621, 0.5...	4	4	4	444
2	76	31	1797.24	[76.0, 31.0, 1797.24]	[0.7666770565075198, 0.13279953535357542, 0.21...	2	2	4	422
3	19	73	1757.55	[19.0, 73.0, 1757.55]	[0.19166926412687996, 0.31272148647777437, 0.2...	3	3	4	433
4	311	17	334.40	[311.0, 17.0, 334.4]	[3.1373232180768245, 0.0728255516455091, 0.040...	1	1	2	211

Fig7. RFM scores for each customer

	Recency	Frequency	MonetaryValue	features	ScaledFeatures	recency_score	Frequency_score	MonetaryValue_score	RFM_Score
694	3	224	4404.44	[3.0, 224.0, 4404.44]	[0.030263568020033674, 0.9595837393290612, 0.5...	4	4	4	444
1752	10	216	9451.54	[10.0, 216.0, 9451.54]	[0.10087856006677892, 0.9253128914958804, 1.14...	4	4	4	444
1770	5	102	1625.97	[5.0, 102.0, 1625.97]	[0.05043928003338946, 0.4369533098730546, 0.19...	4	4	4	444
2221	9	137	2213.11	[9.0, 137.0, 2213.11]	[0.09079070406010102, 0.5868882691432205, 0.26...	4	4	4	444
4064	5	696	4204.10	[5.0, 696.0, 4204.1]	[0.05043928003338946, 2.9815637614867256, 0.50...	4	4	4	444

Fig8. Sorted RFM scores

	Recency	Frequency	MonetaryValue	recency_score	Frequency_score	MonetaryValue_score	RFM_Score
count	4330.000000	4330.000000	4330.000000	4330.000000	4330.000000	4330.000000	4330.000000
mean	90.880139	93.902309	1919.929551	2.487529	2.511778	2.504388	278.044111
std	99.129091	233.434552	8255.902202	1.110007	1.118308	1.118568	120.949182
min	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	111.000000
25%	17.000000	18.000000	300.947500	1.000000	2.000000	2.000000	211.000000
50%	50.000000	42.000000	656.655000	2.000000	3.000000	3.000000	311.000000
75%	139.000000	102.000000	1624.170000	3.000000	4.000000	4.000000	411.750000
max	374.000000	7983.000000	279489.020000	4.000000	4.000000	4.000000	444.000000

Fig9. Data frame description

3.1.3 Selecting Targeted Customer Groups

The third step involves the selection of the segmented customer group with high customer value. We assigned titles to segments of interest, which are Champion, Loyal, Potential, and Churn as shown in Fig10.

Frequency	MonetaryValue	features	ScaledFeatures	recency_score	Frequency_score	MonetaryValue_score	RFM_Score	RFM_Segment_new
224	4404.44	[3.0, 224.0, 4404.44]	[0.030263568020033674, 0.9595837393290612, 0.5...	4	4	4	444	Champion
216	9451.54	[10.0, 216.0, 9451.54]	[0.10087856006677892, 0.9253128914958804, 1.14...	4	4	4	444	Champion
102	1625.97	[5.0, 102.0, 1625.97]	[0.05043928003338946, 0.4369533098730546, 0.19...	4	4	4	444	Champion
137	2213.11	[9.0, 137.0, 2213.11]	[0.09079070406010102, 0.5868882691432205, 0.26...	4	4	4	444	Champion
696	4204.10	[5.0, 696.0, 4204.1]	[0.05043928003338946, 2.9815637614867256, 0.50...	4	4	4	444	Champion
...	...	...	...	...	...	...	...	...
11	207.74	[141.0, 11.0, 207.73999999999998]	[1.4223876969415827, 0.04712241577062354, 0.02...	1	1	1	111	Churn
5	120.00	[142.0, 5.0, 120.0]	[1.4324755529482607, 0.02141927989573797, 0.01...	1	1	1	111	Churn
16	280.55	[235.0, 16.0, 280.55]	[2.3706461615693044, 0.06854169566636151, 0.03...	1	1	1	111	Churn

Fig10. Groups labels

3.2 Machine Learning Model

3.2.1 Choosing the model

There are many machine learning algorithms, each suitable for a specific type of problem. Our goal is to assign each customer to one group, then the K Means algorithm will be a perfect choice**.** K Means is a popular method of unsupervised machine learning methods that finds different clusters and group them together so we end up with the most possible customer segments to interpret.

3.2.2 Finding optimal numbers of clusters

Having the optimal number of clusters increases the reliability of the analysis. However, the difficulty when running the K Means clustering arises when choosing the optimal number of clusters, the algorithm might converge when given way too many clusters as well, but that just would not make sense. There are some methods that help estimate the needed number of clusters. We used the average silhouette method to find the number of clusters in our project.

The average silhouette method is a measure of how well each data point fits its corresponding cluster. This method evaluates the quality of clustering. As a general rule, a high average silhouette width denotes better clustering output.

We plotted the values of K on the x-axis against the corresponding values of the silhouette scores on the y-axis. Whenever the silhouette score is larger the K value will be better. As the Fig11. shows the best K value to go with is 3 with a silhouette score of 0.7775112012960033.

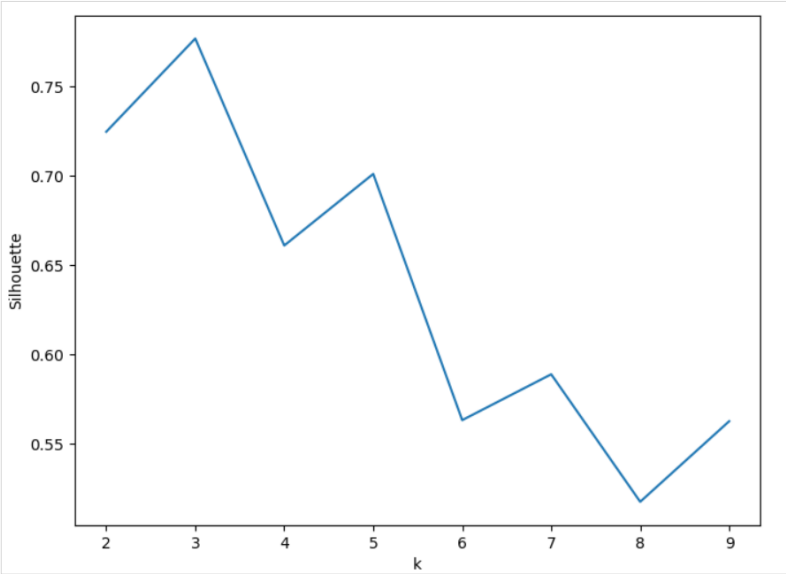

Fig11. K values against silhouette scores

Visualization of clusters of data points is very important. Various edges of the graph provide a quick view of the complex input data set so we visualized a plot for customer segmentation clusters in three dimensions as shown in Fig12.

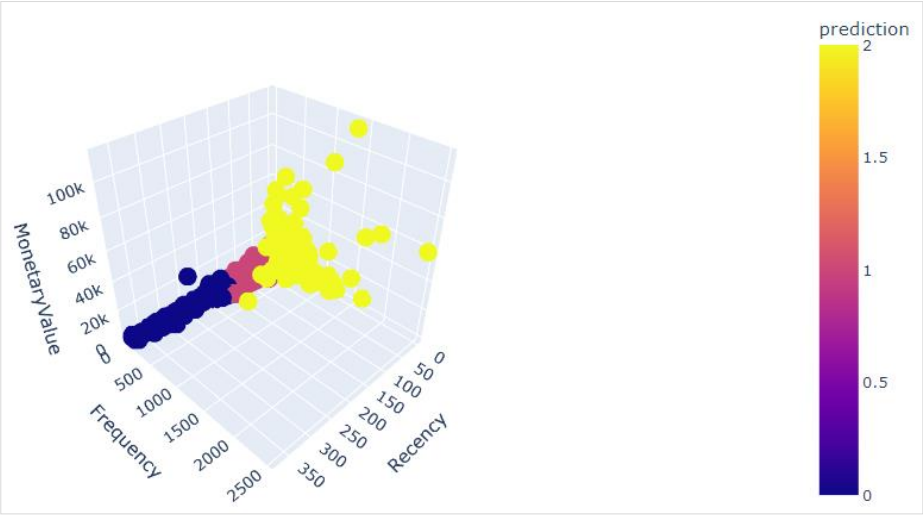

Fig12. Customer segmentation clusters in 3D

3.2.3 Training the model

After we found the K value, we trained the k means cluster model with the optimal number of clusters we found. Then we predicted to which cluster each customer belongs, and counted the number of customers in each group. As Fig13. shows the customers got grouped into 0, 1, and 2 clusters based on their similar features.

Recency	Frequency	MonetaryValue	features	ScaledFeatures	prediction
326	2	0.0	[326.0,2.0,0.0]	[3.28864105817699...	0
3	182	4310.0	[3.0,182.0,4310.0]	[0.03026356802003...	1
76	31	1797.24	[76.0,31.0,1797.24]	[0.76667705650751...	1
19	73	1757.55	[19.0,73.0,1757.55]	[0.19166926412687...	1
311	17	334.4	[311.0,17.0,334.4]	[3.13732321807682...	0
37	95	1545.41	[37.0,95.0,1545.41]	[0.37325067224708...	1
205	4	89.0	[205.0,4.0,89.0]	[2.06801048136896...	0
233	58	1079.4	[233.0,58.0,1079.4]	[2.35047044955594...	0
215	13	459.4	[215.0,13.0,459.4]	[2.16888904143574...	0
23	59	2811.43	[23.0,59.0,2811.43]	[0.23202068815359...	1
34	131	6207.67	[34.0,131.0,6207.67]	[0.34298710422704...	1
2	19	1168.06	[2.0,19.0,1168.06]	[0.02017571201335...	1
8	254	6245.53	[8.0,254.0,6245.53]	[0.08070284805342...	1
53	129	2662.06	[53.0,129.0,2662.06]	[0.53465636835392...	1
288	10	189.9	[288.0,10.0,189.9]	[2.90530252992323...	0
4	274	5154.58	[4.0,274.0,5154.58]	[0.04035142402671...	1
110	23	552.0	[110.0,23.0,552.0]	[1.10966416073456...	1
8	85	1313.1	[8.0,85.0,1313.1]	[0.08070284805342...	1
292	23	320.69	[292.0,23.0,320.69]	[2.94565395394994...	0
5	11	168.9	[5.0,11.0,168.9]	[0.05043928003338...	1

Fig13. Prediction Column

Fig14, 15, and 16 show plots of features against each other

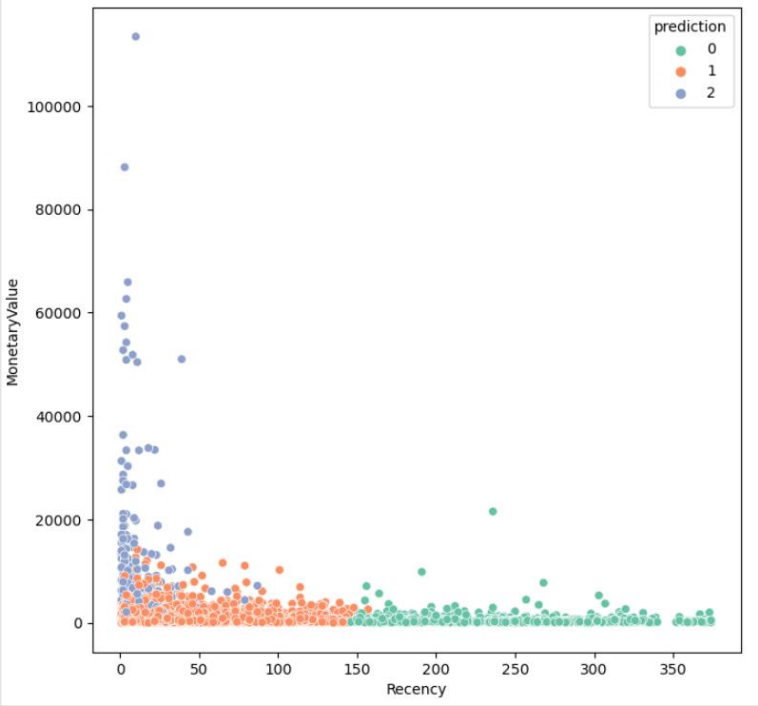

Fig14. Recency against Monetary value

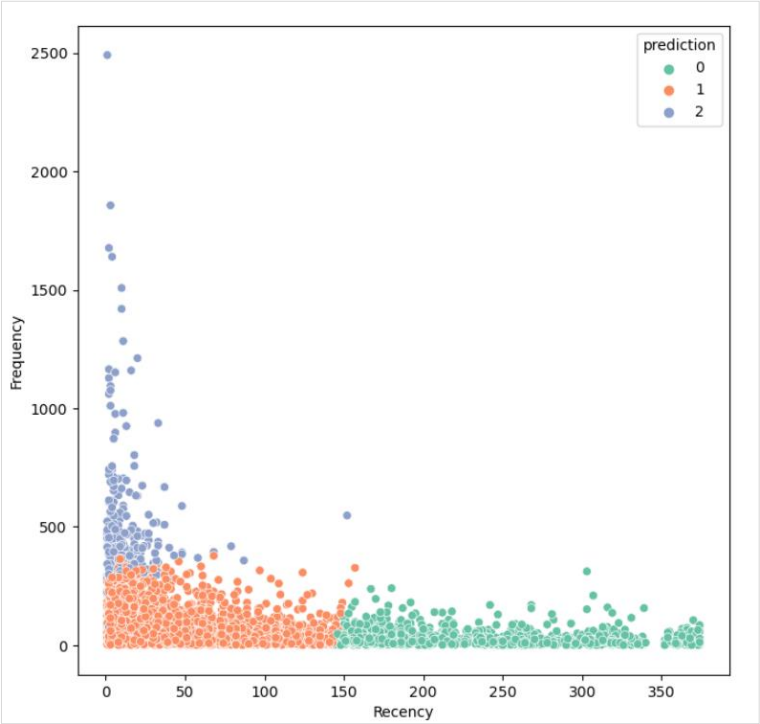

Fig15. Recency against Frequency

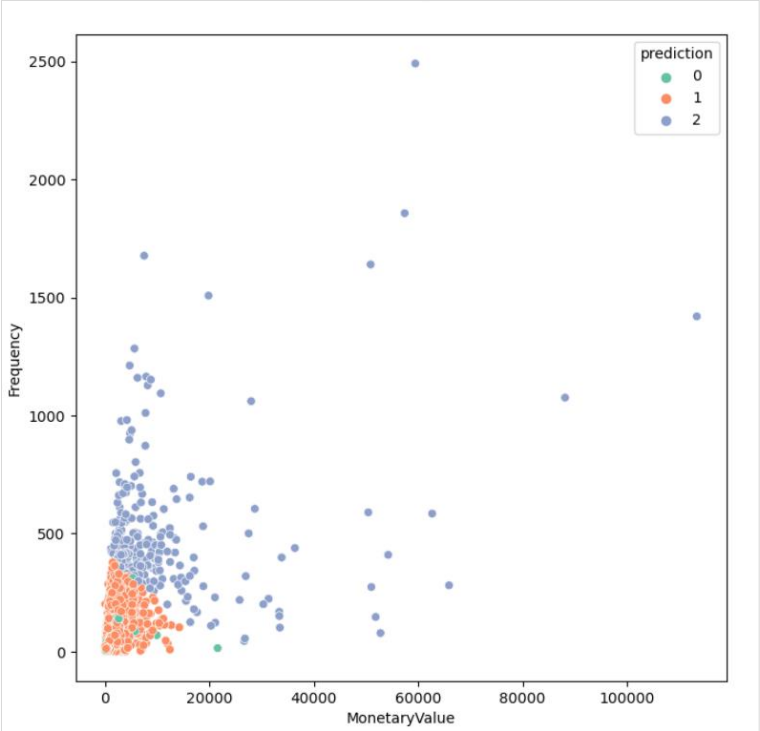

Fig16. Monetary value against Frequency

After prediction, we needed to understand the nature of each cluster and the range of the features by calculating the mean for the three features in each class. Fig17. shows the average, centroids of each feature sorted descending by MonetaryValue and Frequency and sorted ascending by recency, respectively.

prediction	Recency	Frequency	MonetaryValue	group_index
2	12.848101	498.987342	11410.694979	2
1	43.458731	75.163762	1308.157929	1
0	247.611697	28.455417	491.751286	0

Fig17. Mean of each cluster

3.2.4 Segmentation

We assigned titles to segments of interest, which are Champion, Loyal, Potential, and Chun based on their features. As we can see from Fig18. The Champion segment has the highest monetary value and frequency and the lowest recency. While the Potential segment has the highest recency and the lowest monetary value and frequency.

prediction	Recency	Frequency	MonetaryValue	group_index	K-means Segmentation
2	12.848101	498.987342	11410.694979	2	Champion
1	43.458731	75.163762	1308.157929	1	Loyal
0	247.611697	28.455417	491.751286	0	Potential

Fig18. Segments Labels

4. Results

4.1 Results of RFM model

By using the traditional RFM model we got four clusters. As we can see in Fig19. the number of customers in each group is 1083 in Champion, 1069 in Loyal, 998 in potential, and 1180 in Churn group.

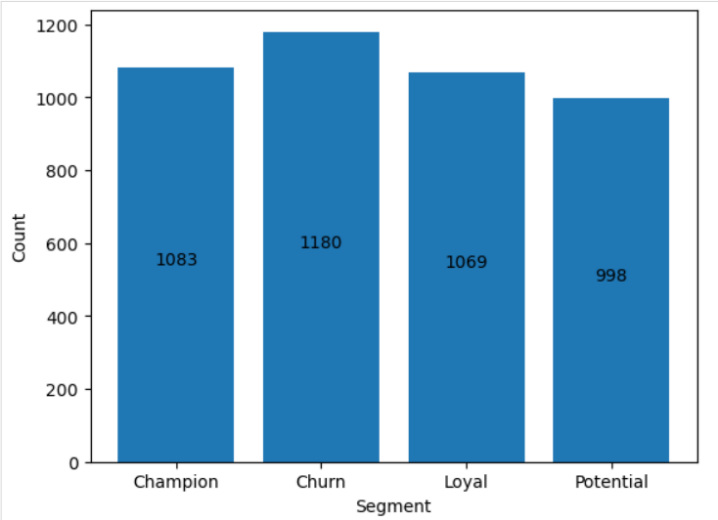

Fig19. Number of customers in each segment in RFM model

Fig20. shows the average of the features in each cluster and we can notice that there is a very small difference between Churn and potential clusters which means we could group them. As well as, there is a huge difference between the average of the monetary value of Champions cluster and the next cluster which is the Loyal.

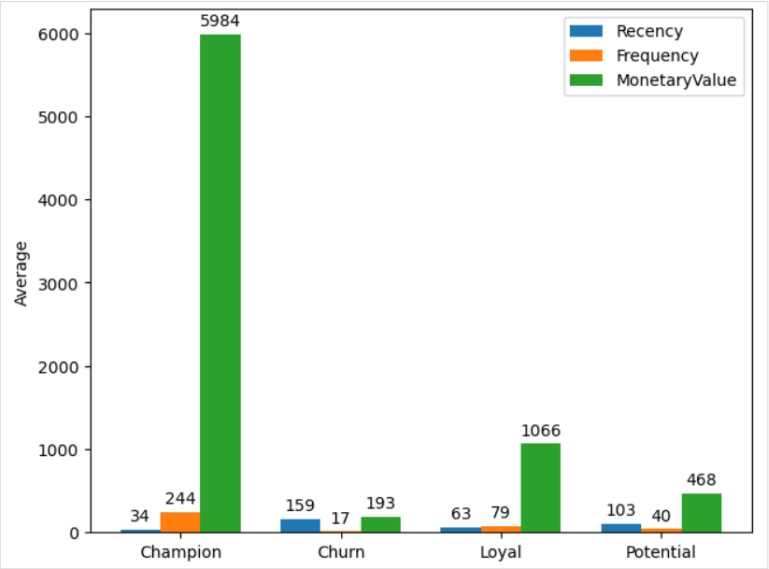

Fig20. Mean of the features in each segment using the RFM model

4.2 Results of K-mean model

By using K-means model we got three different clusters. As shown in Fig21. the number of customers in each group is 237 in Champion, 3040 in Loyal, and 1043 in potential. Fig22. shows the average of the features in each cluster.

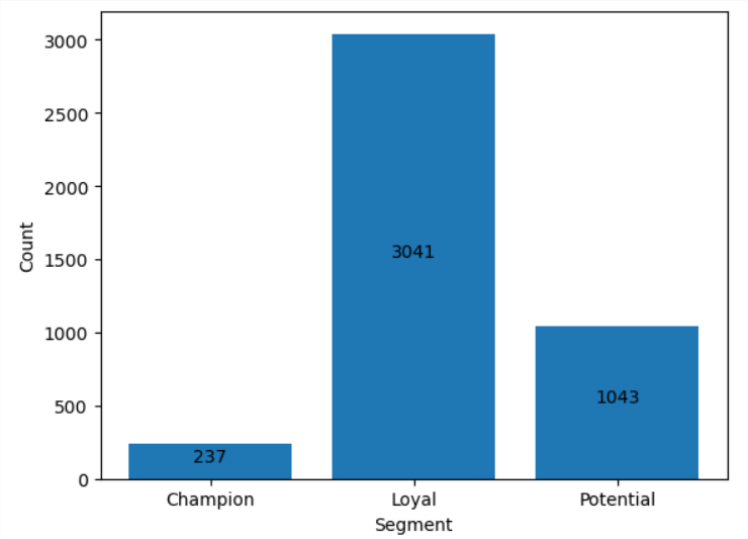

Fig21. Number of customers in each segment in k-means model

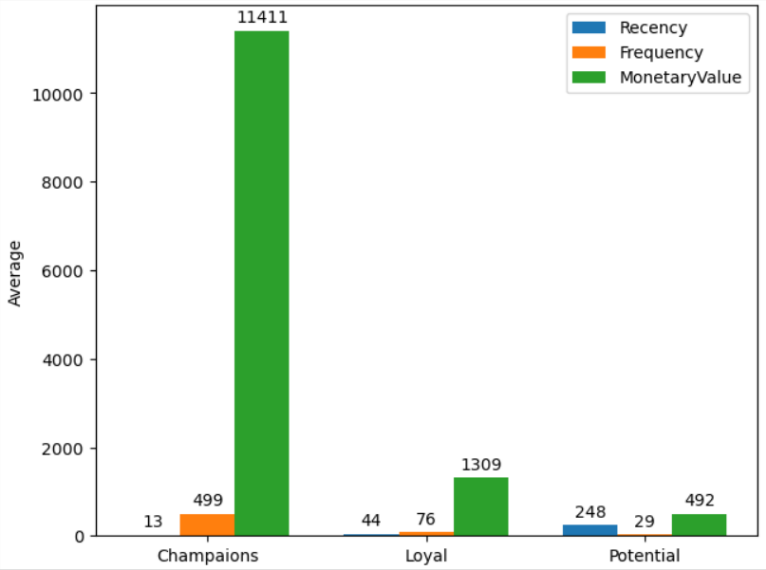

Fig22. Mean of the features in each segment using the k-means model

Using the K-means model and selecting the number clusters by methods such as silhouette analysis is faster than the traditional model. Moreover, the deviations between the clusters are more reasonable, and the overall results are in favor of the K-means model, so we went with it.

5. Conclusion

It’s not wise to serve all customers with the same product model, email, text message campaign, or ad. Customers have different needs. A one-size-for-all approach to business will generally result in less engagement, lower-click through rates, and ultimately fewer sales. Customer segmentation is the cure for this problem.

Finding an optimal number of unique customer groups will help us understand how our customers differ and help us give them exactly what they want. Customer segmentation improves customer experience and boosts company revenue. That’s why segmentation is a must if we want to surpass our competitors and get more customers. Doing it with machine learning is definitely the right way to go.

Workload

<table>
<tr><th colspan="2">Rule</th><th>Name</th></tr>
<tr><td colspan="2">Data Collection</td><td rowspan="2">Mohamed Malek</td></tr>
<tr><td rowspan="3">Data Preprocessing</td><td>Feature Selection and Extraction</td></tr>
<tr><td>Feature Distribution Visualization</td><td rowspan="2">Mohamed Adel</td></tr>
<tr><td>Data Manipulating and scaling</td></tr>
<tr><td colspan="2">Modeling</td><td>Mustafa Shalata</td></tr>
<tr><td colspan="2">Reporting</td><td>Manar Salah</td></tr>
<tr><td colspan="2">Handling Errors</td><td>Mohamed Adel</td></tr>
</table>