Customer Segmentation

A Project Report

Presented to

The Faculty of the College of Engineering

San Jose State University

CMPE-255 Section 49

Data Mining

Submitted by

Team Data Marvels

Priyanka Math	015240134
Jithesh Kurungote Balakrishnan	014654679
Priti Sharma	014561274
Tharun Mukka	014640496

Instructor: Vijay Eranti

ABSTRACT

Data mining can play an important role in marketing. Data mining techniques cannot substitute the significant role of domain experts and their business knowledge. Yet many small online retailers and new entrants to the online retail sector are keen to practice data mining for consumer-centric marketing in their business. This helps provide more customized, personal service addressing individual customer's needs, instead of mass marketing. Altogether the combination of business domain expertise with the power of data mining techniques can help organizations gain a competitive advantage in their efforts to optimize customer management.

Among the various Data mining techniques available, this project showcases the RFM methodology for customer segmentation. This study will use both classic and advanced machine learning methodologies to investigate Customer Segmentation. Through Customer Segmentation companies can identify the various segments of customers allowing them to target the potential user base. Based on the Recency, Frequency, and Monetary model, customers of the ecommerce business have been segmented into various meaningful groups using the k -means clustering algorithm. This Data mining Clustering algorithm is the most commonly used to segment data sets according to their similarities.

INTRODUCTION

Customers are the most important property of an organization. There cannot be any business prospects without satisfied customers who remain loyal and develop their relationship with the organization. That is why an organization should employ a certain strategy for treating customers.

The main goal of every industry is to understand each customer individually and use that to make it easier for the customer to do business with them rather than with competitors.

Customer segmentation is the process of dividing customers into groups or segments with respect to common characteristics. This helps target each group of customers to improve their contribution to the business and even provide them with a better customer experience. Businesses use this to target customers who are less in contribution to their businesses and improve their contribution using offers or discounts.

RELATED WORK

Data mining can play a significant role in customer segmentation and help retailers in identifying the right customers to be contacted. Methodological framework plays an important role for the successful implementation of any Data mining projects. Following are the CRISP-DM process model steps that we followed for the project.

BUSINESS UNDERSTANDING

The data mining project should start with an understanding of the business objective and an assessment of the current situation and also problems. The project‟s parameters should be considered, including resources and limitations. The business objective should be translated into a data mining goal. Success criteria should be defined and a project plan should be developed.

DATA UNDERSTANDING

For this project, we are going to use a publicly available online retail transaction dataset from Kaggle, which includes the transaction information of each customer from all over the world. It includes information such as invoice number, invoice date, customer id, description of the product, purchased quantity, and country where the customer lives.

DATA PREPARATION

In order to conduct the required RFM model-based clustering analysis, the original dataset needs to be pre-processed. Following are the main steps and relevant tasks involved in the data preprocessing.

1. Select appropriate variables of interest from the given dataset. In our case the following variables have been chosen: InvoiceNo, StockCode , Description, Quantity , Price , CustomerID, InvoiceDate and Country
2. Impute missing values and duplicate records from the dataset.
3. Filter out any negative values from the data.
4. The dataset contains more customers in the UK region. So we will be analyzing data only for the county “UK”.
5. Calculate an aggregated variable “Total Price”, by multiplying Quantity with Unit Price. Which gives the total amount of money spent per product/item in each transaction.
6. Filter out any transaction that is not associated with the count “United Kingdom”.
7. Sort out the dataset by Country and create three essential aggregated variables Recency, Frequency and Monetary. We will calculate the values of these variables for the country United kingdom.

METHODS

We used RFM technique and K-mean Clustering as modeling methods for the analysis.

RFM stands for recency, frequency, and monetary, and this is a highly flexible managerial customer segmentation model. With the prepared dataset we intended to identify whether customers can be segmented meaningfully with respect to recency, frequency, and monetary values. Also we employed the K-mean clustering algorithm for this purpose.

EXPERIMENTS AND RESULTS

Raw data before processing

```
customer_data.head()
```

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

After preprocessing United Kingdom Customer data for a specific customerId

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

RFM Analysis

Converting the Date column to DateTime before performing RFM

	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

Added the total price column

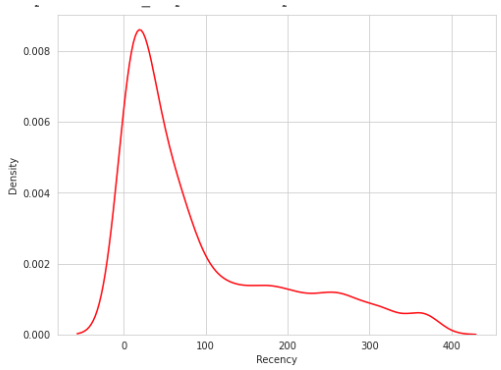
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS 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

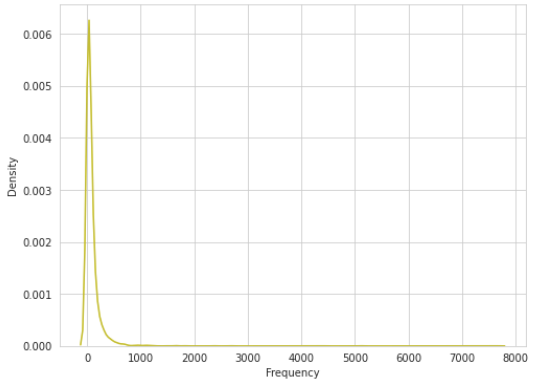
RFM data

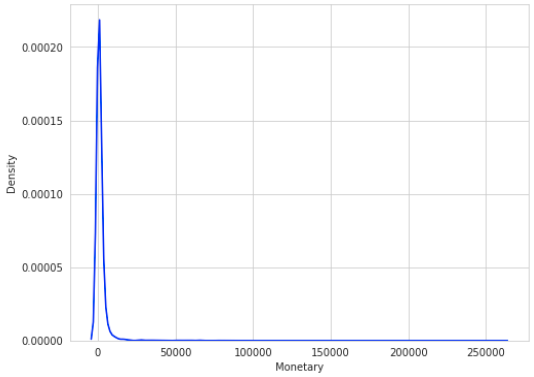
	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12747.0	2	103	4196.01
2	12748.0	0	4413	33053.19
3	12749.0	3	199	4090.88
4	12820.0	3	59	942.34
5	12821.0	214	6	92.72
6	12822.0	70	46	948.88
7	12823.0	74	5	1759.50
8	12824.0	59	25	397.12
9	12826.0	2	91	1474.72

Distribution plot for Recency, Frequency and Monetary

These plots show that the values are skewed to the right. Therefore we will have to normalize data before the K-mean algorithm.







RFM score for the Customers

	CustomerID	Recency	Frequency	Monetary	R	F	M
0	12346.0	325	1	77183.60	4	4	1
1	12747.0	2	103	4196.01	1	1	1
2	12748.0	0	4413	33053.19	1	1	1
3	12749.0	3	199	4090.88	1	1	1
4	12820.0	3	59	942.34	1	2	2
5	12821.0	214	6	92.72	4	4	4
6	12822.0	70	46	948.88	3	2	2
7	12823.0	74	5	1759.50	3	4	1
8	12824.0	59	25	397.12	3	3	3
9	12826.0	2	91	1474.72	1	2	2

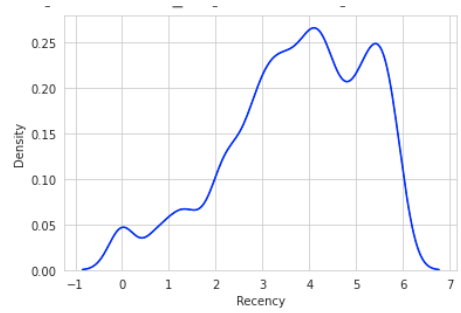
RFM Customer Segmentation showing RFM score and Four Membership Group

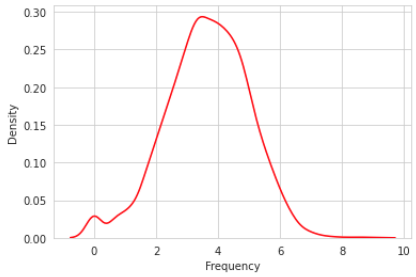
	CustomerID	Recency	Frequency	Monetary	R	F	M	Group	Score	Membership
0	12346.0	325	1	77183.60	4	4	1	441	9	Silver
1	12747.0	2	103	4196.01	1	1	1	111	3	Platinum
2	12748.0	0	4413	33053.19	1	1	1	111	3	Platinum
3	12749.0	3	199	4090.88	1	1	1	111	3	Platinum
4	12820.0	3	59	942.34	1	2	2	122	5	Platinum
5	12821.0	214	6	92.72	4	4	4	444	12	Bronze
6	12822.0	70	46	948.88	3	2	2	322	7	Gold
7	12823.0	74	5	1759.50	3	4	1	341	8	Gold
8	12824.0	59	25	397.12	3	3	3	333	9	Silver
9	12826.0	2	91	1474.72	1	2	2	122	5	Platinum

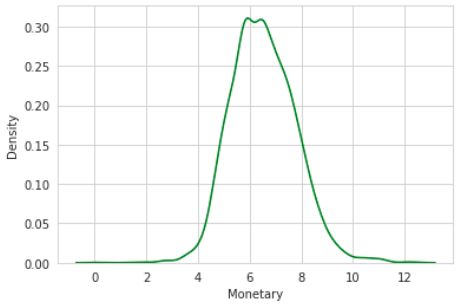
K-Mean Clustering

Applied normalization technique to scale our data using log transformation.

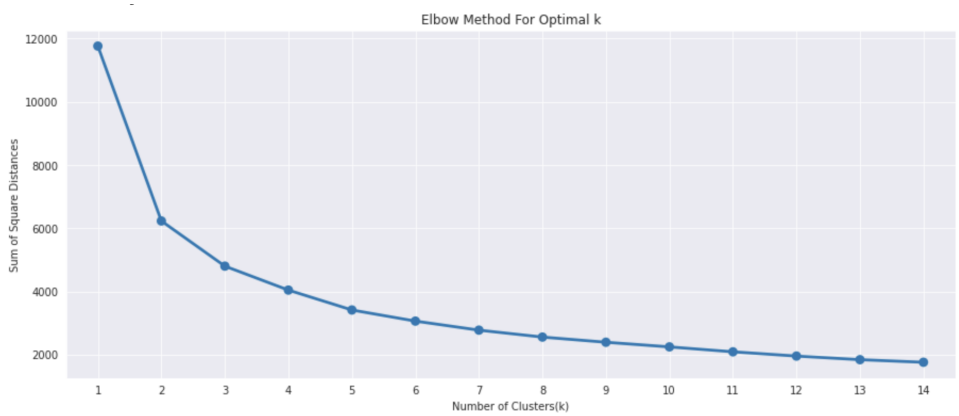
Distribution plot for Recency, Frequency and Monetary on the scaled data







Elbow method to find the value of k



Result from K-means clustering

	CustomerID	Recency	Frequency	Monetary	R	F	M	Group	Score	Membership	Cluster
0	12346.0	325	1	77183.60	4	4	1	441	9	Silver	1
1	12747.0	2	103	4196.01	1	1	1	111	3	Platinum	2
2	12748.0	0	4413	33053.19	1	1	1	111	3	Platinum	2
3	12749.0	3	199	4090.88	1	1	1	111	3	Platinum	2
4	12820.0	3	59	942.34	1	2	2	122	5	Platinum	2

As a result, K-Mean segmented customers into loyal, promising and slipping customer groups.

Higher the RFM score means the most loyal customer.

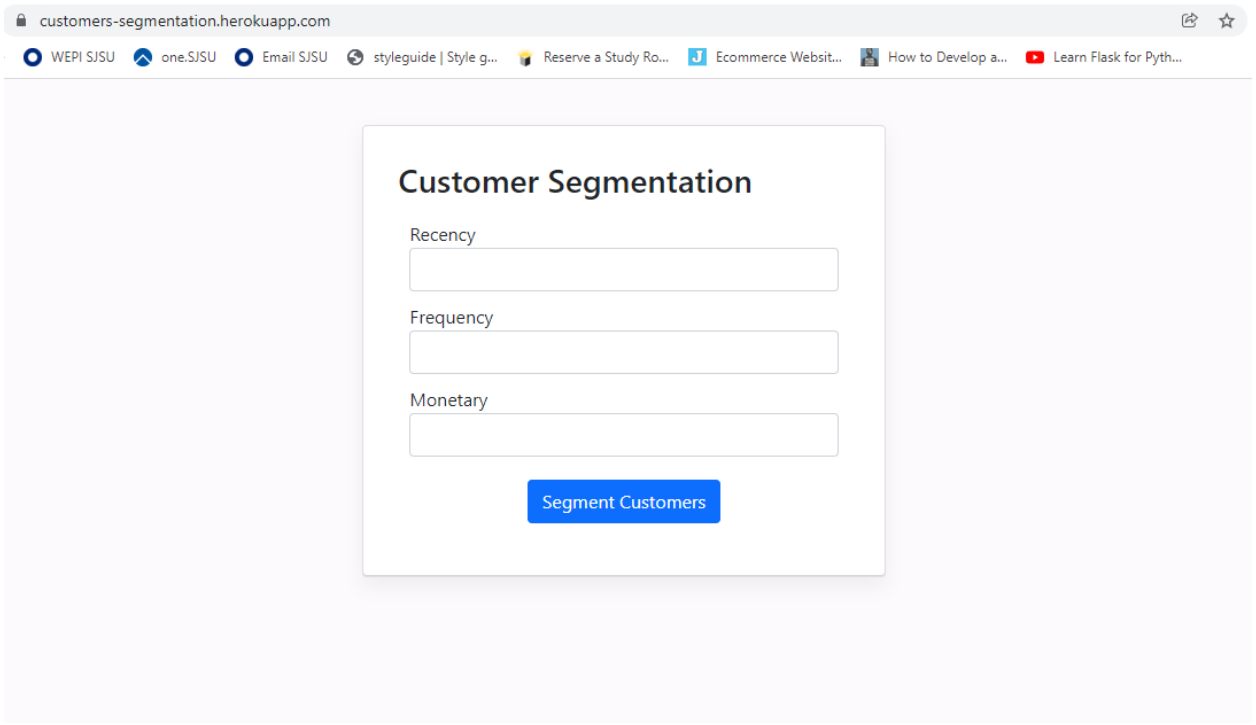
	CustomerID	Recency	Frequency	Monetary	R	F	M	Group	Score	Membership	Cluster	Segments
0	12346.0	325	1	77183.60	4	4	1	441	9	Silver	1	slipping_Customers
1	12747.0	2	103	4196.01	1	1	1	111	3	Platinum	2	Loyal_Customers
2	12748.0	0	4413	33053.19	1	1	1	111	3	Platinum	2	Loyal_Customers
3	12749.0	3	199	4090.88	1	1	1	111	3	Platinum	2	Loyal_Customers
4	12820.0	3	59	942.34	1	2	2	122	5	Platinum	2	Loyal_Customers

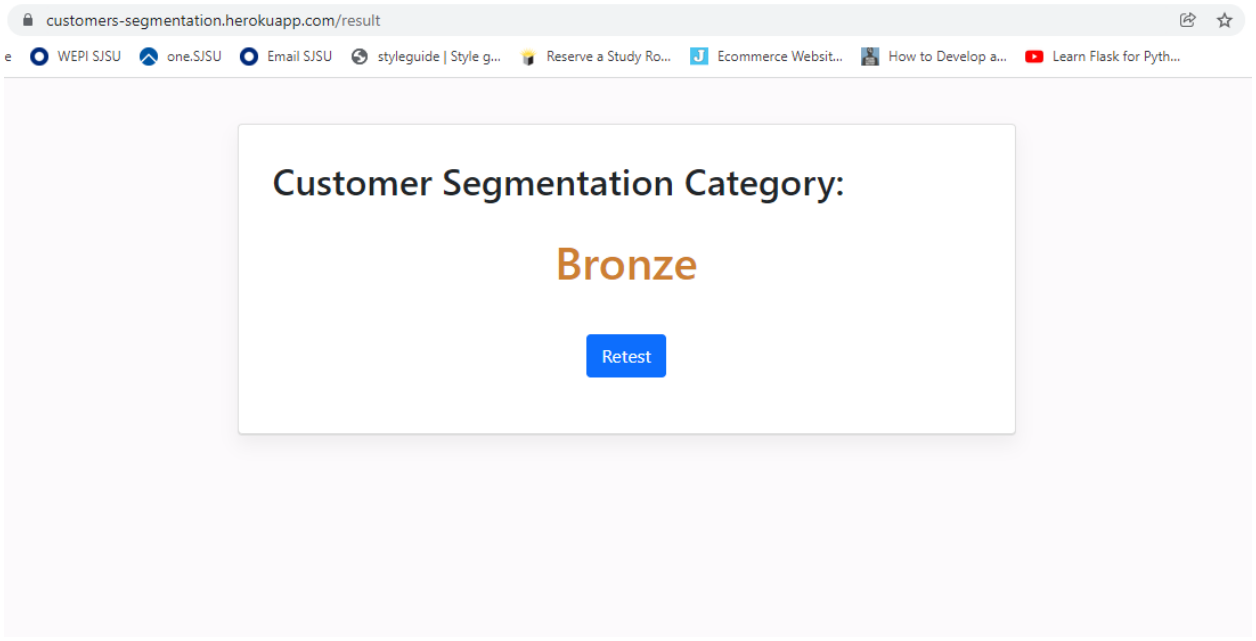
Pickle and Load

Prepared a model to load the best model for future reference.

DEPLOYMENT

Built and deployed a Flask Application on Heroku.





SUPPLEMENTARY MATERIALS

Colab Link:

https://github.com/jithesh9539/Customer_Segmentation_Final/blob/main/Customer_Segmentation_CMPE255.ipynb

Source Code:

https://github.com/jithesh9539/Customer_Segmentation_Final

Presentation slides:

https://docs.google.com/presentation/d/1I5YKlHeuYH2PUvzDs9w-dFgN_ZmU2pOK/edit?usp=sharing&ouid=104997974892965810457&rtpof=true&sd=true

Project Demo (Presentation Video):

https://drive.google.com/file/d/1XfkgutaQSQeq-omNQfEWxmKSAEdUDtKI/view?usp=sharing

REFERENCES

[1] Derya Birant. Data Mining Using RFM Analysis.

[2] RFM analysis for Customer Segmentation