Online Retail Customer Segmentation

SUNANDA DEBNATH, Ajay Tiwari,

Data science trainees,

Alma Better, Bangalore

Abstract: -

These days, you may customize everything. There's no one-size-fits-all approach. But, for business, that is in reality a fantastic thing. It creates numerous areas for wholesome competition and possibilities for businesses to get innovative approximately how they gather and hold customers.

One of the essential steps towards higher personalization is customer segmentation. This is where personalization starts, and the right segmentation will assist you're making choices concerning new features, new products, pricing, advertising, and marketing strategies allows us to better understand our customers helping us target these customers in a more efficient manner and improve the customer experience.

.

1. Problem Statement

This project aims to identify major customers Customer segmentation is the process of segments on a transnational (extending or dividing your customers into sub-groups going beyond national boundaries based on shared features. transnational corporation.) data set which contains all the transactions occurring Because you use on-site data to optimize between 01/12/2010 and 09/12/2011 for advertising off-site, segmentation happens UK-based and registered non-store online after the fact, unlike customization and retail. The company mainly sells unique all-targeting.

Occasion gifts. Many customers of the because you need to build triggers so that company are wholesalers. Your consumers see the advertisements Ecommerce customer segmentation divides when they arrive, you need to do your clients into smaller groups who share targeting and personalization before they have a common interest, making it easier to up with offers and calls to action.

Data Description:(Attribute Information)

1. **Invoice No:** Invoice number. Nominal, is a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
2. **Stock Code:** Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct
3. **Description:** Product (item) name. Nominal.
4. **Quantity:** The quantities of each product (item) per transaction. Numeric.
5. **Invoice Date:** Invoice Date and time. Numeric, the day and time when each transaction was generated.
6. **Unit Price:** Unit price. Numeric, Product price per unit in sterling.
7. **Customer ID:** Customer number. Nominal, is a 5-digit integral number uniquely assigned to each customer.
8. **Country:** Country name. Nominal, is the name of the country where each customer resides.

2. Introduction

When a customer makes a purchase, there is some information that is now stored. Such

information is given below:

In e-commerce, customer segmentation InvoiceNo, StockCode, Description, refers to the use of customer data to divide Quantity, InvoiceDate, UnitPrice, customers into groups that share the same CustomerID, Country. behavior and characteristics such as gender, taste or shopping patterns, interests, and more. Segmenting the customer base helps in better understanding the customers and thus personalizing marketing and communication for each segment. This is very beneficial because people tend to respond better and be

of greater value to your business when they

feel their needs and interests are being specifically addressed.

3. Steps involved:

- **Exploratory Data Analysis**

(EDA) is utilized by statistics scientists to investigate and inspect statistics units and summarize their major characteristics, frequently using statistics visualization methods. It enables deciding how exceptional to govern statistics reasserts to get the solutions you need, making it less complicated for statistics scientists to find out patterns, spot anomalies, take a look at a hypothesis, or take a look at an assumption

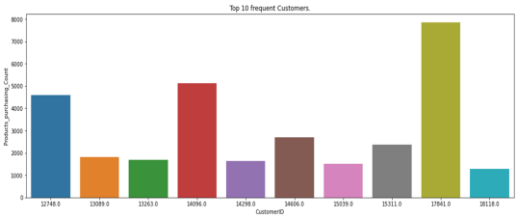
After loading the dataset we performed. This process helped us figure out various aspects and relationships among independent/ feature variables. It gave us a better idea

- **Null values Treatment**

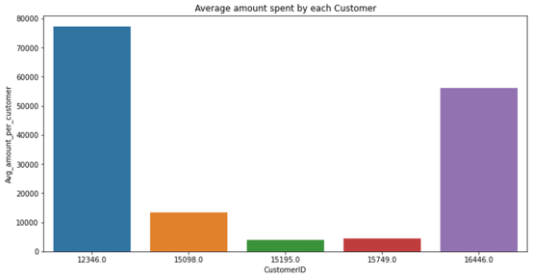
Our dataset contains some null values like customer_id, AND Description which may tend to our project so will come good at the beginning of our project in order to get a better result.

Variable	# of nulls	% of nulls
InvoiceNo	0	0
StockCode	0	0
Description	1454	0.27
Quantity	0	0
InvoiceDate	0	0
UnitPrice	0	0
CustomerID	135080	24.93
StockCode	0	0

- **Data pre-processing and transformation**



in this graph we can observe that based on purchasing count which are the top 10 most frequent customers.



The above graph represents average amount spend by each customer

- **Standardization of features**

Create the RFM model (Recency, Frequency ,Monetary value) for clustering made easy

Recency, frequency, and monetary value are marketing evaluation tools used to identify a company's or an organization's quality customers by the use of sure measures. The

RFM model is primarily based totally on 3 quantitative factors:

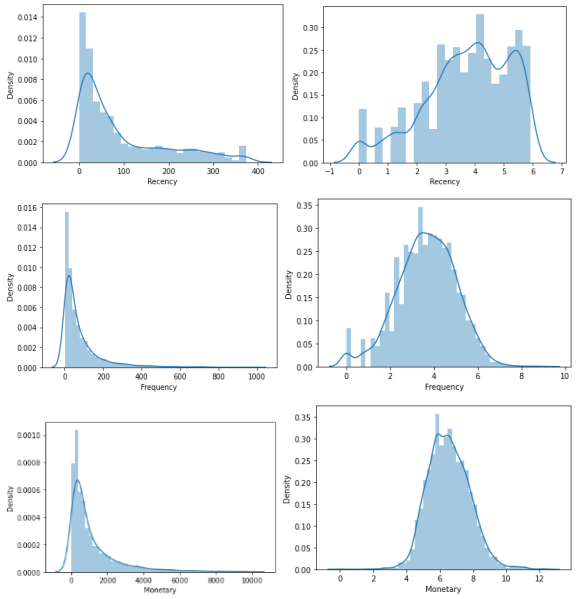
Recency: How many days before the customer had made the purchase.

Frequency: How often a customer makes a purchase.

Monetary Value: How much money a customer spends on items

Performing RFM Segmentation and RFM Analysis Step by Step

The first step in building an RFM model is to assign Recency, Frequency, and Monetary values to each customer. The second step is to divide the customer list into tiered groups for each of the three dimensions (R, F, and M).



Calculating RFM scores

The number is typically 3 or 5. If you decide to code each RFM attribute into 3 categories, you'll end up with 27 different coding combinations ranging from a high of 111 to a low of 444. Generally speaking, the lower the RFM score, the more valuable the customer.

	CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level
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	4596	33719.73	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

- **Model building, Predictions, and Forecasting**

K-Means Clustering

The *k*-means algorithm searches for a pre-determined number of clusters within an unlabeled multidimensional dataset. It accomplishes this using a simple conception of what the optimal clustering looks like:

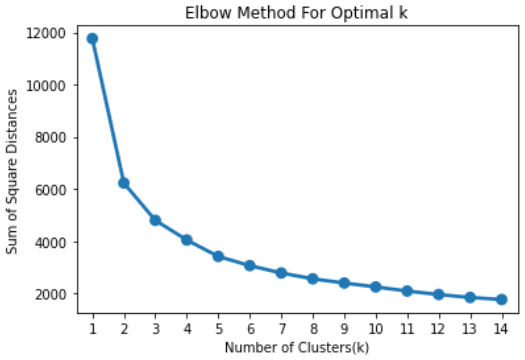
The "cluster center" is the arithmetic mean of all the points belonging to the cluster.

Each point is closer to its own cluster center than to other cluster centers.

Those two assumptions are the basis of the *k*-means model.

The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-15) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

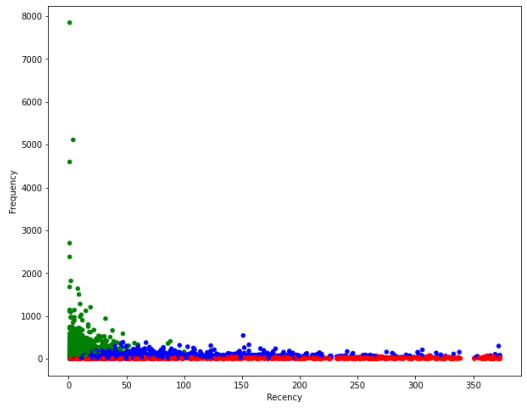
K-Means with Elbow Method and Silhouette Method on RFM




```
For n_clusters=2, the silhouette score is 0.39562494982602087
For n_clusters=3, the silhouette score is 0.30546474589014033
For n_clusters=4, the silhouette score is 0.2980445141762671
For n_clusters=5, the silhouette score is 0.2785167472954507
For n_clusters=6, the silhouette score is 0.27726461404219555
For n_clusters=7, the silhouette score is 0.26066118271648536
For n_clusters=8, the silhouette score is 0.2583990050322177
```

From the plot above, 3 looks like the optimal k for the cluster. Now let's visualize for 3 clusters

visualization of Recency, Frequency, and Monetary



CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level	Cluster
12346.0	325	1	77183.60	4	4	1	441	9	Silver	2
12747.0	2	103	4196.01	1	1	1	111	3	Platinum	1
12748.0	1	4596	33719.73	1	1	1	111	3	Platinum	1
12749.0	3	199	4090.88	1	1	1	111	3	Platinum	1
12820.0	3	59	942.34	1	2	2	122	5	Platinum	1

4. Challenges:

Identify a highly imbalanced data set and manage it carefully.
Identify the highly imbalance large dataset and manage it carefully.
Ensure model is treating all groups fairly.
Lot of NaN values.
Before deploy model set the proper number of clusters.
After deploying understand model behavior on real data.
Tough to make prediction analysis to end users.

5. Conclusion:

That's it! We have come to the end of this analysis. Throughout the evaluation, we went via diverse steps to carry out customer segmentation. We commenced with information wrangling in which we attempted to address null values, and duplicates and accomplished function adjustments. Next, we did a few exploratory information evaluations and attempted to attract observations from the capabilities we had in the dataset. Next, we formulated a few quantitative elements consisting of recency, frequency, and monetary referred to as RFM models for every one of the customers. We applied the KMeans clustering set of rules to those features. We additionally accomplished silhouette and elbow technique evaluation to decide the ideal no. of clusters turned into 3. We noticed clients having excessive recency and low frequency and financial values have been a part of one cluster and clients having low recency and excessive frequency, and financial values have been a part of any other cluster. However, there may be greater adjustments to this evaluation. One can also additionally pick out to cluster into greater no. relying on corporation targets and preferences. The classified function after clustering may be fed into classification-supervised machine learning algorithms that might are expecting the lessons for a brand new set of observations. The clustering also can be accomplished on a brand new set of features consisting of the kind of products every customer decide to shop for often, locating out client lifetime value (clv), segmenting on the premise of the term they visit, and lots greater. As gadget mastering has grown to be greater of an ART, there's not anything consisting of proper or wrong. We best try and get nice results which can in shape our very last targets. There is, and usually will be, a want to improve, going forward. we see that Customers are nicely separated whilst we cluster them with the aid of using Recency, Frequency, and Monetary and the ideal number of clusters is equal to 3

6. <u>References</u>

1. Stack Overflow!
2. GeeksforGeeks
3. Analytics Vidhya
4. Almabetter
5. GitHub
6. Towards data scienc

Conclusion:-

- Due to the Response variable's value 1 being much lower than its value 0, the provided dataset is an imbalanced dataset.
- Compared to their female counterparts, male consumers own a little bit more vehicles and have a higher likelihood to get insurance.
- Customers between the ages of 30 and 60 are the most likely to get insurance whereas Vehicle insurance is not interesting to anyone under the age of 30. The lack of involvement, a lack of knowledge about insurance, and possibly the lack of expensive vehicles are potential causes.
- Customers with driving licenses are more likely to purchase insurance
- Compared to consumers with vehicles less than one-year-old, those with vehicles between one and two years old are more interested in purchasing insurance.
- Due to their personal experience with the costs associated with vehicle repairs, customers with vehicle damage are more likely to purchase insurance.
- The variable such as Age, previously insured, and Annual premium is more affect the target variable.
- We used different types of algorithms to train our model like Logistic Regression, Random Forest model, Decision tree, and XGB Classifier. And Also, we tuned the parameters of the XGB Classifier and Random Forest model. Comparing the model on the basis of precision, recall, accuracy, and F1 score we can see that the XGB Classifier model performs better. Even comparing the ROC curve XGB Classifier performed better because curves closer to the top-left corner indicate