CAPSTONE PROJECT – **ONLINE RETAIL**

Table of Contents

1. Problem Statement
2. Project Objective
3. Data Description
4. Data Pre-processing Steps and Inspiration
5. Choosing the Algorithm for the Project
6. Motivation and Reasons For Choosing the Algorithm
7. Assumptions
8. Model Evaluation and Techniques
9. Inferences from the Same
10. Future Possibilities of the Project
11. Conclusion
12. References

Problem Statement

An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence based insights to provide the same.

1. Using the data, find useful insights about the customer purchasing history

that can be an added advantage for the online retailer.

1. Segment the customers based on their purchasing behaviour.

Project Objective

1. To find which month has the highest sales.
2. To look for correlation between Quantity, Unit price, and Total Amount.
3. To find out Busiest Month of the Year.
4. To find out Busiest Day of the Week.
5. Best customers throughout the Year.
6. Customers who aren’t enjoying our service.
7. Top selling products
8. Top buying countries
9. Most bought item for the year 2010 and 2011 for top buying countries.
10. Grouping customers based on their purchases.

Data Description

* The data consists of 8 columns. Each carrying their own characteristics.
* Following are the features in the dataset:

1. Description
2. Quantity
3. Invoice date
4. Unit Price
5. Customer ID
6. Country
7. Total Amount (Added by me)
8. MonthYear (Added by me)
9. Day (Added by me)

* Data Visualization and Insights:

1. Total purchase by month –

Highest sales has occurred in the month of November in 2011. We only have data till 9th of December 2011 which is the reason for sudden drop in the sales for the month of December.

1. Heatmap:

This shows how much of these three columns Unit Price, Total Amount, and Quantity are correlated to each other.

1. Busiest Month of the Year:

From the graph we can say that November is busiest month of the year. And reason can be showcased as on the occasion of various festivities in catholic countries.

1. Busiest Day of the Week:

Evidently we can prove that Thursday’s are the busiest day in the week.

1. Best Customers throughout the year:

Here are the customer ID’s of the best and loyal customers - 14646, 18102, 17450, 14911, 12415, 14156, 17511, 16684, 13694. We can ask these customers to share their opinions on our product and service which will help us attract more customers

1. Customers who are not satisfied by our service:

Here are the customer IDs of unsatisfied customers - 17448, 15369, 14213, 17603, 12503, 15823, 13154, 15802, 16252, 12666. The customers present in this list can help us improve our service by conducting a survey and reviewing their feedback.

1. Popular products:

Following are the most sold products Dotcom Postage, Regency Cakestand 3 Tier, White Hanging Heart T-Light Holder, Party Bunting, Jumbo Bag Red Retrospot, Rabbit Night Light, Postage, Paper Chain Kit 50's Christmas, Assort-ed Colour Bird Ornament, and Chilli Lights. This data can help us manage our inventory.

1. Top buying countries:

United Kingdom, Netherlands, EIRE, Germany, France, Australia, Switzerland, Spain, Belgium, and Sweden are the most active

Countries.

1. Most bought items in the year 2010 and 2011 in the UK:

Most bought item in the year 2010(only December) in the UK: ‘World War 2 Gliders Asstd Designs’ with 4571 pieces sold

Most bought item in the year 2011 in the UK: ‘World War 2 Gliders Asstd Designs’ with 43659 pieces sold

1. Most bought items in the year 2010 and 2011 in the Netherlands:

Most bought item in the year 2010(only December) in the Netherlands: ‘Wall Tidy Retrospot’ with 300 pieces sold

Most bought item in the year 2011 in the Netherlands: ‘Rabbit Night Light’ with 4801 pieces sold

Data Processing Steps:

1. Firstly we check data type of all the columns.
2. We found that all columns have data types as per their characteristic except for ‘Date’ column. The data type for ‘Date’ column is ‘object’.
3. Changing the data type of ‘Date’ column to ‘datetime’.
4. There are a total of 136534 missing values and 5268 duplicate values in the data.
5. Out of all the missing values 135080 are from Customer ID column. We’ll replace all the null values from Customer ID column with ‘Unknown’. And all the null values from Description column will dropped.
6. Now dropping all the duplicate values from the data.
7. Adding a column and naming it as ‘Total Amount’ which will be a product of Unit Price and Quantity columns.
8. Passing df.describe() function to check properties of numeric columns.
9. Year, MonthYear, Day columns were added for getting more information with the help of visualization tools.

KMeans Clustering

Kmeans clustering is a type of unsupervised machine learning. The method searches for data points which have similar combinations and groups them into one single cluster. Without any previous data training the machine’s job in this case is to organize unsorted data according to parallels, patterns, and variations.

KMeans Clustering can be used in the following sectors:

* Banking
* Customer segregation
* Recommendation Engine
* Cyber security
* Image segmentation

**Elbow Method**

I chose Elbow method for customer segmentation and the reasons are as shown below:

1. It is Easy to compute.
2. Doesn’t require high end RAM for computation.
3. Easily adapts new data.
4. Can form clustering in any shape and size.
5. And I didn’t have enough RAM to adopt other methods. I tried to compute with other methods as well but the system was unable to fetch results.

Assumptions

1. First assumption is that the clusters are spherical:

This assumption helps in separating the data when algorithm works on the data and forms cluster.

1. Second assumption is that the clusters are of same size:

This assumption of the size helps in deciding the boundaries of the cluster. It helps in calculating number of data points in each cluster. With this assumption, one can start with centers of cluster anywhere. This will help in keeping each cluster centers away from each other.

Model Evaluation and Technique

1. As mentioned above that we will be creating a model using Elbow method.
2. For this we need to import KMeans library for segregating data and import plotly library for visualization purpose.
3. We will create a function named ‘try\_diff\_cluster’ and inputs of the function will be ‘data’ and ‘k’. The term ‘k’ represents number clusters.
4. But before using this function we will first apply one hot encoding to ‘object’ data type. ‘Description’ and ‘Country’ columns are to be treated using above technique.
5. Now we pass the function by giving the following inputs.

data = df

k= 12

Here I have given value of k as 12 because I’ll calculate sum of squared distances for 12 iterations by increasing one cluster per iteration.

1. From the graph we can k=3 is the optimal value for clusters but for cross validation we’ll check the same by using kneed library.
2. Kneed library can help us find optimal value from the graph.
3. We just need import KneedLocator package from kneed library.

Evaluation Report:

* Here we can witness that the optimal value of k is 3. Now we’ll Predict the our data set. Here the model gives us the datapoints present in each cluster.

Inferences from Project

* We can segregate customers based on their purchase pattern.
* And with the help of this output we can recommend products to customers which are bought by other customers in the same cluster.
* This analysis can help us manage our inventory in an effective manner.
* As we saw above that the UK holds the highest sales value over all other countries. We need to look forward to make their experience even better.
* It can be said that all other nations are doing the same in terms of total sales. But it would be great if we take their feedback and get to know why their not enjoying our services, we need check any flaws.
* As we know our customers are from different countries, it becomes our responsibility to know their culture, tradition, habits, etc. so that it will help us reach their demands.

Future Possibilities

* This model can be used for different dataset as well. This will help in keeping a track on demand and supply.
* It is better to know other factors like CPI, natural or manmade calamities, scarcity, etc. which can affect overall cost of the product.

Conclusion

This project helped retrieve customers which have similar interests. The model was trained using KMeans method. There are a lot of clustering methods but due to less complexity I chose to go forward with KMeans. This method can not only help us group customers but also recommend products to customers in the same cluster.

References

* <https://medium.com/data-and-beyond/customer-segmentation-using-k-means-clustering-with-pyspark-unveiling-insights-for-business-8c729f110fab>
* <https://towardsdatascience.com/customer-segmentation-with-kmeans-e499f4ebbd3d>
* <https://neptune.ai/blog/customer-segmentation-using-machine-learning>