READ ME

Customer segmentation is considered the backbone of marketing as it allows companies to tailor their marketing message to different subsets of customers. This project aims to segment customers based on limited sales data from a UK based company.

This is Capstone 2 of Springboard's Data Science career track.

1 - Data

Online Retail II dataset was available from the UC Irvine Machine Learning Repository as an Excel file. It contains all online sales transactions for a UK based company between 01/12/2009 and 09/12/2011. The dataset includes 1,067,371 line items across 53,628 transactions and the following eight columns: 'Invoice', 'StockCode', 'Description', 'Quantity', 'InvoiceDate', 'Price', 'Customer ID', and 'Country'.
link to data: https://archive.ics.uci.edu/dataset/502/online+retail+ii

2 - Problem Statement

Identify three to seven customer segments from the transaction data to focus ongoing marketing efforts while answering the below Key Questions about segment characteristics.
Key Questions:

- How often do each segment make purchases?
- How much have they spent over time?
- How recently have they made a purchase
- What does their basket look like?
- Are there differences in seasonality of purchases between

3 - Data Wrangling

The data was split between two sheets in the downloaded Excel file. Using Jupyter Notebook, the data was combined into a single data frame. Cancelled orders and entries with negative or zero values for ‘Price’ or ‘Quantity’ were removed. ‘Sales’ were calculated and added by multiplying ‘Price’ by ‘Quantity’. Known and unknown customers were split into two data frames based on whether or not they had values in ‘Customer ID’. At the end of the initial EDA (next section), the data frame with known customers was used to calculate RFM values.

4 - Initial EDA and RFM calculations

The transaction data explored to find trends or anomalies. ‘StockCode’ and ‘Description’ didn’t provide any insights other than top products sold as there was no associated information on a product type category. ‘Price’, ‘Quantity’, and ‘Sales’ were all heavily right skewed. ‘Customer ID’ combined with ‘Sales’ showed several very large buyers but many small buyers. ‘Country’ showed that the majority of sales came from the UK. ‘InvoiceDate’ showed seasonality with a spike in the months leading up to the holiday gift giving season and a lull in sales after.

RFM values used in modeling were calculated by grouping entries by ‘Customer ID’ and calculating their respective RFM values. ‘Recency’ represents days since last

purchase using the day after the last sale in the dataset as the reference date. ‘Frequency’ represents the number of purchases made. ‘Monetary’ represented total sales. ‘Monetary’ and ‘Frequency’ were both right skewed with ‘Monetary’ having extreme values.

5 - Modeling

RFM data was preprocessed to account for the right skew and different magnitudes in the data. Sklearn’s PowereTransformer and StandardScaler were used for the preprocessing. Various combinations of features and models were explored but K-means using RFM data with k=2 was found to be best. However, k=2 is not best for marketing segmentation so the second best option was used which was K-means using RFM data with k=4. The silhouette score for k=2 was 0.44 and for k=4 was 0.38 which indicated moderate clustering strength with overlap of the clusters.

6 - Post Segmentation EDA

Post segmentation EDA confirmed the moderate clustering and overlap of clusters but was able to develop characteristics of the four clusters. ‘Recency’ values used clustering were inverted to make directionality consistent with ‘Frequency’ and ‘Monetary’ values. Cluster 0 had high ‘Recency’, ‘Frequency’ and ‘Monetary’ values. Cluster 1 had low ‘Recency’, ‘Frequency’ and ‘Monetary’ values. Cluster 2 had low ‘Recency’ but moderate ‘Frequency’ and ‘Monetary’ values. Cluster 3 had high ‘Recency’ but low ‘Frequency’ and ‘Monetary’ values.

7 - Customer Segments

A more detailed profile was given for each of the clusters and the segments were labeled as:

- Cluster 0: Loyal, Frequent Customers
- Cluster 1: One time customers
- Cluster 2: Occasional Customers / Churn Risk
- Cluster 3: Seasonal or New Customers

8 - Recommendations and follow up

It was recommended to focus on Cluster 0 because they are frequency, repeat customers who represent 70% of sales and on Cluster 3 because they provide the best opportunity for growth.

To follow up, it was recommended to explore customers who purchase year round to help mitigate sales lull after the holiday gift season. Additionally it was recommended to explore a DBSCAN model to account for the wide variability in the data and hopefully provide better clusters.

9 - Thank you

Thank you to my mentor Karthik Ramesh for his patience and guidance in this first solo project.