Retail - PGP

Course-end Project 1

DESCRIPTION

- It is a critical requirement for business to understand the value derived from a customer. RFM is a method used for analyzing customer value.
- Customer segmentation is the practice of segregating the customer base into groups of individuals based on some common characteristics such as age, gender, interests, and spending habits.
- Perform customer segmentation using RFM analysis. The resulting segments can be ordered from most valuable (highest recency, frequency, and value) to least valuable (lowest recency, frequency, and value).

Steps Undertaken to Complete the Tasks:

1. Prior to performing any Data Cleaning, it is important to know how big the data we need to explore and handle, and what type of data was provided, and it was found out that:
    a. It is composed of over 541,909 rows and 8 Columns, and
    b. 4 Columns (Attributes) are object, 1 Column is datetime, and 3 integers.
2. Data Cleaning
    a. Two columns contain nulls values. These columns are categorical (Description) and unique identifier(Customer ID). Hence, these null values cannot be replaced or imputed with mean, median or mode.
    b. Recommended approach is to check if there are corresponding Invoice numbers with the Customer ID null values, after checking:
        - None of the elements in the 'CustomerID' column of df 'retail_df' match any customer ID in the same column.
        - No element in the 'InvoiceNo' column match any value in the df 'custid.null_treat'
        - No corresponding Invoice number contain the CustomerID null values.
    c. Decided to retain the column Description, with 0.27% null values, it is not significant to remove.
3. The dataset has a total of 5227 duplicate values which need to be removed, therefore, we are left with 541909 sample, with 8 Features.
4. Data Transformation
    a. After creating the cohorts on a “Month wise basis”, we have the greatest number of Active customers from 2011-09 until 2011-11, it will be observed based on the 12-month data, that there was a considerable decrease in the number of customers on 2011-12 compared to the previous year 2010-12, a 27.74% drop.
    b. The number of unique products purchased by active customers based on column Description are comparable within the range of 2300-2900, it will be best to streamline the product description based on a broader category example” product line” or product department to investigate further on the decline on the affected cohorts.
    c. In addition, based on the customer retention rate comparison per cohort, the biggest drop in customer retention was on 2011-12 at -59.91%
5. Building RFM Segments
    a. Using the RFM Analysis Technique, we designated the variables(columns) as follows: InvoiceDate: Recency

InvoiceNo.: Frequency

TotalPrice: Monetary Value

b. After calculating the RFM metrics, and creating the distribution plots, all metrics were observed as positively skewed.
c. Based on the concatenated or merged RFM scores, we created RFM segments as follows:
    - High-Value Active = VIP (>10) : Highly engaged customers who have purchased most recently, most often and generated the most revenue
    - High-Value Lapsed = Loyal(8-10) : Customers who have generated high revenue, but may have recently purchased less and less frequent
    - Low-Value Active = Potential(7) : Customers who do not generate high revenue, but who purchases often, and frequent
    - Low-Value Lapsed = Need Attention(4-6): Customers who do not generate high revenue, but who have not spent much in the past and have not purchased anything recently
    - Require Activation= (0-3): Inactive, customers who generate low revenue, and have not purchased recently and for a long time.
    - The number of customers falling on each segment were plotted, and the Top 2 Customer Types are:
        - Need Attention-approx. 1400 customers
        - Loyal- approx. 1350 customers
d. Based on the RFM_score range, computed the value counts under Recency, Frequency and Monetary Value aggregated as mean and median values to see if they can be really treated as 'Need Attention, as ' Low-value lapsed' customers:
    - Customers classified as 'Need Attention' have not ordered or spent much in the past (non-Frequent buyer), and not within the past 3 months, in terms of monetary value, have purchased at around 50%-75% less than Loyal Customers

6. Data Modeling
    a. Outliers are predominant in Frequency and Monetary Value, outliers were dropped by creating a new cluster dataframe, segregating data based on thresholds. Positive Skewness is also obvious for both metrics or attributes; therefore, data transformation is necessary
    b. It was important to standardize the features thru Standardscaler prior to performing a distance-based K-Means Algorithms, to avoid bias in the clustering.
    c. Prior to clustering, it is important to first check if there are null values in the scaled dataframe.
        - There were 41 null values under Monetary Value which were imputed using the median value.
    d. Attempted to run K-Means using n-cluster=4, using the Elbow curve plotting and silhouette scores to determine the optimal number clusters:
        - Although the inertia decreases as the number of clusters increases via the Elbow Method's Elbow curve plotted, the Elbow bend was not defined or clear to gauge whether that cluster, in this case, 4 clusters is optimal.
        - Performed silhouette score Method computation, to avoid choosing a cluster which is not optimal and producing overfitting results, this yielded a score of 0.39 at 2 clusters, decided to use 2 as the optimal cluster for the model.

e. Checked for outliers in the scaled data frame after the first K-Means clustering, the Outliers, though were minimal, were still showing from the boxplot charts created, attempted to perform the Clip or Winsorize Method to cap extreme values to predefined thresholds.
f. After winsorsizing the scaled data, outliers were lesser for all Features, data points are better distributed.
g. Re-clustered, re-fitted and recomputed the silhouette score, this yielded a score of 0.46 at 2 clusters.
h. After creating Labels, 0 and 1 representing the clusters as defined below:
   0 = customer who have purchased less or not recently, and not frequently and with low monetary value, wherein customer engagement is required
   1 = customers who have purchased recently, and frequently and with high monetary value for the period of 1 year
   We have the following results and Recommendations:
   - VIP and Require Activation were clustered accordingly, therefore, the model predicted well for these two RFM segments.
   - For the others, we see,
     - Loyal- 10% (134 customers) out of the total Loyal customers were clustered at 0
       - As defined, they fall under 'High Value Lapsed' Segment, they should generate a high revenue, but may not so or so frequently purchase and may have recently purchased lesser.
       - Most of the customers from this group of 134 customers may be a mix of low value and high value purchasers, who have purchased less frequently and not much recently, which may have influenced the results.
       - Recommendation: To continue to re-market to them across other channels. Customer surveys may help provide an opportunity to correct poor experiences and better understand lapsed customers. Did they lapse because of a poor experience, seasonal products, or an isolated event?
     - Potential- 11.67% out of the total Potential Customers, 51 Potential Customers Clustered 1:
       - As defined, they fall under Customers who do not generate high revenue, but who purchases often, and frequent.
       - There are customers from the list who generated high revenue based on high MV scores, perhaps they have purchased more expensive products at some or one point in time, this could be a factor to be clustered under 1, who needs less attention or engagement.
     - Need Attention- 1.45% 2 Customers who were clustered at 1
       - As defined, they fall under Customers who do not generate high revenue, who have not spent much in the past and have not purchased anything recently.
       - One of the customers has purchased expensive products but has been inactive for 235 days. The other has low MV, but have purchased very recently the past 3 days.
       - These are the factors to consider for the two customers.

- Based on the percentages of clustering on 0 and 1, the company may focus its customer engagement activities more on the three segments:
    - Potential, Need Attention and Require Activation
- For the 10% outliers under RFM segment- Loyal which were clustered at 0 (needs customer engagement) and 11.67% outliers under RFM segment- Potential which were clustered at 1 (needs less or no customer engagement), the said outliers compose around 3% of the total data samples, this could be minimal or negligible, and the need to investigate further depends on stakeholder decision.
- In addition, it is also recommended to look into the reasons why there was a big drop in customer retention rate on 2011-12, at -59.91%.