Problem Set Analysis

INTRODUCTION

The requirement to explore sales data from a customer retention perspective has been analysed in detail and has been presented in the form of a report. A structured exercise was undertaken and accordingly this report is presented. The report has been divided into multiple sections and presented below:

1. Problem Set Description and Goal
2. Assumptions and Methodology used
3. General Data Observations
4. Analysis and Results
5. Recommendations and conclusions

PROBLEM SET DESCRIPTION AND GOAL:

The given Dataset contains information critical to the business of an E-Commerce company. It contains 3 files:

1. TranscationData.csv- The main dataset that contains information related to customers' purchasing history overs the years 2013-2017 such as the order number, Order Date, quantity of the product purchased/returned, sales associated with the sold/returned product, Channel used (whether the product was purchased online or offline), Customer Category, etc. There are also product-specific information such as the SKU(Stock Keeping Unit) Number of the product, name of the product supplier, and the Category of the product(Root or leaf Category).
2. RootCategory.csv- It contains the SKU number of the products and the Leaf Category ID of the products. Many products can be associated with a Leaf Category.
3. RootCategory.csv- It contains the root category IDs and the SKU number of the associated products.

Generally, in an e-commerce market, brands are arranged in various levels of a tree based on their categories. There are many broad categories that exist at the root level and the depth of the tree increases when the specificity of a product, that is requested, increases. Such a hierarchical structure to establish categories of various brands is a massive aid to a company at the time of matching product listings with user searches. This not only improves customer satisfaction, but also the brand's visibility(product discovery), which are crucial for sales. However, this part of the provided dataset is not considered for this analysis.

With the given data, the task is to find customers who are most likely to churn. At first glance of the provided data, it is seemingly hard to derive insights that are customer-centric in deciding whether he or she might churn or not. This report explains my approach to accomplish the given task by making use of derived attributes, that reflect a customer's purchasing characteristics, to perform customer segmentation using clustering that give insights on a customer's interaction with the company, and hence the possibility of seeing the customer continue his/her patronage with the company or not.

ASSUMPTIONS AND METHODOLOGY USED:

Churn rate of a customer is dependant on both the brand's quality of service as well as the customer's level of interest in buying and using the brand's products.

From my perspective, I feel that the data provided is more customer-focused than brand-focused. Attributes such as the Order Number, the order amount of the products purchased/returned, and the associated sales values (at the SKU level) are subtle indicators of the customer's interests towards the products. Assuming that all products of all brands are treated with the same level of craze and interest, I decided to focus majorly on the customer side of the data to come up with features/attributes that would be significant in my segmentation analysis. SKU level data, item categories and supplier data have not been considered, as the main focus is on the customers. In fact SKU level data would have focused the analysis more towards the products and or brands and associated suppliers. Since the key mandate for this analysis is customer churn, all attention was towards customer-specific data items only.

My methodology for this task consists of the following in order:

1. Data sorting based on customer's 'order date (Chronological order)
2. Creation of Customer Information Card that contains info such as the customer's ID, the earliest purchase date, the most recent purchase date, the summation of sales per month, and the number of orders he/she placed per month. This in a way represents the extraction,transformation and loading (ETL) of the data set provided.
3. Creation of derived attributes such as Recency, Frequency and overall sales contribution of each customer
4. Conducting a RFS(Recency-Frequency-Sales) based cluster analysis to identify the nature of the statistical values associated with the 3 measures and how it varies from a more active and interested customer to a less active customer.
5. Visualizing and storing the overall scores of each type of cluster (Recency, Frequency and Sales) to determine the threshold as a deciding factor in predicting who might churn or not.

1. The data contains Information from January 2013 to May 2017. It is sorted so as to carry out the subsequent steps as mentioned above..

2. Customer Information card is created to serve as the main lookup. This facilitates the ease in carrying out step 3.

3. Determining the Recency, Frequency and Sales values for each customer throughout the duration (January 2013-May 2017):

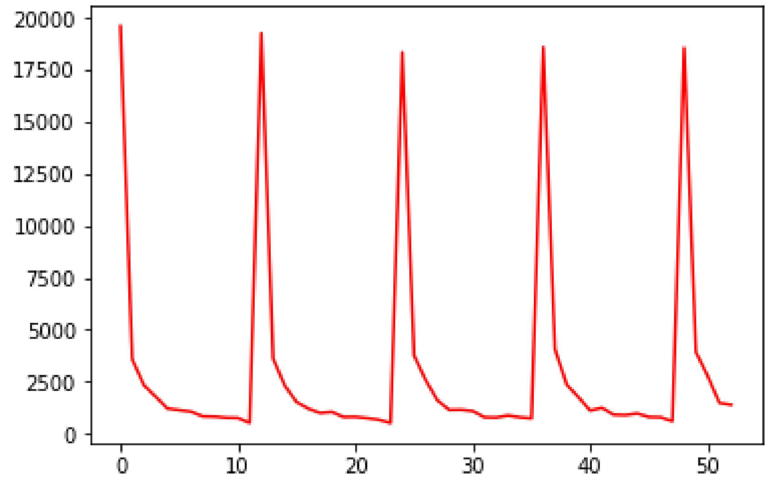
Recency denotes the difference between the last date for reckoning(in this case, May 31st, 2017) and the most recent purchase date of the customer. This difference represents how long the customer has been inactive. The bigger the difference, the more inactive the customer has been.

Frequency denotes the number of orders the customers has placed in a month. This is calculated by counting the number of unique order dates in a month. High values of frequency might indicate the the customer is an active shopper.
Sales is actually revenue and is the summation of the sales of products purchased less value of items returned (sales returns) by a customer in a month. Sales can be negative and high negative values indicate the customer returns more than purchasing and might possibly get dissatisfied with the products or services (another reason could be wrong orders on the customer's part)

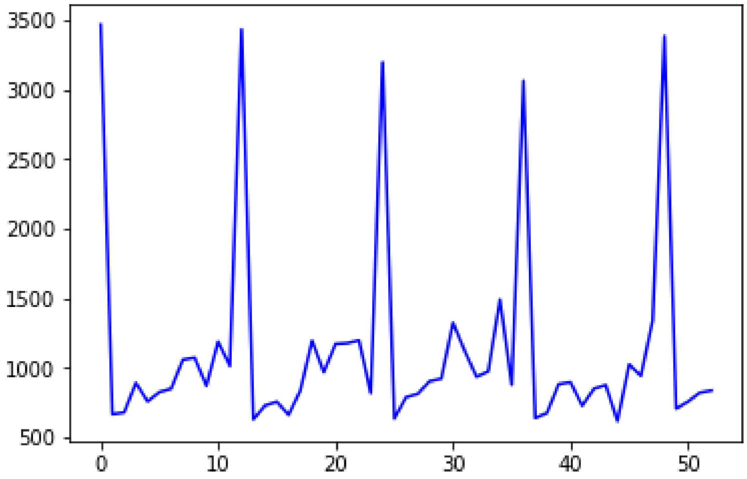
4. RFS Cluster Analysis: The appropriate number of clusters(k) needed for this task is determined by the elbow method, wherein the SSE(Sum of Squared Error) is plotted with different values of k. As the value of k increases, the value of sse decreases and eventually becomes constant. The inertia graph is plotted and visualized to decide the optimal value of k. Based on the value of k, recency, frequency and sales clusters are built, and the statistical properties of the clusters based on each feature are analyzed to compute the overall score for each customer.

GENERAL DATA OBSERVATIONS

From the Information card obtained, I plotted the number of active customers(customers who made a purchase or returned a product) per month, as well as the summation of sales of active customers per month and noticed that both graphs clearly exhibit a consistent trend throughout the year. The number of customers are shown to be high at the beginning of each year and the total sales are high as well at the beginning of each year. Both graphs show a decline in the number of active customers and total sales respectively from February. However, the number of active customers continues to decrease and is at the lowest by the end of each year. As for the plot depicting the total sales made month-wise per year, the sales see a decrease, after january, then go through a series of ups and downs before coming off as increasing overall by the end of each year.



The graph above denotes a plot of the number of active customers per month. Note that the months (year-wise) have been number coded along the x-axis (0 as Jan 13', 1 as Feb 13', etc till 53 as May 17'). The graph below is a plot of the summation of total sales of active customers per month.



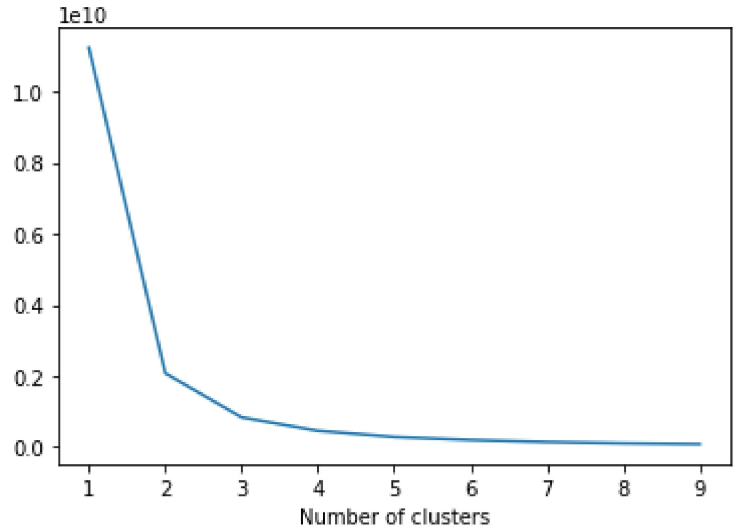
Another observation I made is with regards to the incremental additions of new customers per year. The master data set reveals that there has been very gradual addition of customers only. The customer count has been derived based on the customer id (maximum value) placing order in each month. This reveals the following pattern:

2013-01	**2014-01**	**2015-01**	**2016-01**	**2017-01**	**2017-05**
46207	46981	48171	49553	50857	51234

The above table indicates an average addition of 90-100 customers per month.

ANALYSIS AND RESULTS

After the creation of the information card and the calculation of the RFS values of each customer, an inertia graph was plotted to choose an optimal value of the number of clusters. From my observations, I found the optimal value to be at K=2.



At K=2, the recency clusters, frequency clusters and the sales clusters were built. With regards to the recency cluster, the lesser the value, the better the retention rate of a customer. The statistical report of the different clusters(based on the 3 features) is given below:

Note: The statistical attributes in the table are as follows:
Count-: The number of customers in a cluster
Mean-: The mean of the feature(can be either recency, frequency, or sales) of each cluster
Std-: Standard deviation of the feature of each cluster
Min-: The minimum value of the feature in each cluster
25%-: The 25th percentile value of the feature in each cluster
50%-: The 50th percentile or. the median, of the feature in each cluster
75%-: The 75th percentile of the feature in each cluster
Max- The maximum value of each feature in each cluster.:

Recency cluster	count	mean	std	min	25%	50%	75%	max
0	12393	1106.226902	294.200545	614	847	1104	1359	1611
1	38841	119.35862	160.950439	0.0	8.0	35	177	612

From the table, Recency cluster 1 has lower recency values and is better than cluster 0.

The statistical report of the frequency clusters is given below:

Frequency cluster	count	mean	std	min	25%	50%	75%	max
0	51216	58.018178	163.334016	1	4	14	43	3875
1	18	8343.61111	6468.469517	4233	5022.5	6557	8370	32161

Frequency cluster 1 has better frequency values than frequency cluster 0

The statistical report of the sales clusters is given below:

Sales cluster	count	mean	std	min	25%	50%	75%	max
0	51043	5778.020435	15427.21999	-28101.26	315.505	1146.73	4092.3	197804.09
1	191	390982.129895	265125.253747	199861.80	237771.820	296036.48	428724.105	2294536.17

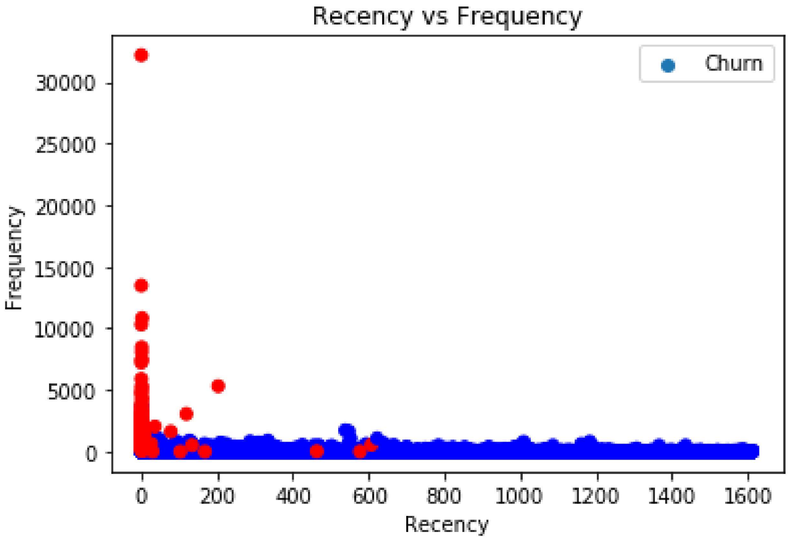
After gathering the statistical reports of the 3 feature-based clusters (Recency, Frequency and Sales), an Overall Score for each customer was calculated as follows:
'Total Score' is a summation of the cluster group number value in the fields Recency Cluster, Frequency Cluster, and the Sales cluster. The minimum value is 0 and the maximum value is 3. Hence, there are 4 possible values for the TotalScore attribute.
The mean of Recency, Frequency and Sales features were calculated and grouped by the TotalScore.
Note: TotalScore value is different from the recency/frequency/sales cluster value. The former is an indicator of the customer's ranking in terms of the interest to purchase. But the latter is only a value of the assigned cluster number as segmented (values of cluster numbers can be any whole number between 0 and k-1, (in this case, 0 and 1)). The resulting table is as follows:

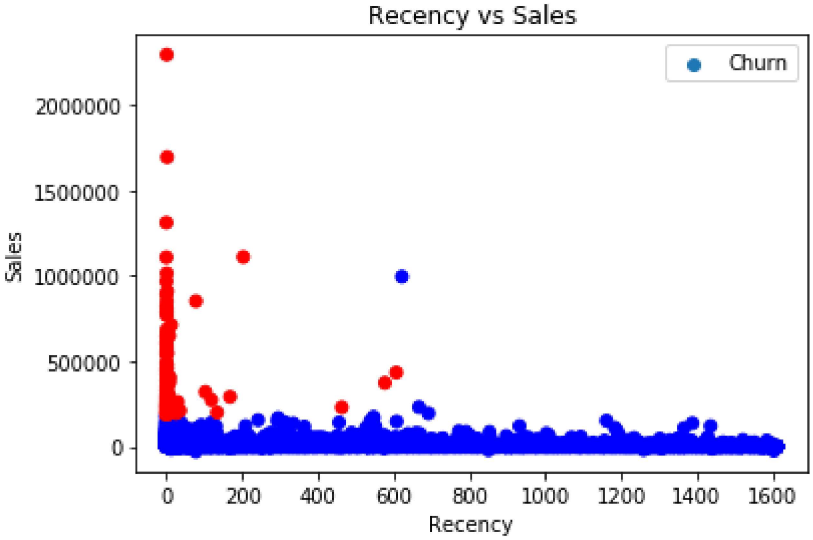
TotalScore	Recency	Frequency	Sales
0	1106.301428	11.103946	1628.948729
1	119.897811	66.764940	7139.495404

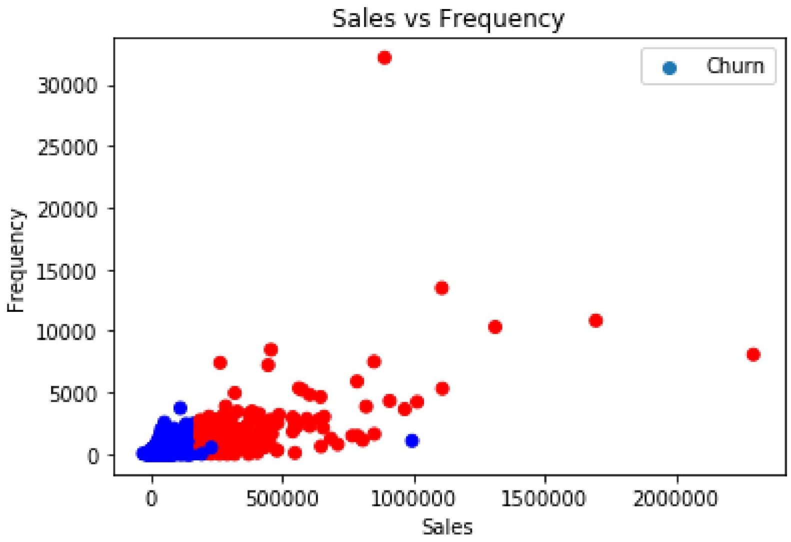
2	14.774566	1512.624277	344023.052948
3	13.250000	8852.437500	870921.043125

From the above table, The row with a Total Score of 3 seems to have the desired set of values (lowest recency value, and highest frequency & sales values). As we move up the table, the quality of thel set of values across the features decreases, hence the reduced TotalScore. The first row (TotalScore=0) has the least desired set of features.

Based on the observations in the above table, I decided to place a threshold on the TotalScore to determine who might and might not churn. Customers having a TotalScore of less than 2 are considered to churn (assigned Label 1) and those with TotalScore>1 are considered to not churn.(assigned label 0). The following scatter plots display how the different classes are distributed across Recency-Frequency, Recency-Sales and Sales-Frequency vector spaces. The red dots indicate customers who might not churn and blue indicates otherwise.







RECOMMENDATIONS AND CONCLUSIONS

From the analysis and results obtained, the number of customers most likely to churn are about 99.6% (since 51045 out of the 51234 customers appear to churn) while <1% won't churn (about 189 out of the 51234 customers).But this conclusion is subject to the following:

- One key aspect is the low frequency observed for most of the customers from the results of the frequency cluster table. However, the other 2 clusters gave a better distribution. This is a significant variation between the 3 clusters.
- The entire time window of 4 years and 5 months have been considered as the reckoning period for this analysis. A shorter time-window (say 2 years) might provide a different set of results.
- I have not considered the significance of the trend (high sales in january and very low sales observed for the rest of the months) as one of the factors in my analysis. It is also possible that customers who might be inactive for 2-3 years might make a comeback and place orders. Also, this analysis was carried out with the intention to keep it customer-focused.
- Another way to extend and improve upon this analysis is to consider observations and tasks from a seller's and brand's perspective. It will be ideal to conduct a seller-centric cluster analysis so that the insights obtained will come from both sides (customer and seller) and might give unbiased results on the whole.

ATTACHMENTS

The following files were created and used for this analysis:

1. .py files:
    a. ON_SEMI_FINAL_1.py-: Code to gather data in chunks and sort them in Chronological order

b. ON_SEMI_FINAL_2.py-: Code to create Customer Information Card, which is used as a main lookup for feature engineering and cluster analysis\
c. ON_SEMI_FINAL_3.py-: Code to obtain derived features and conduct clustering analysis to identify potential churners.

2. .csv files:
    a. TransactionData.csv: The provided dataset
    b. LeafCategory.csv: The dataset containing information of leaf categories
    c. RootCategory.csv: The dataset containing information of root categories
    d. TransactionData_sorted.csv:The Transaction Data in chronological order.
    e. TransactionData_consolidated_final.csv: The Customer Information card used for cluster analysis
    f. TransactionData_rfs_churn_results: The file containing the customer's RFS values, TotalScore and the Churn Label.

Other provided files for preliminary tasks are Transactions.RData and Product_Info.RData.

REFERENCES

[1] https://www.listsmart.io/article/explore-ebay-categories-to-be-a-top-seller
[2]https://towardsdatascience.com/data-driven-growth-with-python-part-1-know-your-metrics-812781e66a5b