Retail Sales Data Analysis

December 19, 2019

The success mantra of retail business is to gauge the consumer needs, behavior and proactively equip for future demands. Making use of retail sales hard data rather than guesswork enables to make smarter decisions toward higher profits, better customer satisfaction, and having a more awesome store overall. Retailers must create effective promotions and offers to meet sales and marketing goals, without forgoing the major opportunities that the current market offers. Data science comes in handy to use prior year's data to better forecast and predict the coming year's sales. It also enables retailers with valuable and analytical insights, especially determining customers with desired products at desired time in a store at different geographical locations. We make use of Apriori algorithm to deduce the product embeddings which can be further utilized to push promotional recommendations. Additionally, we have done the customer segmentation using RFM analysis to discover the valuable customers and suggest tips to improve to hold on to those customers.

1 Introduction

Business needs to be able to see their progress and the factors affecting their sales by analyzing large scale data.Data is worthless if it cannot be analysed, interpreted and applied in context.A picture speaks a thousand words and business analytics would help paint a picture through visualization of data to give the retailers insights on their business. With these insights the businesses can make relevant changes to their strategy for the future to maximize profits and success.The retail stores sell products and gain profit from it.There are a lot of subsidiaries of the stores network which are scattered on various geographical locations. As the network of stores is huge and located at different geographical locations, the company would not fully understand the customer needs and market potentials at these various locations.In this work, we have analyzed We have used Costello's Ace data-set for 2015 to 2018 to find out the product embeddings to build a recommendation system around it. Moreover, we have discovered different valuable types of customers through RFM analysis.

2 Background

[1] Tells about the value of big data analytics in the retail industry. It also explains that Customers' individual behaviors can be segmented using big data analytics and customer behavior at each touchpoint can be collected. By analyzing customers, product recommendations can be personalised to increase customer satisfaction. Big data analytic tools can help improve inventory management, Price optimization and In-store behavior and customer sentiment analysis.

[2] Explains that Retailers can benefit immensely from a structured analytics-driven approach that will help them understand how their customers are using their products and services, how their operations and supply chain are performing, how to manage their workforce and how to identify key risks - insights that they can act upon. The pace and the dexterity with which micro data is collected, gives the retailers immediate insights on the shopping trends.

[3] Provides an analysis of the data sets of Walmart Store to determine the business drivers and predict which departments are affected by different scenarios such as temperature, fuel price and holidays and their impact

on sales at stores of different locations.

[4] Rie Gaku et.al. proposed a procedure for a big data-driven service-level analysis for a real retail store using simulation technology.

[5] Provides the method to realize online analytical processing (OLAP) based data analysis from Web pages through bedding office Web components and realizes the OLAP multidimensional data analyses.

[6] Constructs a sales prediction model for retail stores using deep learning approach.

[7] Firstly selected the profit ratio which was on behalf of commodity profit element and several other key sale attributes including the season ratio and the sale volume to establish the SPV Model, secondly does commodity sale state segmentation based on the SPV Model with ID3 decision tree algorithm and predicted the sale state of the commodity at some future time.

[8] Discusses the limitations of the original Apriori algorithm and presents an advanced version of Apriori algorithm to increase the efficiency of generating association rules.

[9] The method for high-tech products' consumer base analysis based on the RFM-analysis application involving classification methods is presented. Instead of the standard partition of the customer base into quintiles the analytical method of ABC-analysis is applied.

[10] In this study, customers' behaviors are determined by detecting natural clusterings using existing reservation and customer data. They have also customized their services and sales strategies according to these behaviors. The basic characteristics that provide these existing heuristics have been extracted by the decision tree approach after the K-means is implemented.

[11] This research aims to perform clustering and profiling customer by using the model of Recency Frequency and Monetary (RFM) to provide customer relationship management (CRM) recommendation to middle industrial company. The method used in this study consists of four steps: data mining from transaction history data of customer sales, data mining modeling using RFM with K-Means algorithm and customer classification with desicion tree (J48), determination of customer loyalty level and recommendation of customer relationship management (CRM) on the medium-sized industry.

[12] In this paper a comparative analysis among agglomerative, k-means and advanced version of k-means are carried out for RFM based market segmentation approach. The experimental outputs show that the agglomerative clustering needs a long processing time for large dataset in comparison with k-means and advanced version of k-means clustering.

3 Dataset

The Costello's Ace data-set contains over 30 million product selling records. The products can be classified into a tree structure. In that tree, first level contains Department Names, second level contains Class Names, Third level contains Fine-line Name, and leaf level contains Item Description. Moreover, each record of the data-set has a Receipt Number and a product. Different products having the same receipt number can be grouped to obtain a list of products bought under single receipt number. Every record has a Net Sales Units value which corresponds to the number of units sold, positive for selling and negative for returning. The Item type - Sale or Return signifies whether a product was bought or returned.

4 Data Preprocessing

- In order to create Product embedding model, we have sliced the original data set to get out few feature columns as the subset i.e. Receipt Number, Net Sales Units, Fine-line Name, Item Description.

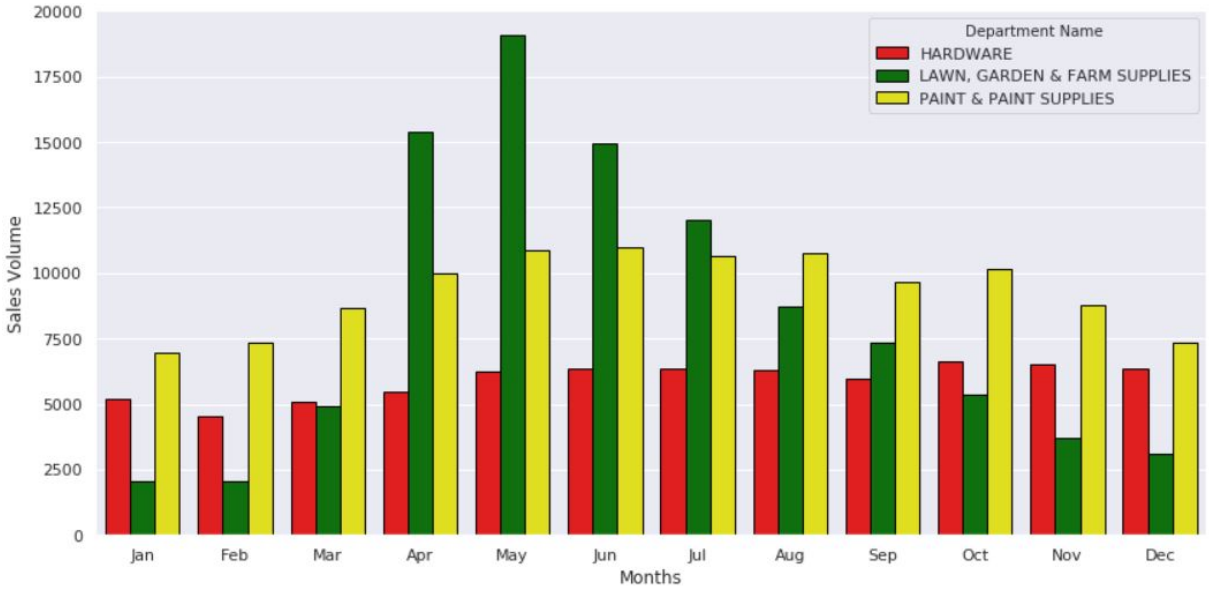

Figure 1: *Sales Volume of Top 3 Departments for each month*

- We have changed the datatype of few columns. For example, We have changed Net sales units datatype from Object to Numeric float.
- We have dealt with Not a Number(NaN) Records by replacing those with zeros or by removing those records.
- In the Fine-line Name column, there were many BLANK fields. To deal with them, we have replaced those BLANKs with their corresponding leaf node values(Item Description).
- We have removed the records which have Net Sales Units -ve in order to get rid of products returned corresponding to a Recipt Number.
- We have grouped the data using Receipt Number, Fine-line Name in order to get a data-frame with Receipt Number and its corresponding products as a list.

To understand the variation of sales volume over time, we plotted sales volume of top 3 departments(such as Hardware, plumbing and Garden supplies) across different months.Figure-2 shows the result. The sales volume from “Paint and Paint Supplies”, and “Hardware” department remain constant throughout the year. However, the sales from “Lawn, Garden and Farm supplies” follows a trend of going up during the period of Mar to May and comes down in the rest months of the year. This can be explained by the fact that during summer (Mar, Apr, and May) people get involved with outdoor gardening activities which cause an increase in the sales volume of these commodities.

To understand how the sales value varies across different stores, we plotted the sales value for different stores. Figure-3 shows the result. This suggests that the sales value at store numbered ‘B’ is significantly larger than that of other stores. So, if the retailer plans to find the store where to stock for future or invest, this store is probably the best choice.

5 Finding Product Embeddings using Apriori

5.1 Methodology

The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules. The key idea of the Apriori algorithm is to make multiple passes over the database. It employs an iterative approach known as a breadth-first search (level-wise search) through the search space, where k-itemsets are used to explore (k+1)-th itemset. The working of the Apriori algorithm fairly depends upon the Apriori property which states that” All nonempty subsets of a frequent itemsets must be frequent”. It also describes the anti-monotonic property which says if the system cannot pass the minimum support test, all its supersets will fail to pass the test.

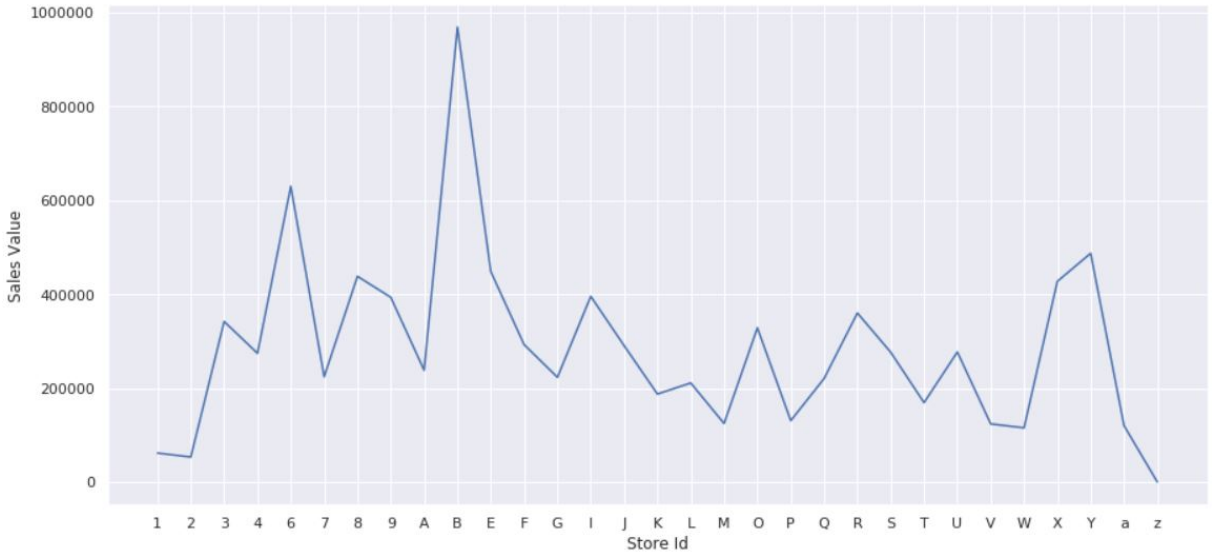

Figure 2: *Sales Value for different stores*

Therefore, if the one set is infrequent then all its supersets are also infrequent and vice versa. This property is used to prune the infrequent candidate elements.

Support

The support of an itemset X, Support(X) is the proportion of transaction in the database in which the item X appears. It signifies the popularity of an itemset.

$$Support(X) = \frac{\#TransactionsHavingX}{\#Transactions}$$

If the sales of a product (item) above a certain proportion have a meaningful effect on profits, that proportion can be considered as the support threshold called min support. Furthermore, we can identify itemsets that have support values beyond this threshold as significant itemsets.

Confidence

Confidence of a rule is defined as follows:

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$$

It can give some important insights, but it also has a major drawback. It only takes into account the popularity of the itemset X and not the popularity of Y. If Y is equally popular as X then there will be a higher probability that a transaction containing X will also contain Y thus increasing the confidence. To overcome this drawback there is another measure called lift.

Lift

The lift of a rule is defined as:

$$Lift(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X).Support(Y)}$$

This signifies the likelihood of the itemset Y being purchased when item X is purchased while considering the popularity of Y. If the value of lift is greater than 1, it means that the itemset Y is likely to be bought with itemset X, while a value less than 1 implies that itemset Y is unlikely to be bought if the itemset X is bought.

	Antecedents	Consequent	Support	Confidence	Lift
0	AEROSOLS	INSECT BAITS/TRAPS	0.010230	0.263973	2.646267
1	BEVERAGES	SNACKS	0.008906	0.199836	2.657760
2	BIRD SEED/NECTAR/SUET	BIRD FEEDERS	0.007193	0.619285	5.993907
3	CONSUMR SYNTHETIC BRISTLE	INTERIOR PAINT	0.005554	0.229560	2.540827
4	TRAYS AND FRAMES	CONSUMR SYNTHETIC BRISTLE	0.005091	0.210440	4.030446
5	CYLINDER - PAINTED THEMES	CYLINDER	0.007449	0.462472	2.900450
6	CYLINDER	KEY RINGS	0.012668	0.489663	3.070978
7	CYLINDER	KEY TAGS	0.010500	0.696317	4.367035
8	EXTERIOR EXTENSION CORDS	INDOOR EXTENSION CORDS	0.005625	0.176494	4.492535
9	GUN	GARDEN HOSE	0.006192	0.209197	4.780062

Figure 3: *Associations rules of few products*

5.2 Number of Association Rules

Given there are n unique items in the dataset.
Total number of itemsets = 2^n
Number of ways of selecting k items from n unique items = $\binom{n}{k}$
Number of ways of selecting rest of the items = $\sum_{j=1}^{n-k} \binom{n-k}{j}$
Total number of possible association rules =

$$\sum_{k=1}^{n-1} \left[\binom{n}{k} * \sum_{j=1}^{n-k} \binom{n-k}{j} \right] = 3^n - 2^{n+1} + 1$$

Algorithm 1 Apriori Algorithm

function GET-FREQUENT-ITEMSET()
  C_k : Ceginandidate item-set of size k
  L_k : Frequent item-set of size k

  L_1 : {Frequent items}
  for $k \leftarrow 1$ to $L_k \neq \Phi$ **do**
    C_{k+1} = candidates generated from L_k
    for $transaction\ t$ in $database$ **do**
      increment the count of all candidates in C_{k+1} that are contained in t
      L_{k+1} = candidates in C_{k+1} with $min\_support$
    end for
  end for
  return $\cup_k L_k$
end function

5.3 Association Rules

We have applied Apriori algorithm with **minimum support 0.005, minimum confidence 0.15, minimum lift 3** on the grouped list of products under a single Receipt Number. We get the following tabular structure from that as shown in Figure 2, where each row depicts an association rule.
Antecedents are the items which we pick initially and Consequent is the item our system suggests according to the desirability of the consumers. Antecedents can be more than one product which is considered as initial part of a rule and Consequent is the product which is most likely consumed in the presence of Antecedents. Support, Confidence, and Lift are numerical relation between the Antecedents and Consequent as discussed in Approach section.

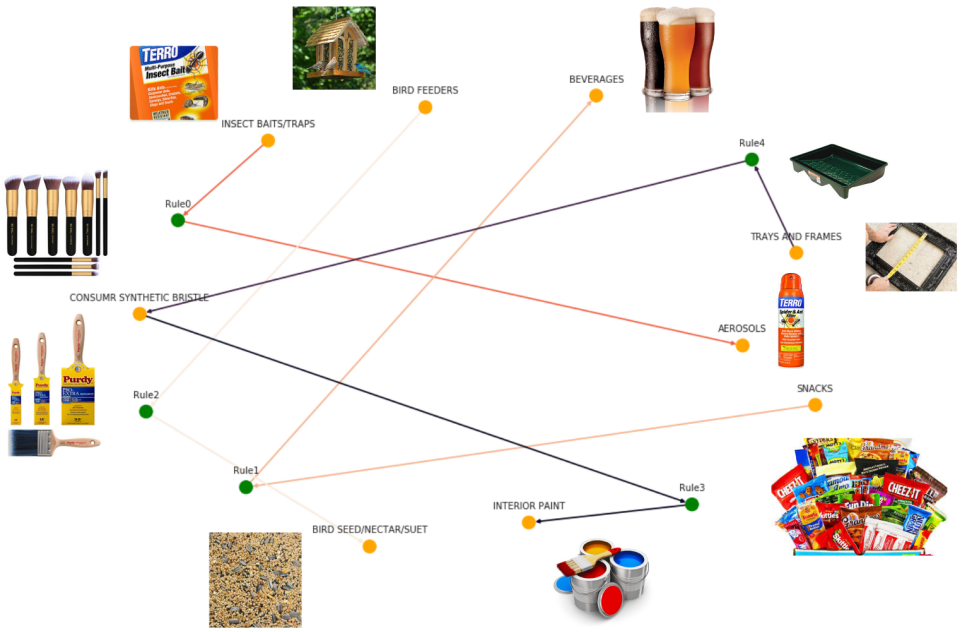

Figure 4: *Graph of few Association Rules*

For example, in the **Rule0(first row)** of Figure 2, **AEROSOLS** is an Antecedent and **INSECT BAITS/TRAPS** is its corresponding consequent.

- Support value for **AEROSOLS**→ **INSECT BAITS/TRAPS** is 0.010230 which means **AEROSOLS,INSECT BAITS/TRAPS** appear together in dataset of over 17 million transactions with 1.023 %.
- Confidence value 0.263973 indicates the accuracy of the rule is 26.3 % of the times a customer buys **AEROSOLS** has also bought **INSECT BAITS/TRAPS**.
- Lift value for this rule is greater than 1, which indicates the degree to which occurrences of **AEROSOLS** and **INSECT BAITS/TRAPS** are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.

we can analyze the association rules more intuitively through Figure 3.

- **Rule0**: From Rule0, it is observed that people bought **Insect Baits/Traps** and **Aerosols together**. A reason behind this may be the fact that typically people buy aerosols and insect baits/traps to get rid of insects.
- **Rule1**: If we analyze Rule1, it is observed that people buy **Snacks** and **Beverages** together, this is because they have get-togethers with their friends and family very often.
- **Rule2**: From Rule2, we see that people bought **BirdFeeders** and **BirdSeeds/Nectar** together, this completely makes sense because feeding wild birds is a wonderful way to add life and color to your backyard, especially during dreary winter months. Bird watching is growing in popularity, and these days almost half of US households keep wild bird feeders on their property. Feeding wild birds also makes their lives easier, especially during colder months when food is scarce
- **Rule3 and Rule4**: Rule3 and Rule4 explain a **transitive** relation between the products bought together. From Rule3, we can see that people buy **ConsumerSyntheticBristle** and **InteriorPaints** together. From Rule4, it is observed that **Trays**, **Frames** and **ConsumerSyntheticBristle** are bought together. This explains that while painting their houses people usually buy InteriorPaints, Synthetic Bristle and Paint Trays together.

5.4 Recommendation System using Association Rules

So, if the confidence value is high enough for an association rule, we can suggest it's implementation in a store for product embedding i.e. **BirdFeeders** and **BirdSeeds/Nectar** could be placed together in a store for better sales of these products.Similarly, **Beverages** and **Snacks** can be grouped together. Even we can suggest to offer a discount on the combined purchase of products part of each association rule.

5.5 Validation

With the help of Market Basket Analysis, we have built a Recommendation System using Association Rules which can recommend products based on current purchased products by a customer. By looking at the Support and Confidence values for a Association Rule, we can validate whether the recommendation engine gives the right suggestion. Let's see another validation method to further corroborate the recommendations provided by the system, The Jaccard index method.
The Jaccard index, also known as Intersection over Union and the Jaccard similarity coefficient, is a statistic used for gauging the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

$$\text{Jaccard Index} = J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$\text{Jaccard Index} = J(A,B) = \frac{The\ number\ of\ baskets\ in\ which\ A\ and\ B\ appear\ together}{The\ number\ of\ baskets\ in\ which\ A\ and\ B\ do\ not\ appear\ together}$$

In case of Market Basket Analysis Jaccard index is defined as the ratio of the number of baskets in which, e.g., items 4 and 5 appear together, divided by the number of baskets in which they do not appear together.

We considered 2 item sets {BEVERAGES,SNACKS} and {GUN,GARDEN HOSE} and calculated the jaccard index with respect to each item set and compared the jaccard index value with the corresponding Support value obtained by Market Basket Analysis. The results are as shown in following table:

Item Set	Support	Jaccard Index
BEVERAGES,SNACKS	0.008906	0.0100157
GUN,GARDEN HOSE	0.006192	0.0066382

From the result, we can infer that the products which are bought together frequently will have higher **Jaccard index** than the product which are not bought together frequently. The same is inferred by analysing the **Support** value.

6 Finding Customer Segmentation using RFM analysis

6.1 Methodology

RFM stands for Recency, Frequency and Monetary value. RFM analysis is a marketing technique used for analyzing customer behavior such as how recently a customer has purchased (recency), how often the customer purchases (frequency), and how much the customer spends (monetary). It is a useful method to improve customer segmentation by dividing customers into various groups for future personalization services and to identify customers who are more likely to respond to promotions.

RECENCY (R): Time since last purchase. How many days ago was their last purchase? Deduct most recent purchase date from today to calculate the recency value. 1 day ago? 14 days ago? 500 days ago?

FREQUENCY (F): Total number of purchases How many times has the customer purchased from our store? For example, if someone placed 10 orders over a period of time, their frequency is 10.

MONETARY VALUE (M): Total monetary value How many $$ (or whatever is your currency of calculation) has this customer spent? Again limit to last two years – or take all time. Simply total up the money from all

Figure 5: *Customer Significance and Valuability using RFM analysis: 1.* ***Color Density:*** *Color density is directly proportional to the Number of Customers with same RFM set values i.e. Darker the colour, more the number of Customers with that set of RFM value. 2.* ***Size:*** *Customer significance to the store is directly proportional to the size of the bubble i.e. Most Valuable Customers will have largest bubble with RFM values set of 5,5,5.*

transactions to get the M value.

6.2 RFM Analysis

RFM analysis involves scaling customers based on each RFM factor separately. The segmentation starts with recency, then frequency, and finally monetary value. We sort customers based on recency, i.e. period since last purchase, in order of lowest to highest (most recent purchasers at the top). The customers are then split into quintiles (five groups), and given the customer with recent transaction under last 4 months have a recency score of 5, the customer with recent transaction under last 8 months have a score of 4 and so on. Customers are then sorted and scored for frequency – from the most to least frequent, coding the top 20% as 5, and the less frequent quintiles as 4, 3, 2, and 1. This process is then undertaken for monetary as well. Finally, all customers are ranked by concatenating R, F, and M values.

Figure 5 depicts the assigned value-scores to each customer on the basis of their past behavior using RFM Analysis. Using the quintile system explained above, at the most, 125 different scores (5x5x5) can be assigned to the total number of distinct customers i.e. 3,49,280. These cells differ in size from one another. A customer's score can range from 555 being the highest, to 111 being the lowest. The best customers are in quintile 5 for each factor (5,5,5) that have purchased most recently, most frequently and have spent the most money. Additionally, the color density depicts the density of the cluster of Customers with same RFM value-score i.e. color density is directly proportional to the number of customers with same RFM value-score.

6.3 Recommendation Sysytem using RFM Analysis

Using the set of value-scores from RFM analysis, we have segmented out the different type of customers involved. From figure 6, **Champions** are most valuable customers with value-scores greater than equal to 4 i.e. $(R \geq 4, F \geq 4, M \geq 4)$. For **Loyal_Customers**, the Frequency flag(F) should be in the quintile 4 or 5, but Recency and Monetary Flag could have any values from 1 to 5 as loyality doesn't depends on them i.e.

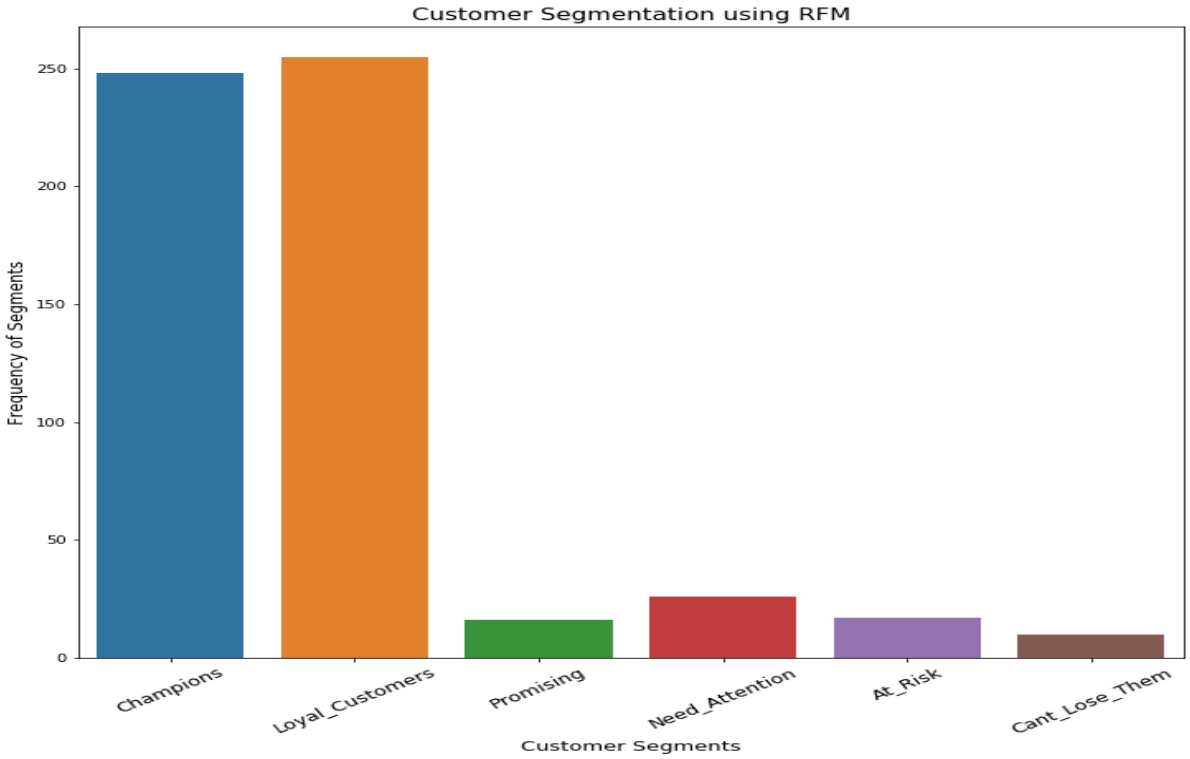

Figure 6: *Customer Segmentation using RFM analysis: Differentitation of Customers on the basis of set of RFM values*

$(R \geq 1, F \geq 4, M \geq 1)$. Similarly, we have categorised different other types of customers on the basis of combination of RFM values.

According to the type of segmentation of a customer, we can suggest different type of Actionable tips for continous involvement or for improving the involvement of that customer. As we can see from the table given below that for **Champions** type of customers, we can have the continuous involvement by providing Rewards. They can also be the early adopters for new products i.e. we can deduce any new product's popularity on the basis of their usage. Similarly, **Loyal_Customers** are frequent customers. so, we can suggest them to buy higher value products and engage them with more products on the basis of their reviews.

Customer Number	Recency_Flag	Freq_Flag	Monetary_Flag
10000	5	5	5
24805	5	1	1

Figure 7: *RFM Scores for two customers*

Customer Number	Recent_Date	Frequency_of_Shopping	Total_Net_sales
10000	09/2018	39117	1117686.91
24805	09/2018	1	12.58

Figure 8: *Dataset Values for two customers*

Customer Segment	RFM Score	Activity	Actionable Tip
Champions	$(R \geq 4, F \geq 4, M \geq 4)$	Bought recently, buy often and spend the most!	Reward them. Can be early adopters for new products. Will promote your brand.
Loyal Customers	$(R \geq 1, F \geq 4, M \geq 1)$	Spend good money with us often. Responsive to promotions.	Up-sell higher value products. Ask for reviews. Engage them.
Promising	$(R = 5, 2 \leq F \leq 4, M \leq 3)$	Recent shoppers, but haven't spent much.	Create brand awareness, offer free trials.
Customers Needing Attention	$(3 \leq R < 5, F \geq 2, M \geq 3)$	Above average recency, frequency and monetary values. May not have bought very recently though. promotions.	Make limited time offers, Recommend based on past purchases. Reactivate them.
At Risk	$(R \leq 3, F \geq 2, M \geq 2)$	Spent big money and purchased often. But long time ago. Need to bring them back!	Send personalized emails to reconnect, offer renewals. Provide helpful resources.
Can't Lose Them	$(R \leq 4, F \geq 3, M \geq 3)$	Made biggest purchases, and often. But haven't returned for a long time.	Win them back via renewals or newer products. Don't lose them to competition, talk to them.

6.4 Validation

To validate our results, we have compared the results of RFM scores of two records as given in figure 7 with Customer Number '10000' and '24805' with the corresponding values of Recent Date of Transaction, Frequency of Shopping and Total Net Sales Amount from the dataset as given figure 8. We have also sorted the dataset with respect to frequency of the shoppings and picked these customers from top and bottom of the dataset.

Recency:

Our dataset contains latest transactions up to 09/2018 for all the customers. **Recent_Date** of transaction for both Customers is from 09/2018 which is under last 4 months. Thus, they should have **Recency_Flag** score 5 which we can validate from the figure 7.

Frequency:

From figure 8, Frequency of shopping for Customer Number '10000' is greater than the Customer Number '24805'. Comparatively, **Freq_Flag** value for Customer Number '10000' should be the highest and for '24805' should be the lowest. Thus, we can validate that from figure 7 as **Freq_Flag** for '10000' is 5 and for '24805' is 1.

Monetary:

Total Net Sales Amount for Customer Number ’10000’ is greater than the Customer Number ’24805’ as shown in figure 8. Comparatively, **Monetary_Flag** value for Customer Number ’10000’ should be the highest and for ’24805’ should be the lowest. Thus, we can validate that from figure 7 as **Monetary_Flag** for ’10000’ is 5 and for ’24805’ is 1.

References

[1] Belarbi, Hamza Tajmouati, Abdelali Bennis, Hamid Mohammed, El Haj Tirari. (2016). Predictive Analysis of Big Data in Retail Industry.

[2] Chandramana, Sudeep. (2017). Retail Analytics: Driving Success in Retail Industry with Business Analytics. Volume 7.

[3] M. Singh, B. Ghutla, R. Lilo Jnr, A. F. S. Mohammed and M. A. Rashid, "Walmart’s Sales Data Analysis - A Big Data Analytics Perspective," 2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), Nadi, 2017, pp. 114-119.

[4] R. Gaku and S. Takakuwa, "Big data-driven service level analysis for a retail store," 2015 Winter Simulation Conference (WSC), Huntington Beach, CA, 2015, pp. 791-799.

[5] C. Ju and M. Han, "Effectiveness of OLAP-Based Sales Analysis in Retail Enterprises," 2008 ISECS International Colloquium on Computing, Communication, Control, and Management, Guangzhou, 2008, pp. 240-244.

[6] Y. Kaneko and K. Yada, "A Deep Learning Approach for the Prediction of Retail Store Sales," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, 2016, pp. 531-537.

[7] J. Zhang and J. Li, "Retail Commodity Sale Forecast Model Based on Data Mining," 2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS), Ostrawva, 2016, pp. 307-310.

[8] K. R. Suneetha and R. Krishnamoorti, "Advanced Version of A Priori Algorithm," 2010 First International Conference on Integrated Intelligent Computing, Bangalore, 2010, pp. 238-245.

[9] M. E. Tsoy and V. Y. Shchekoldin, "RFM-analysis as a tool for segmentation of high-tech products’ consumers," 2016 13th International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE), Novosibirsk, 2016, pp. 290-293.

[10] M. Pakyürek, M. S. Sezgin, S. Kestepe, B. Bora, R. Düzağaç and O. T. Yıldız, "Customer clustering using RFM analysis," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4.

[11] I. Maryani and D. Riana, "Clustering and profiling of customers using RFM for customer relationship management recommendations," 2017 5th International Conference on Cyber and IT Service Management (CITSM), Denpasar, 2017, pp. 1-6.

[12] S. H. Shihab, S. Afroge and S. Z. Mishu, "RFM Based Market Segmentation Approach Using Advanced K-means and Agglomerative Clustering: A Comparative Study," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox’sBazar, Bangladesh, 2019, pp. 1-4.