IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CUSTOMER ANALYTICS BASED ON SEGMENTATION, RETENTION AND FP-GROWTH

[1]Dr. Sunil Bhutada, [2]U. Saran Sri Dath, [3]P. Sai Satya Murthy, [4]M. Naresh
[1]Head of the Department, [2]Student, [3]Student, [4]Student
Department of Information Technology
Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad

Abstract: As the number of rivals and contenders in the market are increasing day by day, it is getting harder and harder for the existing companies to retain customers. The Customer Analytics based on Segmentation, Retention and FP-Growth provides a better way for retention and improving the profit margins. Customer Analytics is the process of analyzing the customer data to gain some valuable and useful information or knowledge that is beneficial to the organization or the store. In this research we will be working on a way to improve the retention of the customers by analyzing the buying patterns of the product consumers. For that first the customers are segmented into various categories based on their interaction with the company using RFM analysis and K-Means Clustering. Then the patterns of purchasing are studied and recommendations are provided using the FP-Growth algorithm. These recommendations are basically used to attract the customers and gain back them back by providing various offers and sales. Finally, this results in the improved commitment of the customers with regard to the organization and also improved profit margin for the organization over long-term.

Index Terms **- Customer Analytics, Segmentation, Retention, FP-Growth Algorithm, K-means, RFM Analysis, Basket Analysis.**

I. INTRODUCTION

Customer Analytics is the most important part of successfully running an organization, as analyzing the customer data gives us knowledge. This knowledge then can be used to gain competitive advantage over other rivals. Segmentation is one of the major parts of customer analytics where an industry tries to characterize its customers into various types based on their purchasing characteristics. Clustering algorithms are commonly the most effective for segmenting the consumers. K-means is one of the most popular clustering techniques, where we try to divide the data points into K number of cluster. Here the value K is usually provided by the user as an input value.

Moving over to the Retention part of customer analytics, Retention is basically an organizations capability to make the customers come back to buy products from them. Retention can generally be defined as a measure of the total number of loyal customers to a particular organization. To improve the retention of customers over a particular store, we use the market-basket analysis, where we study the purchase frequencies, patterns and other attributes of all the purchasers, and try to get the frequent item sets of those patterns. From these frequent items sets we generate the recommendations, upon which various sales and offers can be inaugurated to retain back the customers. The product recommendations are generally based on the basket analysis. And the first algorithm that comes into mind is Apriori. But here we will be using the FP-Growth algorithm. In the FP-Growth algorithm, the customer purchasing dataset is scanned only once to generate an FP tree. By analyzing the FP tree, the frequent patterns can be obtained from which the rules are produced as the overall result.

II. LITERATURE SURVEY

In [1] identified that the K-Means algorithm works best for categorizing the customers into various segments. They have researched on various methods such as K-Means, Fuzzy C-means, RM K-Means which is a slight variation of K-means. The result that was depicted in their research paper was that, out of all these segmenting algorithms, K-Means was faster and takes lesser number of iterations.

In [2] have done analysis on predicting the patterns in which the customers buy the products in and get recommendations out of them. Apriori analysis uses candidate generation at every step and also it scans the database at every step. This way it generates the frequent patterns and analyses them to obtain Association rules based on minimum support and minimum confidence.

In [3] introduced a clustering technique which is again a slight variation of K-means algorithm and also very similar to that of K-medoids algorithms. According to his research paper, the newly depicted algorithm was not able to perform as well as expected and did not give the optimal and good solution as an output. Here, Shah has observed that his own algorithm will be performing better in the case of a greater number of clusters. But while using the optimal number of clusters, the algorithm has failed comparatively.

In [4] researched on the retention of the customers and a method for finding the retention rates of each and every month for a particular given dataset. These retention rates can be very helpful in analysing the before and after results of our application. According to the authors of this article, the retention rates are to be calculated by a method or a type of index known as cohort index, where the cohort index depicts the least or first transaction date of a particular customer.

III. Proposed System And Architecture

3.1 Proposed System:

In this research, we will be using Segmenting the customers based on a combination of K-means clustering algorithm and RFM analysis. The market-basket analysis will be done using FP-Growth algorithm instead of Apriori algorithm for the Basket-Analysis part of customer analytics. As we know, the Apriori algorithm requires frequent scanning of dataset and also needs to generate a lot of candidates in order to get the frequent patterns. This process is a demands lot of processing power and resources.
Using Apriori Algorithm for implementing the Market-Basket analysis on huge data is inefficient as well. The below figure shows the inefficiency of the Apriori algorithm, as it requires more than 52 giga bytes of memory for frequent set generations, when Apriori is performed on a dataset of about 5 lakh entries. The real-world data is comparatively much larger in size and much greater in volume.

```
frequent_itemsets = apriori(basket, min_support=0.01, use_colnames=True)
frequent_itemsets.head()
```

```
C:\Users\usara\anaconda3\lib\site-packages\mlxtend\frequent_patterns\fpcommon.py:111: DeprecationWarning:

DataFrames with non-bool types result in worse computationalperformance and their support might be discontinued in the future.P
lease use a DataFrame with bool type
```

```
---------------------------------------------------------------------------
MemoryError                               Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_14896\518479467.py in <module>
----> 1 frequent_itemsets = apriori(basket, min_support=0.01, use_colnames=True)
      2 frequent_itemsets.head()

~\anaconda3\lib\site-packages\mlxtend\frequent_patterns\apriori.py in apriori(df, min_support, use_colnames, max_len, verbose,
low_memory)
    307                     _bools = _bools & (X[:, combin[:, n]] == all_ones)
    308             else:
--> 309                 _bools = np.all(X[:, combin], axis=2)
    310
    311             support = _support(np.array(_bools), rows_count, is_sparse)

MemoryError: Unable to allocate 52.5 GiB for an array with shape (248160, 2, 14199) and data type int64
```

Figure 1: The Limitations of Apriori Algorithm

So, to overcome these limitations, in this paper FP-Growth algorithm is implemented which is much faster and does not require huge amount of computing resources.

3.2 Proposed Architecture:

The various stages involved in Customer Analytics based on Segmentation, Retention and FP-Growth are Data Cleaning and Pre-processing, Customer Analytics and Segmentation, Retention and Basket-Analysis using the FP-Growth algorithm. As the data in the real world is very dirty and contains many irregularities and anomalies. In the Data Cleaning and Pre-processing stage all the dirty data is removed and the null values, duplicate values and negative values are to be removed from the data. Then data must be organized into the required and workable format on which we can apply various analytical algorithms.
The Retention part is carried out using the cohort index concept where we try to calculate retention rates based on every unique pair of cohort index and cohort month. For segmentation of purchasers, we use the most reliable and efficient K-Means algorithm to divide the consumers into various clusters. Finally, FP-Growth algorithm is applied to obtain the product recommendations as our output.

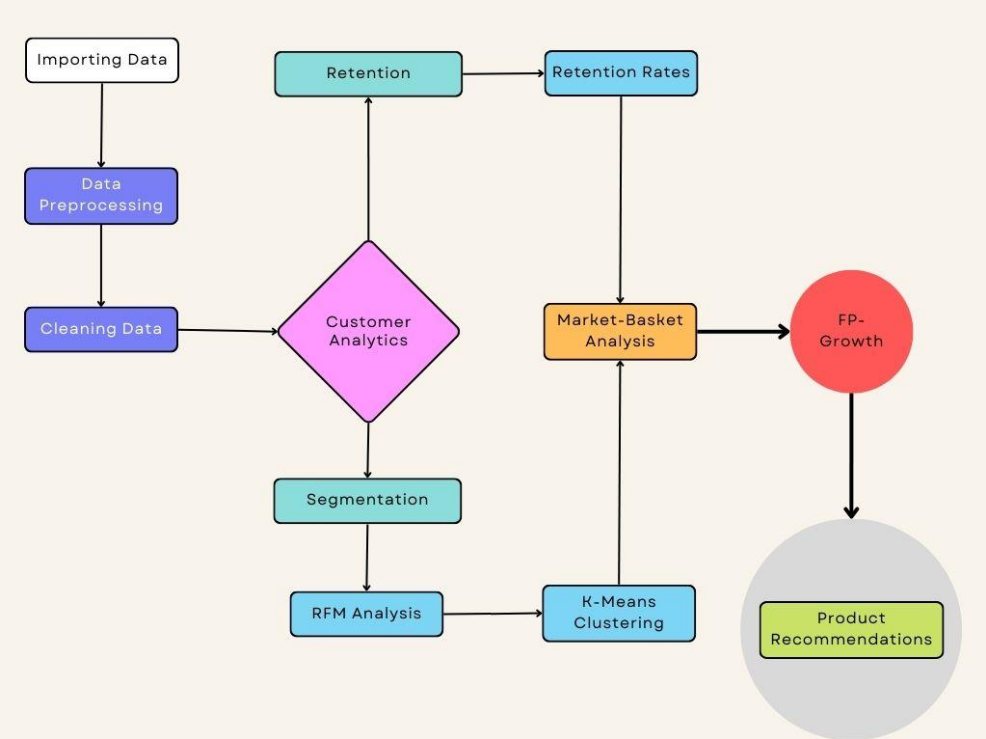

Figure 2: The Architecture of the Proposed System

IV. RESULTS

After successfully clustering the data into 4 clusters using the K-Means method, the result we obtained is as shown in the figure 3. Figure 3 depicts the 2-D version of our cluster graph i.e.; it represents the flattened version of the graph. The various shapes represent the various clusters.

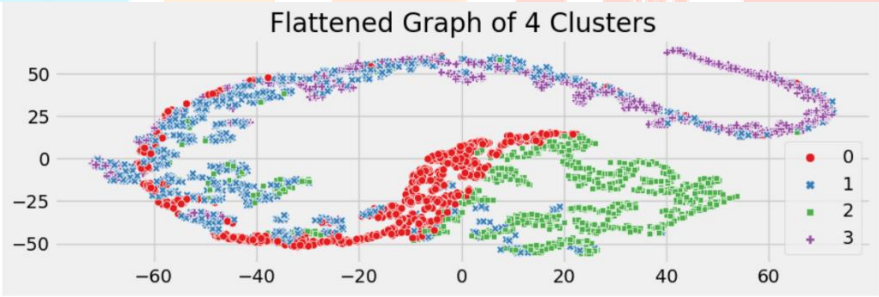

Figure 3: Cluster graph of the data points obtained after clustering the data into 4 clusters

The clustered data is then fed into the FP-Growth model in order to obtain the product recommendations as shown in the figure 3.

```
['ADVENT CALENDAR GINGHAM SACK']
There are no Product recommendations
['AFGHAN SLIPPER SOCK PAIR']
There are no Product recommendations
['AGED GLASS SILVER T-LIGHT HOLDER']
There are no Product recommendations
['AIRLINE BAG VINTAGE JET SET BROWN']
There are no Product recommendations
['AIRLINE BAG VINTAGE JET SET RED']
There are no Product recommendations
['AIRLINE BAG VINTAGE JET SET WHITE']
There are no Product recommendations
['AIRLINE BAG VINTAGE TOKYO 78']
There are no Product recommendations
['AIRLINE BAG VINTAGE WORLD CHAMPION']
There are no Product recommendations
['AIRLINE LOUNGE,METAL SIGN']
There are no Product recommendations
['ALARM CLOCK BAKELIKE CHOCOLATE']
There are no Product recommendations
['ALARM CLOCK BAKELIKE GREEN']
People who bought this also bought: ['SET/6 RED SPOTTY PAPER CUPS', 'SET/6 RED SPOTTY PAPER PLATES']
['ALARM CLOCK BAKELIKE IVORY']
People who bought this also bought: ['CHARLOTTE BAG SUKI DESIGN', 'STRAWBERRY CHARLOTTE BAG']
['ALARM CLOCK BAKELIKE ORANGE']
People who bought this also bought: ['JUMBO BAG RED RETROSPOT', 'JUMBO BAG PINK POLKADOT']
['ALARM CLOCK BAKELIKE PINK']
People who bought this also bought: ['JUMBO BAG RED RETROSPOT', 'JUMBO STORAGE BAG SUKI', 'JUMBO BAG PINK POLKADOT']
['ALARM CLOCK BAKELIKE RED']
People who bought this also bought: ['SET/6 RED SPOTTY PAPER CUPS', 'SET/6 RED SPOTTY PAPER PLATES']
['ALPHABET HEARTS STICKER SHEET']
There are no Product recommendations
['ALPHABET STENCIL CRAFT']
There are no Product recommendations
['ALUMINIUM STAMPED HEART']
There are no Product recommendations
```

Figure 4: Product Recommendations obtained using FP-Growth algorithm

V. CONCLUSION

From the start the primary objective is to increase the retention of the customers so that would be beneficial to the company or organization. To achieve customer retention, we focused on first segmenting the customers into various clusters using the RFM analysis and K-Means clustering. Then we used the FP-Growth algorithm as it is comparatively better than the Apriori analysis as we have discovered in our research. FP-Growth algorithm is faster and also provides better recommendations. These recommendations can finally be utilized in order to retain a greater number of customers and make them come back to the store by providing better offers or sales according the specific recommendations. To conclude, this analysis is very helpful in the perspective of a product-based company to improve their profits over long term either marginally or by huge number.

VI. FUTURE ENHANCEMENT

The scope of this application is exemplary, we can further develop a more visual approach to this Customer Analytics based on Segmentation, Retention and FP-Growth which comes with a user interface, which can be easily understood and used by the company staff. It can be said that using an interactive Interface can be seen as a stepping stone for this research. On the other hand, in the FP Growth algorithm, there is not enough data to generate accurate recommendations candidate as the data utilized here is very compact makes use of lesser memory. We can focus more upon providing the best products to the customers, by making them happy and satisfied as this approach would be carried out among more and more organizations across the globe. The ultimate goal of any product's enhancement is the satisfaction and prosperity of its users

VII. REFERENCES

1. A. Joy Christy, A. Umamakeswari, L. Priyatharsini, A. Neyaa, "RFM ranking – An effective approach to customer segmentation", Journal of King Saud University – Computer and Information Sciences

2. Drashti Shrimal, Dr. Harshali Patil, A Qualitative Approach to Customer Segmentation and Customer ChurnApplication, Mukt Shabd Journal Volume IX, Issue IX, SEPTEMBER/2020 ISSN No: 2347-3150

3. Shah, S., Singh, M., 2012. Comparison of a Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid Algorithm. In: 2012 International Conference on Communication Systems and Network Technologies, Rajkot, pp. 435–437.

4. Drashti, Shrimal and Dr. Harshali Patil, "Consumer purchase patterns based on market basket analysis using apriori algorithms" Journal of Physics: Conference Series, A R Efrat et al 2020 J. Phys.: Conf. Ser. 1524 012109