Business Growth Recommender System For Retail Stores

Purvik S Nukal
Computer Science And engineering
PES University
Bangalore, India
psnukal@gmail.com

Rahul Ranganath
Computer Science And engineering
PES University
Bangalore, India
rahulranganath1102@gmail.com

***Abstract*—The main objective of the project is to build a recommendation system for retail store owners in order to provide them with insights into what may be the necessary steps that need to be taken to increase profits and help them in identifying all the necessary steps and adopt strategies to minimize expenses by identifying the products that are high in demand in the market and take the longest to sell. The model gives recommendations to the end user based on the analysis of customer purchase activity of an item like transaction history, product ratings, frequency of buying and selling a product, and recency of transaction.**

***Index Terms*—Forecasting, Prediction, Recommendation System,**

I. Introduction

Every day a huge number of transactions take place across multiple e-commerce platforms connecting a number of retail stores and a huge amount of data from the customer transaction is generated accordingly. This data generated is not utilized effectively. The data can be utilized to draw meaningful conclusions and predictions about the viability and profitability of a product.

Due to rapid technological improvements, increased customer choices, and product customization, there is a constant fluctuation in consumer demand. This leads to an increase in uncertainty among retail store owners resulting in poor decision-making and a huge loss in revenue.

A. Repercussions of Not implementing the Model

The problem remaining unsolved results in retail store owners being unaware of the customer's liking of a product as customer demands fluctuate constantly, leading to uncertainty and incorrect decision-making with respect to products displayed in the market and a lack of purchasing of specific items by the customers leads to tremendous loss to the owners and wastage of products.

B. Importance of the Model

Using the recommendation system the retail store owner will be able to comprehend the demand for a particular item for a given period of time, based on the customer's activity (transaction history, ratings given to a product, frequency of buying and selling a product, recency of transaction) and based on that make long term strategies about the product's viability in the market. Ultimately, the end user can achieve a big boost in sales and avoid poor decision-making with respect to the selling of products.

II. Techniques Used For Solving The Problem

- Initial Analysis of the dataset is done using **Cohort Analysis**, **Retention Rate** and **Pareto Principle** applied to customers and products.
- **Customer profiling** is done, by segmenting customers into clusters, which exemplify similar behavior based on different parameters, derived from data such as the number of items bought, the value of bought items, the number of items returned, types of items bought, etc.
- Clustering by **K-Means** is performed , which is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters and the optimal number of clusters can be determined using **ElbowMethod** and **Silhouette Score**.
- **RFM analysis** is used to quantitatively rank and group customers based on the recency, frequency, and monetary total of their recent transactions to identify the best customers and perform targeted marketing campaigns. RFM analysis is a way to use data based on 4 existing customer behavior to predict how a new customer is likely to behave in the future.
- **A Priori Algorithm** is used to deal with the recency of a purchased product by a customer, where it explores the state space of possible frequent items and eliminates the rest.
- **The Prophet Algorithm** is used for forecasting the revenue of the retail store owner.The Prophet library is an open-source library designed for making forecasts for univariate time series datasets. It is easy to use and designed to automatically find a good set of hyperparameters for the model in an effort to make skillful forecasts for data with trends and seasonal structure by default.

III. Exploratory Data Analysis

From the dataset, we can get the details of the customer ID, transaction date, quantity of items purchased or sold, type of item, price of the item, and location of the customer during the transaction. The dataset is yet to be cleaned and the missing values present are Missing Completely At Random (MCAR) as the missing values of the columns are independent of the other column values and cannot be derived from them.

A. DataFrame

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
	<chr>	<chr>	<chr>	<dbl>	<chr>	<dbl>	<dbl>	<chr>
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
2	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
3	536365	844068	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850	United Kingdom
8	536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850	United Kingdom
9	536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850	United Kingdom
10	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047	United Kingdom

… with 397,914 more rows

Fig. 1. DataFrame of the model

B. Summary of DataFrame

- A summary of all 8 columns of the DataFrame is shown above. Various observations can be made from this. The mean and Median of the attribute Quantity are 13.02 and 6.00 respectively. Mean and The median of the attribute UnitPrice is 3.116 and 1.950 respectively. We can even observe values of the first and third quartiles of Quantity and Unit price.

```
  InvoiceNo         StockCode         Description          Quantity        InvoiceDate          UnitPrice       
 Length:397924      Length:397924      Length:397924      Min.   :    1.00  Length:397924      Min.   :    0.000  
 Class :character   Class :character   Class :character   1st Qu.:    2.00  Class :character   1st Qu.:    1.250  
 Mode  :character   Mode  :character   Mode  :character   Median :    6.00  Mode  :character   Median :    1.950  
                                                          Mean   :   13.02                     Mean   :    3.116  
                                                          3rd Qu.:   12.00                     3rd Qu.:    3.750  
                                                          Max.   :80995.00                     Max.   :8142.750  

   CustomerID     Country         
 Min.   :12346   Length:397924     
 1st Qu.:13969   Class :character  
 Median :15159   Mode  :character  
 Mean   :15294                     
 3rd Qu.:16795                     
 Max.   :18287
```

Fig. 2. Summary Statistics of Data Frame

C. Outlier Analysis

- The first Boxplot is for Quantity after removing extreme outliers
- The second Boxplot is for Unit Price after removing extreme outliers

D. Bar Graph Analysis

- From the above graph, we can observe that the customers of the United Kingdom have the highest number of transactions.

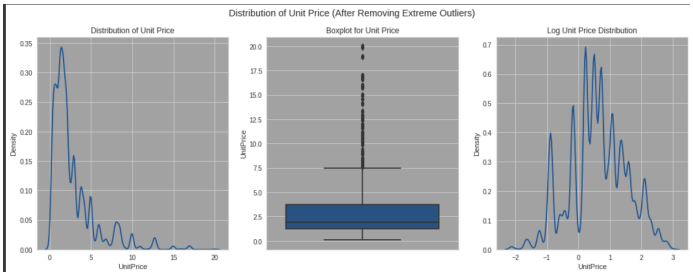

Fig. 3. Box Plot of Quantity

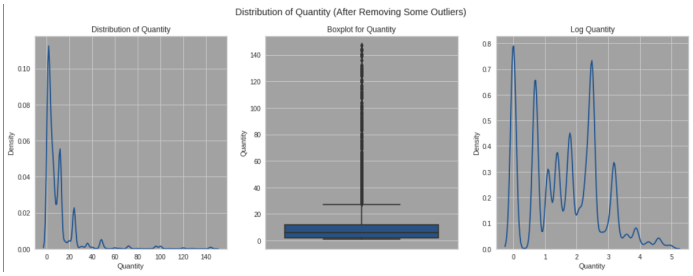

Fig. 4. Box Plot of Unit Price

IV. Implementation

A. Cohort Analysis

- A cohort simply means that a group of people have the same characteristics. Generally, we have three types of cohort analysis:
- Time cohorts or Acquisition cohorts: Groups are divided by first activity.
- Behaviour cohorts or Segment-Based cohorts: Groups are divided by their behaviours and actions about your service.
- Size cohorts: Size-based cohorts refer to the various sizes of customers who purchase a company's products or services.
- Cohort analysis is a subset of behavioural analytics that takes the data from a given e-commerce platform, web application, or online game and rather than looking at all users as one unit, it breaks them into related groups for

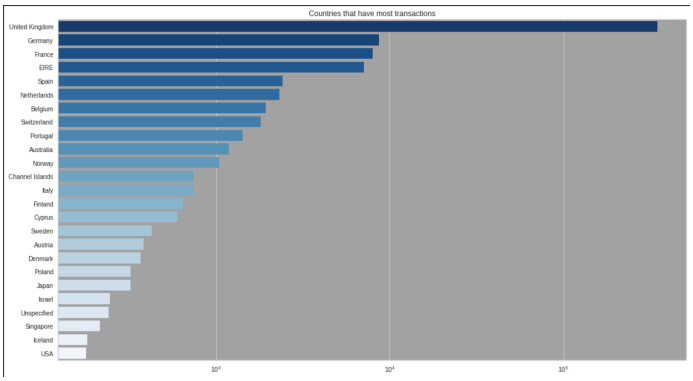

Fig. 5. Country vs Quantity Bar Graph

analysis. These related groups, or cohorts, usually share common characteristics or experiences within a defined time span.

- Cohort analysis is a tool to measure user engagement over time. It helps to know whether user engagement is actually getting better over time or is only appearing to improve because of growth. It proves to be valuable because it helps to separate growth metrics from engagement metrics as growth can easily mask engagement problems. In reality, the lack of activity of the old users is being hidden by the impressive growth numbers of new users, which results in concealing the lack of engagement from a small number of people. Cohort analysis is a better way of looking at data. Its application is not limited to a single industry or function. For example, e-commerce companies can use cohort analysis to spot products that have more potential for sales growth. In digital marketing, it can help identify web pages that perform well based on time spent on websites, conversions or sign-ups. In product marketing, this analysis can be used to identify the success of feature adoption rate and also to reduce churn rates.
- Our data set contains invoice records for more than one year. Let's apply cohort analysis. We can create monthly cohorts. We will group customers for the first invoice record. The cohort index will be the number of months since the first transaction.

CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
CohortMonth													
2010-12-01	873.0	319.0	279.0	329.0	317.0	350.0	316.0	302.0	304.0	345.0	327.0	439.0	230.0
2011-01-01	409.0	91.0	112.0	94.0	131.0	119.0	103.0	100.0	123.0	135.0	151.0	48.0	NaN
2011-02-01	376.0	68.0	69.0	103.0	100.0	91.0	94.0	100.0	92.0	116.0	26.0	NaN	NaN
2011-03-01	450.0	66.0	111.0	90.0	101.0	75.0	119.0	104.0	124.0	38.0	NaN	NaN	NaN
2011-04-01	297.0	64.0	61.0	61.0	56.0	68.0	65.0	77.0	22.0	NaN	NaN	NaN	NaN
2011-05-01	280.0	55.0	48.0	49.0	59.0	65.0	75.0	26.0	NaN	NaN	NaN	NaN	NaN
2011-06-01	240.0	42.0	38.0	64.0	54.0	78.0	23.0	NaN	NaN	NaN	NaN	NaN	NaN
2011-07-01	189.0	34.0	39.0	42.0	50.0	21.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08-01	170.0	32.0	41.0	41.0	22.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09-01	297.0	70.0	90.0	34.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10-01	356.0	84.0	41.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-11-01	322.0	36.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-12-01	39.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fig. 6. Cohort Analysis

- The above data is our cohort table. Its interpretation is simple. For example, We have 873 unique customers whose first transaction is in 2010-12. Its cohort month is 2010-12 and its cohort index is 1. Go to the one right cell, it is 319. It means, 319 unique customers retain their customer ship for next month.

B. Retention rate

- Retention tables show a group of people that visited your site or used your app for the first time during a certain time frame. They also display the progressive 'drop-off' or decline in activity over time for that particular group (a cohort).
- It is an important metric that calculates the percentage of users who continue using your product or service over a given time period. A high retention rate means your current customers value your product and are providing a sustainable source of revenue. A low retention rate means you have a leaky bucket. Some attrition is inevitable. Existing customers stop using your product for reasons beyond your control. However, calculating your product's retention rate is the first step toward turning saveable churn-destined customers into delighted brand enthusiasts. Marketers can use Retention tables to analyse the quality of users brought by a marketing campaign and compare it to other sources of traffic.

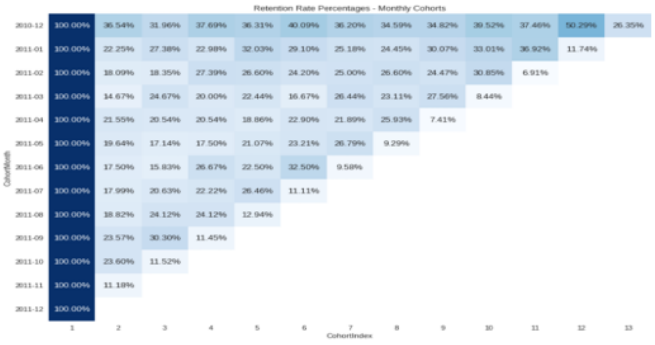

Fig. 7. Retention Rate Percentages - Monthly cohorts

- Let's look at the above retention rate chart and interpret it. 40.05 per cent of customers that made their first shopping in January 2011, use this company after five months.

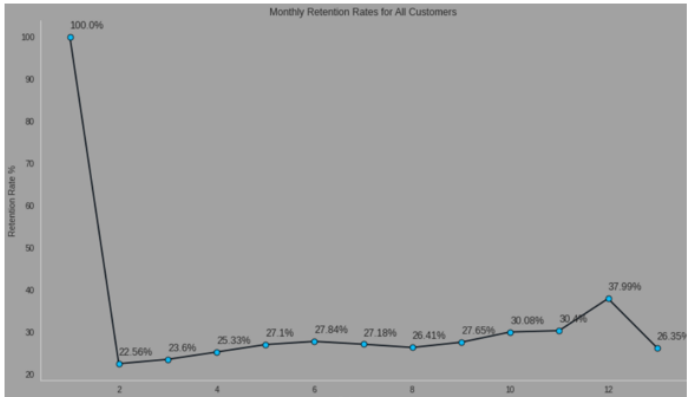

Fig. 8. Monthly Retention Rate for all customers

- Retention rate increases significantly for the last months of the year. Probably, Thanksgiving, Black Friday and Christmas cause it. There are lots of special days at the end of the year.

C. Pareto Principle

- The Pareto principle states that for many outcomes, roughly 80 percent of consequences come from 20 percent of causes (the "vital few"). Other names for this principle are the 80/20 rule, the law of the vital few, or

the principle of factor sparsity. This principle serves as a general reminder that the relationship between inputs and outputs is not balanced. Unlike other principles, the Pareto Principle is merely an observation, not a law. Although broadly applied, it does not apply to every scenario. The Pareto Principle can be applied in a wide range of areas such as manufacturing, management, and human resources. The Pareto Principle can be applied especially in those businesses that are client-service based.

- We have two hypotheses:
  1) 80 percent of the company's revenue comes from 20 percent of total customers.
  2) 80 percent of the company's revenue comes from 20 percent of total products.
  Two functions for calculation and visualization has been used.
- prepare-pareto-data() - finds individual revenue per customer/product and calculates the cumulative percentage of them.
- create-pareto-plot() - takes the output from these data and visualizes it.

Pareto Chart for Customers:

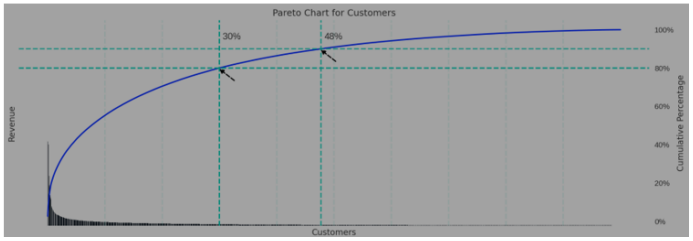

Fig. 9. Pareto chart for customers

We can see that 80 percent of the company's revenue comes from the top 30 percent of customers. Also, 90 percent of the company's revenue comes from the top 48 percent of customers.

Pareto Chart for Products:

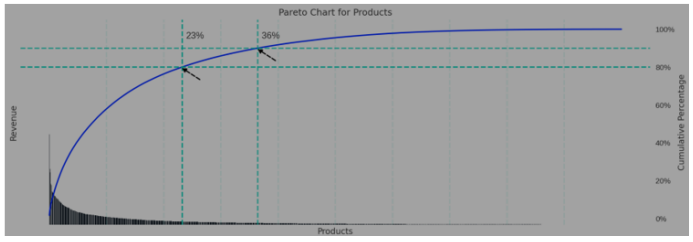

Fig. 10. Pareto chart for products

- We can see that 80 percent of the company's revenue comes from the top 23 percent of products that have the most revenue. Also, 90 percent of the company's revenue comes from 36 percent of the products that have the most revenue. Maybe, if the company reduces half its variety of items, revenue couldn't decrease significantly.

D. Customer Segmentation using RFM Analysis

- RFM analysis is a marketing technique used to quantitatively rank and group customers based on the recency, frequency and monetary total of their recent transactions to identify the best customers and perform targeted marketing campaigns. The system assigns each customer numerical scores based on these factors to provide an objective analysis. RFM analysis ranks each customer on the following factors:
- Recency - How recent was the customer's last purchase? Customers who recently made a purchase will still have the product on their mind and are more likely to purchase or use the product again. Businesses often measure recency in days. But, depending on the product, they may measure it in years, weeks or even hours.
- Frequency - How often did this customer make a purchase in a given period? Customers who purchased once are often are more likely to purchase again. Additionally, first time customers may be good targets for follow-up advertising to convert them into more frequent customers.
- Monetary - How much money did the customer spend in a given period? Customers who spend a lot of money are more likely to spend money in the future and have a high value to a business. RFM analysis numerically ranks a customer in each of these three categories, generally on a scale of 1 to 5 (the higher the number, the better the result). The "best" customer would receive a top score in every category.
  The Segments in RFM Analysis are:
- Champions: Bought recently, buy often, and spend the most
- Loyal customers: Buy on a regular basis. Responsive to promotions.
- Potential loyalist: Recent customers with average frequency.
- New customers: Bought most recently, but not often.
- Promising: Recent shoppers, but haven't spent much.
- Needs attention: Above average recency, frequency, and monetary values. May not have bought it very recently though.
- About to sleep: Below average recency and frequency. Will lose them if not reactivated.
- At risk: Some time since they've purchased. Need to bring them back! Can't lose: Used to purchase frequently but haven't returned for a long time.
- Hibernating: Last purchase was a long back and low number of orders. May be lost.
- 15 percent of customers are considered Champions. These customers are responsible for a big share of your revenue so we can put a lot of effort into keeping improving their experience. What we can do: Give them something extra that the regulars do not get, for example, limited series of products or special discounts to make them feel valued. Use communication similar to the "Loyal" segment. For example, making them ambassadors, giving

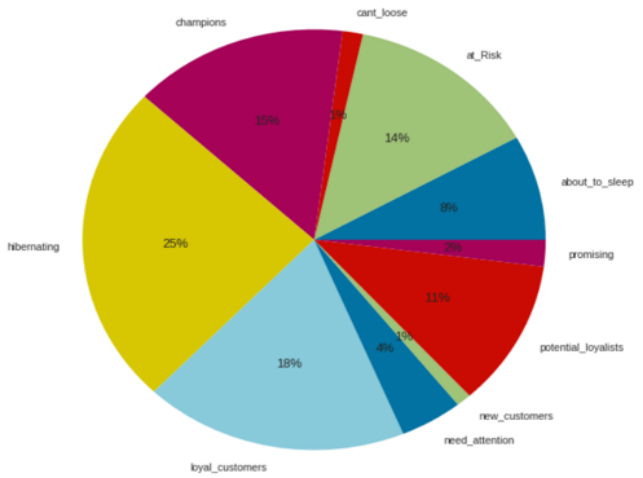

Fig. 11. Customer Segmentation using rfm analysis

them a margin of your profits for bringing you, new customers. Ask them for feedback as they might know the products and services very well.

E. Customer Segmentation using K-Means

- K-Means Clustering is an unsupervised learning algorithm that is used to solve clustering problems in machine learning or data science. It allows us to cluster the data into different groups and is a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

Algorithm 1 k-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

Fig. 12. KMeans Algorithm

- To determine the optimal number of clusters, We will use the elbow method and the silhouette method.
- Elbow method only takes euclidean distance into account whereas the silhouette score takes variance, skewness, and high-low differences (multiple criteria) into account. Hence silhouette method is more accurate
- From the KMeans clustering, we can sort every customer into 4 different clusters that seem to have different behaviors.
  Cluster 0: "Punctual customers" - Those are the customers that buy more punctual items on the website.
  Cluster 1: "Hibernating customers" - Those are the

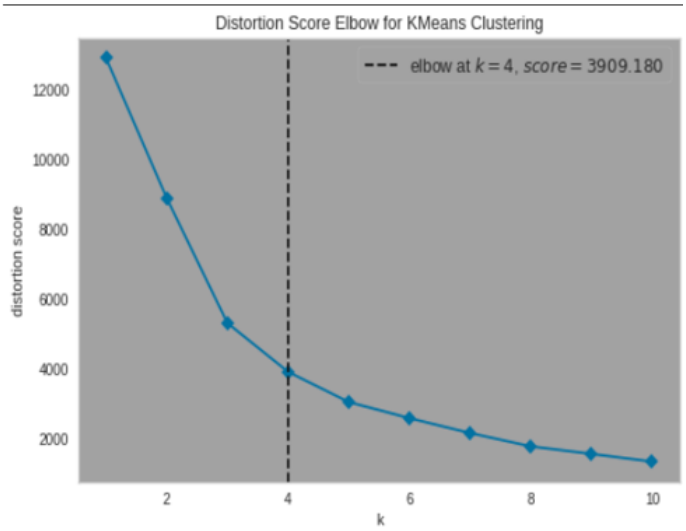

Fig. 13. Finding optimal k using Silhouette's method

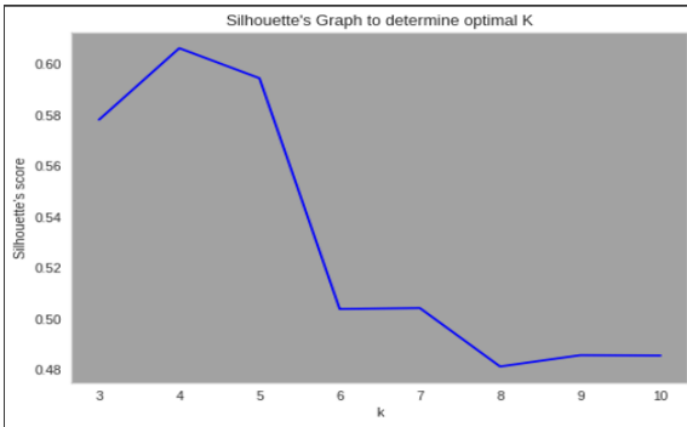

Fig. 14. Finding optimal k using Silhouette's method

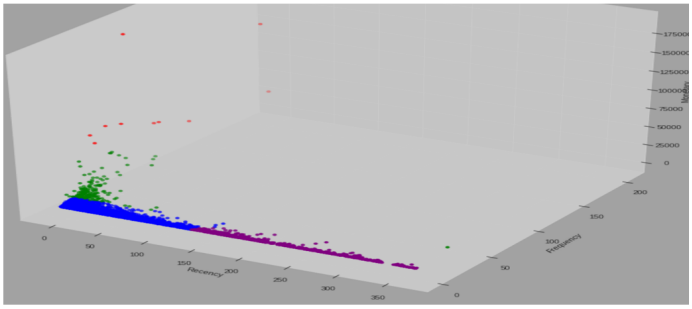

Fig. 15. Customer segmentation using KMeans

customers that buy at the lowest frequency, the lowest recently and that spend the least money.
Cluster 2: ”Exceptional customers” - Those are the customers that we want to keep, that buys at the highest frequency, the most recent and that spend the most money.
Cluster 3: ” Recent customers” - Those are customers that have been active quite recently that might be interesting to keep stimulated.

F. Apriori Algorithm

- Apriori Algorithm usually contains or deals with a large number of transactions. For example, customers buy a lot of items from a retail store, by applying this algorithm the retail store owner can enhance their sales performance and could work effectively. It is also very effective in the fields of education and healthcare. To perform this algorithm, one should know about association rules because it is the most important and well-explored method for knowing the weak or the strong relationships among variables in a huge database and information. Apriori Algorithm has the property that helps to improve the efficiency level by reducing the search space.
- The algorithm involves selecting the frequent itemsets purchased based on a minimum support threshold value and once we obtain the largest frequent itemset we convert them into association rules.
- Association Rules tell you that if you buy product X, how likely you are going to buy product Y.
- There is an additional measure called confidence. The confidence tells you a percentage of cases in which this rule is valid. 100 percent confidence means that this association always occurs; 50 percent for example means that the rule only holds 50 percent of the time.
- Once you have obtained the rules, the last step is to compute the lift of each rule. According to the definition, the lift of a rule is a performance metric that indicates the strength of the association between the products in the rule.
- If the lift of a rule is 1, then the products are independent of each other. Any rule that has a lift of 1 can be discarded.
  If the lift of a rule is higher than 1, the lift value tells you how strongly the consequent(righthand side product) depends on the antecedent (left-hand side).

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(60 teatime fairy cake cases)	(pack of 60 pink paisley cake cases)	0.035644	0.035521	0.015206	0.426598	12.009794	0.013940	1.682028
1	(pack of 60 pink paisley cake cases)	(60 teatime fairy cake cases)	0.035521	0.035644	0.015206	0.428076	12.009794	0.013940	1.686162
2	(60 teatime fairy cake cases)	(pack of 72 retrospot cake cases)	0.035644	0.051404	0.017422	0.488774	9.508552	0.015590	1.855531
3	(pack of 72 retrospot cake cases)	(60 teatime fairy cake cases)	0.051404	0.035644	0.017422	0.338922	9.508552	0.015590	1.458763
4	(alarm clock bakelike green)	(alarm clock bakelike ivory)	0.042231	0.026040	0.015267	0.361516	13.882899	0.014167	1.525425
...	...	...	...	...	...	...	...	...	...
211	(lunch bag cars blue, lunch bag pink polkadot)	(lunch bag red retrospot)	0.024378	0.068887	0.015637	0.641414	9.311109	0.013957	2.596625
212	(lunch bag red retrospot, lunch bag pink polka...	(lunch bag cars blue)	0.028995	0.054235	0.015637	0.539278	9.943285	0.014064	2.052789
213	(lunch bag cars blue)	(lunch bag red retrospot, lunch bag pink polka...	0.054235	0.028995	0.015637	0.288309	9.943285	0.014064	1.364362
214	(lunch bag red retrospot)	(lunch bag cars blue, lunch bag pink polkadot)	0.068887	0.024378	0.015637	0.226988	9.311109	0.013957	1.262105
215	(lunch bag pink polkadot)	(lunch bag cars blue, lunch bag red retrospot)	0.052204	0.025425	0.015637	0.299528	11.780963	0.014309	1.391313

Fig. 16.

G. Revenue Forecasting using Prophet Algorithm

It is an algorithm to build forecasting models for time series data. Unlike the traditional approach, it tries to fit additive regression models a.k.a. ‘curve fitting’.
The cool thing about this algorithm is that it is very flexible when it comes to the data that is fed to the algorithm. You can have missing values and don’t need to have all the dates and times lined up.
And, it works pretty reasonably by default, without setting any parameters explicitly.

The Prophet algorithm is an additive model, which means that it detects the following trend and seasonality from the data first, then combines them together to get the forecasted values.
1)Overall Trend
2)Yearly, Weekly, Daily Seasonality
3)Holiday Effect
From the trend and seasonality detected by the model, we can gain quite a lot of insights from the model.

The Prophet model has the form y(t)=g(t)+s(t)+h(t)+e, where:
g(t) is the trend function.
s(t) is the periodic component (seasonalities).
h(t) represents holidays/events which occur on a potentially irregular basis.
e is the error term (which is often assumed to be normally distributed)

Stationarity of the series is tested using Augmented Dickey-Fuller Test and based on that differencing is done to convert it from a non-stationary to a stationary series.

This is followed by the decomposition of the series into trend, seasonal and residual components.

The forecasting metrics used in our model are R-squared score, Mean Square Error, Root Mean Square Error, and Pearson’s Correlation Coefficient.

Initially a simple model was built but due to its failure to capture huge peaks in the graph, more parameters like holidays and seasonality components are added to forecast more accurately.

Initially only weekly seasonality is enabled (yearly seasonality is not present as our data set is of one year).

80 percent of the data set was used for training and the rest for validation. The model was evaluated using the accuracy metrics mentioned earlier and the best model (after some modifications) was used for testing.

Yearly seasonality is enabled when forecasting future sales and we get the graph as shown below.

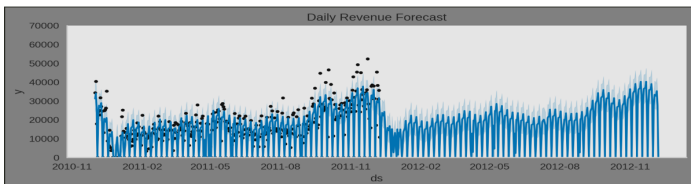

Fig. 17. Revenue forecasting using Prophet Algorithm for next year

In the above graph the Black dots represent actual revenue, and the blue lines represent forecasts.

We can infer that in the last months of the year, our predictions are on the border of the upper confidence interval, or sometimes outside of it. If we look at next year's forecast, we can see minor peaks in the middle of the year, and also we have a positive trend late in the year.

One possible reason for this is there are more holidays at the end of the year and customers buy more items for their loved ones from the retail store.

REFERENCES

[1] P. P. Pramono, I. Surjandari and E. Laoh, "Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method," 2019 16th International Conference on Service Systems and Service Management (ICSSSM), 2019, pp. 1-5, doi: 10.1109/ICSSSM.2019.8887704.

[2] N. M. N. Mathivanan, N. A. M. Ghani and R. M. Janor, "Analysis of K-Means Clustering Algorithm: A Case Study Using Large Scale E-Commerce Products," 2019 IEEE Conference on Big Data and Analytics (ICBDA), 2019, pp. 1-4, doi: 10.1109/ICBDA47563.2019.8987140.

[3] Feixiang Gong, Ninghui Han, Dezhi Li and Shiming Tian, "Trend Analysis of building Power consumption model based on the Prophet Algorithm," China Electric Power Research Institute Co., Ltd Institute of Power Consumption and Energy Efficiency Beijing, China