School of Computer Science & Electronic Engineering

Topic:

Customer Life Time Value (CLTV) Integrated with Recommendation System with Customer Segmentation and Migration

Submitted as a part of:

CE901-7- MSc Project and Dissertation

Author:
KEERTHAN BORAIAH

(Registration: 2200420)

Supervisors:
Dr. Ian Daly

Dr. Luca Citi

Finished on August 29, 2023

TABLE OF CONTENTS

Page-No: -

GITLAB LINK OF CODE: -

https://cseegit.essex.ac.uk/22-23-ce901-ce902-su/22-23_CE901-CE902-SU_boraiah_keerthan//blob/6a19d3e68c8b5e2cc4b8b9438a678d432311bacd/Customer_Life_Time_value__CLTV__Integrated_with_Recommendation_system_and_Customer_Segmentation_and_Migration.ipynb

Customer Life Time value (CLTV) Integrated with Recommendation system and Customer Segmentation and Migration.ipynb

I. ABSTRACT

The project customer life time value (CLTV) Integrated with Recommendation system with customer segmentation and Migration is a machine learning project based on supermarket's customer movement, sales improvement, understanding the customer purchase behaviour, items or products sales behaviour, their spending, revenue increase and recommendation of items to customers. The aim of the project is to advance the supermarket/vendor's technique and tactics to identify their customers, how to maintain their customers, how to increase their sales, how to keep customers and make them regular aka loyal customers and consequently improve the business methods and scheme based on statistical and machine learning approach. The project also aims to nurture low value customers aka (Customers who are not regular or customers who don't spend more to increase the business revenue) and also improve the customer behaviour who are high value customers (Customers who are loyal or regular or spend more to improve the business revenue).

II. INTRODUCTION

The project consists of three different sub category or three different working models

1. Customer Life Time Value (CLTV)
2. Recommendation System
3. Customer Segmentation and Migration

- **CUSTOMER LIFE TIME VALUE (CLTV)**

Customer life time value (CLTV) as the name suggest, it is a machine learning approach to identify what's the customer ranking, what's the customer total spending, what was their previous spending, what is the revenue they generated till now and what revenue will they generate in the future for the business.

There are many type of customers who purchases and shops in a super market or a small vendors, each having different item purchases, each having different money spend and all of them have unique features and behaviours. Let's say, Customer 1 have a purchase history of 1 month and spend over 1000 pounds, customer 2 have a purchase history of 10 months and spend over 150 pounds per month, Customer 3 have purchase history of only 5 days and spend over 50 pounds in these 5 days. Each have different between first purchase and last purchase, each having difference between their spending on the business. If these trends could be identified and make the machine learning approach and algorithm to identify the trends, we can predict what's the customer total life time value, what the customer will spend next and how much the customer will spend in the next days on the business etc. This problem statement can be solved by recording the multiple data of customer behaviour, their spending, their total tenure of purchases and their items purchased. Let's say a customer have 10 months spending, their items purchased, total revenue generated each month. This data could be used to segregate and make the machine learning algorithm learn the pattern and trends in the purchase which is a time series problem to predict or forecast what will be the next tenure revenue this customer will generate for the business.

Here CLTV will predict the next 3 months revenue generated of each customer in the supermarket.

- **RECOMMENDATION SYSTEM**

Recommendation system is a sort of information filtering system that offers users individualized suggestions or recommendations that are based on their interests, activities, and previous interactions. It also helps in consumers find related to objects, products, material, or services that they might be interested in. The recommendation system algorithms usage is widespread across the online platforms, including social media sites, streaming services, and e-commerce websites. In order to anticipate what products a user might like or find useful or interesting to purchase or interact, they utilize a variety of algorithms and approaches to assess user data and item properties. Recommendation system entirely depends on the quality of the data recorded/available, algorithms used to solve them because these systems play crucial role in user experience; increasing their engagement and helping business increase their revenue. Let's say for example, customer 1 purchases milk along with bread, customer 2 purchases the same thing, so the frequency of the items, confidence score and the support score between these two items increases which in turn will be used as an example to recommend bread to a new customer if he have milk in his basket. Recommendation system has been put to use in streaming service, you tube video recommendations, and can be used for supermarket items/grocery recommendation which will benefit and increase the business revenue.

Here in this project the recommendation system will work by recommending only the items which are frequently purchases with other items.

- **CUSTOMER SEGMENTATION AND MIGRATION**

Customer segmentation is a technique proposed to identify the type of customer based on their spending, and segmenting them into groups/clusters. The purpose of customer segmentation is to help the businesses understand the customers better in various demands, preferences, and heir behaviours of various consumer groups so that they can strategize and develop marketing plans, merchandise, and services that are specifically targeted to each segment.

Customer migration is a technique an extension of customer segmentation, the working of customer migration is what's the number of customer shifted from one cluster to another cluster. Let's say for example, There is a high value customer and low value customer group based on their spending on supermarket, and some of the customer who are in low value group starts to spend more and move to high value customer cluster which vice versa happens that means many transitioned from high to low and low to high, this data will tell the customer movement and how it is affecting the business revenue.

The Customer segments that are named in this project are

- Low-Value Customer who are spending between 2 pounds to 50 pounds over a year period
- Upper-Low-Value Customer who are spending between 50 pounds to 100 pounds
- Average Customer who are spending between 100 pounds to 150 pounds
- High-Value Customer who are spending between 150 pounds to 250 pounds
- Upper-High-Value Customer who are spending between 250 pounds to 550 pounds

III. UNDERSTANDING THE DATASET

A single dataset was used to implement for all the three different models but the attributes in the dataset which are used in the models are always different.

Dataset Shape: - the dataset contains 38765 recorded samples rows of all the customers shopping behaviour and 5 columns

The structure of the dataset compromises of 5 different features/attribute

1. **Member Number**
2. **Date**
3. **Item Description**
4. **Sales**
5. **Year**

1. **Member Number: -** Member number as the name suggests is the unique number given to individual customer to identify each and every customer. This feature is crucial for customer life time value.
2. **Date: -** Date is a feature where on what date did the unique number for customer shopped and purchased the items.
3. **Item Description: -** This feature is mainly used for recommendation system where all the details and what products or items where purchased from the supermarket for specific customers
4. **Sales: -** The feature records total amount spend by the customer purchasing the products from the supermarket and this feature is mainly used for Customer life time value and customer segmentation.
5. **Year: -** Year is a new attribute created for my personally to solve some problem statements like this yearly sales customer segmentation, and yearly migration. Mainly used for customer segmentation and migration models.

IV. MACHINE LEARNING APPROACH

Details of What is machine learning approach, what type of problem is this project, what can machine learning approach solve this problem statement, how can machine learning help us solve these problem statements.

- **What is machine learning approach?**

Machine learning approach involves using algorithm, statistics, modelling to establish the computer to under the data, its pattern and its specific task means the machine learning can be used to make it possible for computers to become better at certain specific task by learning from examples and data rather than just using explicitly or manually programming it.

- **What type of problem is this project?**

Machine learning approaches can encompass various types of algorithms, including:-

- **Supervised Learning**: - Algorithm learns from already derived dependent and independent variables.

- **Unsupervised Learning**: - Algorithm learns from patterns and structures instead of derived dependent and independent variables.
- **Semi-Supervised Learning**: -A combination of labelled and unlabelled data is used to building the model.

This project is a supervised machine learning as we already know the dependent variable and feeding the algorithm the independent variable.

- **What is the type of machine learning problem?**

Machine learning approach can solve various types of problems across the domain and some of the common problem categories are:

- **Regression: -** predicting a continuous numerical value based on the provided input data.
- **Classification: -** predicting non continuous categorized data based on the input data.
- **Time Series Analysis:** - Predicting /forecasting data points based on the ordered over time.

Customer life time Value is a Time series analysis problem statement whereas recommendation system is a supervised learning algorithm and Customer segmentation and migration is a unsupervised machine learning.

- **How can machine learning help us solve these problem,?**

Time series data are observations that are recorded at regular intervals with the intention of drawing conclusions, patterns, or predictions from the data where machine learning algorithm can help is predicting and forecasting based on previous historical data recorded

Machine learning algorithm helps in analysing the large dataset, to identify the hidden patterns, and to recommend users with suggesting the products, consequently increasing the user engagement and increasing the business outcomes.

Clustering is a unsupervised learning where ML helps clustering by combine the data points into clusters or segments based on similarities. Clustering helps in the identifying the patterns, and extract insights from the large datasets.

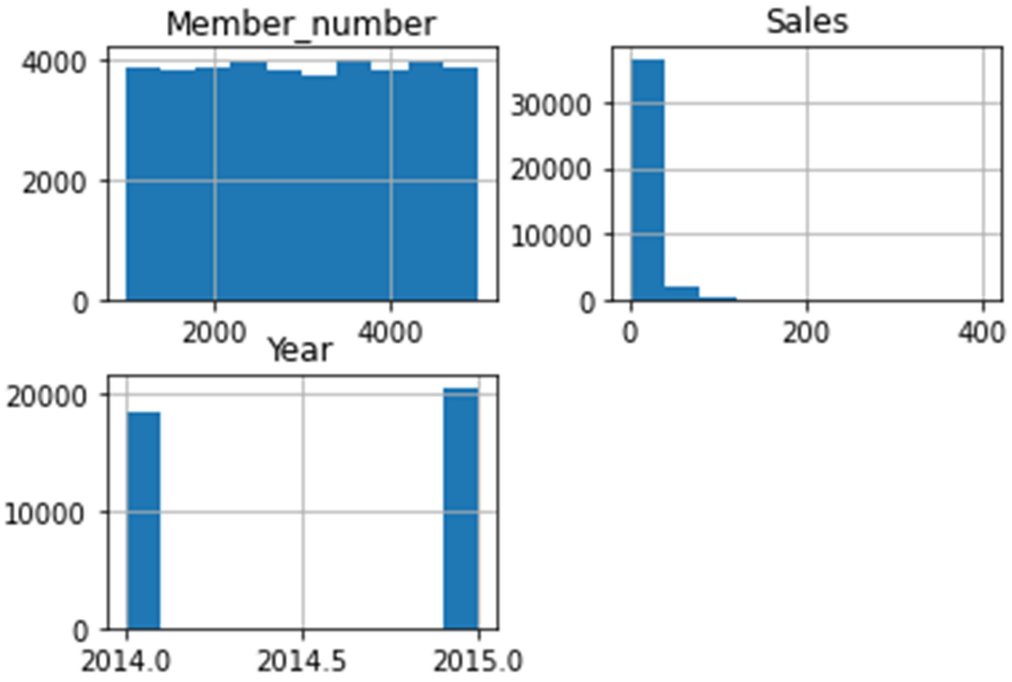
V. DATA PREPROCESSING

Some of the basic Data Pre-processing steps to statistical tests were performed on the time series dataset like

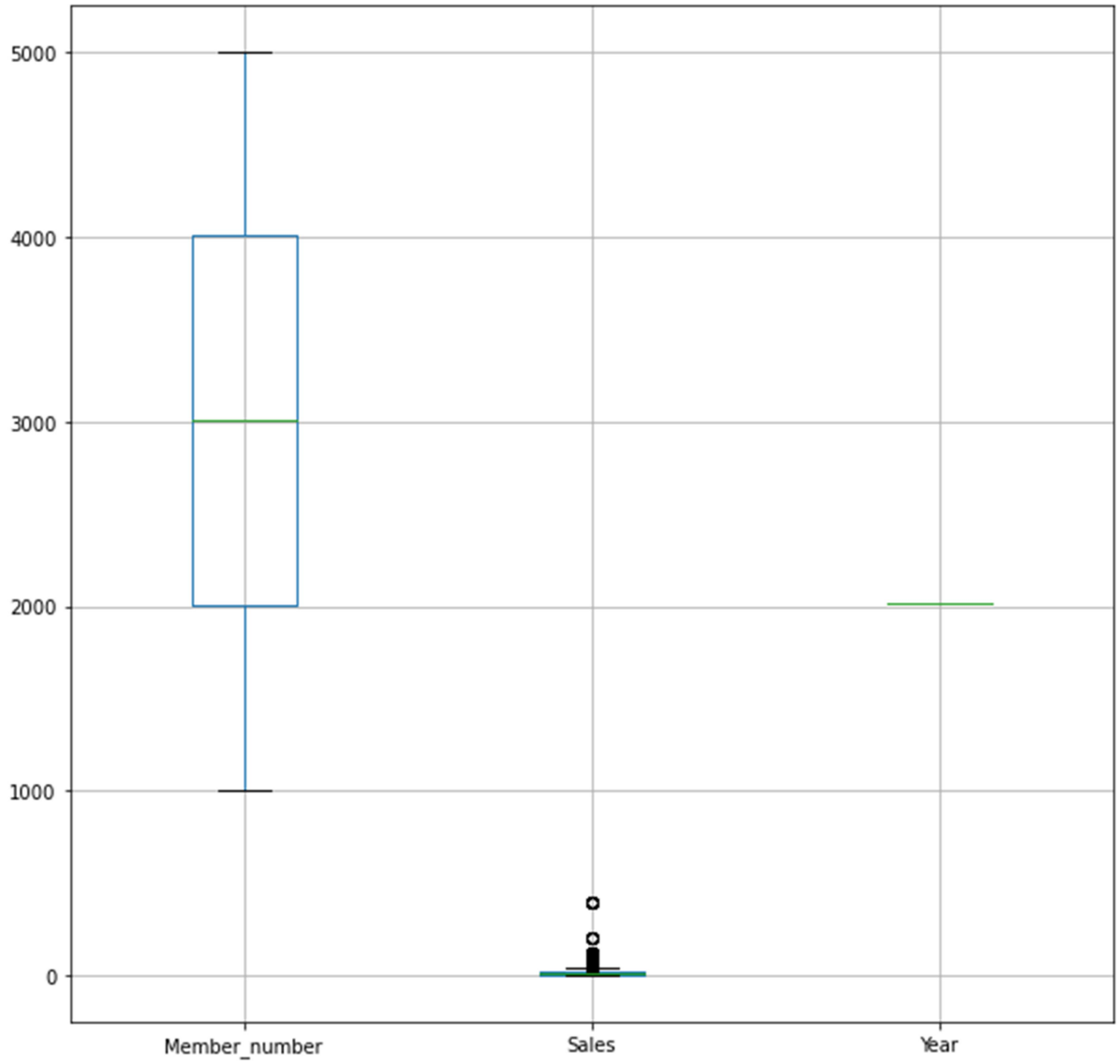
- **Handling Null value: -** Null values are empty data rows or empty recorded data where it can be handled by removing rows or columns with null values, also some methods like imputing them with linear regression or with values like mean, median, and mode, or using advanced techniques like predictive modelling to estimate missing values. Although these can be implemented, it was not necessary as the dataset was complete with no null values.
- **Handling Duplicates:** - Duplicates are repeated values and it is technically wasting processing power to fit the same value twice to models hence Duplicates can be removed to ensure that each rows are unique. However, duplicates sometime contains important information like timeseries dataset, such as repeated measurements. Before removing duplicates, it is very important to

consider the context. In this case we decided not to remove duplicates as it is the time series dataset and removing it would cause temporal aggregation.

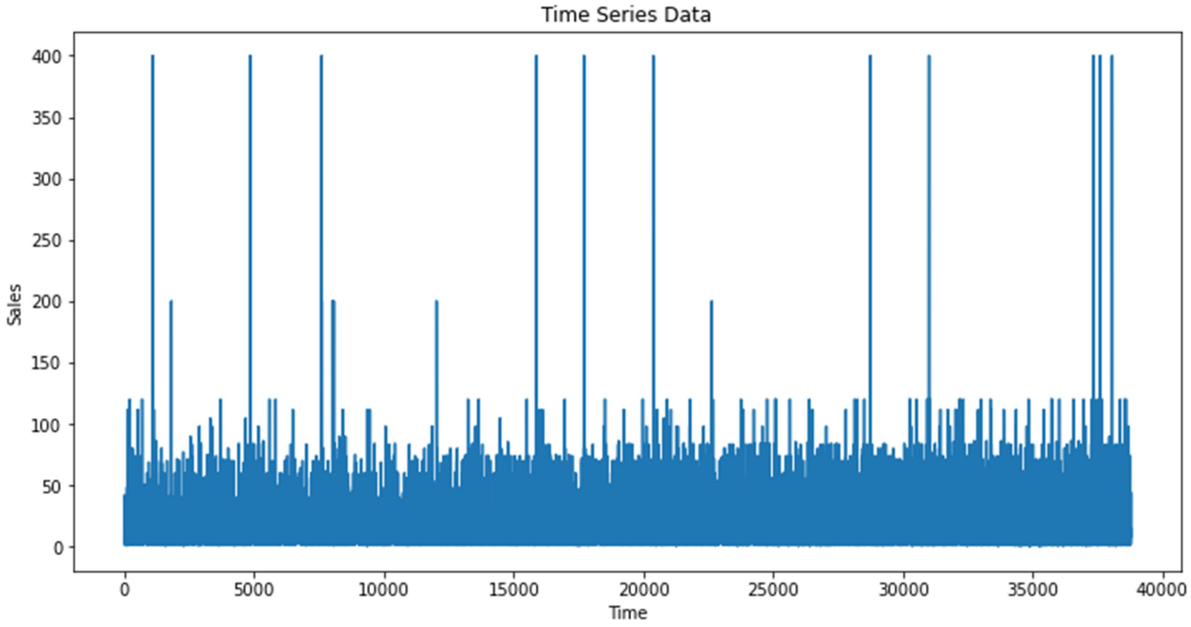
- **Handling Different Data type: -** There are multiple Different data types like (numeric, categorical, text, datetime) and all recorded dataset does not have variable that are its attribute. Encoding categorical variables (one-hot encoding, label encoding), scaling numeric features, and handling text data via tokenization and vectorization area also some of the examples. In our case the datetime attribute was an object type which needs to be changed to datetime.
- **Handling Skewed Data: -** Each and every attribute has to be normally distributed aka the (Bell Curve) The asymmetry in the data distribution is called as to as skewness. if its not distributed correctly then Transformations such as logarithmic, square root, or Box-Cox transformations can be used to make data more normally distributed. The dataset contains only few attribute and the dependent variable should not be converted and left as it is for the model to identify.



- **Handling Outlier: -** Outliers are data points that have significance change from the rest of the data set. Outliers can be removed, transformed, or handled using robust statistical methods that are less sensitive to extreme values. In our case the only outlier are the sales which can be ignored because some items may costs more than other grocery items.



- **Checking Data stationary: -** Stationarity is a time series data constraints in which statistical properties (such as mean and variance) remain constant over time. Time series data can be made stationary using techniques such as differencing. In our case the data was stationary and hence we had to reject the null hypothesis.

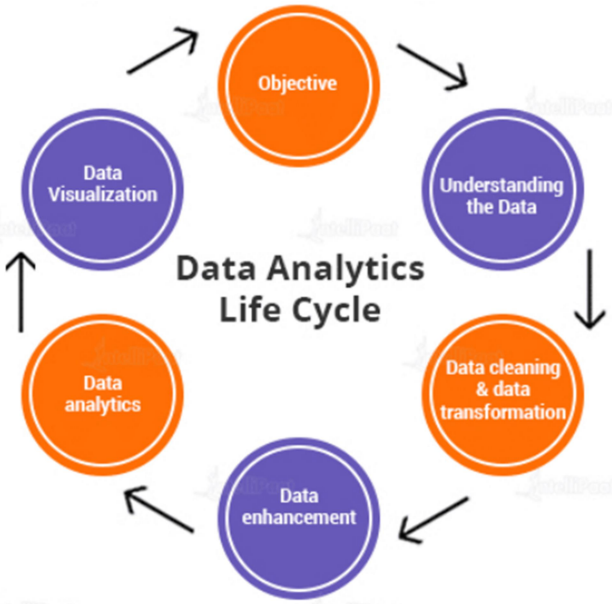


- **Feature Selection: -** Feature selection identifies the most relevant features the modelling while discarding irrelevant ones. This can improve model performance and reduce overfitting. In our case due to all the attributes are necessary for all 3 models but tried to implement and check correlation to know if it's gives any insights or not.

- **Data Conditioning: -** Some of the customers over a period of 1 year have only recorded samples of less than 10 transactions which will give less accuracy if we feed it to the algorithm. Hence data conditioning and constraints come in place where I selected only customers who have more than 20 transactions in a year so the algorithm can identify the pattern in the data for predicting and forecasting.

VI. ARCHITECTURE (MACHINE LEARNING LIFE CYCLE)

I have followed the same architecture life cycle which are been followed in the industrial standards for making this project.



1. Objective
2. Understanding the data
3. Data cleaning and data transformation
4. Data enhancement
5. Data analytics
6. Data visualization

- **Objective: -** This life cycle process involves defining the specific goals and objectives of the data analysis of the project during this phase, any issue are you attempting to resolve or any problem statement needs to be solved, Having specific goals helps to guide the entire lifecycle.
- **Understanding the data: -** In this life cycle, investigating the raw data to gain a better understanding of its structure, content, and quality helps us in identifying potential issues such as missing values, duplicates, and outliers, etc.
- **Data cleaning and data transformation: -** In this life cycle we address data quality issues. Dealing with missing values by imputing or removing them, duplicates, and inconsistencies and the process of converting and standardizing data into a consistent format for analysis and making data consistent.
- **Data enhancement: -** In the life cycle, the process of adding value to the dataset with additional information that can improve analysis and insights is referred as data enhancement. This could entail combining datasets, incorporating external data sources, or creating a new features from existing ones.
- **Data analytics: -** In the life cycle, the main analysis stage is performed where extracting patterns, relationships, and insights from the cleaned and transformed data using statistical methods, machine learning algorithms, or other analytical techniques like hypothesis tests, statistical tests etc are performed.

- **Data visualization: -** In the life cycle, we perform Data visualization where it is one of the easiest or effectively way to explaining and communicating your findings, either by using graphs, charts or visual representations of data, you can communicate complex findings in an understandable and actionable way there.

VII. ALGORITHM USED (ALGORITHM EXPLANATIONS)

- **CUSTOMER LIFE TIME VALUE (ALGORITHM USED)**
  1. ARIMA (Auto Regression Integrated with moving Average)
  2. SES (Simple Exponential Smoothing)
  3. HWES (Holt Winter's Exponential Smoothing)
  4. XGBOOST
- **RECOMMENDATION SYSTEM (ALGORITHM USED)**
  1. APRIORI
  2. ASSOCIATION RULES
- **CUSTOMER SEGMENTATION AND MIGRATION (ALGORITHM USED)**
  1. K-MEANS CLISTERING

o ARIMA (Auto Regression Integrated with moving Average)

ARIMA (Auto Regressive Integrated Moving Average) is one of the popular time series forecasting models that combines autoregressive (AR) and moving average (MA) components while taking in the differencing (I) to make time series data stationary. ARIMA models are particularly effective at capturing linear relationships in time series data for predicting and forecasting.

Auto Regressive (AR) Component: - The autoregressive component in a time series represents the relationship between a current value and its previous values. The current value is referred by the previous values with a certain lag. The number of lag terms included in the model is denoted by AR(p).

Moving Average (MA) Component: - The moving average model has a the relationship between the current value and the previous white noise or error terms. The number of past error terms considered in the model is denoted by MA(q).

Integrated (I) Component: - To achieve stationarity, the integrated component involves differencing the time series. Because many time series models, including ARIMA, assume that statistical properties remain constant over time, stationarity is necessary. The differencing order, denoted by d, represents the amount of times the data must be differenced in order to achieve stationarity.

#REFERENCED FROM WIKIPEDIA FROM HERE

The ARIMA model is denoted as ARIMA(p, d, q), where:

p: Order of the autoregressive component (AR order).

d: Order of differencing (integration order).

q: Order of the moving average component (MA order).

#REFERENCED FROM WIKIPEDIA TILL HERE

- **SES (Simple Exponential Smoothing)**

Simple Exponential Smoothing (SES) is a basic and widely used time series forecasting method that is useful for generating short-term forecasts for time series data that have less or no clear trend or seasonality. SES is based on the concept of exponential smoothing, which assigns decreasing weights to old observations as they move further back in time.

Single Smoothing Parameter (α): -

Single Smoothing Parameter (): The SES method uses a single smoothing parameter, denoted as (alpha), which controls the weight given to the most recent observation as well as the rate at which the influence the observations diminishes. The value of ranges between 0 and 1.

Initialization: -

The forecasting process an initial value or starting point must be selected. This can be done through various methods, such as taking the average of the first few observations.

Recursive Process: -

The process is recursive, means that the forecast for each successive time period is calculated using the forecast for the previous time period and the actual observation for the current time period.

#REFERENCED FROM WIKIPEDIA FROM HERE for formula

Forecast Calculation: -

The forecast for the next time period (t+1) is calculated using the following formula:

α * Actual(t) + (1 - α) * Forecast(t)

Forecast(t) is the forecast for the current time period (t).

Actual(t) is the actual observation for the current time period (t).

The formula combines the weighted average of the most recent observation (Actual(t)) with the previous forecast (Forecast(t)).

#REFERENCED FROM WIKIPEDIA TILL HERE

Updating Smoothing Parameter: -

The value is determined by the responsive data in the recent changes. Smaller values of weight past observations are more, while larger values make the forecast more sensitive to recent observations. The best value is can often be identified through the trial and error or optimization techniques.

Forecasting Horizon: -

Because SES does not account for trends or seasonality in the data, it is primarily used for short-term forecasting. It is only suited to data with few variations.

Evaluation and tuning: -

Metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are used to assess the performance of the SES model. These metrics quantify the forecasting accuracy.

- **HWES (Holt Winter's Exponential Smoothing)**

Holt-Winters Exponential Smoothing (HWES) is a method that is similar to the Simple Exponential Smoothing (SES) this method includes additional components to capture trend and seasonality in time series data. HWES incorporates three forecasting components: level (α), trend (β), and seasonality, additive and multiplicative.

Level (α): -

The level component represents the time series data's underlying baseline value. It is determined, as in SES, by the smoothing parameter, which controls the weight given to the most recent observation versus previous observations.

Trend (β): -

The trend component captures the time series data's overall direction and rate of change. A trend smoothing parameter is used for the model. The forecasted value is influenced by both the current level and the change in level over time.

Seasonality: -

The seasonality component accounts for data patterns that repeat at regular intervals. A seasonality smoothing parameter is used for the model. Seasonality can be additive or multiplicative (the difference between periods is expressed as a percentage).

- **XGBOOST**

XGBoost is a machine learning algorithm that can be used for classification and regression tasks. XGBoost's base learners are an ensemble of decision trees based on the features, Decision trees are created to predict the target variable, and XGBoost combines these trees to make final predictions. age).

- **Apriori Algorithm: -**

The Apriori algorithm works similarly to a detective. It examines all of your store's customers' shopping baskets and attempts to identify items that are frequently purchased together. However, it only focuses on things that are frequently purchased together.

It begins by determining how frequently each item (such as bread or milk) is purchased on its own. Then it groups similar items and examines how frequently they are purchased together. When a pair of items is purchased together far more frequently than you'd expect by chance, the algorithm takes notice. It will continue to do this for larger groups of items until it is certain about the ones that are frequently purchased together.

- **Association Rules: -**

Now that the Apriori algorithm has discovered these frequently occurring pairs or groups of items, it creates rules based on what it has discovered.

If the algorithm discovers that many people who buy chips also buy soda, it creates an association rule that states, "If you buy chips, you might also want to buy soda." This way, you can place chips and soda near each other, and when customers see chips, they'll remember the rule and might buy something.

- **K-MEANS CLISTERING**

K-Means is a data sorter that takes a large amount of data and divides it into groups based on similarities. It begins by guessing where the groups should be, then adjusts those guesses until the data are in the correct groups and the groups no longer change significantly. It aids in the organization of similar items into their own small groups.

VIII. RESULTS

CUSTOMER LIFE TIME VALUE (CLTV) RESULTS

- **ARIMA: -**

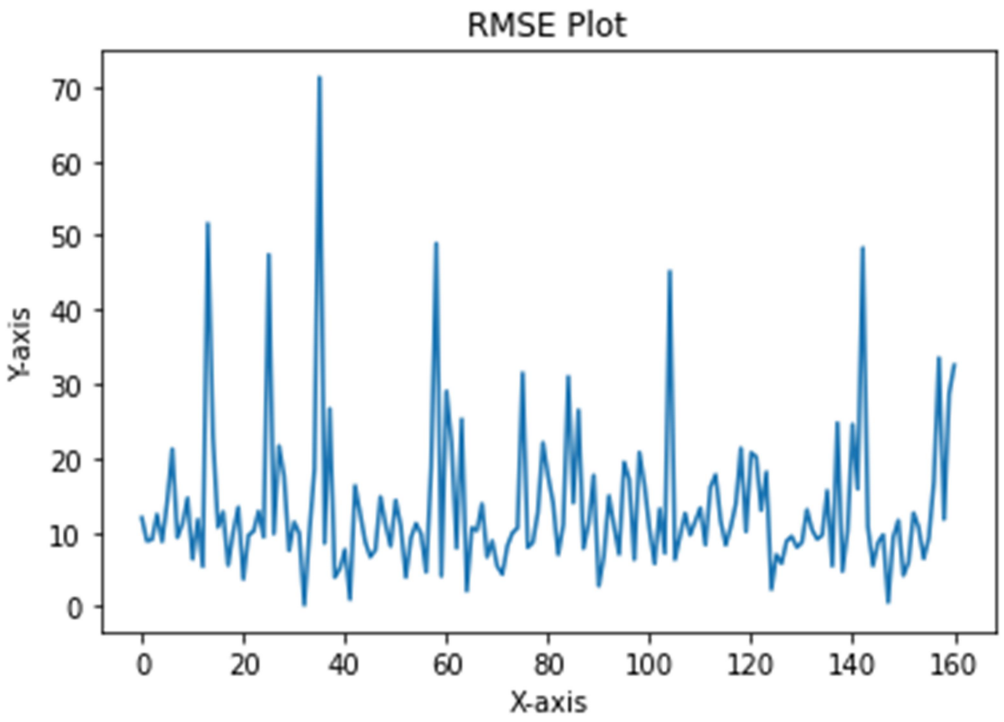
Each customer forecasted and RMSE is calculated for each and every customer

Some examples of each customer forecasted is: -

```
The RMSE for Forcasted Customer 4941 is 11.974430366258867
The RMSE for Forcasted Customer 2193 is 8.8362389285867
The RMSE for Forcasted Customer 1997 is 9.01953765523751
The RMSE for Forcasted Customer 2421 is 12.428152640300464
The RMSE for Forcasted Customer 1905 is 8.766082323130163
The RMSE for Forcasted Customer 4783 is 14.1647100718013
The RMSE for Forcasted Customer 3709 is 21.24330907945537
The RMSE for Forcasted Customer 4272 is 9.305697413959948
```

The Total RMSE is:-

```
The Total RMSE of All Customer is for ARIMA: - 13.422907825203396
```



- **SES: -**

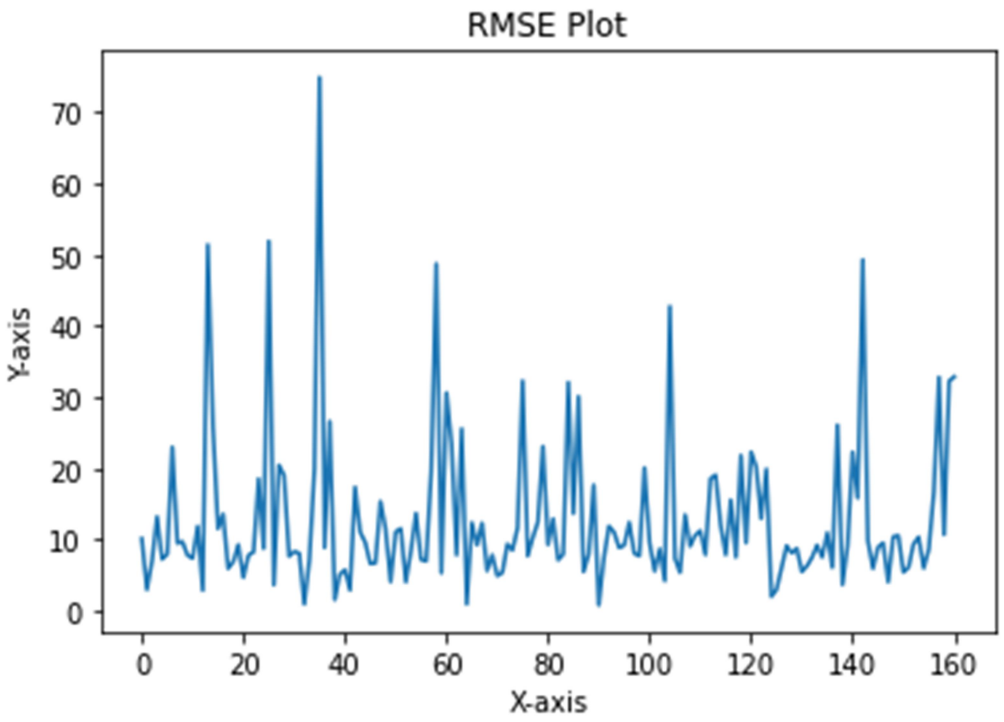
Each customer forecasted and RMSE is calculated for each and every customer

Some examples of each customer forecasted is: -

```
The RMSE for Forcasted Customer 4941 is 10.1501094387984
The RMSE for Forcasted Customer 2193 is 2.995538265146454
The RMSE for Forcasted Customer 1997 is 7.002921758509045
The RMSE for Forcasted Customer 2421 is 13.235110371306062
The RMSE for Forcasted Customer 1905 is 7.273773856754359
The RMSE for Forcasted Customer 4783 is 7.96547397699096
The RMSE for Forcasted Customer 3709 is 23.006277440809296
```

The Total RMSE is:-

```
The Total RMSE of All Customer is for SES- 12.806686303856837
```



- **HWES: -**

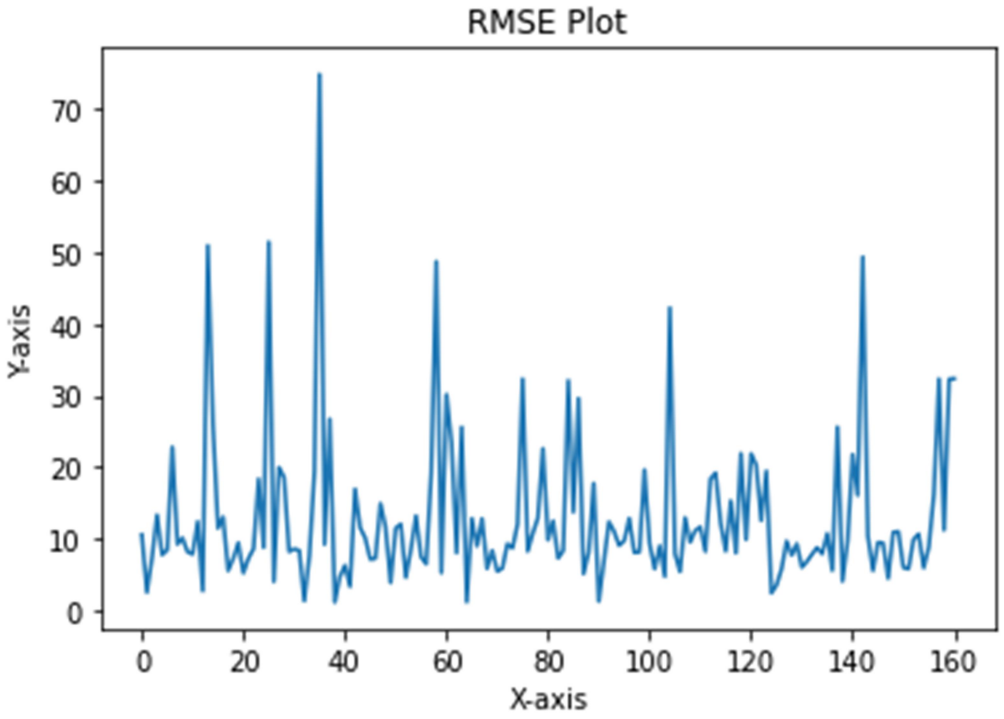
Each customer forecasted and RMSE is calculated for each and every customer

Some examples of each customer forecasted is: -

```
The RMSE for Forcasted Customer 4941 is 10.584447049690256
The RMSE for Forcasted Customer 2193 is 2.5419616176156894
The RMSE for Forcasted Customer 1997 is 7.3555242842974415
The RMSE for Forcasted Customer 2421 is 13.341816103032686
The RMSE for Forcasted Customer 1905 is 7.790450196447677
The RMSE for Forcasted Customer 4783 is 8.440073901261504
The RMSE for Forcasted Customer 3709 is 22.796517111130623
```

The Total RMSE is:-

```
The Total RMSE of All Customer is for HWES- 12.896545359347424
```



- **XGBOOST: -**

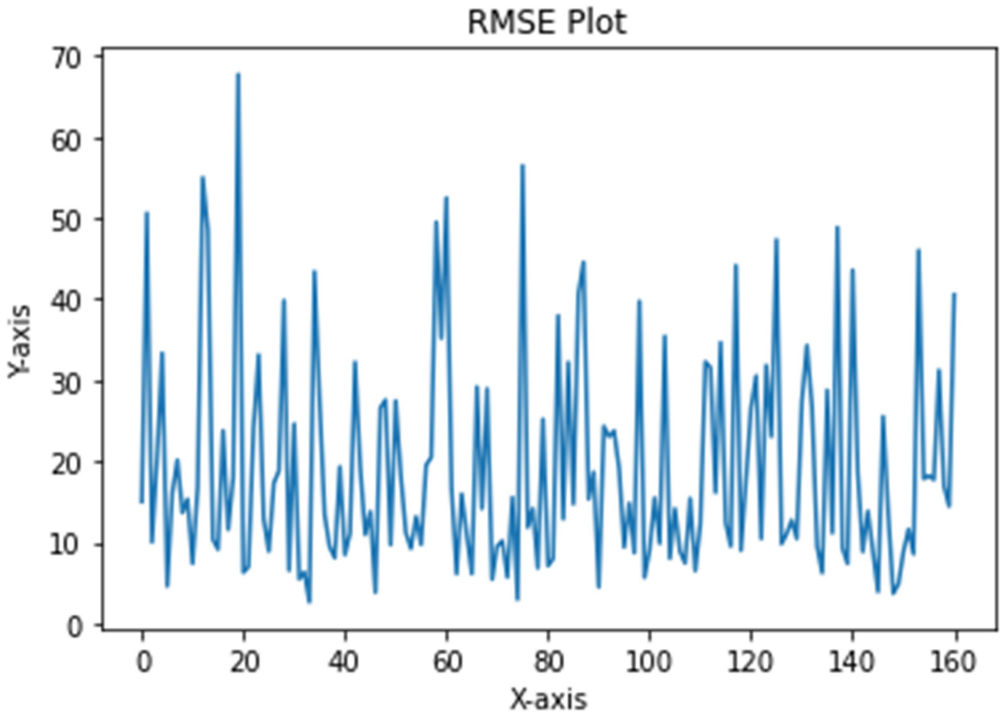
Each customer forecasted and RMSE is calculated for each and every customer

Some examples of each customer forecasted is: -

```
The RMSE for Forcasted Customer 4941 is 15.025671295948323
The RMSE for Forcasted Customer 2193 is 50.62559162164386
The RMSE for Forcasted Customer 1997 is 10.007808442981279
The RMSE for Forcasted Customer 2421 is 20.466368478343757
The RMSE for Forcasted Customer 1905 is 33.342296918295375
The RMSE for Forcasted Customer 4783 is 4.600656373693438
The RMSE for Forcasted Customer 3709 is 16.155274680781886
The RMSE for Forcasted Customer 4272 is 20.154777495871073
```

The Total RMSE is:-

```
The Total RMSE of All Customer is for XGBoost:- 19.14011987155644
```



- **RECOMMENDATION SYSTEM: -**

All the items confidence and support were calculated which results in identifying items which have both good support and confidence score to be recommended together if one item is purchased.

Example for support score

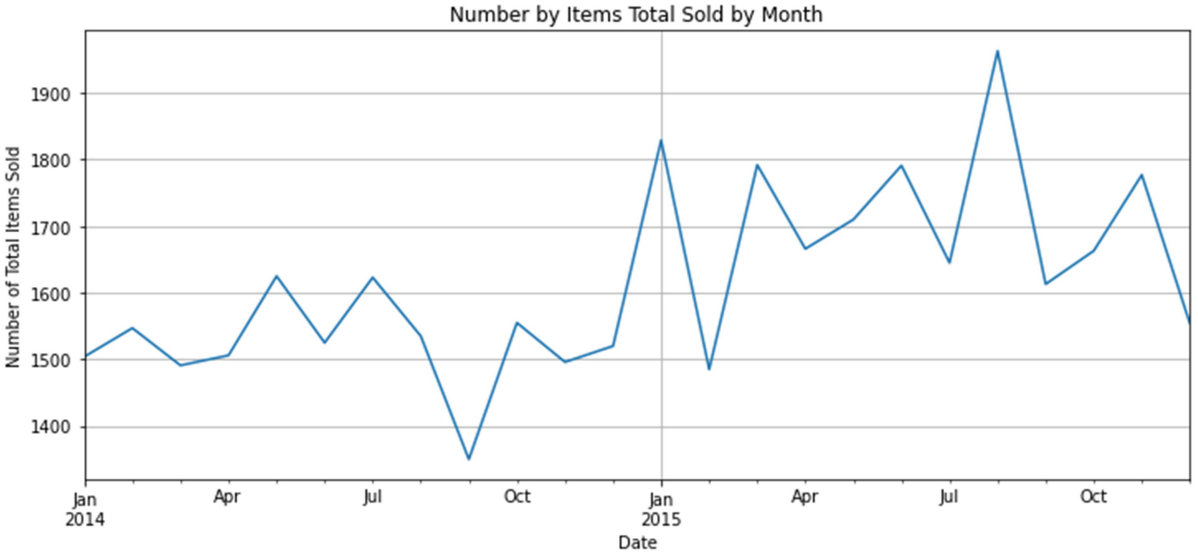
```
Processing 28 combinations | Sampling itemset size 4e 3
```

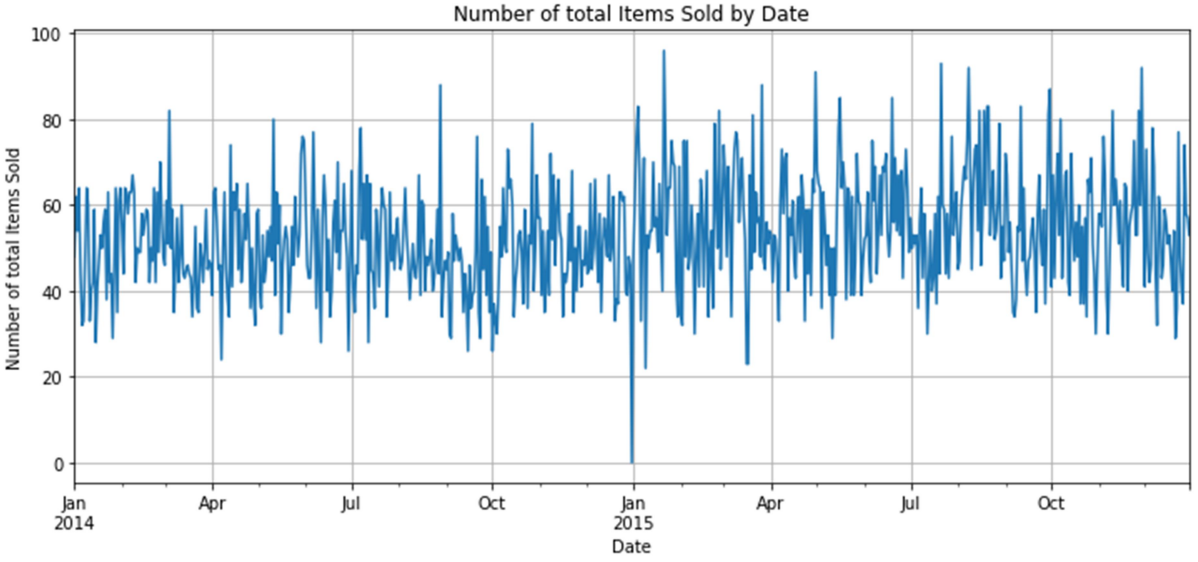
	support	itemsets
0	0.004010	(Instant Food Products)
1	0.021386	(Uht-Milk)
2	0.001470	(Abrasive Cleaner)
3	0.001938	(Artif. Sweetener)
4	0.008087	(Baking Powder)

Example for both support and Confidence Score

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric	
0	(Bottled Water)	(Uht-Milk)	0.060683	0.021386	0.001069	0.017621	0.823954	-0.000228	0.996168	-0.185312	

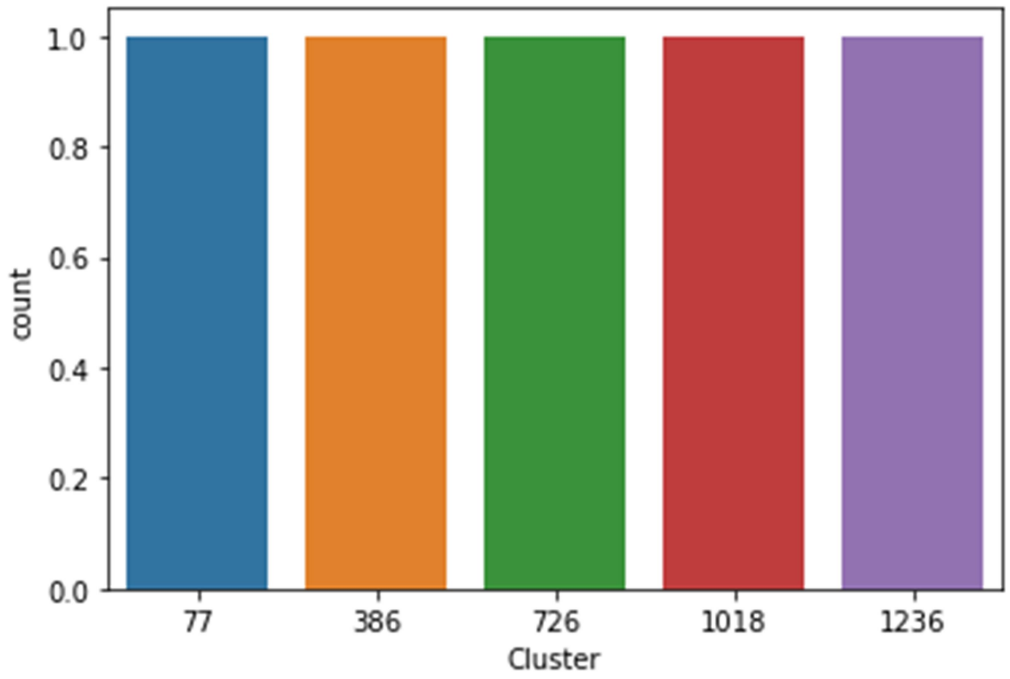
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric	
1	(Uht-Milk)	(Bottled Water)	0.021386	0.060683	0.001069	0.050000	0.823954	-0.000228	0.988755	-0.179204	
2	(Other Vegetables)	(Uht-Milk)	0.122101	0.021386	0.002139	0.017515	0.818993	-0.000473	0.996060	-0.201119	
3	(Uht-Milk)	(Other Vegetables)	0.021386	0.122101	0.002139	0.100000	0.818993	-0.000473	0.975443	-0.184234	
4	(Rolls/Buns)	(Uht-Milk)	0.110005	0.021386	0.001804	0.016403	0.767013	-0.000548	0.994934	-0.254457	



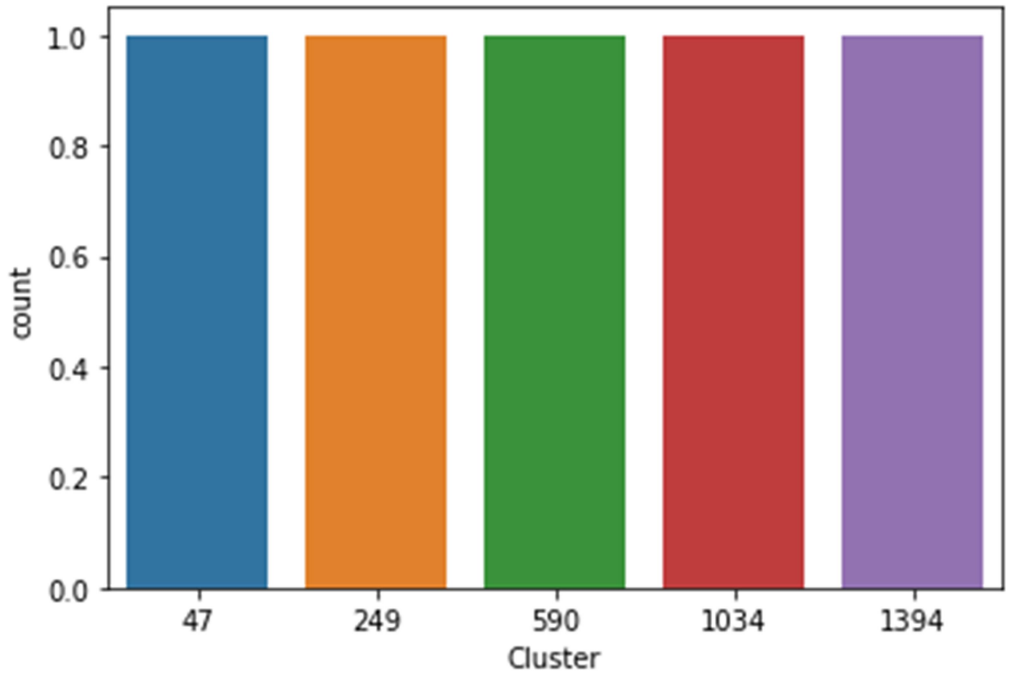


- **CUSTOMER SEGMENTATION AND MIGRATION: -**

Segmentations of customer were created into five different cluster



2014 Cluster counts



2015 Cluster counts

Cluster 0 (Average-Customer):- Customer who are spending above 100 and Maximum 150

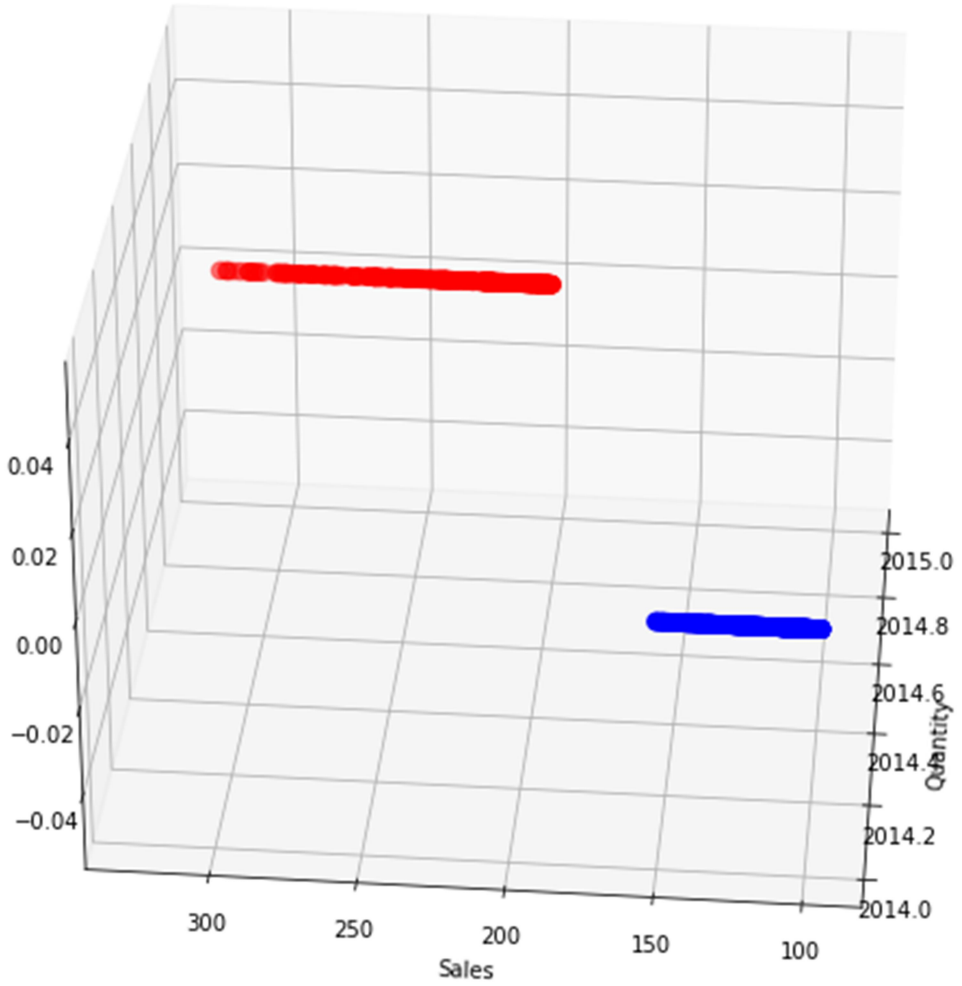
Cluster 1 (Low-Value-Customer):- Customer who are spending above 2 and Maximum 50

Cluster 2 (Upper-High-Value-Customer):- Customer who are spending above 250 and Maximum 550

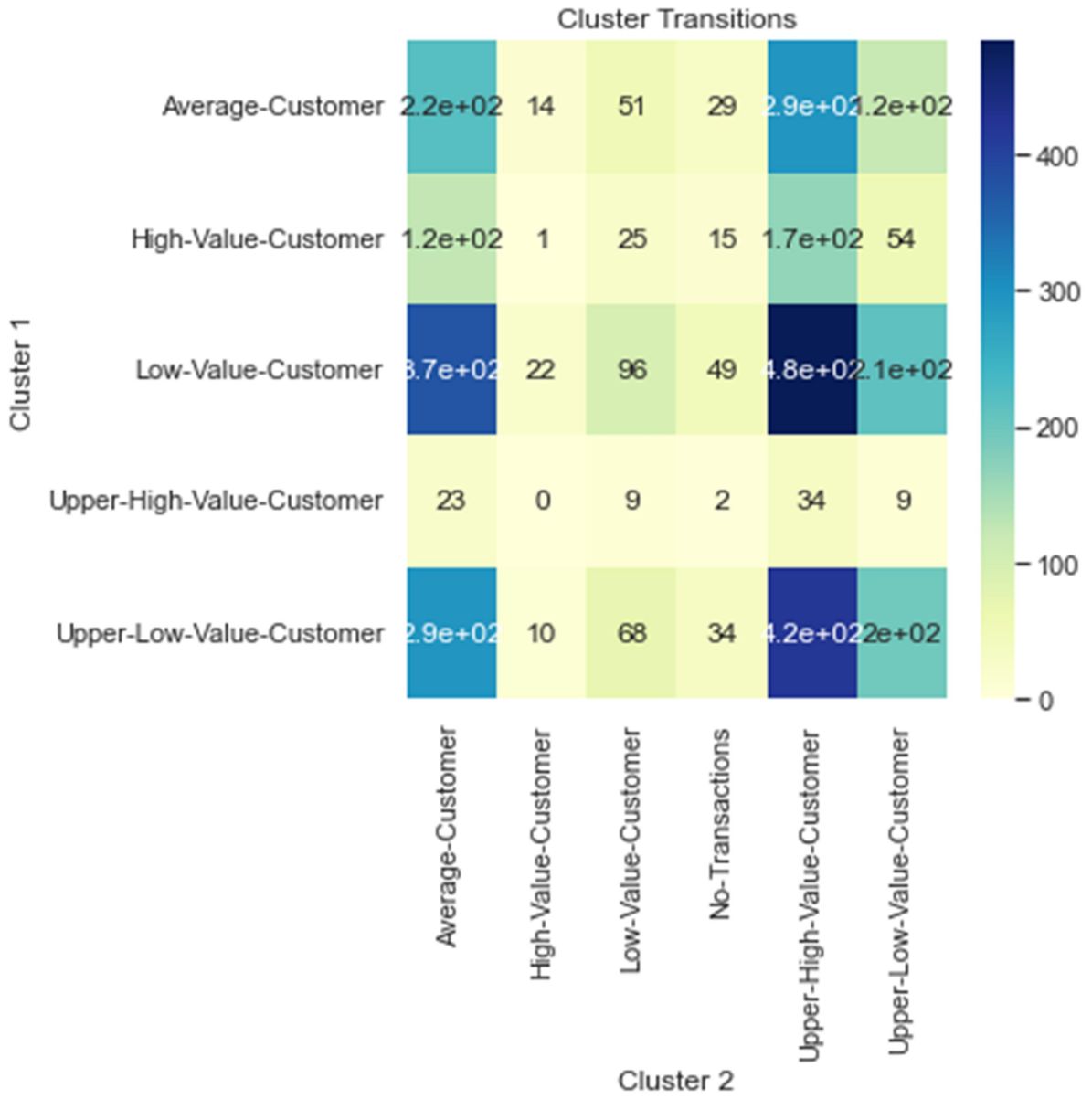
Cluster 3 (High-Value-Customer):- Customer who are spending above 150 and Maximum 250

Cluster 4 (Upper-Low-Value-Customer):- Customer who are spending above 50 and Maximum 100

Furthermore, their migration from 2014 to 2015 were identified as shown in the below figure.



Clustering Way based on sales explanation



Cluster Migration

IX. CONCLUSION

- **CUSTOMER LIFE TIME VALUE (CLTV)**

The model was able to identify the pattern in the and forecast the next sales prices generated by each customers. This was concluded on the basis on random data generated model where, a completely random data which is similar to the original data was generated and fitted to the model, and the result was the original data models got better RMSE than the random data which is similar to the original data had lesser RMSE than the original data, This hypothesis was proved by using the STUDENTS PAIRED T-TEST where To Accept or Reject the Hypothesis We use Students paired T-Test to compare the means, Standard Deviation of between two groups to calculate the critical value "T" to check with the Probability of "P" = 0.05 in the T-table to accept or reject the Null Hypothesis. The Hypothesis was: - There are no significant difference between the "Random" Sales generated compared to the "Original" Sales Data. And the hypothesis was accepted because when calculated the T-value was (0.13) is less than the Critical Value (1.960) so We will accept the Null Hypothesis. There is no Significant difference between the "Random Sales" Generated and the "Original Sales" dataset. The hypotheses proved that the algorithm was able to identify the pattern in the original data and was able to forecast the next coming sales.

Furthermore, ANNOVA test was perform to check weather to get the best performing algorithm by comparing the performance metric (e.g., accuracy, F1 score, RMSE) for each model on the same dataset or set of datasets. The Hypothesis was

Null Hypothesis (H0): There is no significant difference in the performance of the models.

Alternative Hypothesis (Ha): There is a significant difference in the performance of the models.

And the result was The P-value for four different models by comparing the MSE and measure the ratio hence Fail to reject the null hypothesis: There is no significant difference in model performance.

- **RECOMMENDATION SYSTEM**

Recommendation system was able to identify the frequency and the support and confidence score of each and every items resulting in the algorithm able to suggest items when a single items is purchased

- **CUSTOMER SEGMENTATION AND MIGRATION**

Customer segmentation and migration was a unsupervised algorithm where the algorithm were able to identify the pattern and cluster and segment them correctly which resulted in analysing and identifying the different type of customer and their migration from one segment to another segment.

X. REFERENCES

1. Gupta, Sunil, Dominique Hanssens, Bruce Hardie, Wiliam Kahn, V. Kumar, Nathaniel Lin, Nalini Ravishanker, and S. Sriram. "Modeling customer lifetime value." *Journal of service research* 9, no. 2 (2006): 139-155.
2. Jain D, Singh SS. Customer lifetime value research in marketing: A review and future directions. Journal of interactive marketing. 2002 May;16(2):34-46.
3. Malthouse, Edward C., and Robert C. Blattberg. "Can we predict customer lifetime value?." *Journal of interactive marketing* 19, no. 1 (2005): 2-16.
4. Satheesan, Pranavi, Prasanna S. Haddela, and Jesuthasan Alosius. "Product Recommendation System for Supermarket." In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 930-935. IEEE, 2020.
5. Tatiana, Kutuzova, and Melnik Mikhail. "Market basket analysis of heterogeneous data sources for recommendation system improvement." *Procedia Computer Science* 136 (2018): 246-254.
6. Christodoulou, Panayiotis, Klitos Christodoulou, and Andreas S. Andreou. "A real-time targeted recommender system for supermarkets." (2017).
7. Cooil, Bruce, Lerzan Aksoy, and Timothy L. Keiningham. "Approaches to customer segmentation." *Journal of Relationship Marketing* 6, no. 3-4 (2008): 9-39.
8. Kim SY, Jung TS, Suh EH, Hwang HS. Customer segmentation and strategy development based on customer lifetime value: A case study. Expert systems with applications. 2006 Jul 1;31(1):101-7.
9. Yao, Zhiyuan, Peter Sarlin, Tomas Eklund, and Barbro Back. "Combining visual customer segmentation and response modeling." *Neural Computing and Applications* 25 (2014): 123-134.