Indian Institute of Technology Indore
Discipline of Computer Science and Engineering
Minor Project in the course “Computational Intelligence”
Spring 2022-2023

Title: Product Recommendation and Market Basket Analysis

Evaluation Report

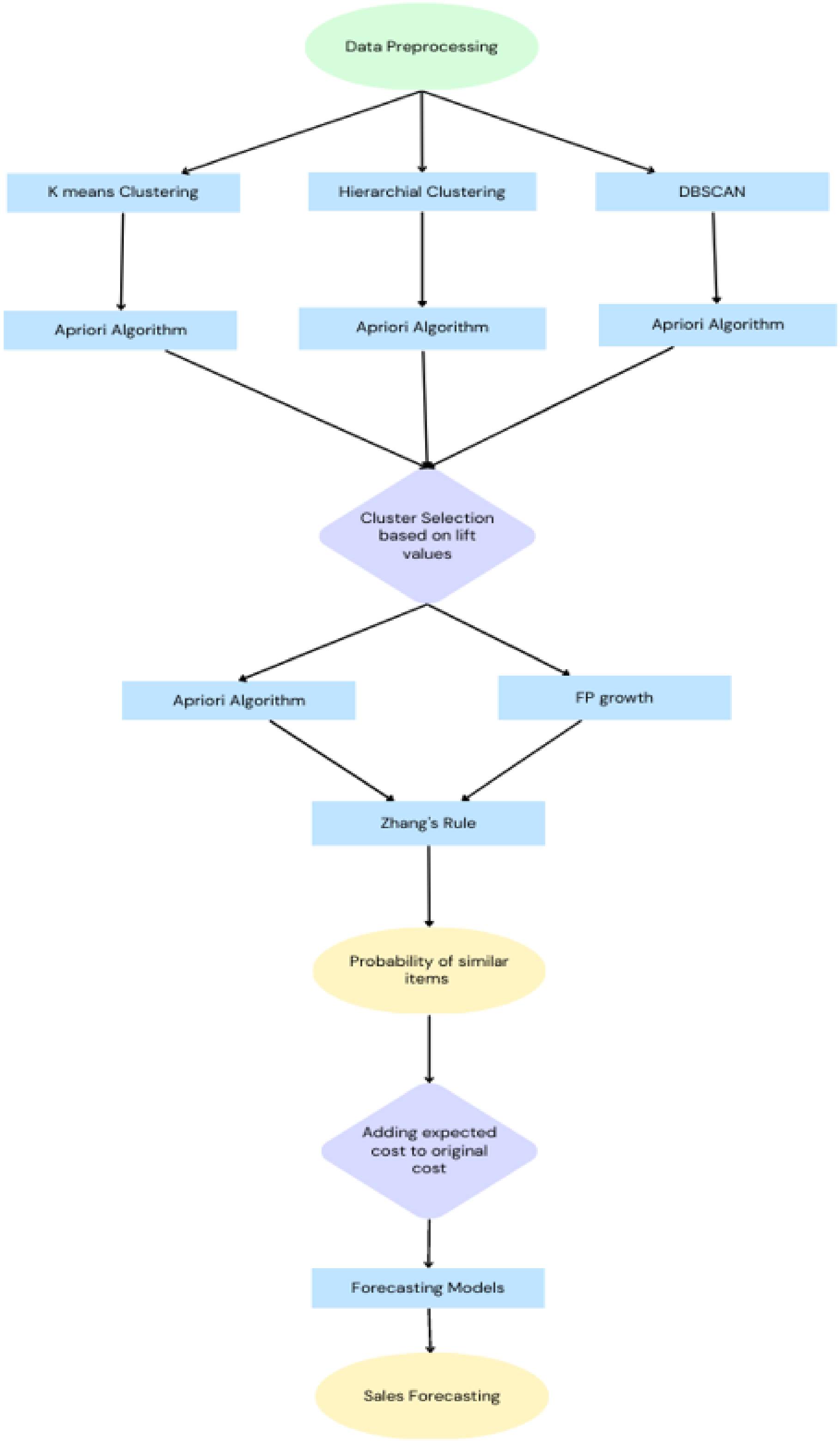
Title and Problem Definition

Our problem is to analyze customer purchase behavior, recommend products based on their purchase history, and forecast sales for an ecommerce company.

Analysis and Design of the Problem

- **Data Understanding**: We understand and explore the e-commerce dataset in this step. This also involves cleaning and preprocessing of data.
- **Customer Purchase behavior pipeline**: To find the items which are bought together, First, we make customer segments depending upon the frequency and amount they spent using clustering algorithms like K means Clustering, Hierarchical Clustering, and DBSCAN. Then we will apply the apriori algorithm on the clusters obtained from different clustering algorithms to find the best clustering algorithm. This can be done by checking the lift value of some pairs of an itemset. After finding the best clusters, we apply different association rule mining techniques like the apriori algorithm and FP growth algorithm, and then we calculate zhang rules to find the best possible set of items bought together.
- **Sales forecasting**: Using the association rules, we can find the probability of buying a similar item in a given basket, then add the expected amount in a basket. Now, using a given set of daily and weekly sales, we can apply time series forecasting models such as ARIMA to predict future daily and monthly sales.

Pipeline


Data Preprocessing
K means Clustering
Hierarchial Clustering
DBSCAN
Apriori Algorithm
Apriori Algorithm
Apriori Algorithm
Cluster Selection based on lift values
Apriori Algorithm
FP growth
Zhang's Rule
Probability of similar items
Adding expected cost to original cost
Forecasting Models
Sales Forecasting

Data Collection and Preprocessing

1. Dataset Description

Number of Rows	541909
Number of Columns	8
Start Date	01-12-2010
End Date	09-12-2011

2. Data Preprocessing for the Clustering task

- Drop the “StockCode”, “InvoiceDate”, and “Description” columns, as they are not required for clustering.
- Drop the rows with negative “Quantity” or “UnitPrice” values, which are considered erroneous.
- Calculate the total amount spent in a transaction by multiplying the “Quantity” and the “UnitPrice”. This value is stored in the “TotalPrice” column.
- Drop the rows with “null” values.
- We use the “groupby” and “sum” aggregate functions on the “CustomerID” and “TotalPrice” columns to calculate the total amount spent by a customer. This value is stored in the “Amount” column.
- We use the “groupby” and “count” aggregate functions on the “CustomerID” and “InvoiceNo” columns to calculate the total number of transactions for a customer. This value is stored in the “Frequency” column.
- If we do not handle anomalous data, the results provided by clustering algorithms may be substantially impacted. So, to eliminate anomalies, we use the "Isolation Forest" technique.
- Isolation Forest is an unsupervised machine learning approach that isolates observations using decision trees to identify abnormalities (outliers) in data. High-dimensional datasets can be used with it, and it is quick and scalable.
- Scaling is required for clustering algorithms to achieve distance values unaffected by scale differences and ensure that all variables are on a

comparable scale so that no variable dominates the analysis due to its higher magnitude.

- We use the “StandarScaler” preprocessing class from “scikit-learn”, which scales the inputs by subtracting the mean and dividing by the standard deviation. This step ensures that the dataset has zero mean and unit variance.

3. Data Preprocessing for the Associate-Rules Task

- Drop the rows with “null” values.
- We use the Customer_Clusters Dataset generated by the previous step, which contains the ClusterID for each CustomerID.
- We combine the Customer_Clusters and E-Commerce dataset using inner join by CustomerID column.
- We divide the dataset into different datasets having unique ClusterID.
- We make a Pivot table such that the items become column values, InvoiceNo becomes rows, and each cell represents the quantity bought of a particular item in a particular invoice.

4. Data Preprocessing for the Sales Forecasting Task

- Drop the rows with negative “Quantity” or “UnitPrice” values, which are considered erroneous.
- Calculate the total amount spent in a transaction by multiplying the “Quantity” and the “UnitPrice”. This value is stored in the “TotalPrice” column.
- Drop the rows with “null” values.
- Using the “InvoiceDate” column, calculate fields such as the “Year”, “Month,” “Week”, “Day” and “Quarter”

Study and Understanding of the Algorithms

1. **K-Means Clustering**

   K-Means clustering is an unsupervised algorithm that clusters the input data points into K different clusters groups based on direct or indirect patterns present in the data. We start with K random cluster centroids, and in every iteration, we assign the data points to the closest centroid and then recalculate the cluster centroids. The algorithm runs until the centroids are unchanged across multiple iterations or a certain number of iterations has been reached. The algorithm's performance depends on the initial selection of cluster centers, and the number of clusters K must be given beforehand.

2. **DB Scan**

   DBSCAN primarily uses two metrics, the neighborhood size and density threshold. Based on these, it clusters data points with many neighbors by first constructing a neighborhood around each data point based on the neighborhood size. It has several advantages over the K-Means algorithm, such that it can locate clusters of any shape and does not need the number of clusters to be predetermined. Also, it outperforms K-Means in situations where clusters have different forms and densities.

3. **Hierarchical Clustering**

   An unsupervised machine learning approach, hierarchical clustering, combines or divides clusters based on a similarity score to group comparable data points together. The algorithm produces a dendrogram that displays the order and distance of merges or splits, and there are two forms of hierarchical clustering: agglomerative and divisive. It has many applications but can be computationally expensive and sensitive to the chosen connection criterion and similarity metric.

4. **Apriori Algorithm**

   Apriori algorithm is used to discover associations and patterns in transactional data. It generates a set of frequent itemsets, which are sets of items that occur together in a certain percentage of transactions. The algorithm uses a level-wise approach, starting with single itemsets and progressively building more significant itemsets by combining smaller frequent itemsets. Then we select subsets of frequent itemsets and calculate their support and confidence. Then we can find items bought together frequently based on passing the minimum threshold value of support, confidence, and lift.

Support - Number of times an item set appears in the total number of transactions.

Confidence - For {X -> Y}, it is the proportion of transactions with itemset X in which itemset Y also appears.

Lift - It shows the association between 2 items.

$$\text{Lift (Item 1, Item 2)} = \text{Confidence (Item 1, Item 2)} / \text{Support (Item 2)}$$

5. FP Growth Algorithm

A data mining method called FP-growth identifies recurring trends in transactional data. It functions by building an FP-tree, a compact version of the transaction database that effectively captures all frequent itemsets. It then builds conditional FP trees for each prefix path and iteratively mines the tree to produce frequent itemsets. This makes it more scalable and effective than the Apriori approach, particularly for big datasets with many elements.

6. Zhang's Metric

A data mining method, Zhang's metric algorithm, is used in market basket analysis to quantify the relationship between items based on how frequently they appear together in transactions. It employs a statistical method that considers both the frequency of the co-occurrence of items and their presence or absence in transactions. As a result, it can accurately evaluate associations and capture interactions between things that are more complicated. Zhang's metric algorithm has been demonstrated to perform better in accuracy and efficacy than other conventional measures of association, such as support, confidence, and lift

$$\text{Zhang}(A \to B) = \frac{\text{Support}(A\&B) - \text{Support}(A)\text{Support}(B)}{\max[(\text{Support}(AB)(1 - \text{Support}(A)), \text{Support}(A)(\text{Support}(B) - \text{Support}(AB))]} .$$

7. **ARIMA Model**

The prominent time series forecasting technique known as the ARIMA (Autoregressive Integrated Moving Average) model is used to forecast future sales using existing sales data. It uses three techniques to detect patterns and trends in the data: moving average, differencing, and autoregression. The moving average component models the error or residual term, the differencing component corrects for trends and seasonality, and the autoregressive component models the link between the current and previous observations. The model can be trained using previous sales data and used to predict sales in the future.

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$$

If d=0: $y_t = Y_t$

If d=1: $y_t = Y_t - Y_{t-1}$

If d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$

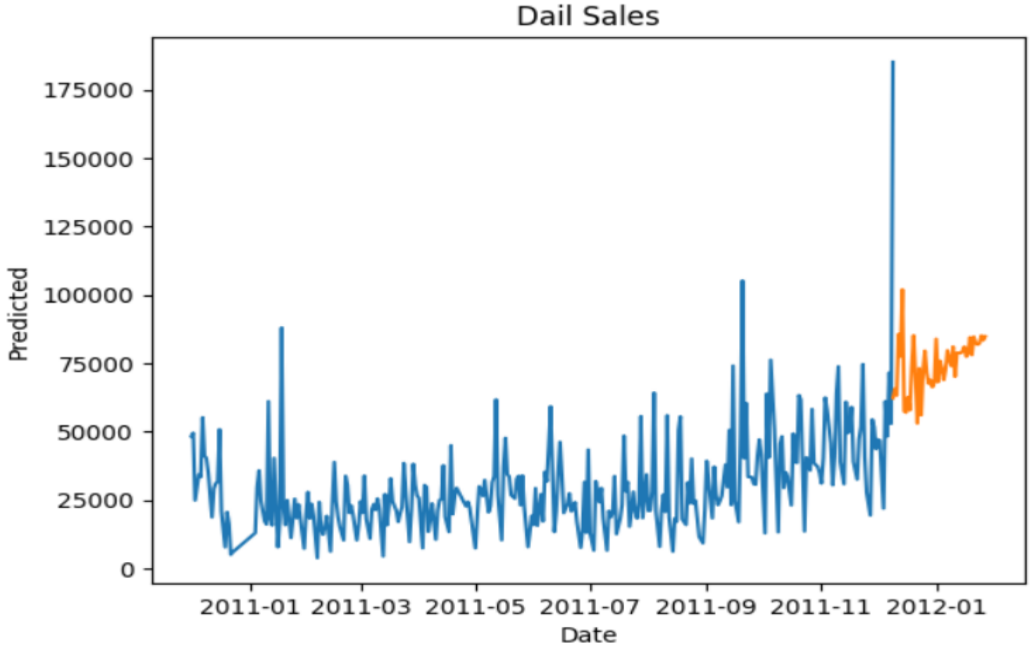
Performance Measurement Criteria

1. To choose the best from the results of the clustering algorithms, we pass the clusters formed by each one of them to the Apriori algorithm. Now, we check the lift values for common item sets. Higher lift values validate that the clusters were

created appropriately. Finally, clusters formed from **K means Clustering** are choosen.

2. To choose the best from the results of the association algorithms, we find the zhang metric for all common association rules formed by Apriori and fp growth algorithm. Higher the zhang value, means items are more associated. Finally, association rules given by **apriori algorithm** is chosen.
3. Metrics are essential for judging how well time series forecasting models work. Mean Squared Error (MSE) is the average of squared mistakes, whereas Mean Absolute Error (MAE) indicates the average magnitude of errors. It's common to practice assessing performance between various models using the Root Mean Squared Error (RMSE), which computes the square root of MSE. Mean Absolute Percentage Error (MAPE), a different statistic that shows the percentage difference between actual and anticipated values, is another one.

Test RMSE: 6456.943



References:-

Dataset

Apriori Algorithm - AN IMPROVED APRIORI ALGORITHM FOR ASSOCIATION RULES Mohammed Al-Maolegi1 , Bassam Arkok2

Time series forecasting - A superior alternating for market basket analysis, Swee Chaun Tan and Jess Pei San Lau

Clustering

Team Members:

Name1: Amit Kumar Makkad	Reg. No: 200001003	Sign. Amit
Name2: Mihir Karandikar	Reg. No: 200001044	Sign.

Under the Supervision of

Dr. Aruna Tiwari

Professor, CSE