INVENTORY MANAGEMENT BY PREDICTING INTENT FOR ONLINE SHOPPING

LM Data Science Group Project (32252)

Group MARVEL

Madhurima Sarkar : 2438650

Arjunkumar Wandiwash Krishnakumar: 2457663

Ramanathan Kathiresan : 2290101

Vignesh Venkataraman : 2330385

ABSTRACT

The retail industry has undergone a transformation because of e-commerce, which produces difficulties for retailers in predicting demand and controlling inventory. Retailers are increasingly utilizing cutting-edge methods to identify correlations between purchased goods, classify customers according to their buying habits, how frequently a product is purchased, calculate the approximate number of products that customers might purchase in the future and raise awareness of products that are similar in nature but have different names. Our study proposes effective inventory management using the features Market basket analysis, Customer Segmentation, Demand Forecasting and Product Compatibility. Market basket analysis searches for trends and connections among the goods that individuals buy in groups. This technique can be applied to find products that are frequently purchased together and to make better product recommendations to clients based on their purchasing patterns. Customer segmentation categorizes the customers into groups based on their characteristics and behaviors to enhance marketing campaigns and improve individual customer experiences. Demand forecasting helps the company to make accurate supply decisions by projecting future sales and revenue. Integrating this feature would help retailers reduce waste, guarantee the right products are available when they're required and optimize their inventory levels. The product compatibility study aims to identify analogous products and induce joint promotion to boost sales. Utilizing these innovative methods and tools, retailers may better understand consumer behavior, enhance sales and customer satisfaction, and optimize their inventory levels. According to our results, we highlight the advantages of inventory control using machine learning methods like clustering and regression models. Theoretical algorithms and experimental results are provided to showcase the effectiveness of the proposed algorithm on online retail data.

Table of Contents

Table of Figures

1 INTRODUCTION

Over the years, the emergence of the online shopping industry has attracted the attention of researchers and practitioners in the world to obtain a better understanding of its applications with respect to buying online by region. Customers can utilize information technology to conduct shopping activities, as well as to determine how simple the system is to use and what other people think of it thanks to the expanding online retail industry. It provides a broad range of benefits such as timesaving, great promotions, wide product ranges and lower and competitive prices, which significantly provoke online purchase intentions.**[1]**

E-commerce's explosive rise in recent years has significantly strengthened the retail sector. Numerous individuals are using online channels to make their purchases due to the convenience of online shopping and the growing popularity of mobile devices. Online retailers should be able to anticipate demand and manage inventory levels, which presents additional challenges. In this study, we discuss the impact of inventory management by projecting online shoppers' intentions. We examine several methods that can be employed to categorize customers, predict customer behavior and manage inventory.

Identifying the connections between the products that customers have purchased will be useful in determining which product combinations are most popular and used in conjunction for marketing campaigns. Analysis of transaction patterns could be used to spot seasonal trends in consumer behavior and modify marketing strategies accordingly. Utilizing historical sales data, the model would segregate the customers. This would make it simpler to provide better marketing services and comprehend customers across locations.

Even though it is more typical to predict demand at the product level, demand forecasting should be done at the store level. By concentrating on factors like store location and demography, retailers may enhance their forecasting and inventory management. It may be difficult to match products that are sold in various categories or departments since they may have different names or traits. Retailers can still locate products that are presumably good fits despite the existence of various categories.

The proposed work provides a summary of e-commerce and the reasons influencing its development. After discussing the significance of forecasting online shopping intention and inventory management and how they affect an online retailer's capacity for success, the exploratory data analysis section will be covered. The paper will also explore the benefits and challenges. We'll conclude by summarizing the key findings and their significance for e-commerce firms. The ultimate objective is to offer practical guidance to business owners looking to improve their e-commerce operations as well as educational material on the significance of inventory management.

The data used as part of the research involves Invoice Number uniquely assigned to each transaction, Stock Code uniquely assigned to each distinct product, Description as the Product (item) name, Quantity: The quantities of each product (item) per transaction, Invoice Date as The day and time when a transaction was generated, Unit Price as the Product price per unit in sterling (£), Customer ID being the Customer number uniquely assigned to each customer and Country where a customer resides.

2 BACKGROUND RESEARCH

Several factors have contributed to the expansion of e-commerce, such as the rising popularity of mobile devices, the ease of online shopping, and the growing global market. In recent years, the value of global retail e-commerce sales is expected to double, according to a forecast by eMarketer. However, the expansion of e-commerce has also brought about new difficulties for retailers, especially when it comes to predicting demand and controlling inventory levels. Based on previous sales data and patterns,

demand can be predicted in conventional retail settings. Forecasting demand can be quite difficult in the internet context, because customer behavior might be more erratic and unexpected.**[2]**

Retailers have used a variety of strategies for predicting online shopping intention to meet this difficulty. These consist of predictive analytics, market trend analysis, customer behavior analysis, and machine learning algorithms. Retailers can better understand consumer behavior and forecast future demand for products by utilizing these techniques. Another crucial element of e-commerce is inventory management, as businesses must control their stock levels to fulfil client demand while reducing waste. To make sure that the proper products are available when they are needed, this requires anticipating demand for various products and optimizing inventory levels. Retailers can utilize a variety of techniques for managing their inventory, including supply chain management systems, demand forecasting models, and inventory management software. The success of e-commerce merchants depends on their capacity to predict customers' intentions to shop online accurately and to control inventory levels. Retailers may streamline their processes, enhance customer service, and beat the competition by employing the proper methods. There are many challenges associated with online shopping prediction, such as data quality, scalability, and privacy concerns. However, with the advancement of machine learning and data analysis techniques, these challenges can be addressed, and online shopping prediction can become an increasingly valuable tool for e-commerce businesses.

Customers are often concerned about their personal data being collected and used for predictive analysis. Online retailers must ensure that their data collection and analysis methods comply with data privacy regulations. Online retailers may collect data from various sources, such as social media, customer reviews, and purchase history. Integrating these data sources can be a challenge, as they may be in different formats and require different analysis techniques.

Customer preferences and behavior can change over time, making it difficult to make accurate predictions. Retailers must continuously update their predictive models to adapt to changing customer behavior. Predictive models rely on historical data to make accurate predictions. However, for new products or services, there may be limited historical data available, making it challenging to make accurate predictions. Overall, online shopping prediction is a complex and challenging field that requires careful consideration of data quality, privacy concerns, and changing customer behavior. However, with the right tools and techniques, retailers can use predictive analysis to improve their marketing strategies, enhance customer experience, and optimize their inventory management. The growing availability of data and improvements in machine learning techniques have led to several recent developments in online shopping prediction. Some of the notable developments include:

Recommendation systems are employed to provide product recommendations to clients based on their previous actions and preferences. Deep learning algorithms and the incorporation of several data sources, such as social media and consumer evaluations, are recent developments in recommendation systems.

Real-time prediction enables merchants to make forecasts and suggestions in real-time. To deliver individualized suggestions in real time, this method needs high-speed data processing and sophisticated algorithms.

Deep Learning involves teaching artificial neural networks to predict the future using enormous data is known as deep learning. This method is becoming more and more popular for predicting online purchases. It has proven particularly successful in image and speech recognition.

AI is a new advancement in machine learning that makes it possible for people to comprehend how a model arrived at a certain prediction. This strategy is especially helpful for forecasting online purchases since sellers must be able to justify their suggestions and projections.

Natural Language Processing teaches computers to comprehend human language. To understand consumer behavior and preferences, NLP has a wide range of applications in online shopping forecasting, including the analysis of customer reviews and social media data.

Retailers may now offer more precise predictions and recommendations, improve the consumer experience, and improve inventory management thanks to recent developments in online shopping prediction.

3 QUESTION DEVELOPMENT

There are several important aspects to consider when formulating cutting edge features when attempting to forecast online shopping intent. It's critical to comprehend the consumer journey to effectively estimate online purchase intent. We would have to investigate how consumers learn about a product, how they investigate it, and how they eventually choose to buy it. It is possible to create machine learning models that would accurately forecast online purchase intent by studying the various consumer journey stages. Several questions must be addressed to manage the inventory efficiently like:

1. How can online purchase behaviour be predicted using machine learning models?
2. What kinds of data may be utilised to train machine learning models for predicting online purchases?
3. Which machine learning techniques are most effective in forecasting consumers' online purchase habits?
4. How accurate can machine learning models anticipate consumers' behaviour when they purchase online?
5. What are some of the challenges faced in predicting online purchases using machine learning?
6. Can a user's online shopping experience be customised using machine learning models?
7. How could machine learning models assist online retailers enhance the products they offer and the marketing strategies they employ?

The major variables and analytical strategies chosen by consolidating all these inquiries reveals our study focus, which is on effective "**Inventory Management by Predicting Intent for Online Shopping**" by incorporating Market Basket Analysis, Customer Segmentation, Demand Forecasting and Product Compatibility.

4 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an approach to analyzing data, the analyst explores data to find patterns, connections. EDA is the first stage of data analysis, used to gain understanding of the data and create hypotheses that may be tested using advanced statistical methods. To understand the distribution of the data, spot outliers and other deviations, and spot patterns and relationships between variables, EDA often entails visualizing the data using graphs, charts, and other tools. EDA can be used to spot any biases or inaccuracies in the data as well as to direct the choice of the best statistical methods for further analysis. EDA is a crucial phase in the data analysis process because it establishes the groundwork for more complex statistical modelling and gives analysts a better knowledge of the data they are working with. The process of EDA typically involves four stages which are Data Collection, Data Preprocessing, Data Visualisation and Data Transformation. In Data Collection stage, the necessary raw data is gathered from various sources such as customer profiles, product information, web server logs and legacy storage. The Data is called raw because it has not already had any treatment. In Data Preprocessing stage, Data is cleaned, transformed and prepared for analysis for the further stages because it might have corrupt, incorrect , null records or information in different formats. In Data Visualisation stage, the data is explored and visualized to gain various insights and understanding from the patterns present in the data. Finally, Data Transformation stage involves selecting relevant features and transforming them to improve the model's performance.

4.1 DATA COLLECTION

After extensive research, the data was collected through the below website https://archive.ics.uci.edu/ml/machine-learning-databases/00553/

The data was collected in the year 2009-2011. This "Online Retail II.xlsx" data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011.The company mainly sells unique all-occasion giftware. Many customers of the company are wholesalers. The data obtained is a part of the retail store using software programs. Moreover, the data subjects cannot be identified using their names and they exist in data sets in the form of ID's. In the following Figure 1 reflects a snapshot of the data as it was collected as part of the study.

Out[4]:

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country	Unique Invoice Stockcode
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom	48943485048
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom	48943479323P
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom	48943479323W
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom	48943422041
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom	48943421232
...	...	...	...	...	...	...	...	...	...
525456	538171	22271	FELTCRAFT DOLL ROSIE	2	2010-12-09 20:01:00	2.95	17530.0	United Kingdom	53817122271
					2010-12-09			United	

Figure 1 : Snapshot of the collected Data before processing

4.2 DATA PREPOCESSING

Preprocessing the data is the next stage in making sure it is correct and comprehensive. Identifying and transforming missing values, controlling outliers, and resolving any discrepancies or errors in the data

4.3 DATA VISUALISATION

Data visualization, the following stage, entails using graphs, charts, and other tools to display the data. Finding patterns, trends, and connections in the data is made simpler as a result. Figure 2 depicts the trend of sales for the top 20 sold items by quantity.

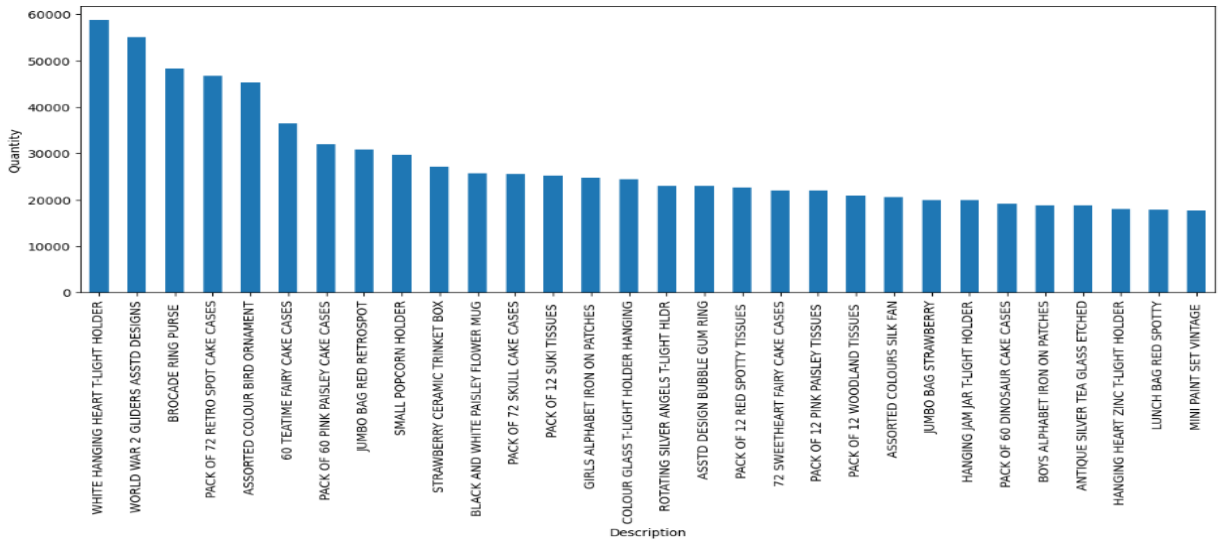

Figure 2 : Top 20 sold Items

4.4 DATA TRANSFORMATION

Data may occasionally need to be converted before processing. This might need scaling, normalization, or other processes to ensure the data is in a suitable format for analysis. Figure 3 shows a heatmap that quantitatively visualizes all item sales for each month to provide a clear picture of the best performing months relative to the average and the stagnant months. Dark green represents high sales and dark red represents lower sales.

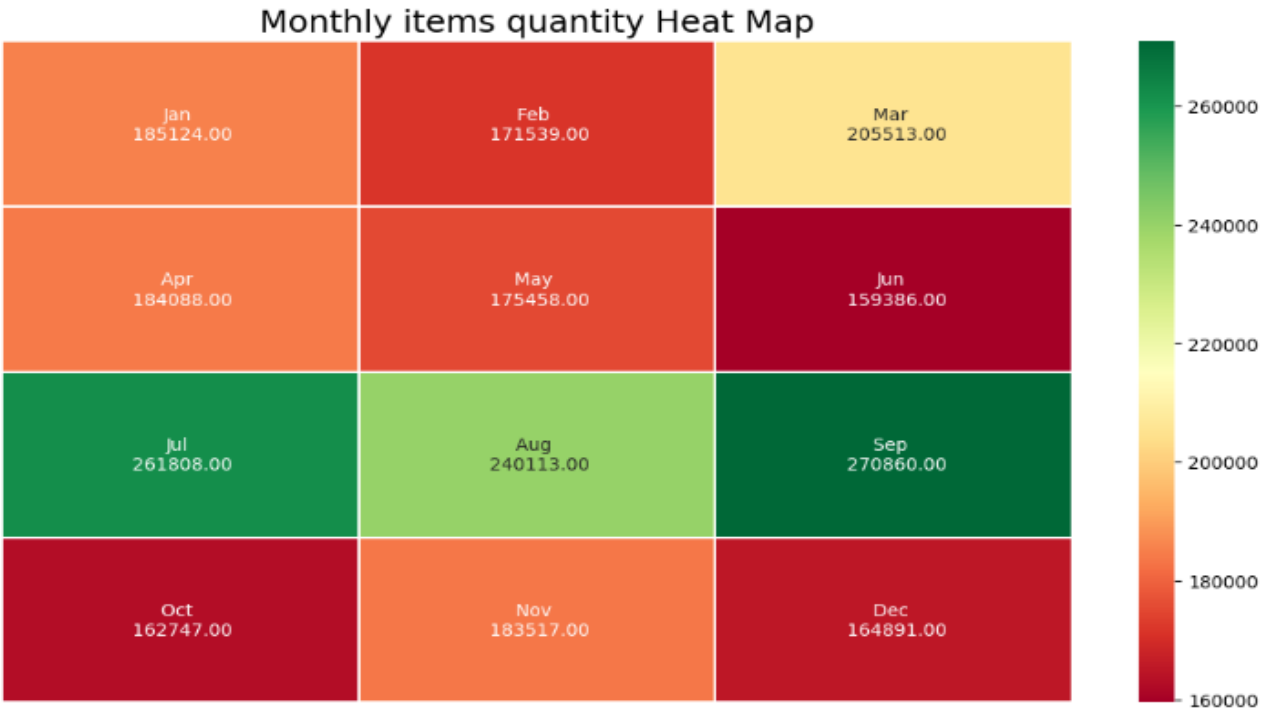

Figure 3 : Heat Map visualising the monthly sales of all items quantitatively over the entire year

Figure 4 shows a boxplot that quantitatively visualizes all item sales for each month to provide a clear picture of the best performing months relative to the average and the stagnant months.

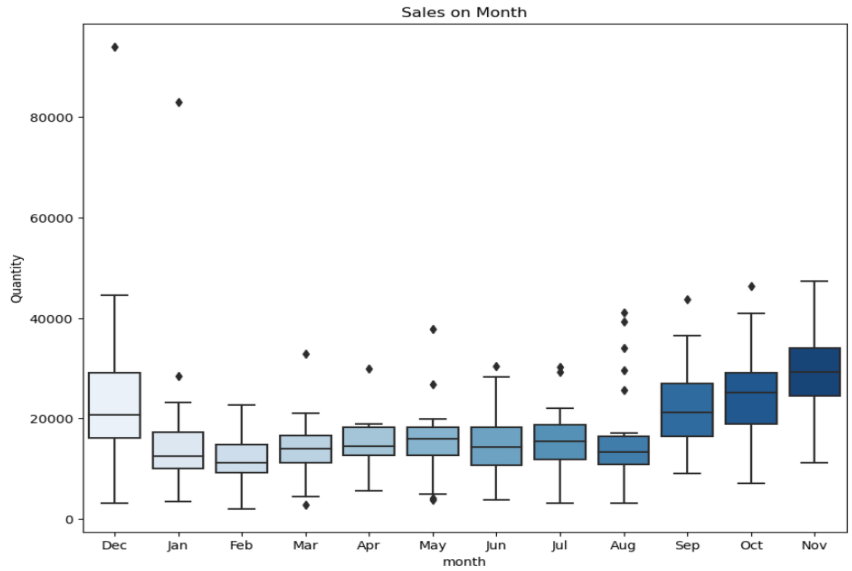

Figure 4 : Box-Plot of Monthly Sales for all items quantitatively

EDA is a flexible strategy that may be tailored to fit the specific needs of a given analysis. EDA seeks to fully comprehend the data so that decisions about how to proceed with additional analysis or modelling can be made with the knowledge of the situation.

5 RATIONALE FOR DATA MODELLING/EXPERIMENTATION

5.1 MARKET BASKET ANALYSIS

Market basket analysis is a method of data analysis used to identify the relationships between products which buyers frequently buy together. Identifying product linkages, maximising product offerings and promotions, and enhancing inventory management procedures are all frequent uses for MBA in e-commerce enterprises.**[3]** Market Basket Analysis can assist e-commerce enterprises to understand customer behaviour and preferences better, find cross-selling opportunities, and optimise their inventory management. Businesses can determine which products are usually bought together and which products aren't by looking at previous transaction data. This data can be utilised to find potential new product offerings as well as to improve product placement and advertising.

Association rules, which are declarations describing the connections between products, are frequently used to illustrate the results of market basket analyses. For instance, an association rule might specify that "Customers who purchase product A are also likely to purchase product B." These guidelines can be used to inform additional data modelling techniques, like consumer segmentation and demand forecast, as well as data-driven decisions about product placement, marketing, and inventory management. In conclusion, market basket analysis is an effective method for this question that can assist e-commerce companies in better understanding customer behaviour and preferences, optimising their product offerings and promotions, and enhancing their inventory management procedures.

For effective inventory control and better sales prediction, we needed to analyse how different groups of customers interact with the different clusters of products and predict which targets demographic is more likely to purchase which group of products and for this analysis we take motivation from the following research paper. Market Basket Analysis (MBA) also known as association rule learning or affinity analysis, is a data mining technique that can be used in various fields, such as marketing, bioinformatics, education field, nuclear science etc. The main aim of MBA in marketing is to provide the information to the retailer to understand the purchase behaviour of the buyer, which can help the

retailer in correct decision making. There are various algorithms are available for performing MBA. This paper discusses the data mining technique i.e., association rule mining and provide a new algorithm which would be helpful to examine the customer behaviour and assists in increasing the sales.

For Example, Let's consider a scenario where an online retailer aims to streamline its inventory control procedures for a specific product category (ex. Beverage). The company possesses transactional data that details which products were bought in tandem during each transaction. The company can determine which products are more frequently and less commonly bought together using market basket analysis. The following are the example of association rules that could be produced: Customers are 75% more likely to purchase Milk after purchasing Coffee.

Utilising these association rules will improve product placement, promotions, and inventory control. For instance, the company might decide to list milk and coffee side by side on the internet. For Market Basket Analysis, we have trained the data using two algorithms Apriori and FP Growth.

5.2 METHODOLOGIES

5.2.1 APRIORI MODEL

For locating frequently occurring item sets in transactional databases, such as Market Basket Analysis, the Apriori model is a widely used technique. The approach is based on the idea that if a set is frequent overall, then all of its subsets must also be frequent.

Support, Confidence, and Lift are the three fundamental metrics that are employed in association rule learning, and we may make use of them. Support is simply the likelihood or probability that an event will occur. The percentage of transactions that contain an item set is used to calculate it.

Support (Item A) is the number of transactions which includes item A divided by the total number of transactions.

Confidence is the Conditional probability can be used to express the confidence in an occurrence given its antecedent. It is, to put it simply, the likelihood that item B will occur given that item A has already occurred.

Lift is the observed to anticipated ratio. Lift accounts for the popularity of the two items in its measurement of the likelihood that one will be purchased when the other is. It may be determined by multiplying the likelihood that each of the events will occur separately by the likelihood that both events will occur independently as if there were no correlation between them.

The Apriori algorithm has two stages of operation: The technique generates a set of frequent 1-itemsets, or things that appear frequently in transactions, in the first phase by scanning the transactional database. A user-specified minimum support threshold determines the frequency of an item. The algorithm then joins the frequent (k-1)-itemsets discovered in the first iteration to produce candidate k-itemsets or sets of items that appear together in transactions**.[4]** The candidate itemsets that do not satisfy the minimal support criterion are subsequently pruned by the algorithm.

As it generates candidate itemsets and analyses their frequency iteratively, the Apriori algorithm is known as an iterative algorithm. It avoids checking every possible itemset in the database, which can be computationally expensive for large datasets which makes the algorithm efficient. The Apriori algorithm produces a list of frequently occurring itemsets that can be used to spot interesting trends in transactional data. Applications for these patterns include targeted marketing campaigns, inventory management, and product recommendations.

Figure 5 depicts the performance metrices such as support, confidence and lift generated as part of the Apriori algorithm implemented to perform market basket analysis feature.

MARKET BASKET ANALYSIS

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric	MBA Value(%)
2	(SWEETHEART CERAMIC TRINKET BOX)	(STRAWBERRY CERAMIC TRINKET BOX)	0.047776	0.078274	0.037992	0.795205	10.159228	0.034252	4.500720	0.946802	79.520480
1	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.050878	0.158219	0.036798	0.723265	4.571294	0.028749	3.041826	0.823123	72.326454
3	(STRAWBERRY CERAMIC TRINKET BOX)	(SWEETHEART CERAMIC TRINKET BOX)	0.078274	0.047776	0.037992	0.485366	10.159228	0.034252	1.850293	0.978130	48.536585
0	(WHITE HANGING HEART T-LIGHT HOLDER)	(RED HANGING HEART T-LIGHT HOLDER)	0.158219	0.050878	0.036798	0.232579	4.571294	0.028749	1.236768	0.928084	23.257919

Figure 5 : Snapshot of Market Basket Analysis depicting various performance metrics generated by Apriori algorithm

Businesses may create hundreds of data-driven strategies to increase their sales and profitability by adopting the Apriori Algorithm and studying the association measurements. The analysis of consumer purchasing behaviour through data mining is dependent on these association rules. Association rule mining may provide useful insights for several of a retailer's most crucial tactics, including Customer analytics, Market Basket analysis, and Product Clustering.

5.2.2 FP GROWTH MODEL

A data mining approach called FP-Growth (Frequent Pattern Growth) is used to locate frequent item groupings in transactional databases. The FP-Growth technique builds an FP-Tree (Frequent Pattern Tree), which is a compressed representation of the input database. The entire dataset is scanned while keeping track of each item's frequency to build the FP-Tree.**[5]** The algorithm then arranges the items according to decreasing frequency before adding each transaction as a path in the tree and creating a tree from it. The edges between nodes in the tree show how frequently certain things occur together in the database, and each node in the tree represents a single item.

After building the FP-Tree, the algorithm traverses the tree in depth-first fashion to locate all frequent item sets. By beginning at the bottom of the tree and working your way up, you can find the frequently occurring itemsets. Each connection between a leaf node and a root node represents a common itemset. The algorithm keeps track of the support count for each frequently encountered itemset as it moves through the tree. Divide-and-conquer strategy is employed by the FP-Growth algorithm to iteratively mine the FP-Tree. The technique first identifies all frequent 1-itemsets, after which it combines frequent (k-1) itemsets to produce candidate itemsets. By removing infrequent itemsets, it prunes the tree and creates a conditional FP-Tree for each frequent item.

Apriori algorithm provided promising results with the confidence ranging from 20% to 80% compared to FP Growth which provided 50% confidence on the same data. After thorough evaluation we suggest Apriori algorithm as the preferred algorithm to achieve accurate results.

6 CUSTOMER SEGMENTATION

Customer segmentation is segregating a client base into distinct groups or segments according to their specific traits, behaviours, and preferences **[6]**. By utilising this technique, firms can derive a better insight of their target markets and develop marketing plans that are tailored to meet the customer demands. It can be used to identify groups of customers who share comparable buying habits, tastes, and demographics. Understanding the various shopping habits of customers, figuring out the most lucrative client segments, and developing focused marketing strategies can help boost customer acquisition and retention.

Customer segmentation, for instance, can involve classifying customers based on their past purchases, frequency of visits, demographics, and product preferences in an online retail setting. The various segments can then be specifically targeted with product promotions, discounts, offers, product recommendations, and marketing advertisements that are more likely to meet their unique interests and requirements. This may result in improved client satisfaction, higher revenue, and effective inventory control.

We propose the Recency-Frequency-Monetary Value (RFM) method to group the customers based on:

- How recent was their last transaction?
- How frequently do they purchase?
- How much money have they spent with us?

These three criteria—how recently a client made a purchase, how frequently they placed an order, and how much money they spent—are a strong indicator of their value and a forerunner of their future behaviour. The best customers, new customers, frequent customers, and high value customers are identified by the RFM scores. Customers with high value are the most current, have made the most purchases, and have made the most orders. Given their high value and status as the best customers, the big spenders would receive offers/promotions as an effort to retain them.

One of numerous techniques, including Business Rule, Magento, Customer Profiling, Quantile Membership, RFM Cell Classification Grouping, Supervised Clustering, Customer Likeness Clustering, Purchase Affinity Clustering, and Unsupervised Clustering, can be used to process these data. These techniques were categorised into Simple technique, RFM technique, Target technique, and Unsupervised technique in this study. The procedure was generalised to include setting the business aim, gathering the data, preparing the data, analysing the variables, processing the data, and assessing performance.

6.1 METHODOLOGY

6.1.1 K-MEANS CLUSTERING

Customers can be segmented using K-means clustering and the RFM (Recency, Frequency, Monetary) framework. **[7]** RFM is a popular technique for examining consumer behaviour and is based on three crucial metrics:

- Recency: how recently a customer has made a purchase (ideally should be low)
- Frequency: how often a customer makes purchases (should be high)
- Monetary: how much a customer spends on purchases (should be high)

Figure 6 i - iii displays the box plots representing the recency, frequency and monetary value obtained from the dataset used as part of the study.

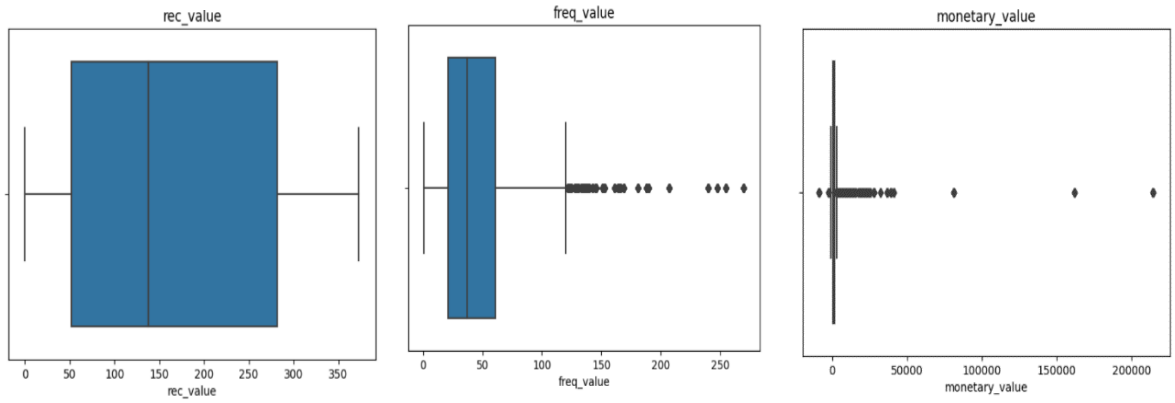

Figure 6 (i ,ii ,iii) : Recency, Frequency, Monetary values obtained from Dataset from left to right

Calculating the RFM values for each client is the first step in applying K-means clustering to RFM data. Then, using the RFM values, a matrix of customer is created, with each row denoting a different customer and each column representing a different RFM value. K-means clustering is used to divide similar customers into groups after the RFM matrix has been constructed. A variety of techniques, including the elbow approach and silhouette analysis, can be used to estimate the number of clusters.

RFM-based K-means clustering provides information that businesses can utilise to better comprehend the behaviour of various client clusters and develop segment-specific marketing strategies. Customers in the "Low-Value" cluster (low monetary value, low frequency, and distant purchase), on the other hand, could be targeted with incentives to increase their engagement and purchasing frequency. As an illustration, customers in the "High-Value" cluster (high monetary value, high frequency, and recent purchase) could be targeted with personalised offers and promotions to maintain their loyalty and retention.

- Inventory allocation: Different customer segments may have different preferences and purchasing habits. Inventory managers may allocate inventory more effectively to make sure that the proper products are available in the right quantities to satisfy consumer demand by being aware of these distinctions.
- Product mix: Inventory managers can choose which products to stock and in what amounts. They may determine which goods are more popular with each sector by evaluating client data, and they can then modify their product mix accordingly.

Figure 7 represents a scatter plot representing the cluster of customers based on the quantity and price. Some of the benefits of RFM Customer Segmentation include Boosting remarketing strategy, obtaining more loyal customers, reducing churn rate and Increasing sales.

Figure 7 : Scatterplot representing Customer Clusters based on quantity and price of items purchased

7 DEMAND PREDICTION

An important aspect of inventory management is demand prediction to estimate the volume of goods that customers would buy during a specific time frame **[8].** Accurate demand forecasting aids companies to maximise inventory levels and preventing stockouts or overstocking, which can result in lost revenues and increased expenses. Demand prediction can be done in several ways, such as through time series analysis, regression analysis, and machine learning algorithms. These techniques are based on historical sales, consumer behaviour information, and external variables like seasonality and marketing efforts.

To forecast future demand, time series analysis examines patterns and trends in historical sales data. Identifying the correlation between demand and additional factors like price, promotions, and season is the goal of regression analysis. Demand may be predicted using a variety of criteria using machine learning methods like random forests and neural networks. Effective inventory management depends on accurate demand forecasting, which may also help firms cut costs, boost sales, and enhance customer satisfaction.

7.1 METHODOLOGY

7.1.1 RANDOM FOREST

Random Forest is widely used for applying classification and regression modelling to structured data sets. Classification is the forecast of the class label with the majority of the vote across the decision trees, whereas Regression is the average prediction across the decision trees.

Features of Random Forest: Compared to other trees, each tree has a distinct trait, variation, and qualities. Different types of trees exist. Trees are exempt from the dimensionality curse since they are conceptual concepts and don't need to take features into account. As a result, the feature space is smaller. Building random forests may be done in parallel since each tree is independently constructed from various data and attributes.**[9]**

Splitting the data into train and test is not necessary with a Random Forest because only 30% of the data is ever seen by the decision tree. The outcome is based on Bagging, which means that it is determined by the majority vote or the average. Using the Random Forest Algorithm has several advantages, but one of the most important ones is that it decreases the possibility of overfitting and the amount of time needed for training. It also provides a high degree of precision. Hyperparameters are utilised to either speed up the model or improve its performance and predictive ability. *n_estimators* : Number of trees the algorithm constructed before averaging the results. *max_features* : Before considering splitting a node, random forest uses a maximum number of features. *mini_sample_leaf* : Minimal number of leaves necessary to separate an internal node is determined by the function. Figure 8 displays a plot of actual vs predicted demand for a particular product.

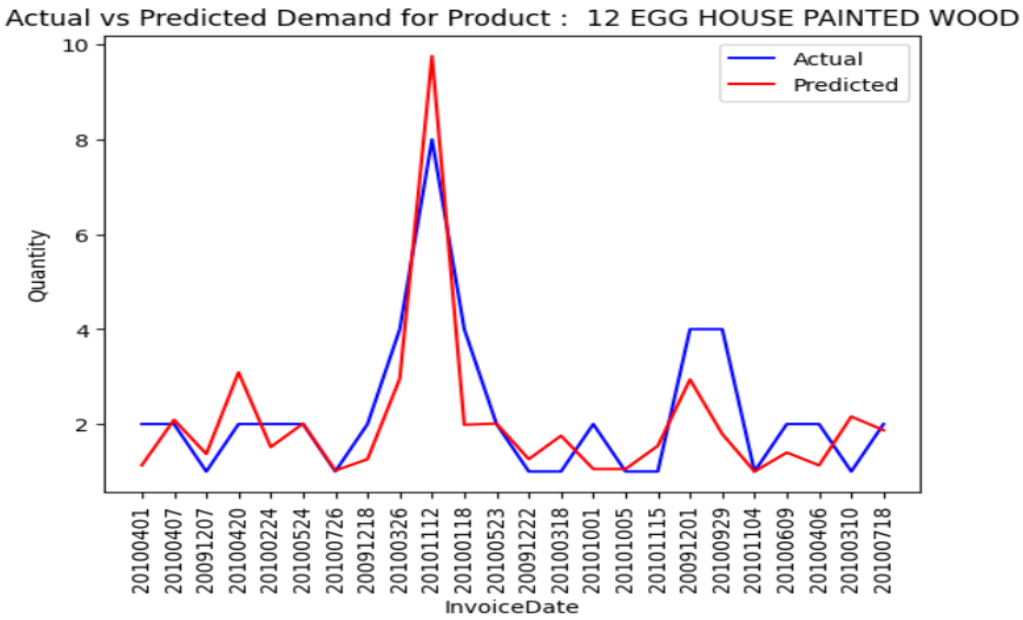

Figure 8 : Line plot comparing the actual vs predicted demand for a particular product

Random Forest Algorithm is beneficial as it can effectively handle large data sets, provides better result prediction accuracy than the decision algorithm, can carry out both classification and regression tasks, makes accurate forecasts that are simple to understand. Some of the domains which uses Random Forest Algorithm are Banking, Healthcare, Stock Market and E-Commerce.

7.1.2 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machines (SVM) is a popular machine learning algorithm that can be used to estimate demand. The supervised learning method SVM can be applied to both classification and regression applications. The study investigates the fundamentals of Support Vector Machine (SVM) in retail industry to create an SVM model that would forecast future demand.**[10]**

SVM models are created utilising the sigmoid and radial basis kernel functions. In developing the model, several variables that influence the demand for the product, such as the number of produced units, inventory, sales costs have been considered.

The algorithm finds a hyperplane that divides the data into two classes in this case, periods of high and low demand and then divides the data along this hyperplane. Historical sales data, combined with external characteristics like marketing campaigns, seasonality, and economic indicators, can be used as input variables for SVM to anticipate demand. The algorithm can then be trained using this information

to discover how the input variables and output variable (i.e., the number of products sold) are related. It is beneficial to business for maximising their inventory levels and preventing stockouts/overstocking.

Through evaluation, our study concludes that Random Forest algorithm makes better prediction than SVM model as the MSE and RMSE values produced in Random Forest algorithm were minimal and worked for a large range of items in the dataset.

8 PRODUCT COMPATIBILITY

Product compatibility is the technique where products which are similar are grouped together and recommended in the online shopping industry based on their traits, names in different languages, past customer buying behaviours and their images. By using this technique, we can recommend better and more tailored products to the customer to help them in their shopping decisions. This will improve the revenue because even if the customer is reluctant to buy one product, they might buy a different but similar product based on these suggestions.

Product Compatibility, for example, recognise biscuits in America and the cookies in United Kingdom to be the same product but with different names in different countries and will group them together and will show the popular items to the customers across different countries to assist in their buying decisions and improve revenue**.[11]** Similarly, identical items with different names in different languages will also be grouped together based on product compatibility and this will allow the company to sell their products across different geographic regions. This could result in better customer satisfaction, higher revenue, and better inventory control.

8.1 METHODOLOGY

Text Clustering using Term Frequency-Inverse Document Frequency and K-means Clustering-With clustering, data scientists can discover intrinsic grouping among unlabelled data.**[12]** Though there are no specific criteria for a good clustering and it completely depends on the user, how they want to use it for their specific needs , clustering can be used to find unusual data points/outliers in the data or to identify unknown properties to find a suitable grouping in the dataset. K-means is a centroid-based clustering algorithm, where we calculate the distance between each data point and a centroid to assign it to a cluster.**[13]** The goal is to identify the K number of groups in the dataset.

Equation 1 represents the Term Frequency-Inverse Document Frequency which quantifies the significance of a word within a corpus. Term Frequency is just the ratio of the current word to all the other words in the document, string, etc.

$$tf(t,d) = \frac{n_t}{\sum_k n_k} \tag{1}$$

Frequency of term t_i, where n_t — the number of t_i in current document/string, the sum of n_k is the number of all terms in current document/string.

Equation 2 shows the Inverse Document Frequency which is a log of the ratio of the number of all documents/string in the corpus to the number of documents with term t_i.

$$idf(t,D) = \log \frac{|D|}{|\{\, d_i \in D \mid t \in d_i \,\}|} \tag{2}$$

Equation 3 depicts the tf-idf (t, d, D) which is the product tf(t, d) to idf(t, D).

$$tf\text{–}idf(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D) \quad (3)$$

8.1.1 K-MEANS CLUSTERING WITH TF-IDF

Since we know how TF-IDF work, we will use clustering with TF-IDF weights. Here we use scikit-learn implementation of TF-IDF and K-Means in this project.

The clusters can be interpreted through the output shown in the below snapshot where the similar "coat" items are grouped together in Cluster 8 and the "bag" products are grouped in Cluster 9. For instance, when a customer buys a "Red Coat Rack Paris Fashion" and is recommended another item such as "Cream Cupid Hearts Coat Hanger" from the same category, it makes the shopping experience better. This is how the concept of product recommendation to the customers is enabled through Product Compatibility feature. Figure 9 reflects a snapshot of the various clusters for the feature product compatibility.

```
-------------------------------------------
Cluster  8
-------------------------------------------

2     CREAM CUPID HEARTS COAT HANGER
7              HAND WARMER UNION JACK
8           HAND WARMER RED POLKA DOT
10        RED COAT RACK PARIS FASHION
11     YELLOW COAT RACK PARIS FASHION
Name: Description, dtype: object

-------------------------------------------
Cluster  9
-------------------------------------------

86              JUMBO BAG PINK POLKADOT
88      JUMBO BAG CHARLIE AND LOLA TOYS
89             STRAWBERRY CHARLOTTE BAG
103              JUMBO STORAGE BAG SUKI
104      JUMBO BAG PINK VINTAGE PAISLEY
```

Figure 9 : Sample of some clusters for Product Compatibility

8.1.2 REDUCING NUMBER OF CLUSTERS USING PCA

Principal component analysis is also used to reduce the number of clusters formed. TF-IDF is used to tokenize the documents learn the vocabulary and inverse the document frequency weightings and allow to encode new documents. for e.g. A vocabulary of 8 words is learned from the given documents and each word is assigned a unique integer index in the output vector.TF-IDF will transform the text into meaningful representation of integers or numbers which is used to fit machine learning algorithm for predictions. PCA will then reduce the number of clusters formed which in our case is 20 from K means to fit out data model. Figure 10 represents the cluster of products based on similarity using PCA and TF-IDF and is visualised using a scatterplot.

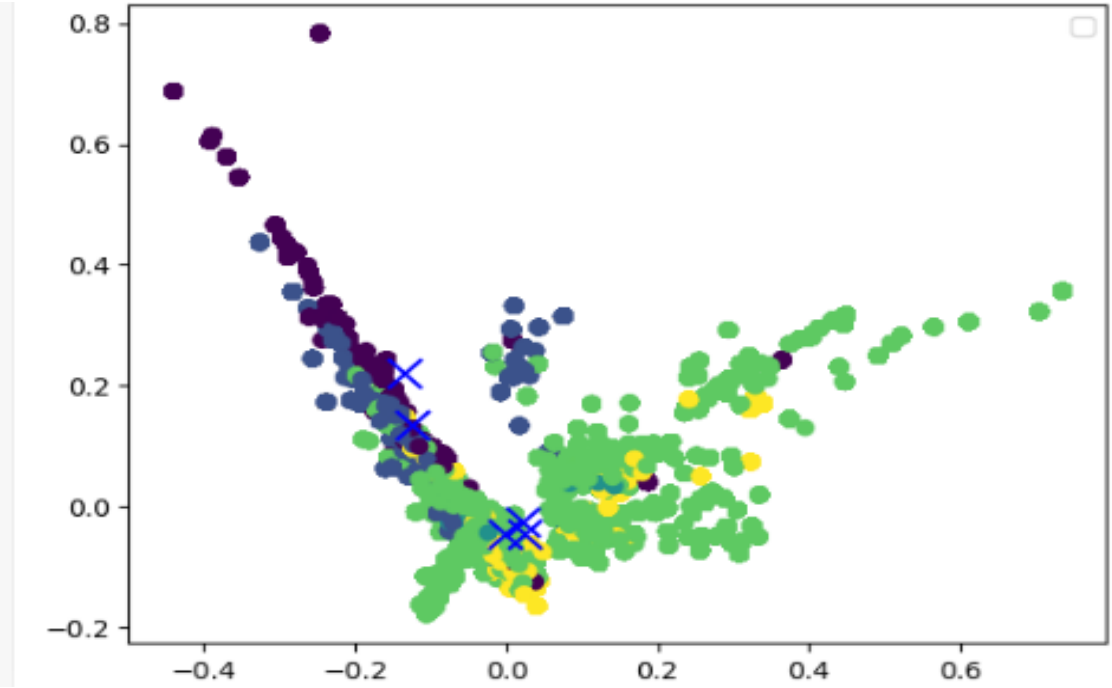

Figure 10 : Representation of products based on their similarity using PCA and TF-IDF

9 RESULTS LEADING TO ANSWER THE QUESTION

Market basket analysis using Apriori model provides promising results which enables businesses understand the behaviour and purchase habits of their customers more efficiently. Business can identify frequently bought products and correlations between products by analysing transactional data, which aids in inventory management and marketing strategies. Apriori algorithm is more efficient when compared to current state of art models for performing market basket analysis. Figure 11 depicts the results of Market Basket Analysis.

MARKET BASKET ANALYSIS

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric	MBA Value(%)
2	(SWEETHEART CERAMIC TRINKET BOX)	(STRAWBERRY CERAMIC TRINKET BOX)	0.047776	0.078274	0.037992	0.795205	10.159226	0.034252	4.500720	0.946802	79.520480
1	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.050878	0.158219	0.036798	0.723265	4.571294	0.028749	3.041826	0.823123	72.326454
3	(STRAWBERRY CERAMIC TRINKET BOX)	(SWEETHEART CERAMIC TRINKET BOX)	0.078274	0.047776	0.037992	0.485366	10.159226	0.034252	1.850293	0.978130	48.536585
0	(WHITE HANGING HEART T-LIGHT HOLDER)	(RED HANGING HEART T-LIGHT HOLDER)	0.158219	0.050878	0.036798	0.232579	4.571294	0.028749	1.236768	0.928084	23.257919

Figure 11 : Market Basket Analysis screenshot showcasing major metrics

Customer segmentation is a crucial approach for online merchants to understand their client base and target specific consumer groups with their marketing initiatives effectively. Retailers can utilise the K-means clustering approach to discover distinct consumer categories based on their purchasing preferences and behaviours and they can use this data to develop focused marketing campaigns and raise customer satisfaction. The RFM framework, a prominent method for customer segmentation that

has been successful in numerous studies which considers the recency, frequency, and monetary value of customer transactions. Retailers must constantly update their segmentation models to account for developments in consumer behaviour and preferences as well as changes in the market and the amount of competition. Figure 12 reflects the optimum number of clusters created by the elbow method used in customer segmentation.

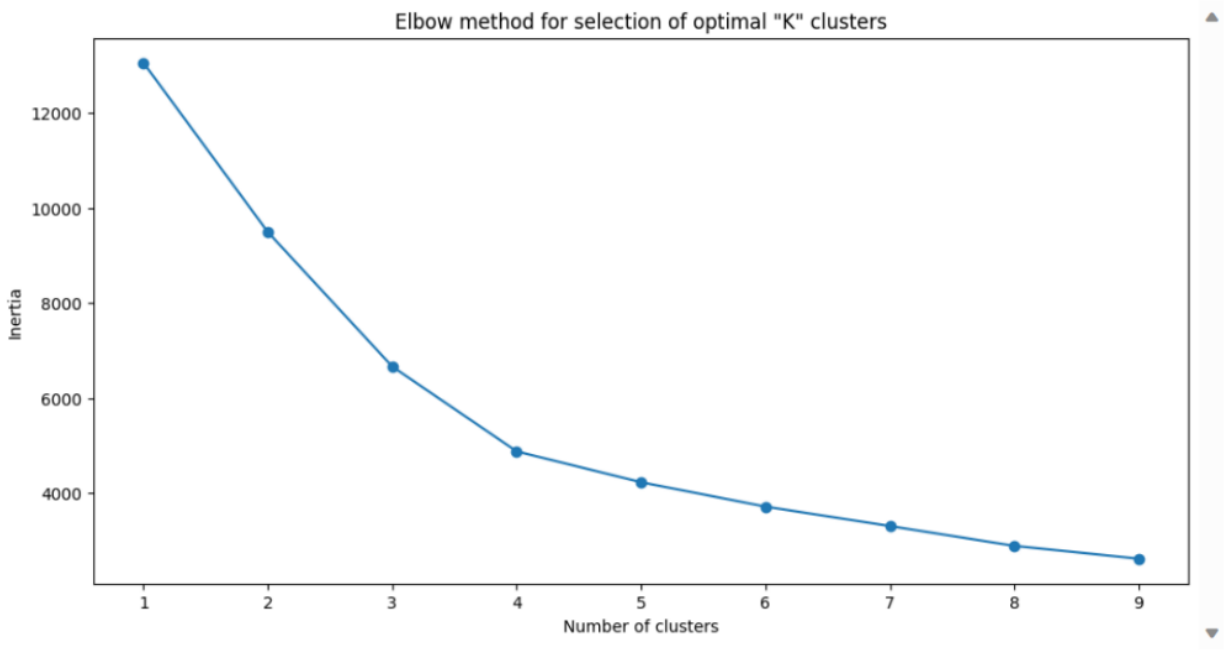

Figure 12 : Depiction of optimum number of clusters using Elbow method

Figure 13 i, ii and iii depicts the clusters of customers created based on the recency, frequency and the monetary value.

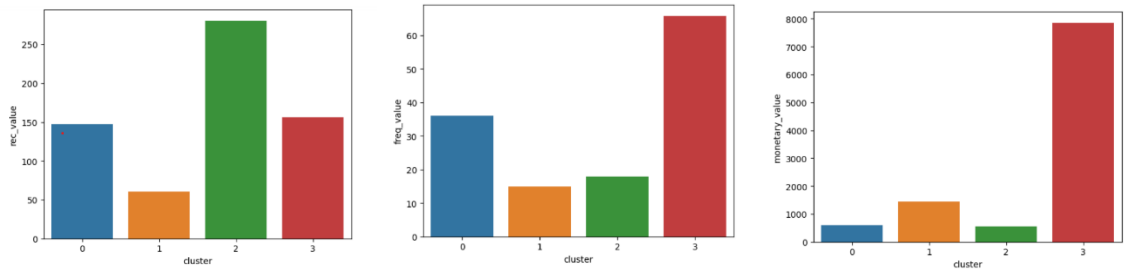

Figure 13 (i, ii, iii) : Bar-charts with clusters of customers on recency, frequency and monetary value from left to right

Demand forecasting is an important parameter of inventory control for e-retail. Retailers would estimate demand effectively by using machine learning algorithms like random forest. As a result, merchants can increase profitability and customer satisfaction by optimising inventory levels, reducing stock-outs, and avoiding surplus inventory. However, reliable data and the application of pertinent feature engineering techniques are necessary for accurate demand prediction. To reflect changes in consumer behaviour, tastes as well as changes in the market and competition, retailers should also regularly monitor and modify their models. Overall, demand forecasting can help merchants manage their inventories be more competitive.

Figure 14 displays the box plot of monthly sales based on quantity of items sold where the random forest predicts the monthly sales of the products.

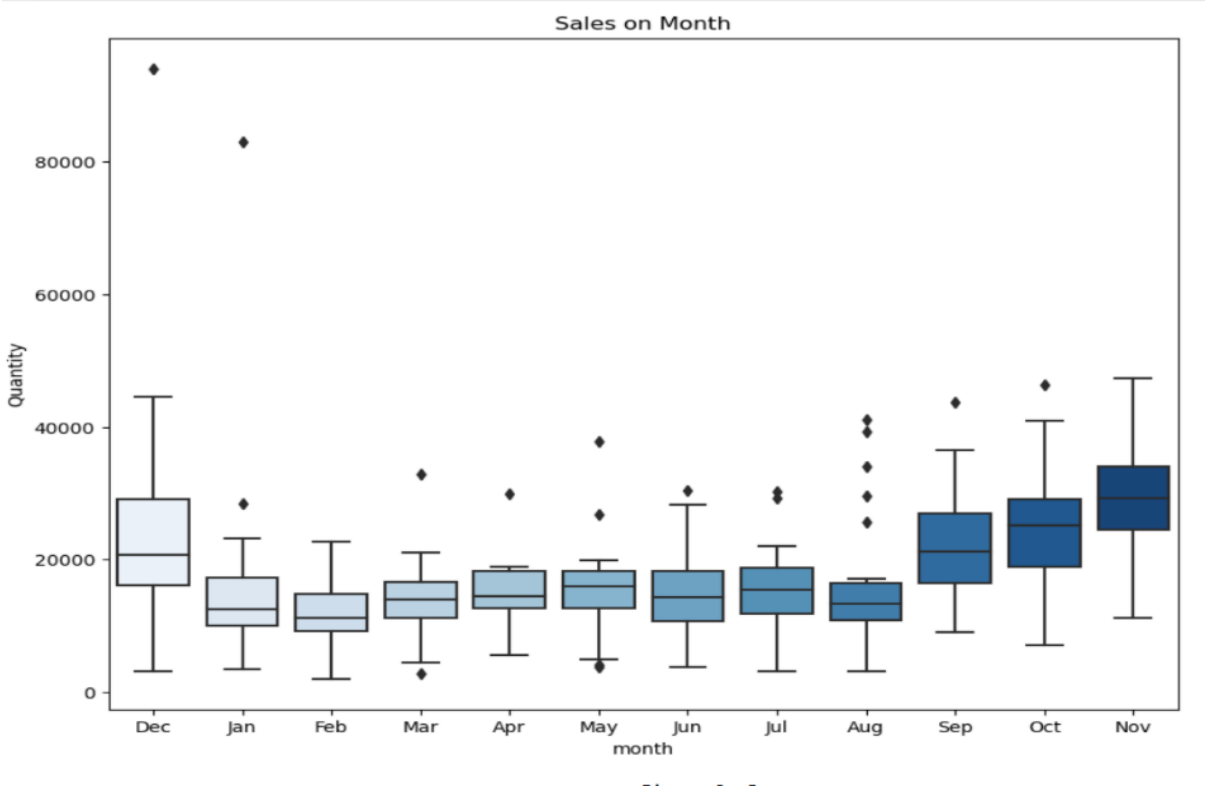

Figure 14 : Box plot displaying monthly sales on quantity

The research demonstrates how product compatibility analysis using k-means clustering provides online retail companies meaningful insights about the data. Business can uncover potential cross-selling possibilities and improve their inventory management by segmenting products based on their similar functionalities. This strategy helps in developing tailored recommendations for customers, enhancing their purchasing experience and eventually boosting client loyalty and overall experience. It is crucial to remember that the quality and quantity of the data provided, as well as the suitable choice of characteristics for clustering, have a significant impact on how accurate the clustering model is. To ensure the success of the clustering model, appropriate data analysis and feature selection should be performed. Figure 15 shows the clusters of the products based on similar functionality.

```
Cluster  0
----------------------------------------

33     PARTY CONE CHRISTMAS DECORATION
83     PARTY CONE CHRISTMAS DECORATION
92             BIRD DECORATION RED SPOT
129            BIRD DECORATION RED SPOT
130    PARTY CONE CHRISTMAS DECORATION
Name: Description, dtype: object

----------------------------------------
Cluster  1
----------------------------------------

292    GREY FLORAL FELTCRAFT SHOULDER BAG
293    PINK FLORAL FELTCRAFT SHOULDER BAG
374                JUMBO STORAGE BAG SUKI
828                    SKULL SHOULDER BAG
832                JUMBO STORAGE BAG SUKI
Name: Description, dtype: object
```

Figure 15 : Product Clusters formed based on similar functionality

10 DISCUSSION

The findings of this study indicate that few variables play a significant role in predicting online shopping intention. The most important indicators of intention to shop online, according to our data, are date of purchase, product diversity, and quantities purchased. These results emphasise the need of offering a convenient and varied online purchasing experience. Our findings and those of earlier studies did, however, differ noticeably in a few important ways. For instance, despite some research that security is a strong predictor of online purchasing intent, we were unable to detect a significant link between security and online shopping intent in our sample. This can be because of variations in our sample's features, or the specific products being sold on the online marketplace.

Our research's ramifications are twofold. First, our findings imply that when developing their online shopping platforms, online retailers should give user experience and product diversity top priority. Second, our study emphasises the need of considering the distinctive qualities of a target market or product when forecasting online purchasing intent. Retailers might be able to increase online buying intent and boost sales by customising online shopping platforms to a particular population's requirements and tastes. Of course, our study has few limitations that need to be addressed in further investigation. For instance, our study was restricted to a particular group and geographic area, and future research should investigate how generalizable our results are to other populations and areas. Future studies should also investigate the influence of additional variables, including social influence or perceived risk, in predicting online purchase intent. The challenge which we found to be prevalent while implementing Machine Learning models is that each models produce results which varied depending on the type of dataset used. Hence a validation of results of each model was needed through visualization and cross validation. The integration of all the four features was necessary to create a basic system for an organization to be more coherent. While implementing customer segmentation, we divide the customers based on recency and frequency of items bought but in truth the monetary value which takes the price of the item sold into consideration must have the highest weightage while calculating the RFM value since the customers with highest monetary value yield high profits to the organization. The inclusion of attributes that don't provide any meaningful information about the product being

searched is a flaw in the TF-IDF. Other words like "nice" and "small" are also worthy of argument. Although they can be used for matching, these low-information words have the potential to cause an unusual ranking in the search results. The IDF places emphasis on a term's rarity while ignoring its usefulness.

While building the product compatibility feature for a large dataset, a scatter plot was initially attempted but concrete inference couldn't be gathered from it and hence we decided to reflect the feature based on product clusters instead of cluster plots. And as far as the MBA is concerned, assuming that the transaction database is kept in memory, the MBA requires numerous database scans. The dataset for the Apriori method has a huge number of item sets and a low minimum support for some of the items. E-processing of our work is a matter of time and if the work is output wrong, the traders will have to face a lot of losses. As a result, organizations must operate after thorough examination. Our study sheds light on the variables that influence online shoppers' intent to purchase and emphasises the value of user experience and product diversity. Our findings imply that online merchants give priority to these characteristics when developing their online shopping platforms, and that further study should focus on the potential contribution of other variables to the prediction of online shopping intent.

In relation to other works, "Market Basket Analysis: Identify the changing trends of market data using association rule mining" **[3]** results have influenced greatly by the manual threshold values for score, so it is needed to automate the threshold values for better recognition of outliers. This paper also outlines data mining algorithms have been developed and applied on variety of practical problems.
In the paper, "A practical yet meaningful approach to customer segmentation" **[14]** Market segmentation is clearly one of the most important concepts in marketing. A-priori and specially created post-hoc assessments are two of the most crucial strategies that a firm should consider.
Researching the paper, "Application of machine learning techniques for supply chain demand forecasting" **[15]**, they investigate the applicability of advanced machine learning techniques, including neural networks, recurrent neural networks, and support vector machines, their findings suggest that while recurrent neural networks and support vector machines show the best performance, their forecasting accuracy was not statistically significantly better than that of the regression model.

In the paper, "Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms" **[16]** the goal was to identify the optimum procedure and strategy for achieving the highest accuracy with the best generalisation. To improve the classification accuracy, various feature engineering techniques like term frequency-inverse document frequency (TF-IDF), bag of words, global vectors for word representations, and Word2Vec are used in conjunction with hyperparameter tweaking of the classification models. According to experimental findings, the SVM achieves an accuracy of 89.55% when combined with TF-IDF features.

11 SUMMARY

Monitoring and managing a company's product flow is what inventory management entails. Tracking inventory levels, predicting demand, controlling storage and distribution, and maintaining correct record-keeping are just a few of the many tasks involved in effective inventory management. Businesses can increase customer happiness, decrease waste and associated expenses, and provide products when needed by optimising inventory levels. The inventory can be effectively managed by incorporating features Market Basket Analysis, Customer segmentation, Demand Forecasting and Product compatibility in business.

MBA gives retailers insight into which products are frequently bought together and is particularly helpful in inventory management because it allows them to update their inventory as necessary. Here are a few ways that market basket analysis is crucial for inventory control. Analysing market baskets helps to determine which related products are frequently bought together. For instance, if customers frequently buy coffee and milk together, a merchant could wish to place these products close to one another or provide them as part of a package deal. Retailers may optimise the effectiveness of their inventory management and boost sales by discovering complementary items. MBA can assist retailers in identifying long-term sales trends, such as which products are more well-liked throughout times of the year or occasions. To ensure that the merchant has adequate product on hand to match demand, this information can be utilised to alter inventory levels. Retailers can identify underperforming products and change their inventory levels by analysing purchasing trends. This can lessen waste and prevent companies from storing excess inventory of items that are not in high demand.

Because it enables retailers to comprehend the needs and preferences of various customer groups and modify their inventory strategies accordingly, customer segmentation is crucial for inventory management. Retailers can modify their inventory levels to satisfy demand while preventing overstocking by analysing sales patterns for each consumer category. It can identify which goods are frequently bought in tandem by specific customer categories. Utilising this data, inventory management strategies can be modified to take advantage of cross-selling opportunities. Retailers can improve client loyalty by customising their merchandise to the wants and needs of various consumer segments. Customers are more inclined to shop at a store again if they can find products that suit their individual wants and tastes.

In our research, we were able to predict customer shopping intentions and estimate the number of items needed, both of which will aid the organisation in managing its inventory. To enhance sales and increase profits, an organisation will raise the supply of a highly demanded commodity in its inventory. An organisation raising the supply or amount of winter jackets in November because historical data suggests that this is the month when winter jackets are sold the most would be a simple real-world example of demand prediction.

With the help of the product compatibility function, the business links related items so that, if one in high demand runs out of supply, the other connected product can be replenished and recommended to clients. These elements are crucial in helping a business manage its inventory using the intentions of the consumer.

12 GROUP WORK

The Group had performed intensive research in deciding the project topic and the specific problem statement, once the specific research question was finalized, we started doing research on previous projects under this topic, decided on the project's end goal and objective to achieve. Following this, we started with searching and collecting the data. We used the webserver logs of an online retail website and did data cleaning and preprocessing on it to change it to our required format. Data was analyzed and visualized to get a better look at the underlying trends and patterns. We completed Exploratory data analysis to get better insight of the data and decide on the various features and models we would implement for the project. We started with the model building for the first feature which is Market Basket Analysis before the midterm evaluations. The project was being worked on steadily with each milestone mentioned in the project's Gantt chart being achieved without fail.

Subsequently, we started with model building iteratively for the other features of the project and evaluating them to figure out the best performing models. We uploaded our code to GitHub and Deepnote for ease of access, version control and to showcase them for reproducing the results. The results were recorded and documented in the report for the project. We studied various papers on the different techniques used across our project and cited them in the report wherever appropriate. The report was made while following all the necessary formats and conventions and was completed while showcasing all the results achieved. Every individual supported each other to bring the best out of one another and supported the group in completing its milestones.

13 INDIVIDUAL WORK

Madhurima Sarkar, has contributed to the below areas as part of this study:

- Implementation of MARKET BASKET ANALYSIS from end to end starting from organizing and transforming the transactional data to a format which would fit the MBA algorithms. Performed intensive research in selecting the best algorithms such as Apriori or FP-Growth to analyse the data and implementing them. Adjusting algorithm parameters to achieve the desired level of accuracy and efficiency. Finally creating the visualizations to help stakeholders understand the results of the analysis, such as association rules, item-sets, or lift charts.
- Complete implementation of PRODUCT COMPATIBILITY which involves examining the compatibility requirements for various systems or products and detecting any potential compatibility problems and then performed a thorough study of the current state of the art algorithms that would provide best possible solution to the feature implemented to the sample dataset in the study. Based on the metrices calculated found out that TF-IDF vectorizer method provided promising results to group the products with similar functionality but different names. Created meaningful representations from the algorithm which would aid to effective inventory management.
- Collaborated with the team and ensure smooth integration of these features with the remaining features in the scope of study.
- Have provided the domain and technical expertise to the team in visualization and documenting the details.

Arjunkumar Wandiwash Krishnakumar, has contributed to the below areas of this study:

- Data collection, Researched and collected for the appropriate data used for the project across many data sources and then transformed them into usable format. Did the Data collection, pre-processing and various analysis on the data.

- Implementation of the exploratory data analysis of the data was done and various checks and pattern recognition was completed. The complete visualisation and reporting of the various findings were completed and showcased.
- Worked on the report and implemented the various conventions required.
- Researched the various literature used across this project and included their findings wherever necessary.
- Collaborated with other group members and provided support in the various coding challenges and ensure smooth progress.
- Did the Visualisation and documentations of the project and did the reporting.
- Implemented the version controls of the code and the setup of GitHub and Deepnote implementations.
- Provided the technical expertise and domain knowledge for the smooth progress of the project.

Vignesh Venkataraman, was responsible for the following sectors as part of this project:

- Feature of Customer Segmentation: Performed extensive research to implement the K-means clustering and RFM model in the second feature Customer Segmentation.
- Debugging: Fixed the errors in the demand forecasting algorithms and TF-IDF models.
- Data Engineering: As a data engineer, was responsible for designing and implementing the data infrastructure, also ensure that dataset is properly collected, stored, and processed for analysis. Executed the transformation part of EDA by converting the dataset into a dictionary so that the values can be accessed for all the features modelled.
- Decision making: From the options determined at the determining alternatives stage, picked the best choice through evaluation of each choice satisfying the requirements.
- Communicating the findings to the team: Responsible in communicating the findings to the rest of the team in a clear and concise manner by Using data visualization tools to make the findings easy to understand.
- Testing: Tested and refined the models for all the four features till desired level of accuracy and efficiency is achieved.

Ramanathan Kathiresan, have contributed to the following sectors as part of this project:

- Implementation of the feature “Demand Prediction” using Support Vector Machine and Random Forest Regressor to provide accurate prediction of future demand based on the product feature and the past data.
- EDA: Collaborated with the team in Exploratory Data Analysis in Data Collection and Data Pre-processing.
- Data Analyst: Analysing data sets to make decisions and identify trends that might help business outcomes.
- Problem solving: Responsible in Understanding the problem. Identify, Evaluate and Implement the solution and to ensure that the team is able to tackle problems effectively and achieve its objectives.
- Report: Collaborated with the team in documentation of the Project Report.

REFERENCES

[1] Stanelytė, G., 2021. *Inventory Optimization in Retail Network by Creating a Demand Prediction Model* (Doctoral dissertation, Vilniaus Gedimino technikos universitetas).

[2] Stanelytė, G., 2021. *Inventory Optimization in Retail Network by Creating a Demand Prediction Model* (Doctoral dissertation, Vilniaus Gedimino technikos universitetas).

[3] Kaur, M. and Kang, S., 2016. Market Basket Analysis: Identify the changing trends of market data using association rule mining. *Procedia computer science*, *85*, pp.78-85.

[4] Agrawal, R. and Srikant, R., 1994, September. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

[5] Borgelt, C., 2005, August. An Implementation of the FP-growth Algorithm. In *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations* (pp. 1-5).

[6] Cooil, B., Aksoy, L. and Keiningham, T.L., 2008. Approaches to customer segmentation. *Journal of Relationship Marketing*, *6*(3-4), pp.9-39.

[7] Anitha, P. and Patil, M.M., 2022. RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), pp.1785-1792.

[8] Cohen, M.C., Gras, P.E., Pentecoste, A. and Zhang, R., 2022. *Demand prediction in retail: A practical guide to leverage data and predictive analytics*. Springer.

[9] Stanelytė, G., 2021. *Inventory Optimization in Retail Network by Creating a Demand Prediction Model* (Doctoral dissertation, Vilniaus Gedimino technikos universitetas).

[10] Yue, L., Yafeng, Y., Junjun, G. and Chongli, T., 2007, August. Demand forecasting by using support vector machine. In *Third International Conference on Natural Computation (ICNC 2007)* (Vol. 3, pp. 272-276). IEEE.

[11] Wang, Q., Chen, Y. and Xie, J., 2010. Survival in markets with network effects: Product compatibility and order-of-entry effects. *Journal of Marketing*, *74*(4), pp.1-14.

[12] Qaiser, S. and Ali, R., 2018. Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, *181*(1), pp.25-29.

[13] Prasetyo, V.R., 2018, August. Searching cheapest product on three different e-commerce using k-means algorithm. In *2018 International Seminar on Intelligent Technology and Its Applications (ISITIA)* (pp. 239-244). IEEE.

[14] Marcus, C., 1998. A practical yet meaningful approach to customer segmentation. *Journal of consumer marketing*, *15*(5), pp.494-504.

[15] Carbonneau, R., Laframboise, K. and Vahidov, R., 2008. Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, *184*(3), pp.1140-1154.

[16] Naeem, M.Z., Rustam, F., Mehmood, A., Ashraf, I. and Choi, G.S., 2022. Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms. *PeerJ Computer Science*, *8*, p.e914.