National College of Ireland

Retail Manufacturing Analysis using Machine Learning Techniques.

MSc Research Project
Data Analytics

Meet Sangoi
Student ID: X21207526

School of Computing
National College of Ireland

Supervisor: Arjun Chikkankod

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	MEET DEEPEN SANGOI.
Student ID:	X21207526
Programme:	MSc in Data Analytics **Year:** …2023……
Module:	MSc Research Project
Supervisor:	Arjun Chikkankod
Submission Due Date:	14-08-2023
Project Title:	Retail Manufacturing Analysis using Machine Learning techniques.
Word Count:	6951 **Page Count**……24……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Meet Deepen Sangoi
Date:	13-08-2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Retail Manufacturing Analysis Using Machine Learning Techniques.

MEET DEEPEN SANGOI
Student ID – X21207526

Abstract

It is easy to make confident business decisions by picking up understanding customer into customer behaviour through predictive investigation. The objective is to focus customer obtaining propensity. Having numerous competitors to discover unused customers and keep existing clients, has come about in incredible bargain of pressure between competing businesses. As a result, offering the leading customer benefit and continuously keeping the stock more stock in advance may advantage companies in all ways. The method of gathering a customer base into a few categories based on demands, conduct, cash, etc. is known as Customer Division or customer categories. While item division is the method of classifying items into diverse categories. The major objective of our project is to separate customers and items into different clusters based on distinctive criteria to procure comes about. There are a few calculations which will parcel customers based on different measurements. These are utilized to find covered up designs in data, discover important, steadfast customer, get it consumer obtaining propensities, and more. To move forward decision-making and make the foremost exact show conceivable, cluster examination is done. This research work has compared five for products and eleven for customer, clustering calculations and chosen the most excellent of them for advance analysis.

1 Introduction

The organization's fundamental objective is to get it to its retail division clients. Since it is presently troublesome for businesses to dodge having a relationship with their clients, having a framework input to oversee that relationship is vital. Client relationship administration, too alluded to as CRM, could be a strategy of controlling client relationships with endeavors and is broadly utilized within the cutting-edge world.

Segmentation of customers is a critical ingredient of customer relationship management that helps businesses successfully market to their target audience. Businesses must understand their customers and show this understanding by engaging with them only in a pertinent and targeted manner. Customers want to be valued and treated uniquely, but all but the tiniest businesses are unable to achieve this degree of customer awareness. Segmentation also makes it possible for businesses to allocate resources wisely. Customer psychological and demographic data analysis offers a range of insights that support the creation of new goods and services as well as the forecasting of consumer wants.

As a result, this enables marketers to target the customers or prospects who are most likely to be fascinated with them. Businesses can use a variety of marketing methods to each

subdivided group by using customer segmentation. Client happiness and expected earnings for a corporation are enhanced by customer segmentation.
This study suggests a technique for determining ideal customers based on past purchases. The suggested approach groups customers to identify secret trends and knowledge that can advance the organization's commercial endeavours. If you can determine the optimal number of different consumer groups, it becomes simple to understand the variations between your clients and meet their wants. Customer segmentation improves customer experience while increasing company revenue. Segmentation is essential if you want to outperform your competitors and attract more customers.

1.2 Research Question

- How can a predictive model be developed to forecast the future purchasing behaviour of new customers, specifically estimating their purchases for the upcoming year starting from their initial transaction?

2 Related Work

(V. Mehta, 2021) This paper is a literature survey on customer segmentation using machine learning algorithms to find prospective clients. It discusses the importance of market segmentation in understanding customer needs and improving retailer-consumer relationships. The paper focuses on unsupervised learning, specifically cluster analysis, as a method for grouping customers based on their geographic, demographic, psychographic, and behavioural traits. The paper also analyses and compares various clustering algorithms, such as K-means, DBSCAN, and hierarchical clustering, using metrics like Silhouette and Davies Bouldin. Overall, the paper provides a comprehensive overview of the use of machine learning algorithms for customer segmentation.

These days, most companies utilize buyer conduct mining systems for the good thing about their clients. These models can be utilized to estimate an assortment of deals, foreseeing, and promoting patterns. To create forecasts around client conduct, machine learning and information mining approaches have been connected. The objective of this ponder was to look at the viability of a deep-learning (DL) procedure in distinguishing critical components impacting online buys that are associated to client engagement on stages and clients. Since the DL method's thick organize design makes it conceivable to discover modest designs in datasets, it is fitting for this request. This work coordinating distinctive explanatory strategies, such as both conventional machine learning (ML) methods and the state-of-the-art DL strategy, to address the subject of the think about (Singh, 2023).

(Tawfiq, 2021) Essentially in this paper, they utilised administered machine learning to analyse and anticipate client behaviour. The inquire about centred on making a prescient system that may be utilised in trade and taken after all the forecast stages. Amid the investigate, behavioural information was collected from the members some time recently the

examination factors were shaped. Information preparing and testing were then conducted, and the expectation demonstrate utilized within the inquire about was recognized. The forecast was conducted utilizing discriminant examination and other prescient models. The comes about uncovered that the KNN demonstrate was the foremost precise, taken after by calculated relapse. The choice trees model developed as the slightest exact. Analysing information taken from web patterns is the foremost successful ways of foreseeing client behaviours. The examination uncovered that single univariate models are not great indicators of client behaviour on design patterns.

(Wang, 2013) Before identifying out-of-role behaviours (OCBs), I need to understand the behaviours expected in the role customer support context. expected behaviour of a customer service officer covers a wide range. Key job responsibilities in call centres include meet customer requirements, provide customers with information about products and services, handling and Resolve customer complaints and conduct research necessary information using available sources. According to Call Centre Helper, a popular call centre UK magazine, call centre Employees are evaluated according to some common criteria metrics, including the total number of calls handled, abandoned call rates and average calls Period. All these performance metrics focus on the effectiveness of customer interactions, for example: total number of calls and response time. For an additional role behaviour in this context, I focused on effectiveness of customer service interactions.

(Kumar, 2020) The study has proposed a consumption analysis model that helps telecom operators to predict which customers are most likely to be rejected. The system uses machine learning strategies based on big data. The standard measure of the area under the curve (AUC) was used to evaluate the performance of the model. The dataset used for the study was provided by the telecommunications company Syriatel. The model works with 4 methods: Decision trees, random forests, gradient-enhanced machine trees (GBM) and extreme gradient-enhanced machines (XGBOOST). Hortonworks Data Platform (HDP) has been selected as the big data platform. The Spark engine has been used in almost every stage of the product such as data analysis, feature development, training, and software testing. The hyperparameters of the algorithm were optimized using K-fold cross-validation. Since the target class is unbalanced, the training sample is rebalanced by taking one data sample to balance the two classes. Start training on decision tree algorithm and optimize hyperparameter depth and maximum number of nodes. In Random Forest and GBM, the best results show that the best number of trees is 200 trees. And GBM works better than DT and RF.

(H. Valecha, 2018) In this paper Within the ultramodern age of innovation, expectation of advertise slant is exceptionally vital to watch buyer conduct in this competitive world as patterns are unstable. Building on improvements in machine learning and earlier work within the science of conduct forecast, I built a show planned to foresee the behaviour of Shopper. The point of this term paper is to look at the connection between customer conduct parameters and readiness to purchase. To begin with we examine to discover relationship between buyer conduct to purchase items on changing parameters such as natural figure, organisational calculate, person calculate and interpersonal figure. In this way this paper proposes time-evolving arbitrary timberland classifier that leverages interesting include designing to anticipate the conduct of customer that influence the choice of acquiring the

item altogether. Comes about of irregular woodland classifier are more exact than other machine learning calculation.

(Cirqueira, 2019) This paper shows that advanced retailers are encountering an expanding number of exchanges coming from their customers online, a result of the comfort in buying merchandise through E-commerce stages. Such intelligent compose complex behavioural designs which can be analysed through prescient analytics to empower businesses to get it shopper needs. In this plenitude of huge information and conceivable apparatuses to analyse them, an orderly audit of the writing is lost. In this manner, this paper presents an orderly writing survey of later inquire about managing with client buy expectation within the E-commerce setting. The most commitments are a novel expository system and a inquire about motivation within the field. The system uncovers three fundamental assignments in this survey, to be specific, the expectation of client entombs, buying sessions, and buy choices. Those are taken after by their utilised prescient strategies and are analysed from three perspectives. Finally, the inquire about plan gives major existing issues for advance inquire about within the field of buy behaviour expectation online.

(M. Guan, 2022) The online retail market sees a spike in traffic during holiday sales. The ability to differentiate between customers who are more likely to purchase is critical to driving traffic and offering cost-effective promotions. Only browsing will be considered in this article. Purchasing behaviour of online shoppers during the annual sales event in China, the world's largest online marketplace, based on 31Characterized and identify steps to purchase from millions of behaviour logs collected from vast geographies and precursor. Examine the impact of time (e.g., date, time of day), environment (e.g., platform, category viewed) and action (e.g., action session time, clicks, orders at time of purchase). Action instructions from shopping behaviour can be used for early detection in most cases, The shopper initially has a strong purchase intent, but the moment of ordering he arrives within 30-30 seconds rather impulsively. A few minutes of browsing. The prediction accuracy reaches a high AUC of 0.924. The results of this article provide an understanding of this. Traffic during the Mega Sale event. Help online stores plan upcoming shopping festivals and provide a better user experience.

(S. Peker, 2018) I can see that shopping list prediction is an important task for businesses as it can help provide a specific customer with a personalised product list, while improving customer satisfaction. customer satisfaction and loyalty. To predict customer behaviour, many of the studies in the literature have used methods of modelling customer behaviour at the individual and segment-based levels. However, previous attempts to predict customer shopping lists rarely used these edge methods. In this way, this article introduces a segment-based approach to shopping list prediction, and then presents an empirical comparison between individual and segment-based approaches in this regard. For this purpose, well-known machine learning classifiers and customer purchase history are used, and the comparison is made on a real data set by conducting a series of tests. The results show that there is no clear success in this comparison and the performance of customer behaviour modelling methods depends on the machine learning algorithm used. This study can help researchers and interpreter understand different phase of utilising customer behaviour modelling methods in shopping list prediction.

(Kansal, 2018) The zeitgeist of cutting-edge period is development, where everybody is involved into competition to be way better than others. Today's commerce run on the premise of such development having capacity to enthral the clients with the items, but with such an expansive flatboat of items take off the clients bewildered, what to purchase and what to not conjointly the companies are bewildered around what segment of clients to target to sell their items. Usually where machine learning comes into play, various calculations are connected for disentangling the hidden patterns within the information for superior choice making for the long run. This evade concept of which fragment to target is made unequivocal by applying division. The method of portioning the clients with comparative practices into the same portion and with distinctive designs into diverse portions is called client division.

(M. Aryuni, 2018) In this paper, The Internet banking clients has development exceptionally quick. Client division can be connected based on Web managing an account information. Clustering is unsupervised information mining method that can be utilized for client division. This inquiries about construct clustering models on client profile information based on their utilization of Internet Managing an account in wxyz bank. The clustering strategies utilized K-Means strategy and K-Medoids strategy based on RFM score of customer's Internet Banking exchanges. This inquiries about utilized Information Revelation strategy. The exhibitions of both strategies were measured and compared. The result appears that K-Means strategy berated K-Medoids strategy based on intra cluster (AWC) remove. Whereas based on Davies-Bouldin record, K-Means performs somewhat way better than K-Medoids.

(N. R. Maulina, 2019) Enormous information and progressed data analytics in organizations are overwhelming in customer-centric offices such as showcasing, deals, and client benefit. In this paper, diverse clustering calculations will be compared, particularly centroid-based clustering K-Means, CLARA, and PAM with Fluffy C-Means clustering. The reason of this inquire about is to discover ideal number of clusters utilizing clustering calculation with the most excellent approval degree score. Dataset is obtained from Tech Company in Indonesia that give machine with Point of Deal framework for nourishment and refreshments dealers, since the company in B2B settings. Among three clustering strategies, K-Means have the finest approval degree score. After compared to Fluffy C-Means, K-Means outflanks FCM based on time complexity and quality of clustering. Cluster examination is done to distinguish client data. Hence, this investigates able to convey a shrewd understanding around client characteristics utilizing huge information analytics and give a viable Client Relationship Administration Frameworks.

(Jayasena, 2020) Collaboration between businesses and consumers is increasingly important with the development of technology. It is necessary to manage this relationship for the future growth of the business. The communication defence mechanism between the organisation and its customers is called Customer Relationship Management (CRM). CRM plays an important role in the business field. In this study, six groups were identified based on annual income and expenditure scores. Also define six clusters according to two principal components. After successful integration, lead the organization to make correct decisions and the organization can offer new products and services and can change the existing product offerings as required of customers by identifying the right customers.

(Ozan, 2018) Customer division is vital both in customer relationship administration writing and program. The foremost common way to isolated one client from another is to advance a

bunch of clients as premium and the remaining clients as standard. In this work, a company's physically sectioned client information is analysed. The consider points to illuminate the company's information division issue by utilizing its genuine information with respect to its customers' instalments. Since, machine learning strategies are valuable to fathom issues approximately information administration, the arrangement is looked inside machine learning strategies. Distinctive classification strategies which are utilized to separate between premium and standard clients having a place to a company's database are matched. Two-dimensional instalment data of clients are utilized as input factors (highlights) and the strategies are compared agreeing to their division exhibitions.

(Kavitha, 2021) Online shopping could be an exceptionally much created wonder in numerous nations. Distinctive variables affect the online customer buying behaviour. These days, individuals across nations like to shop online instead of shopping in conventional implies due to their straightforwardness and ease. Online shopping incites real comparison of products and client administrations from shopping websites which is additionally called as business-to-consumer (B2C) e-commerce. This work is based on the respondent input towards online shopping collected employing a survey. The most deliberate is to ponder the relationship between online shopping with age, item inclination, buy recurrence, instalment security with sex, and source of online shopping with age.

(X. Chen, 2018) Clustering of client transaction information is an imperative strategy to analyse client behaviours in retail and E-commerce companies. Note that items from companies are frequently organized as an item tree, in which the leaf hubs are merchandise to offer, and the inner hubs (but root hub) may well be numerous item categories. Based on this tree, we propose the “personalized item tree”, named buy tree, to speak to a customer's exchange records. So, the customers' transaction information set can be compressed into a set of buy trees. At last, the clustering comes about are gotten by allotting each client to the closest agent. We also propose a crevice measurement-based strategy to assess the number of clusters. An arrangement of tests was conducted on ten real-life exchange information sets, and test comes about appear the prevalent execution of the proposed strategy.

(V. L. Narayana, 2022) So, we have appeared our speculative firm as an illustration, and you're attempting to figure out how well a specific item will perform from a marketing perspective. Clients may well be sectioned based on their market behaviour. Take note that the sum of information accessible is colossal, which cannot handle it with our human faculties alone. Machine learning calculations and handling control will be utilized. Unsupervised learning has a few vital employments, one of which is client division. Distinguish client sections to centre on the conceivable client base by utilizing clustering methods (K-means, Agglomerative, and Cruel Move). As a result, they portion clients into bunches based on comparable variables such as sexual orientation and age as well as interface and investing propensities.

(S. De, 2021) Customer churn, a significant issue impacting a company`s revenue and growth, is a top concern for C-level executives, with 91% highlighting its importance according to a Gartner survey. Despite this, only 43% have invested in resources for customer expansion. To foster growth, retaining existing customers is crucial. This paper examines recent machine learning techniques for predicting customer churn. It analyses methods, datasets, and performance, highlighting hybrid and ensemble approaches as

successful. However, it underscores the need for clear evaluation guidelines. While current methods are quantitative, there's a gap in research focusing on information-rich interactions like emails and chats. This paper aids industry awareness of ML trends in churn prediction and guides researchers in their work.

3 Research Methodology

This chapter provides a brief and concise description of the methodology used to conduct this study. The main purpose of this research work is to segmentation of customers using various algorithms in machine learning.

The KDD also referred as Knowledge Discovery in Database method is used in the tests. KDD method is a repetition and sequential process that operates in different phases and is used to discover wisdom in data. The steps are Data selecting, Data preprocessing, data transformation, data mining, interpretation, and evaluation. The KDD method task is illustrated in the figure below:

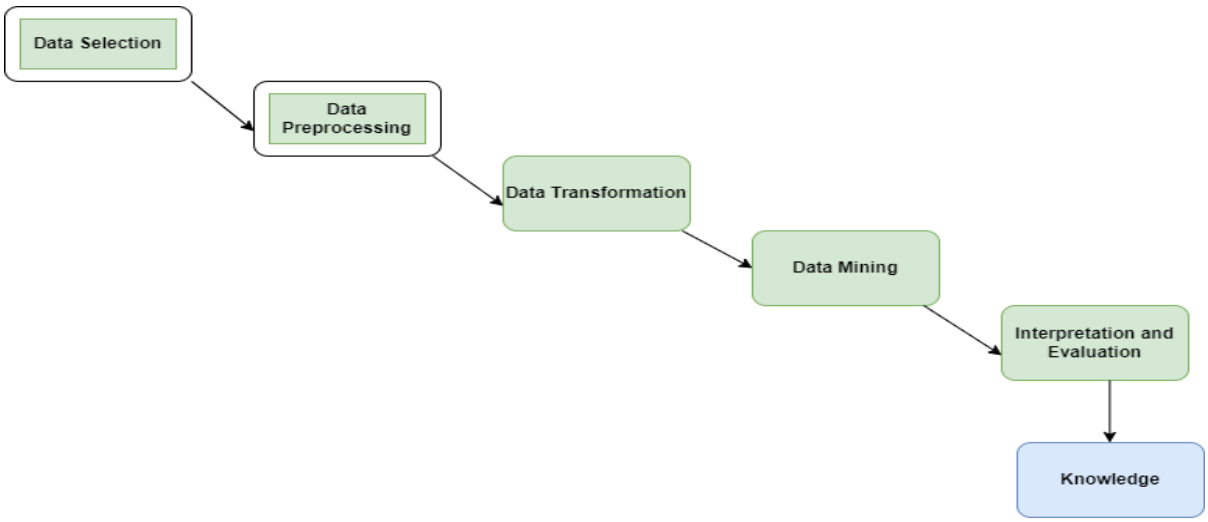

Fig.1. KDD Flowchart

3.1 Data Selection/ Preparation

The dataset Link is provided here: "https://www.kaggle.com/datasets/carrie1/ecommerce-data". The dataset is accessible from Kaggle, and it is open-source data, and does not require consent or approval. This includes all out-of-store transactions made between 1 December 2010 and 9 December 2011 for registered online retailers based in the UK. is a dataset. The company primarily vends antique presents for all occasions. Organisation's numerous clients are wholesalers. It is important to maintain customer relationship with them. My goal is to build a model that predicts future purchases for all new customers by evaluating their first and last transaction.

The Dataset consists of '541909' rows and '8' columns. The detailed description of the 'E-commerce' dataset is given below:

- Invoice Number on Bills: Unique number to identify the customer's order.
- Stock code (in numbers): Each personalized code consists of numbers and/or letters for a specific product.
- Description of each product: The title of the individual product.
- Quantity: The amounts of each item or thing per exchange.
- 'Invoice Date': The time and date on the receipt.
- 'Unit Price': The cost for an individual product.
- 'Customer ID': A special number on your receipt that's utilized to mention your account.
- Countries: The name of the countries.

3.2 Data Transformation

Converts "DataFrame to date/time format". This helps in accurately processing and analyzing date-related data. Once the conversion is complete, you will see a DataFrame with updated content with the "InvoiceDate" column now in date/time format. This transformation is important for tasks such as time-based analysis and sorting within a DataFrame. The visualisation for the same is visible below:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680.0	France

541909 rows × 8 columns

Fig. 2. Date and Time Format

The data cleaning step was carried out, the dimensions of the DataFrame were altered, and the current shape is printed to indicate the change, showing the number of rows and columns in the DataFrame. This action was taken to ensure data quality and consistency for subsequent analysis. Then we again checked the null values for the purity of the data. This step is crucial for assessing data quality and deciding on further data preprocessing steps if needed. We dropped the duplicate data records to maintain accurate and reliable data for meaningful analysis and decision-making. The record was around ‘5225’ rows.

3.3 Exploratory Data Analysis/ Data Preprocessing

Raw data must be processed to build perfect models and make exact predictions. Therefore, data preprocessing is very important for our research. Data preprocessing means transforming, shaping, structuring, and encoding data in ways that are readily predictable to algorithms. The dataset considered in this research contains both numerical and alphabetical variables. Explore the data to find elementary connection between all variables in the data set and assemble the data set into a frame. The library named pandas is carried out here for data preprocessing. It feeds machine for processing, analysing, and manipulating data.

A choropleth map visualizes the distribution of customers based on their country. This map illustrates the number of orders per country using varying colours. "The generated choropleth map provides insights into the geographical distribution of customers. Each country is color-coded based on the number of orders originating from there. The darker shades represent a higher number of orders, while lighter shades indicate relatively lower order counts. The map clearly shows that most customers are in the United Kingdom (UK). The colour scale, ranging from light to dark, helps to visually convey the distribution pattern. The 'Number of orders per country' title emphasizes the map's purpose. The map is created using a Mercator projection to ensure accurate geographical representation." The image of choropleth map is show below:

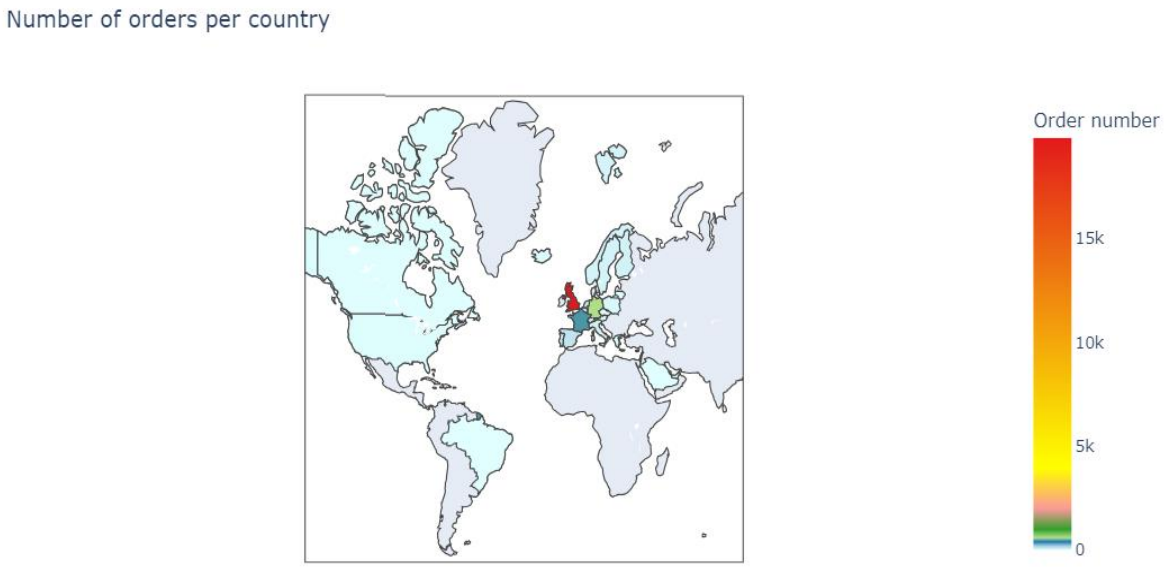

Fig. 3. Choropleth Map

We saw that a significant portion of orders were cancelled. Out of the total number of transactions, a substantial 16% were identified as cancelled orders i.e., “3654/22190”. The invoice number starting with ‘C’ is referred as cancelled products. For more purity of the data, we cleaned entries to remove and doubtful entry.

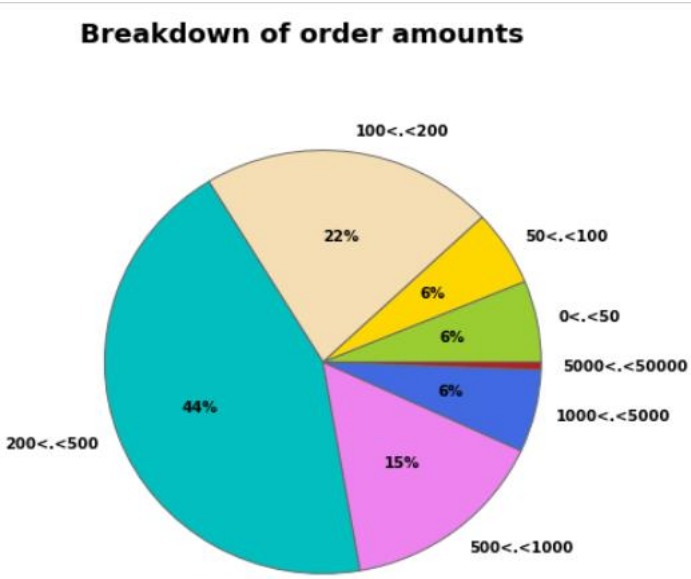

Fig.4. Pie Chart

In the above fig.4. "The generated pie chart shows the distribution of order amounts over a predefined price range. Each slice of the pie corresponds to a specific price range, with the range from £0 to £50 being '0<. <50”. The slice size represents the number of orders within each price range. In particular, the chart shows that a significant percentage of orders (about 65%) correspond to relatively large purchases of over £200. This suggests that a significant

percentage of customers are making higher value transactions. This visualization effectively captures order quantity distribution and provides insight into customer spending behaviour. "

3.4 Data Mining

This list of unique product descriptions can be useful for various purposes such as analysing the range of products available, identifying popular or niche products, or categorizing products for further analysis. "These Keywords collectively offer a comprehensive perspective on the products 'key characteristics', empowering us to perform a more insightful analysis of the dataset." The preserved words are 193. The silhouette scores for each cluster count were then printed, providing valuable insights into the effectiveness of the clustering. This analysis aids us in discovering natural groupings among products, contributing to a deeper understanding of the product landscape and potential business strategies."

It starts with 5 clusters and iteratively increases the number of clusters until the silhouette score reaches a certain threshold (0.145 in this case). The silhouette score is used as a measure of how well-separated the clusters are. The loop terminates when a satisfactory silhouette score is achieved. The norm silhouette score for products is 0.15. This score helps assess how well the data points are grouped into clusters, with higher scores indicating better-defined and well-separated clusters. This visualization assists in understanding how well the clustering algorithm has grouped the data. It ranges from -0.07 to 0.33. The provided parameters include the number of clusters used in the analysis, a range for setting the x-axis limits on the graph, the length of a dataset, an array of silhouette scores calculated for each data point, and the cluster assignments for each data point. This graphical representation helps assess the quality and distribution of silhouette scores across different clusters, aiding in the evaluation of the clustering results:

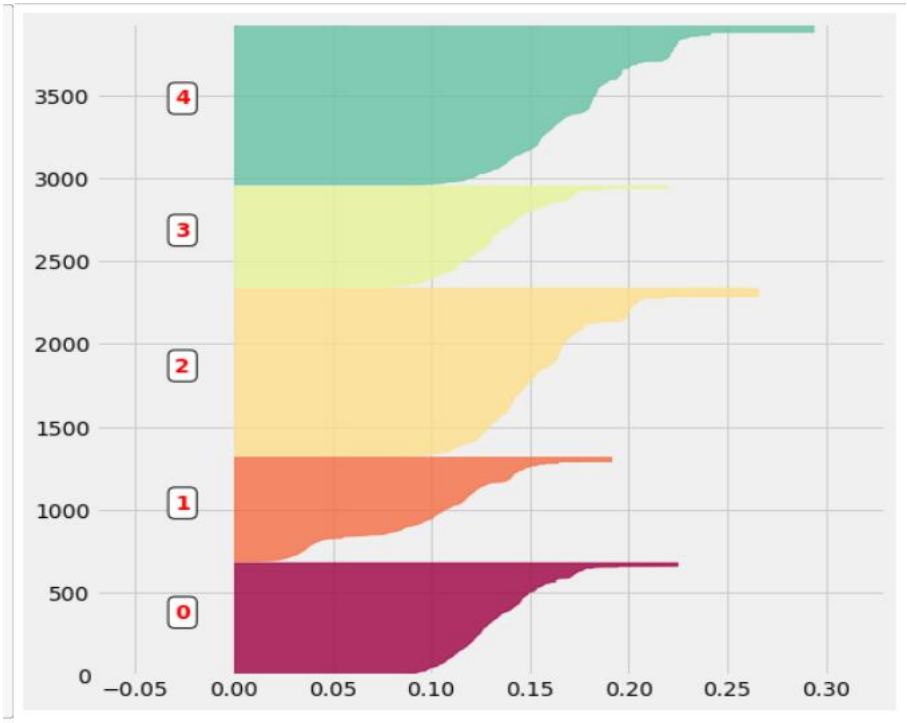

Fig.5. Silhouette Scores for Products

We utilized Principal Component Analysis (PCA) as a statistical technique to better understand our dataset's underlying structure. PCA benefits us by reducing data dimensions, highlighting prominent variances, and enabling visual exploration. This aids in efficient analysis, variance comprehension, and pattern visualization within our dataset. We visualized the explained variance of Principal Components Analysis (PCA) using a step plot and bar plot.

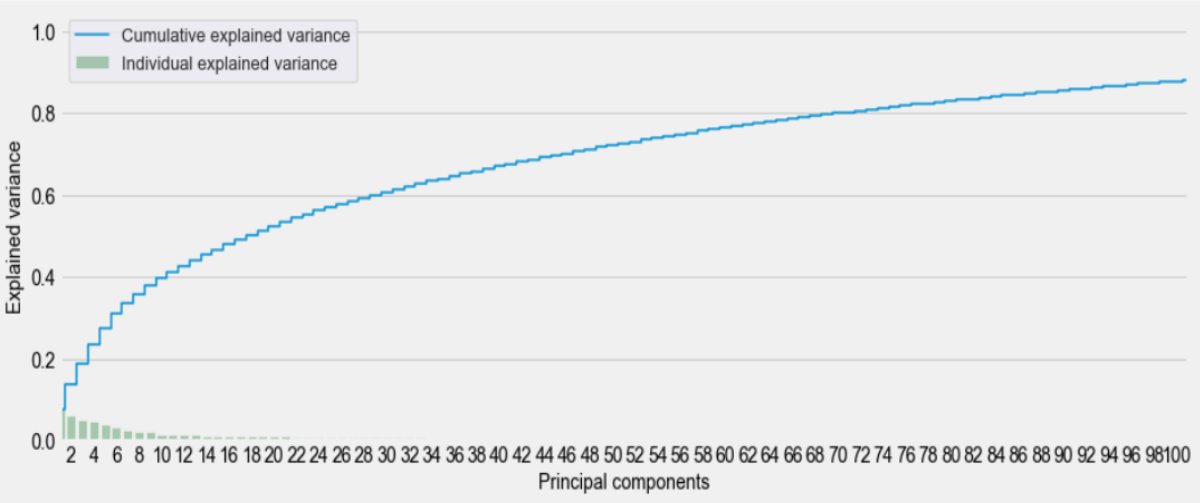

Fig.5. PCA for Products

A clustering analysis on a dataset using the K-Means algorithm. The goal is to create customer categories (or clusters) based on the features in the dataset. We set the number of clusters to 11. This means we want to group the customers into 11 distinct categories. This analysis is important for understanding the natural groupings within the customer data, which can be valuable for targeted marketing, personalized services, and other business decisions. The normal silhouette score for customer categories is 0.22. This visualization is essential for assessing the consistency of the clusters (customer categories) and identifying potential outliers or misclassified data points. It provides a more detailed understanding of the internal structure of the clusters, complementing the overall silhouette score calculated previously.

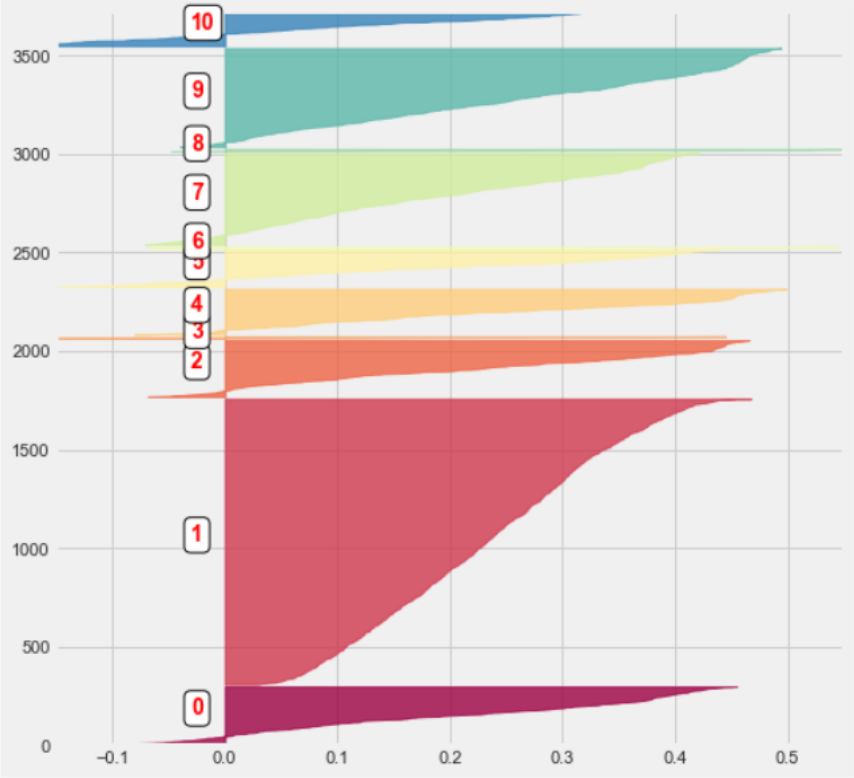

Fig.7. Silhouette score for Customer categories

We are creating a bar plot and a step plot to visualize the explained variance of the principal components obtained from the earlier PCA analysis. This visualization is crucial for understanding how much information each principal component retains from the original data and helps in deciding how many principal components to retain for further analysis. The cumulative explained variance curve typically shows an elbow point where adding more principal components provides diminishing returns in terms of explaining the overall variance.

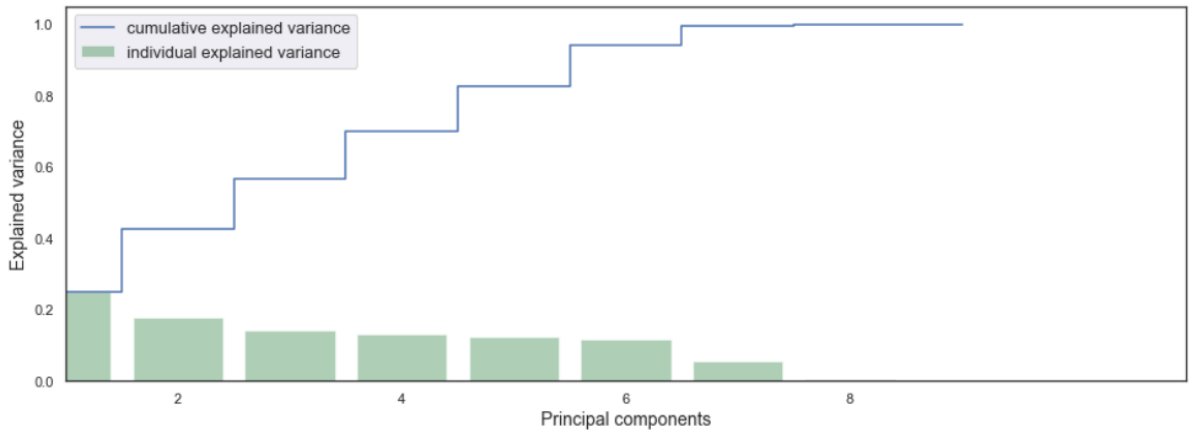

Fig.8. PCA for Customer categories

4 Design Specification

A. Support Vector Machine: The Support Vector Machine, known as SVM, is a powerful machine learning model used in various applications, including retail manufacturing analytics. SVM is particularly effective for classification tasks where the goal is to separate data points into distinct categories. In the context of retail manufacturing analytics, SVM can be used for quality control, demand forecasting, customer segmentation, fraud detection, and consumer shopping cart analysis.

B. Logistic Regression: Logistic regression is a fundamental machine learning technique with important applications in retail manufacturing analysis. Despite its name, it is primarily used for binary sorting tasks, making it valuable for various aspects of retail production. Logistic regression works by modelling the relationship between input characteristics and the probability of a binary outcome. It estimates the coefficients for each feature and combines them to calculate a probability score, which is then converted into a binary prediction based on a predefined threshold.

C. K-Nearest Neighbour: The K-Nearest Neighbors (KNN) model is a simple and intuitive machine learning algorithm used for classification and regression tasks. In KNN, a data point is classified or predicted based on the majority type or the mean of the "k" nearest neighbors in the object space. The "K" in KNN represents the number of neighbors to consider. This is an important parameter that can affect the performance of the model. KNNs are often used in cases where decision boundaries are irregular or when the data is not easily separable by a linear model. It is relatively easy to understand and implement, making it a good starting point for data exploration and for underlying models. However, this may require preprocessing steps (such as scaling) to process objects of different scales, and the choice of "k" must be chosen carefully to balance the offsets and methods. wrong in the model.

D. Decision Tree: A decision tree model is a flexible and easy-to-understand machine learning algorithm, commonly used for classification and regression tasks. It makes decisions based on a hierarchy of rules, like a tree, to divide data into distinct classes or predict continuous values. They are widely used in different fields due to their simplicity, ease of understanding and efficiency in capturing complex relationships in data. When including decision trees in a report, it is essential to discuss techniques to avoid overfitting and the benefits of understanding the model's decision-making process, which can be very important in the context of business and decision making.

E. Random Forest: The Random Forest model is a powerful and flexible synchronous learning algorithm used for both classification and regression tasks. Combining multiple decision trees improves the constraints of a decision tree model, making the model more robust, less redundant, and capable of more accurate predictions. When discussing random forests in the report, it is important to emphasize their ability to handle multidimensional data, their robustness, and their effectiveness in reducing variance and improving the predictive accuracy of individual decision trees. Referencing the feature importance analysis that can be gleaned from the model provides added value in describing the importance of features in the dataset.

F. Gradient boosting: Gradient Boosting is a powerful machine learning technique used for both regression and classification tasks. It sequentially builds a set of decision trees, each of which focuses on correcting the errors of the previous tree, resulting in a robust and highly accurate predictive model. What's important to highlight is Gradient Boosting's ability to handle complex relationships, feature importance analysis, and hyperparameter effects (like learning rate and tree depth). for model performance.

5 Implementation

The implementation of the proposed solution is explained in this area. Python 3 (ipykernal) is used to implement models, with pandas, numpy, seaborn, and matplotlib for data processing, graphing, and visualization.

5.1 Support Vector Machine

The code shown below in fig.9 is using a linear SVM with a Linear SVC classifier, conducting a grid search to optimize the 'C' hyperparameter, and performing K-fold cross-validation which is '5' to assess the model's performance. The goal is to find the best 'C' value that results in the highest accuracy or another suitable performance metric for the given dataset and problem. This approach helps to improve the SVM's ability to classify data points accurately by finding the right balance between maximizing the margin and minimizing the classification error.

```
In [91]: #Support Vector Machine
         svc = Class_Fit(clf = svm.LinearSVC)
         svc.grid_search(parameters = [{'C':np.logspace(-2,2,10)}], Kfold = 5)
```

Fig.9. K-fold Cross Validation, K=5

The dataset is divided in arbitrarily into preparing and test set. Therefore, this result depends on the dividing the data into training set and testing set, we moreover utilized our preparing data, we fit our train data into the show and made a forecast on our test dataset. For this model, we got an accuracy of 76.45 %. The code for this model is shown below:

```
In [97]: svc.grid_predict(X_test, Y_test)

         Precision: 76.45 %
```

Fig.10. SVM Accuracy

learning curves for the best performing SVM model is shown here. The curves show the model's performance (typically accuracy or another relevant metric) on both the training and validation data as the training data size increases. By observing how the curves evolve, we can gain insights into whether the model is underfitting (high bias) or overfitting (high variance), helping to make informed decisions about model complexity and data requirements.

```
In [101]: g = plot_learning_curve(svc.grid.best_estimator_,
                                  "SVC learning curves", X_train, Y_train, ylim = [1.01, 0.6],
                                  cv = 5,  train_sizes = [0.05, 0.1, 0.2, 0.3, 0.4, 0.5,
                                                          0.6, 0.7, 0.8, 0.9, 1])
```

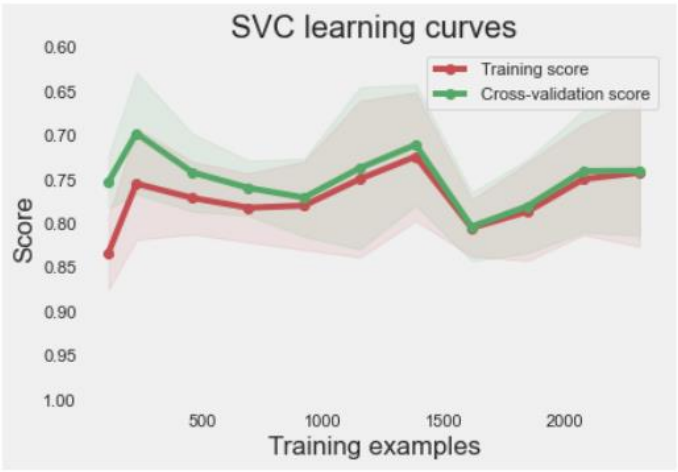

Fig.11. SVM Learning Curves

5.2 Logistic Regression

A Logistic Regression model is being used for the classification task. Logistic Regression is a common algorithm used for binary and multi-class classification. The precision of Logistic regression model is valuated using learning curves. The precision of this model is 89.61% which is shown in fig.12.

```
In [102]: #Logistic Regression
          lr = Class_Fit(clf = linear_model.LogisticRegression)
          lr.grid_search(parameters = [{'C':np.logspace(-2,2,20)}], Kfold = 5)
          lr.grid_fit(X = X_train, Y = Y_train)
          lr.grid_predict(X_test, Y_test)

          Precision: 89.61 %

In [103]: g = plot_learning_curve(lr.grid.best_estimator_, "Logistic Regression learning curves", X_train, Y_train,
                                  ylim = [1.01, 0.7], cv = 5,
                                  train_sizes = [0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1])
```

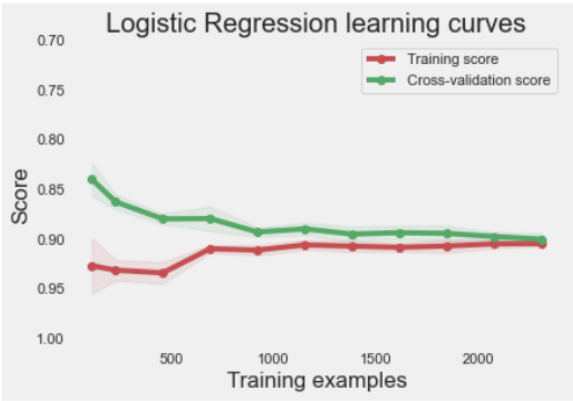

Fig.12. Logistic Regression Precision

5.3 K-Nearest Neighbors

Here, we will check the precision of KNN Model evaluating the training data. So, the precision of this model achieved is 81.72% which is good. The classifier of K-Nearest Neighbors is generated with the number of neighbors as 5.

```
In [104]: #k-Nearest Neighbors

knn = Class_Fit(clf = neighbors.KNeighborsClassifier)
knn.grid_search(parameters = [{'n_neighbors': np.arange(1,50,1)}], Kfold = 5)
knn.grid_fit(X = X_train, Y = Y_train)
knn.grid_predict(X_test, Y_test)

Precision: 81.72 %
```

```
In [105]: g = plot_learning_curve(knn.grid.best_estimator_, "Nearest Neighbors learning curves", X_train, Y_train,
                                 ylim = [1.01, 0.7], cv = 5,
                                 train_sizes = [0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1])
```

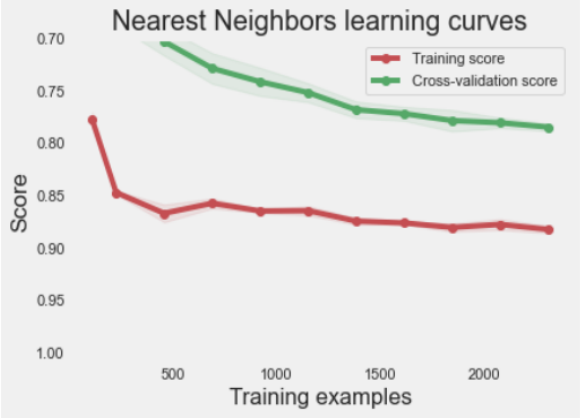

Fig.13. KNN Accuracy and Learning Curves

5.4 Decision Tree

This code focuses on training, tuning, and evaluating a Decision Tree classification model. Decision Trees make decisions based on a hierarchical structure of rules to partition the data into distinct classes. The precision for this model is 83.66%. This model provides a set of proportions of the total training data to use, ranging from a small portion (5%) to the full dataset (100%). The visualisation is shown below.

```
In [106]: #Decision Tree
tr = Class_Fit(clf = tree.DecisionTreeClassifier)
tr.grid_search(parameters = [{'criterion' : ['entropy', 'gini'], 'max_features' :['sqrt', 'log2']}], Kfold = 5)
tr.grid_fit(X = X_train, Y = Y_train)
tr.grid_predict(X_test, Y_test)

Precision: 83.66 %

In [107]: g = plot_learning_curve(tr.grid.best_estimator_, "Decision tree learning curves", X_train, Y_train,
                                  ylim = [1.01, 0.7], cv = 5,
                                  train_sizes = [0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1])
```

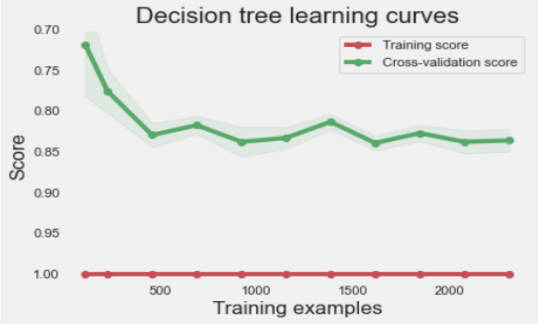

Fig.14. Decision Tree accuracy and Learning curves.

5.5 Random Forest Model

Random Forest classification model, which is an ensemble of decision trees that improves predictive performance and reduces overfitting. Rf model was built using 5 as number of estimators, criterion as entropy and random state set to 0. To check accuracy of the random forest model, learning curves is generated and acquire an accuracy of 90.30%.

```
In [108]: #Random Forest
rf = Class_Fit(clf = ensemble.RandomForestClassifier)
param_grid = {'criterion' : ['entropy', 'gini'], 'n_estimators' : [20, 40, 60, 80, 100],
              'max_features' :['sqrt', 'log2']}
rf.grid_search(parameters = param_grid, Kfold = 5)
rf.grid_fit(X = X_train, Y = Y_train)
rf.grid_predict(X_test, Y_test)

Precision: 90.30 %

In [109]: g = plot_learning_curve(rf.grid.best_estimator_, "Random Forest learning curves", X_train, Y_train,
                                  ylim = [1.01, 0.7], cv = 5,
                                  train_sizes = [0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1])
```

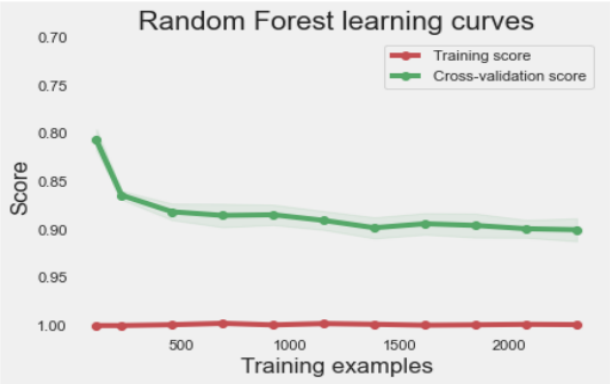

Fig.15. Random forest accuracy and Learning curves.

5.6 Gradient Boosting

This code snippet demonstrates how to use a Gradient Boosting Classifier with a grid search for hyperparameter tuning, fit it to training data, and make predictions on test data to evaluate its performance. The learning curve shows how the model's performance changes as the amount of training data increases. The plot helps to understand if the model is overfitting (high training performance, low testing performance) or underfitting (both training and testing performance are low). The accuracy of Gradient Boosting model is 89.75%.

```
In [112]: #Gradient Boosting Classifier
          gb = Class_Fit(clf = ensemble.GradientBoostingClassifier)
          param_grid = {'n_estimators' : [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]}
          gb.grid_search(parameters = param_grid, Kfold = 5)
          gb.grid_fit(X = X_train, Y = Y_train)
          gb.grid_predict(X_test, Y_test)

          Precision: 89.75 %

In [113]: g = plot_learning_curve(gb.grid.best_estimator_, "Gradient Boosting learning curves", X_train, Y_train,
                                  ylim = [1.01, 0.7], cv = 5,
                                  train_sizes = [0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1])
```

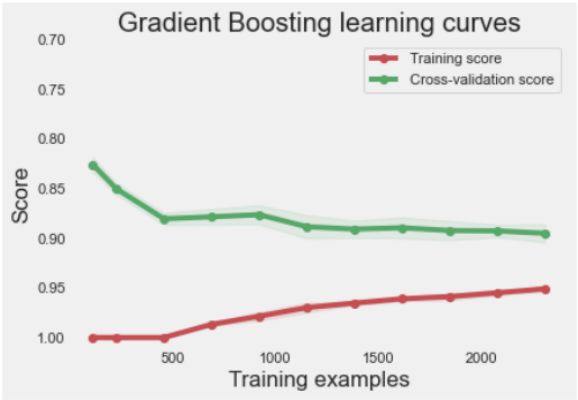

Fig.16. Gradient Boosting Accuracy and Learning Curves

Now, we have six models (one for each algorithm) with the best hyperparameters that were identified during the grid search. These models are ready to be fitted and evaluated on your dataset. The code below will create a Voting Classifier that combines the predictions of the best Random Forest, Gradient Boosting, and k-Nearest Neighbors models using a soft voting strategy. The resulting ensemble classifier can potentially improve the overall predictive performance by leveraging the strengths of these individual models.

```
In [114]: rf_best  = ensemble.RandomForestClassifier(**rf.grid.best_params_)
          gb_best  = ensemble.GradientBoostingClassifier(**gb.grid.best_params_)
          svc_best = svm.LinearSVC(**svc.grid.best_params_)
          tr_best  = tree.DecisionTreeClassifier(**tr.grid.best_params_)
          knn_best = neighbors.KNeighborsClassifier(**knn.grid.best_params_)
          lr_best  = linear_model.LogisticRegression(**lr.grid.best_params_)

In [115]: votingC = ensemble.VotingClassifier(estimators=[('rf', rf_best),('gb', gb_best),
                                                           ('knn', knn_best)], voting='soft')

In [116]: votingC = votingC.fit(X_train, Y_train)

In [117]: predictions = votingC.predict(X_test)
          print("Precision: {:.2f} % ".format(100*metrics.accuracy_score(Y_test, predictions)))

          Precision: 90.03 %
```

Fig.17. Voting Classifier

. Our evaluation indicates a high accuracy of 90.03%. This outcome suggests that the ensemble approach, leveraging the strengths of these individual models, demonstrates strong predictive performance on the test data.

6 Evaluation

In this step, we're planning the testing forecasts by making a duplicate of the 'set_test' DataFrame. This duplicate, named 'basket_price,' is an indistinguishable copy of the first dataset, permitting us to work with the testing information whereas keeping the first intaglio. This ensures that our testing and analysis don't influence the initial dataset, permitting us to preserve data astuteness all through the assessment prepare. In this section, I will study the algorithm and state that which model is best for forecasting of customer's need in the upcoming year.

Name of the Model	Accuracy
Support Vector Machine	62.68 %
Logistic Regression	75.11%
K-Nearest Neighbors	67.19%
Decision Tree	71.23%
Random Forest	74.87%
Gradient Boosting	74.48%

Table.1. Accuracy of all the models.

We can see in the above table 1, metrics ranking is shown for all the models with the precision. The Support vector machine (SVM) model has an Accuracy of 62.68%, K-nearest neighbor has an accuracy of 67.19%, Decision tree has 71.23% of accuracy, Random Forest has 74.87% accuracy, along with this gradient boosting has an accuracy of 74.48%, while the Logistic regression has an accuracy of 75.11% respectively which is genuinely better for forecasting the customer's needs and demands. Here is the visualisation of the testing models.

```
n [124]: classifiers = [(svc, 'Support Vector Machine'),
                        (lr, 'Logistic Regression'),
                        (knn, 'k-Nearest Neighbors'),
                        (tr, 'Decision Tree'),
                        (rf, 'Random Forest'),
                        (gb, 'Gradient Boosting')]

         for clf, label in classifiers:
             print(25*'_', '\n{}'.format(label))
             result = clf.grid_predict_precision(X, Y)
             df_result.loc[len(df_result.index)] = [label, result]

_________________________
Support Vector Machine
Precision: 62.68 %

_________________________
Logistic Regression
Precision: 75.11 %

_________________________
k-Nearest Neighbors
Precision: 67.19 %

_________________________
Decision Tree
Precision: 71.23 %

_________________________
Random Forest
Precision: 74.87 %

_________________________
Gradient Boosting
Precision: 74.48 %
```

Fig.18. Classification of testing predictions models.

In this evaluation for the testing results, I got a voting prediction of 75.62% accuracy. I created a bar plot using Matplotlib to visualize the precision of different algorithms which can be seen clearly in the below figure.

- The advantages of SOTA are enhances performance precision. Precision is one evaluation matrix; it produces the best result in terms of accuracy.
- Precision evaluates how precise a model is predicting positive labels. Precision is a good evaluation metric to use when the cost of a false positive is very high, and the cost of a false negative is low.

```
In [126]: predictions = votingC.predict(X)
          print("Precision: {:.2f} % ".format(100*metrics.accuracy_score(Y, predictions)))

          Precision: 75.62 %
```

```
In [127]: import matplotlib.pyplot as plt

          def addlabels(x,y):
              for i in range(len(x)):
                  plt.text(i,y[i],y[i])

          plt.figure(figsize=(20, 7))
          plt.bar(df_result.Algorithms, df_result.Precision, color='green', width=0.4, align='center')
          addlabels(df_result.Algorithms, df_result.Precision)
          plt.title('Precision for different alogirthms')
          plt.xlabel('Algorithms')
          plt.ylabel('Precision')
          plt.show()
```

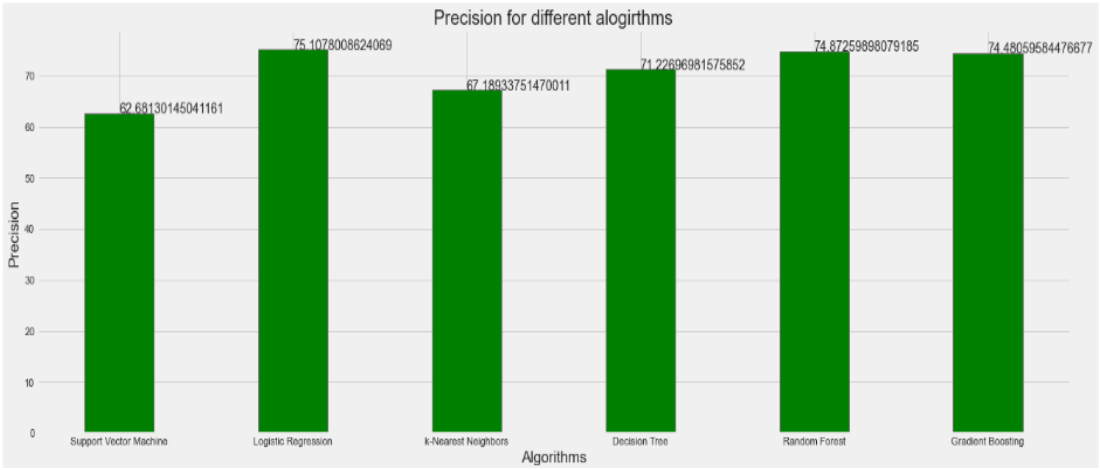

Fig.19. Testing Results with voting classifier and bar plot.

The bar graph clearly shows that logistic regression is the best algorithm for performing my project.

6.1 Discussion

After doing a lot of analysis, I performed a preprocessing step to get a better accuracy score. Looking at the customer category silhouette score chart, I saw that most customers buy his second category out of 11. Therefore, new customers joining the scenario may belong to the same category. At the same time, looking at the Product Silhouette Score graph, when a new customer enters the image, he may be in the 9th category and purchase a product in the 2nd category. It changes according to the taste of the customer. I proved this analysis by comparing all training and testing models. The comparison shows that logistic regression has the highest accuracy score for both training and testing models. Therefore, to prove my analysis, I found logistic regression to be the perfect algorithm. Running this algorithm predicts the percentage of classifications that have occurred and forecast which category the customer belongs to.

Applied Data Transformation-
I have applied Data Aggregation, Data Cleansing, Data Deduplication, Data Filtering and Data Splitting.

7 Conclusion and Future Work

The key elements of a marketing strategy are segmentation, mark, and positioning of customers and products. The ability of marketers to recognize and select the most effective target segments will determine the outcome of marketing initiatives. Customers in this research were divided into 11 groups based on factors like how much they spend on purchases, so we created fields for different types of customers, how often they shop, and so on. considered as the most optimal model, used for this study. The outcome of the project depends on the data considered and the algorithms used, such as support vector machines, logistic regression, k-nearest neighbors, decision trees, random and augmented forests. slope. In general, I divide the cluster into two groups, i.e., the customer category (cluster=11) and the product category (cluster=5) are formed. This work helps to increase customer awareness and allows companies to expand their marketing plans to increase sales. Even maintaining relationships with new and old customers helps generate more profits for the manufacturing company and keeps their forecasted products in demand.

As a part of my future work, If the customer purchases the same products more than 5 times and the invoice will be generated with same customer Id again and again, I'll provide him a 15-20% discount from my side, so that the customer will be more attracted towards my company and will help in maintaining a good bond with them. It will help me in boosting my company's name and profit margin. My main advantage is that no other companies or factories give additional discounts, and I will be at least a better competitor.

References

Cirqueira, D. H. (2019). Customer Purchase Behavior Prediction in E-commerce: A Conceptual Framework and Research Agenda. *NFMCP@PKDD/ECML*.

H. Valecha, A. V. (2018). Prediction of Consumer Behaviour using Random Forest Algorithm. *5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 1-6.

Jayasena, E. Y. (2020). The practical approach in Customers segmentation by using the K-Means Algorithm. *IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, 344-349.

Kansal, T. &. (2018). Customer Segmentation using K-means Clustering. *International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 135-139.

Kavitha, S. R. (2021). Consumer Online Buying Behaviour - A Perspective Study Analysis using R. *International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 1-6.

Kumar, M. R. (2020). Machine Learning Based Customer Churn Prediction In Banking. *4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 1196-1201.

M. Aryuni, E. D. (2018). Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering. *International Conference on Information Management and Technology (ICIMTech)*, 412-416.

M. Guan, M. C. (2022). From Anticipation to Action: Data Reveal Mobile Shopping Patterns During a Yearly Mega Sale Event in China. *IEEE Transactions on Knowledge and Data Engineering*, 1775-1785.

N. R. Maulina, I. S. (2019). Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering. *16th International Conference on Service Systems and Service Management (ICSSSM)*, 1-6.

Ozan, Ş. (2018). "A Case Study on Customer Segmentation by using Machine Learning Methods. *International Conference on Artificial Intelligence and Data Processing (IDAP)*, 1-6.

S. De, P. P. (2021). "Effective ML Techniques to Predict Customer Churn. *Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 895-902.

S. Peker, A. K. (2018). An empirical comparison of customer behavior modeling approaches for shopping list prediction. *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1220-1225.

Singh, N. a. (2023). Customer Behavior Prediction using Deep Learning Techniques for Online Purchasing. *2nd International Conference for Innovation in Technology (INOCON)*, 1-7.

Tawfiq, F. &. (2021). An E-Commerce Recommendation System Based on Dynamic Analysis of Customer Behavior. *Advanced Application of Sustainable Transportation: Intelligent and Autonomous Traffic Monitoring, Control and Management Systems for Smart Cities*, 13-19.

V. L. Narayana, S. S. (2022). Mall Customer Segmentation Using Machine Learning. *International Conference on Electronics and Renewable Systems (ICEARS)*, 1280-1288.

V. Mehta, R. M. (2021). A Survey on Customer Segmentation using Machine Learning Algorithms to Find Prospective Clients. *9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 1-4.

Wang, X. D. (2013). Understanding customer-oriented organizational citizenship behavior in information system support: An exploratory study. *46th Hawaii International Conference on System Sciences*, 4115-4124.

X. Chen, Y. F. (2018). "PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data. *IEEE Transactions on Knowledge and Data Engineering*, 559-572.