Predictive analysis of customer spending patterns using Machine Learning Algorithms

Dinesh Banswal[1], Punam Chaudhari[2], Subhranil Das[3], Mahesh Kumar Ladge [4]

[1,3,4.]School of Business, Vishwanath Karad MIT World Peace University, Pune, India

[2]Sinhgad Institute of Management, Savitribai Phule Pune University, Pune, India

ABSTRACT:

In recent times, e-commerce sector has been extremely competitive for any information relating to consumer behaviour. Data mining can be utilised to locate those obvious data that could be used to lessen competition while also boosting business profit. This study aims to analyse the spending patterns of customers by applying the various predictive analysis models on the data set generated from an ecommerce website. This attempt provides researcher to understand the significance between the demographic factors and their influence in the buying decisions made by the customers. In this empirical study Predictive models has been developed using features of customers demographic. Supervised learning methods like Decision tree classifier, Random forest classifier, Linear Regression & Naïve Bayes Classifier will be used for predictions. The results of this analysis will provide insight into how customer spending habits impact revenue for the ecommerce website and can inform strategies for increasing revenue. Thus, the results of this predictive study helps the ecommerce market players to design the strategies that will give more consumer driven buying experience additionally, the study will also identify key segments of customers that are more tend to buy from the ecommerce website.

KEYWORDS*: Data Mining, Predictive model, Supervised learning methods, Decision tree classifier, Random forest classifier, Linear Regression, Naïve Bayes Classifier.*

1. Introduction

In the contemporary digital era, the landscape of e-commerce is experiencing exponential growth, rendering the online platforms of businesses increasingly competitive. This surge in e-commerce activity generates voluminous data daily, encapsulating valuable consumer insights pivotal for market analysis and forecasting [1,2]. By leveraging data mining techniques and advanced information processing methods, it becomes feasible to extract actionable consumer behavior patterns from the web. This strategic utilization of data not only mitigates competitive pressures but also propels business growth by enhancing online shopping experiences. Predictive analytics emerges as a cornerstone in this context, offering a predictive score or probability for each entity involved—be it customers, employees, healthcare patients, product SKUs, or other units [3,4]. This predictive capability fundamentally transforms decision-making processes across various domains such as marketing, credit risk analysis, fraud detection, manufacturing, and healthcare by providing a probabilistic assessment of outcomes based on historical and transactional data. Such predictive models harness the power of data to

uncover patterns, capturing intricate relationships among variables to forecast risks and opportunities[5,6]. This analytical approach facilitates informed decision-making, influencing the strategic directions of e-commerce entities aiming to augment their market presence and revenue generation [7,8]. This paradigm shift towards data-driven strategies underscores the significance of predictive analytics in sculpting the future of business operations, emphasizing its indispensable role in navigating the complexities of the modern commercial landscape.

The present paper is organized in the following way. Section II highlights some of the most recent literature in this field of research. Section III presents the methodology, dataset. Section IV discussed about data mining process the models discussed in the previous section. Section V talks about the predictive modelling, section VI brief about data analysis and interpretation, section VII illustrate the findings, section VIII states the discussion and section IX is conclude the paper with a focus on future avenues of research.

2. Related Works

Predictive analysis is widely used in e-commerce to improve the customer experience and drive sales [9]. It is frequently used by businesses to manage product inventory and establish pricing policies. Predictive analysis of this kind enables businesses to meet customer demand without overstocking their warehouses. Additionally, it enables businesses to evaluate the price and long-term value of their products. If one part of a given product becomes more expensive to import, companies can project the long-term impact on revenue if they do or do not pass on additional costs to their customer base. Predictive analysis is used to segment customers based on their past behaviours and demographics, which can help e-commerce companies to understand their target audience and personalize their marketing efforts. By studying and analysing large volumes of collected customer data, businesses can improve their marketing decisions based on the customers' preferences [10]. According to [11], maximum profits can be generated for any business entity if the resources are utilized judiciously in order to cultivate the most loyal and useful group of customers once customer segmentation and clustering have enabled the allocation of customers to such groups. The total customer set can be divided and grouped into clusters based on their buying behaviour, frequency, demographics etc.

Extraction of useful information and patterns from massive data sets is known as data mining[12]. Predictive analysis, which employs statistical models to make predictions about future trends and behaviours, is then performed using this information as the starting point. The accuracy, precision, recall, and F1 scores used to evaluate these predictive models are typically taken into account. A specific data mining algorithm must be used to train the data for this purpose, and it must then be tested on a new dataset to determine how well our model fits the new dataset by evaluating accuracy and F1 scores. Data mining generally involve two categories with respect to the data to be mined that includes Description [13,14]. Description mining usually is mining of association rules, frequent patterns, clusters, or correlations [15,16]. Classification and Prediction involves classifying a class label for data using probability equations and predicting any feature using numeric measures accordingly [17,18].

The development of the modules was done in phases that were carefully planned, designed, coded, tested, and integrated using the Software Development Lifecycle [19,20]. Gathering requirements, researching technology, looking over data, and choosing how to work were all part of the planning process. Coding involved the application of predictive models through data training and testing. The paper begins with a topic introduction, then moves on to a literature

review, related work in predictive analysis, and applications to crime data. It also incorporates various tools and techniques for achieving results. In modules, a thorough explanation of the libraries used for this prediction and classification as well as the performance metrics selected, for obtaining the most accurate results, have been covered. Data pre-processing describes the data in its current format as well as the methods used to process it [21]. Many values in the given dataset were missing, null, or incorrect. As a result, the necessary data to be cleaned, transformed, and integrated in order to reduce noise. Three data mining techniques for creating prediction models were used to increase prediction accuracy. The results of these techniques can then be compared to determine which model fits this type of data the best and produces the most accurate predictions. In this paper, each classifier's analysis and implementation outcomes will be explained in detail. In conclusion, a comparison of these algorithms' scores with the most important factors that influence consumer spending.

Decision trees [22] start with a root node, which acts as a starting point (at the top), and is followed by splits that produce branches. The statistical/mathematical term for these branches is edges. The branches then link to leaves, known also as nodes, which form decision points. A final categorization is produced when a leaf does not generate any new branches and results in what is known as a terminal node. Rather than striving for the most efficient split at each round of recursive partitioning, an alternative technique is to construct multiple trees and combine their predictions to select an optimal path of classification or prediction. This involves a randomized selection of binary questions to grow multiple different decision trees, known as random forest [23]. As a mathematical classification approach, the Naive Bayes classifier involves a series of probabilistic computations for the purpose of finding the best-fitted classification for a given piece of data within a problem domain. In this paper, an implementation of Naive Bayes classifier is described [24]. Linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modelling technique where a dependent variable is predicted based on one or more independent variables [25].

3. Data Collection and Preparation

The Analysing Customer Spending Habits (ACSH) dataset [26] was obtained from the Kaggle has been considered for analysis. In this dataset, 34865 total instances are created with 15 different attributes such as ages of customers, age, gender, and country. Here, the variables are considered as multivariate where the information has been stored in the xlsx file and cleaned later. For renaming convention, the original data has been named as dirty data. Further, the data was cleaned where there were no missing values. The objective of considering this dataset is to analyze the different changes in consumer trends and motivates consumers to make purchases both online and offline.

3.1 Data Preprocessing

A data mining technique called data pre-processing [27] entails putting raw data into an understandable format. Data that lacks certain behaviours or trends, is frequently unstructured, inconsistent, has missing values, and produces numerous errors. It must therefore be cleaned, integrated, transformed, and subsequently reduced. Cleaning eliminates noise and fills in the missing values. Data blocks or chunks are combined through integration using multiple databases. Data are normalised and aggregated during transformation, and reduction aids in reducing the amount of data while maintaining similar analytical results. To obtain suitable

consumer spending behaviour data, clean data needed missing values removed. Since the dataset didn't contain any blank values, maintaining the highest level of accuracy was simple. Since NaN values wouldn't have produced accurate results, they were all eliminated. While the feature named Revenue is set to be the Target, features like Year, Month, Customer Age, Customer Gender, Country, State, Product Category, Sub Category, Quantity, and Cost were entered into the variable Inputs. Index, Date, Unit Cost, and Unit Price were removed because they had no bearing on the target. The Label Encoder function was imported from the sklearn pre-processing library and used to encode the features assigned to the variables Inputs and Target. This was done because strings cannot be used in data mining algorithms.

4. Proposed Methodology

Empirical exploration is a sort of examination procedure that utilizes unquestionable proof to show up at research results. All in all, this sort of exploration depends exclusively on proof got through perception or logical information assortment techniques. Empirical exploration can be completed utilizing subjective or quantitative perception techniques, contingent upon the information test, that is to say, quantifiable information or non-mathematical information. Not at all like hypothetical exploration that relies upon assumptions about the examination factors, has exact examination conveyed a logical examination to gauge the trial likelihood of the examination factors.

Finding potential predictors and learning more about how they relate to the target variable are the objectives of empirical research for predictive analysis. With the aid of this data, a more reliable predictive model can be created and future events can be predicted with greater accuracy. The same has been applied by the researcher in the paper. Finding potential predictors and learning more about how they relate to the target variable are the objectives of empirical research for predictive analysis. With the aid of this data, a more reliable predictive model can be created and future events can be predicted with greater accuracy.

4.1 Pattern Discovery

The goal of this step is to uncover any hidden relationships in the data. Classification analysis, association rule discovery, sequential pattern discovery, and clustering analysis are the four main techniques used in pattern discovery. The client profiles are categorised based on the pre-processed data using classification analysis.

After classification, personalised information services and activities will begin. The most popular pages in user sessions are found through association rule discovery in data mining. Utilising data mining techniques based on the group of data on the pages that visitors visit, the design of the website can be improved [28]. The process of sequential pattern discovery is crucial to data mining. The dataset from various sources that contain records of the customers' visit patterns can be utilised to forecast the behaviour of the customers. To promote engagement between the firm and the client and to keep current customers, the business might offer personalised service. The most common technique used in data mining applications is clustering analysis. To aid marketers make their marketing judgements, this technique organises the data that include comparable qualities and attributes. Partition clustering and hierarchical clustering are the two clustering techniques used.

The model pattern discovery algorithm creates a set of patterns for pattern analysis. This algorithm chooses an intriguing pattern in order to discover further useful models. It won't matter about other patterns.

4.2 Decision Tree (DT) model

Decision trees are powerful and intuitive tools in the field of machine learning, facilitating both classification and regression tasks. At the core of a decision tree lies a hierarchical structure comprising nodes, branches, and leaf nodes. Each internal node represents a decision based on a specific feature, while leaf nodes correspond to the final predictions or classifications. We delve into the binary tree structure of decision trees and explain how they recursively partition the feature space to facilitate decision-making. The construction of decision trees involves iteratively selecting features and determining optimal split points to maximize information gain or minimize impurity. Central to the construction of decision trees is the choice of splitting criteria, which determines how feature space is divided at each node. We examine common impurity measures such as Gini impurity, entropy, and classification error, elucidating their roles in guiding the splitting process. By understanding these measures, practitioners can make informed decisions during tree construction. To prevent overfitting and enhance generalization performance, decision trees often undergo pruning and regularization. We discuss techniques such as cost-complexity pruning and minimum sample split, which aim to simplify the tree structure while preserving predictive accuracy. Additionally, we explore the trade-offs between tree complexity and performance in the context of model regularization.

A decision tree can be mathematically represented as a hierarchical binary tree structure *T* comprising nodes and branches, where each node represents a decision based on a feature and each leaf node represents a class label or a numerical value. Let's define the mathematical notation used to represent a decision tree: *T* represents the decision tree, *N* represents the total number of nodes in the tree, *D* represents the training dataset, X_i represents i^{th} feature and Y_i represents the output value for the i^{th} sample. Each node *n* in the decision tree *T* is represented by a mathematical expression as given below.

$$N(n) = (X_j, s, N_{left}, N_{right}) \quad (1)$$

Where X_j is the feature associated with the decision at the node, s is the splitting criterion or threshold for the feature X_j. N_{left} and N_{right} child nodes correspond to the left and right branches, respectively. Each leaf node *I* in the decision tree *T* is represented by a mathematical representation which is given by the following equation.

$$L(l) = y \quad (2)$$

where y is the predicted class label or numerical value associated with the leaf node. The decision tree *T* can be represented as a set of nodes and leaf nodes which is given by Eqn.3.

$$T = N_1, N_2, ..., N_N, L_1, L_2, ..., L_L \quad (3)$$

where *N* represents internal nodes and *L* represents leaf nodes.

The prediction function of the decision tree involves traversing the tree from the root node to a leaf node based on the feature values of the input sample *x*. Once a leaf node is reached, the predicted class label or numerical value associated with that leaf node is returned. Mathematically, the prediction function *f(x)* of the decision tree can be represented as

$$f(x) = \begin{cases} L(l) & \text{if x reaches the leaf node } l \\ f(N_{left}) & x_{Xj} \leq s \\ f(N_{right}) & x_{Xj} > s \end{cases} \quad (4)$$

where x_{Xj} represents the value of feature X_j in the input sample.

4.3 Naïve Bayes (NB) model

Naive Bayes classifiers are a family of probabilistic classifiers based on Bayes' theorem with a strong (naive) assumption of independence between the features. They are among the simplest Bayesian network models but coupled with kernel density estimation, they can achieve higher accuracy levels. Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event which is represented mathematically as follows:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad (5)$$

where P(A|B) is the probability of hypothesis given the data *B*, known as the posterior probability. P(B|A) is the probability of data *B* given that the hypothesis *A* holds. P(A) is the probability of hypothesis *A* being true (regardless of the data), known as the prior probability. P(B) is the probability of the data (regardless of the hypothesis).

In the context of classification, we are interested in finding the most probable class given the feature set. If we consider a classification problem where we want to classify an instance into a class *C*, Bayes' theorem which is given by the mathematical expression as

$$P(C|x) = \frac{P(x|C).P(C)}{P(x)} \quad (6)$$

However, P(x) is constant for all classes, only the numerator needs to be considered to determine the most probable class. This is computationally advantageous as it avoids the need for the calculation of *P(x)* , which can be complex. The naive assumption of feature independence simplifies \(P(x|C) \) to the product of the individual probabilities which is given by the mathematical equation are as follows:

$$P(x|C) = P(x_1, x_2, ..., x_n|\, C) = P(x_1|C)\, P(x_2|C)\ldots..P(x_n|C) \quad (7)$$

Where x_i are the feature values of instance *x* and n is the number of features.

4.3 Linear Regression (LR) model

Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables. The simplest form of linear regression is called simple linear regression, which involves a single independent variable. When there are multiple independent variables, it's called multiple linear regression. In simple linear regression, we predict the value of one variable *Y* based on the value of another variable \(X\). It is assumed that the relationship between *X* and *Y* can be described by a straight line. The mathematical model for this straight line is given by the mathematical expression are:

$$Y = \beta_0 + \beta_1 X + \xi \quad (8)$$

Where Y is the dependent variable for prediction, X is the independent variable we are using to make predictions. β_0 is the y-intercept of the regression line. β_1 is the slope of the regression line, representing the change in Y for a one-unit change in X, ξ is the error term, accounting for the variability in Y that cannot be explained by X.

The goal of linear regression is to find the best-fitting line through the data, which is done by minimizing the sum of the squares of the vertical distances of the points from the line. This method is known as the least squares method.

The coefficients β_0, β_1, ..., β_n are estimated using the least squares criterion, which finds the line (in 2D) or hyperplane (in higher dimensions) that minimizes the sum of squared residuals, where a residual is the difference between an observed value and the value predicted by the model. The assumptions made by the Linear regression is based on several key assumptions:

1. Linearity: The relationship between the independent variables and the dependent variable is linear.

2. Independence: The residuals (errors) are independent.

3. Homoscedasticity: The residuals have constant variance at every level of the independent variables.

4. Normality: The residuals are normally distributed.

4.4 Random Forest (RF) Model

Random Forest stands out as for its robustness and accuracy across diverse domains of application. Introduced by [29], the Random Forest algorithm amalgamates the simplicity of decision trees with the power of ensemble techniques to forge a model that is both interpretable and capable of handling complex data structures. At its core, Random Forest operates by constructing a multitude of decision trees *T* during the training phase, each tree grown from a bootstrap sample of the training data. This process, known as bootstrap aggregating or bagging, allows for the creation of varied trees by sampling with replacement. The algorithm's distinctiveness further manifests in its approach to feature selection for splitting nodes within each tree. Unlike traditional decision tree algorithms that consider all available features when making a split, Random Forest introduces an additional layer of randomness by limiting each split to a random subset of features *m* is approximately $\sqrt{p}$, typically chosen as *m* , where p is the total number of features. This strategy, termed feature bagging, enhances the diversity among the trees in the forest, thereby reducing the correlation between individual trees and mitigating the risk of overfitting. Formally, the prediction for a new sample in classification tasks is determined by a majority voting scheme among the trees which is given by mathematically as

$$Y = y_1, y_2, ..., y_T \quad (9)$$

where Y is the predicted class, and y_i represents the prediction of the i^{th} tree. For regression tasks, the algorithm employs an averaging method which is given as mathematically

$$Y = \frac{1}{T} \sum_{i=1}^{T} y_i \quad (10)$$

where Y is the predicted value. This mathematical underpinning elucidates the algorithm's efficacy in capturing the underlying data patterns without succumbing to the pitfalls of overfitting, a common drawback of single decision trees. Beyond its predictive prowess, Random Forest offers an intrinsic mechanism for evaluating feature importance, a valuable asset for exploratory data analysis. The significance of a feature is gauged based on how much the accuracy of the trees decreases when the feature is excluded, thereby providing insights into the features that contribute most significantly to the prediction outcome.

5. Simulation results and Discussion

In this study, the Python Integrated Development and Learning Environment (IDLE) was employed for carrying out simulation tasks. The computational operations were executed on a workstation equipped with an Intel® Core™ i7-9750H Central Processing Unit (CPU) operating at 2.60GHz and supported by 16 Gigabytes (GB) of Random Access Memory (RAM). For the construction and implementation of the predictive analytics model, the Scikit-Learn library, a widely acknowledged Python module, was utilized. Scikit-Learn, which is developed upon foundational Python libraries such as NumPy, SciPy, and Matplotlib, is recognized for its robustness and applicability in data mining tasks due to its open-source nature and efficient programming capabilities.

The methodology of the research involved partitioning the dataset into two segments: approximately two-thirds (66.67%) of the data was allocated for training purposes, utilizing algorithms provided within the Scikit-Learn library, while the remaining one-third (33.33%) was reserved for the model testing phase. The primary objective of this model is to identify and accurately predict key determinants that affect consumer expenditure behaviour. By pinpointing these significant features, the model aims to furnish e-commerce platforms and other related business entities with insightful data, enabling them to make informed decisions and undertake strategic actions to optimize their operations and marketing strategies. This approach not only underscores the practical application of machine learning techniques in analysing consumer trends but also illustrates the potential of predictive modelling in enhancing business intelligence and decision-making processes.

5.1 Performance Indices

In the domain of predictive modelling, the fidelity of a model in mirroring accurately the phenomena observed in the real world is of paramount importance. This fidelity is assessed using a labelled dataset, which encompasses the precise values that the model aims to predict. Within this context, the construct of the confusion matrix emerges as a critical evaluative tool, as elucidated by [30]. The confusion matrix elucidates the correlation between the predicted classifications and their actual counterparts, thereby offering a quantifiable measure of a model's predictive accuracy.

In the framework established by [31], the terminologies of true positives and true negatives occupy a central role, denoting observations that were correctly identified by the model. Conversely, the concepts of false positives and false negatives represent instances of predictive inaccuracies. Specifically, a false negative denotes an observation where the actual occurrence was positive, yet the model erroneously predicted it as negative. The nomenclature employed here—'true' and 'false'—serves to distinguish between the accuracy of the model's predictions, with 'true' indicating a match between predicted and actual values, and 'false' signifying a discrepancy. Thus, when a designation commences with 'false,' the veracity of its categorization lies in opposition to the term that subsequently follows. This delineation not only aids in understanding the model's performance but also in refining its predictive capabilities by highlighting areas of misclassification. The parameters of the Confusion Matrix are explained by the mathematical equations from Eqn. 11 to Eqn.14.

$$\text{Classification Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

In the context of our analysis focusing on the determinants of revenue generation, the following top ten features were identified based on their significance in relation to feature importance metrics:

- Year (Integer): The specification of the year when the sale occurred serves as a temporal marker, enabling analysts to observe and understand trends over time, including cyclical patterns or growth trajectories in revenue generation.
- Month (Integer): By categorizing sales data by month, it becomes possible to discern seasonal variations or specific periods of heightened activity, which are crucial for planning marketing strategies and inventory management.
- Customer Age (Integer): The age of the customer can significantly impact purchasing habits, preferences, and ultimately, the revenue generated. Understanding age demographics helps tailor product offerings and marketing messages to the most responsive segments.
- Customer Gender (String): Gender can influence product preferences and purchasing decisions. Recognizing gender-based trends allows for more targeted marketing efforts and product development to meet the distinct needs or preferences of different gender groups.
- Country (String): The geographical location of customers at the country level provides insights into market size, economic power, and cultural factors that influence purchasing behavior, enabling the customization of marketing strategies to various international markets.
- State (String): Further refining the geographical analysis to the state level can uncover regional trends, preferences, and economic conditions, allowing businesses to localize their strategies effectively.
- Product Category (String): Identifying which categories of products contribute most significantly to revenue can guide inventory decisions, product development, and promotional efforts towards those areas with the highest return on investment.
- Sub Category (String): Analysis at the sub-category level allows for an even more granular understanding of consumer preferences, highlighting specific products or lines that are particularly successful or identifying areas for expansion or improvement.
- Quantity (Integer): The volume of products sold is directly correlated with revenue. Understanding the relationship between quantity sold and other factors can inform pricing strategies, promotions, and stocking decisions.
- Unit Cost (Float): The cost per unit of product impacts pricing strategies and profit margins. Analyzing how unit costs relate to sales volumes and revenue can help in optimizing pricing to enhance profitability.
- Revenue (Float): As the primary outcome variable, total revenue from sales encapsulates the financial success of transactions. Analyzing revenue in conjunction with other variables provides a comprehensive view of the factors driving financial performance.

These attributes were extracted to facilitate a nuanced understanding of the factors influencing revenue outcomes, thereby enabling a targeted approach towards enhancing revenue streams. Table 1 explains the classification results for Decision Tree classifier model.

Table 1 : Classification Results for Decision Tree Classifier

Performance Metrics Measure	Values
Classification Accuracy	94.924
Precision	97.443
Sensitivity	95.322
Specificity	93.245
Recall	97.2150
F1Score	97.3289

Employing entropy as the criterion for splitting, the model exhibits an accuracy of 94.924%, indicating that this percentage of the data was correctly classified out of the total classifications conducted by the model. The precision rate stands at 97.443%, reflecting the percentage of positive identifications that were accurately correct. The recall rate is observed at 97.3289%, representing the percentage of true positives accurately identified by the model. The F1 score,

Performance Metrics Measure	Results
Classification Accuracy	98.6702
Sensitivity	98.7698
Specificity	96.6547
Precision	97.3635
Recall	97.1866
F1 Score	97.2750

which is the harmonic mean of precision and recall, takes into consideration both false positives and false negatives, providing a balanced measure of the model's predictive accuracy.

Table 2 : Classification Results for Random Forest (RF) Classifier

The RF model demonstrates an accuracy of 98.6702%, indicating that this proportion of data was accurately classified out of all the classifications performed as explained in Table 2. The precision, measured at 97.36350%, denotes the fraction of positive identifications that were indeed correct. The recall rate, at 97.1866%, signifies the fraction of true positives that were accurately detected by the model. The F1 score, calculated as the harmonic mean of precision and recall, incorporates both false positives and false negatives to provide a balanced assessment of the model's predictive performance.

Given the closely aligned accuracy, precision, and recall metrics for the Random Forest Classifier, this model is deemed exceptionally suitable for predicting median values, both in scenarios with complete data and those involving missing values. Moreover, critical features such as 'Customer Age', 'State', 'Country', 'Product Category', 'Quantity', 'Month', 'Year', and 'Unit Cost' are consistently identified across the ensemble of decision trees, further

attesting to the model's robustness and the relevance of these variables in the predictive framework.

Performance Metrics Measure	Results
Classification Accuracy	94.8608
Precision	93.7674
Sensitivity	94.7245
Specificity	92.7351
Recall	92.9852
F1 Score	95.9824

Table 3: Classification Results for Naïve Bayes (NB) Classifier

The classification accuracy of NB model has been quantified at 98.6702%, reflects the percentage of data that was classified correctly out of all the model's classifications as shown in Table 3. Precision, at 97.3635%, indicates the accuracy of positive predictions made by the model. The recall rate, recorded at 97.1866%, measures the model's ability to correctly identify all actual positive cases. The F1 score, representing the harmonic mean of precision and recall, takes both false positives and false negatives into account to provide a balanced measure of the model's predictive accuracy.

For a dataset devoid of any impurities, this Gaussian Naive Bayes predictive model demonstrates commendable performance, as evidenced by the high accuracy rate. The pristine dataset contributes to the highest proportion of documents being identified correctly, as demonstrated by the precision metric. Nevertheless, when evaluating the model's performance by considering both recall and precision, it becomes evident that the features deemed relevant through recall analysis align with those yielding the highest precision. This synthesis underscores the model's effectiveness in accurately predicting outcomes based on the selected features.

Table 4: Classification Results for Linear Regression (LR) Classifier

In our analytical assessment of the predictive model's performance, a suite of key metrics was utilized, culminating in a comprehensive understanding of its efficacy. The model demonstrated a commendable classification accuracy of 94.8608%, signifying a substantial proportion of correct predictions across all evaluated instances. Precision was recorded at 93.7674%, indicating a high reliability of the model in generating positive predictions.

Sensitivity, or the ability to accurately identify true positives, was observed at 94.7245%, highlighting the model's effectiveness in minimizing false negatives within the positive class. Specificity stood at 92.7351%, showcasing the model's aptitude in correctly recognizing true negatives, thereby reducing false positives in the negative class. The recall rate, reflective of the model's capacity to correctly identify all actual instances, was marked at 92.9852%. Furthermore, the F1 score, a harmonic mean of precision and recall, reached an exceptional value of 95.9824%, underscoring a balanced performance with minimal trade-off between precision and the identification of relevant instances. These metrics collectively affirm the model's robust predictive capabilities, characterized by high accuracy, precision, sensitivity, specificity, and an outstanding F1 score, thereby validating its reliability and applicability in practical scenarios. This integrated performance analysis forms an essential part of our assessment, elucidating the model's commendable predictive prowess within the domain of predictive modeling.

5.2 Correlation Heatmap

A correlation heatmap is a visual tool that represents the 2D correlation matrix across two distinct dimensions, utilizing color-coded cells, usually on a single-hue scale, to signify data values. The arrangement in a correlation heatmap is such that the values of the first dimension are mapped along the rows of the table, whereas the values pertaining to the second dimension are aligned along the columns. The intensity or shade of the color within each cell corresponds to the proportion of measurements that align with the specific values of the two dimensions being compared. The utility of correlation heatmaps in data analysis stems from their efficacy in underscoring disparities and variations within the dataset, thereby rendering patterns more discernible and interpretable. This characteristic makes them exceptionally suited for exploring and understanding complex data relationships which is given by Fig.1.

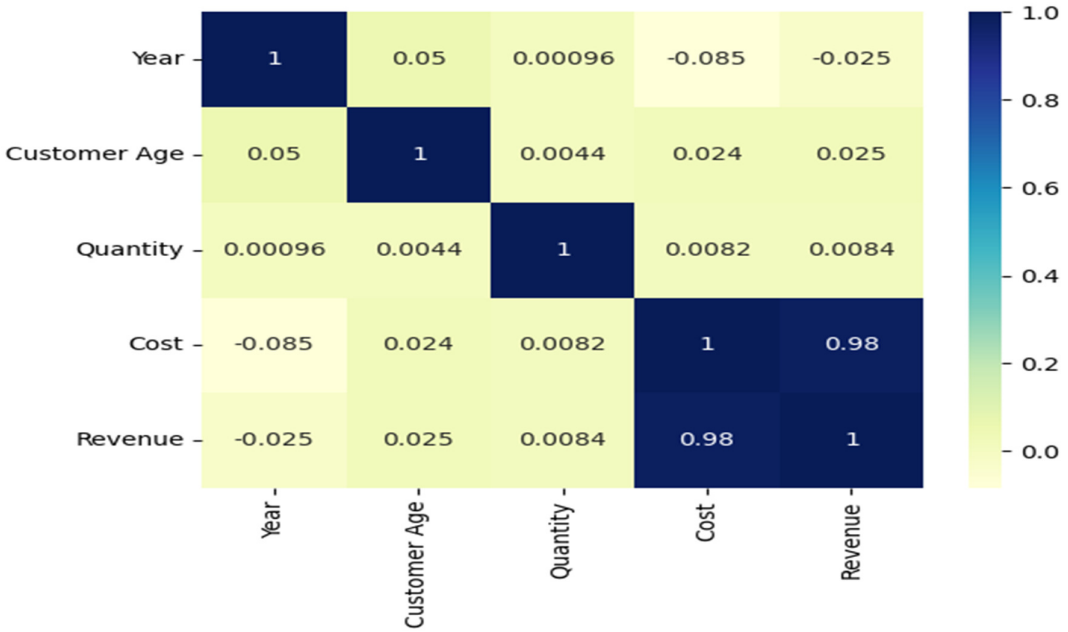

Fig.1 Corelation Heat map of features

Fig.1 presents a correlation heatmap, which is a graphical representation of the correlation matrix between various variables. In this heatmap, we have the variables: Year, Customer Age, Quantity, Cost, and Revenue. Each cell in the heatmap corresponds to the correlation coefficient between two variables. The correlation coefficient is a statistical measure that describes the extent to which two variables change together. It ranges from -1 to 1, where - 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and - 0 indicates no linear relationship between the variables. The color intensity in each cell reflects the magnitude and direction of the correlation between the variables. Darker shades typically represent stronger correlations, whether positive (dark blue) or negative (dark green to yellow). Lighter shades or colors towards the middle of the color scale represent weaker correlations. "Year" and "Customer Age" show a very weak positive correlation, as indicated by the light shade and the value (0.05) close to 0. "Quantity" and "Cost" are strongly positively correlated with "Revenue," shown by the dark blue color and the values (0.98) close to 1. This suggests that as the Quantity and Cost increase, the Revenue tends to increase proportionally. "Year," "Customer Age," and "Quantity" have very low correlations with "Revenue," given the very light shades and values close to 0, implying that these variables don't have a strong linear relationship with Revenue. The diagonal cells, which correlate each variable with itself, are always 1, as shown by the solid blue color, representing a perfect positive linear relationship with itself.

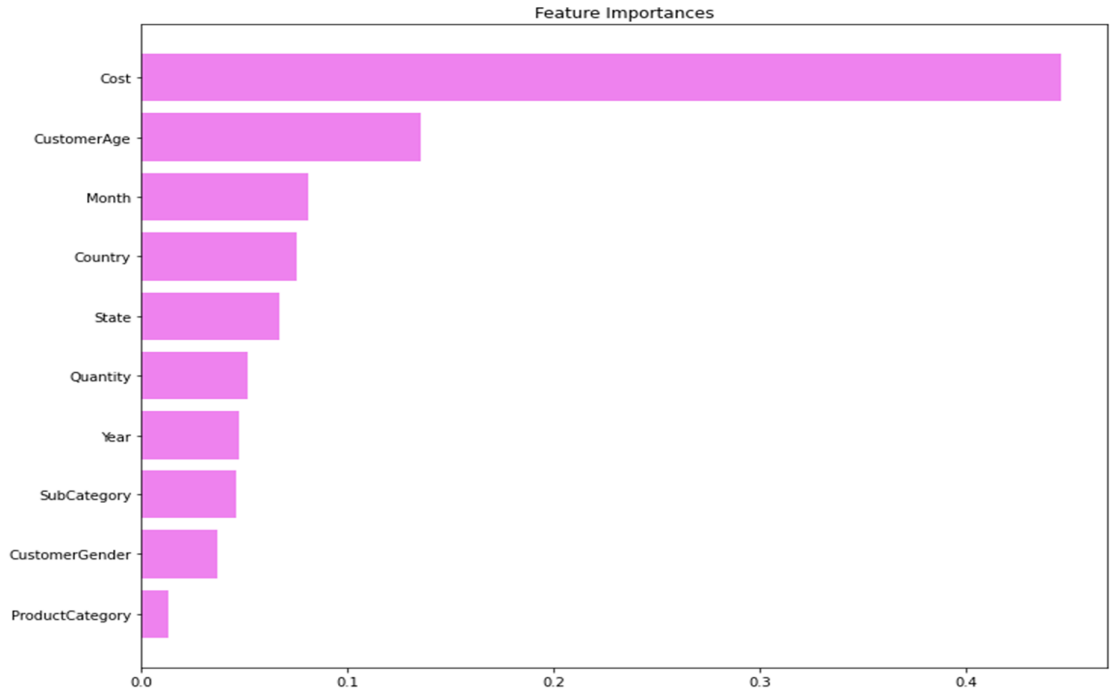

Fig. 2 Graphical representation of the features with respect with Revenue

From the Fig.2, Cost feature emerges as the predominant factor impacting the target variable, demonstrating a robust positive correlation with revenue. This suggests that higher costs are associated with increased revenue generation. Following closely, Customer Age is identified

as a significant predictor, suggesting an upward trend in spending power with advancing age, which in turn translates into greater expenditure on e-commerce platforms. This reflects a demographic trend where older individuals, presumably with more disposable income, contribute more to e-commerce sales. Interestingly, the expected norm that an increase in product price leads to a corresponding rise in sales does not necessarily hold true for older customers, indicating that this demographic's purchasing decisions may be less influenced by price changes.

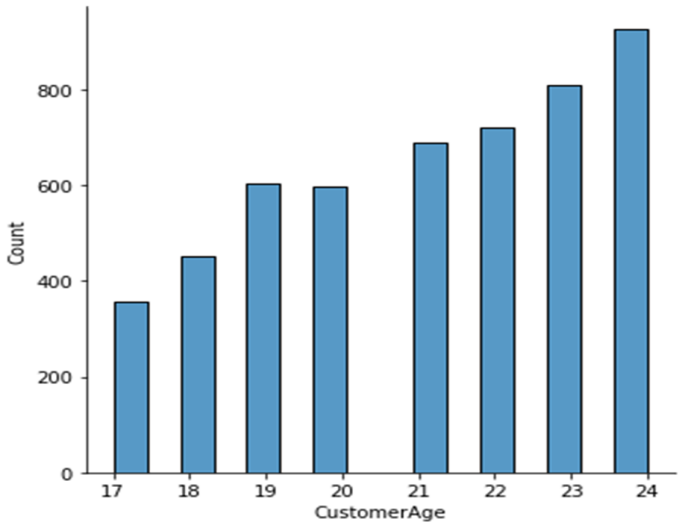

Fig.3 Graphical representation over the customer age with number of counts

In this figure, the demographic aged between 17 to 25 years has been identified as a key revenue source for e-commerce websites. This particular age group is typically characterized by a stage in life where financial obligations are comparatively minimal, often due to the support of guardians or the absence of extensive personal or familial commitments. This situation typically allows for a greater portion of their income or financial resources to be allocated towards discretionary purchases as opposed to essential expenses. The spending behavior of this demographic is further quantified by the observation that the mean transaction value, or the average order price, falls within the range of 350 to 400 Indian Rupees. This price point indicates a purchasing preference for items that are affordable yet significant enough to contribute to the revenue streams of e-commerce platforms. Such insights into the purchasing patterns and the associated financial latitude of this age group provide valuable information for targeted marketing strategies and inventory decisions aimed at maximizing engagement and sales within this segment.

From the Fig.4, it has been observed that the 25-45 age group stands as the most significant revenue driver for e-commerce platforms. This segment, often in the midst of their professional lives, tends to have a higher disposable income, which is the portion of income available for spending and saving after taxes and other obligations have been accounted for. Their financial position is usually more secure due to established careers or advanced professional experience, resulting in a greater capacity and willingness to allocate funds for non-essential items or

services. Correspondingly, the average transaction value recorded for this demographic is 1000 Indian Rupees, suggesting that individuals within this age range are comfortable making relatively substantial purchases online. This could indicate a consumer behavior that leans towards premium products, a diverse range of goods, or a combination of both in a single transaction. The average order value being notably higher than that of younger age groups underlines a greater economic impact on e-commerce revenues, and it also hints at a mature consumer base that potentially values quality, convenience, or a blend of factors that lead to higher spend per purchase. Understanding this behavior is crucial for e-commerce businesses as it assists in tailoring marketing strategies, product offerings, and customer experience enhancements to attract and retain this lucrative age segment.

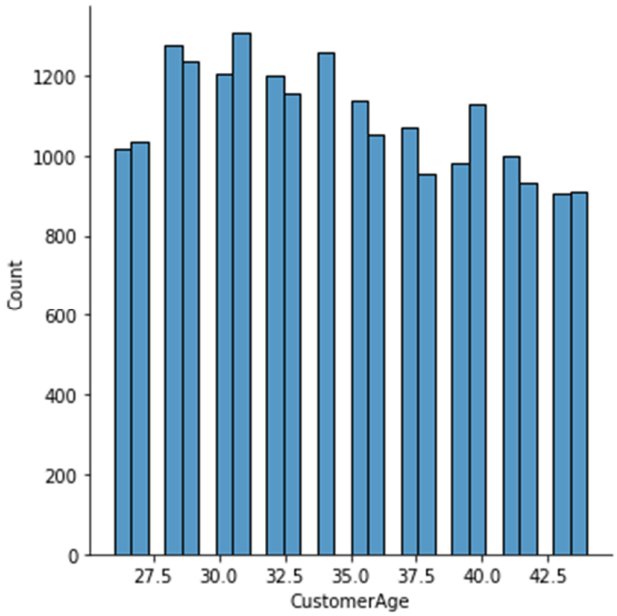

Fig.4 Average counts for the customer age between 25 to 45 years

Table 5: Demographic characteristics of Year, Customer Age, Quantity, Cost and Revenue

	Year	Customer Age	Quantity	Cost	Revenue
Year	1.000000	0.049879	0.000959	-0.084503	-0.025316
Customer Age	0.049879	1.000000	0.004378	0.023634	0.024831
Quantity	0.000959	0.004378	1.000000	0.008221	0.008369
Cost	-0.084503	0.023634	0.008221	1.000000	0.979125
Revenue	-0.025316	0.024831	0.008369	0.979125	1.000000

In this table, a numerical correlation matrix has been presented , which shows correlation coefficients between several variables. Each cell in the table shows the correlation coefficient

between two variables. The variables included in this matrix are Year, Customer Age, Quantity, Cost, and Revenue. The values of the correlation coefficients range from -1 to 1. A value close to 1 suggests a strong positive correlation, meaning that as one variable increases, the other variable tends to also increase. Conversely, a value close to -1 indicates a strong negative correlation, where an increase in one variable corresponds to a decrease in the other. A value near 0 suggests no linear correlation between the variables.

A correlation coefficient of 0.049879 suggests a very weak positive linear relationship. With a coefficient of 0.000959, there is virtually no linear relationship. A negative coefficient of -0.084503 implies a very weak inverse relationship; as the year increases, cost slightly decreases. At -0.025316, the relationship is negligibly negative. With a correlation of 0.004378, there is no significant linear relationship. Similarly, a coefficient of 0.024831 indicates a very weak positive correlation. Very weak positive correlation at 0.008221, suggesting little to no linear relationship. Also, a very weak positive correlation of 0.008369 which exhibits a very strong positive correlation of 0.979125, indicating that as the cost increases, the revenue tends to increase in a proportional manner.

6. CONCLUSION

In the fiercely competitive realm of e-commerce, insights into consumer behavior are invaluable. Leveraging data mining techniques allows for the extraction of critical data points that can be instrumental in mitigating competitive pressures and augmenting business profitability. The primary objective of this research is to dissect customer expenditure trends through the application of various predictive analytic models to datasets originating from an e-commerce platform. This endeavor enables researchers to decipher the relationships between demographic variables and their sway over consumer purchase decisions. The empirical investigation has led to the development of predictive models constructed on the demographic attributes of customers, employing supervised learning algorithms such as Decision Tree Classifier, Random Forest Classifier, Linear Regression, and Naïve Bayes Classifier to facilitate predictions. The insights derived from this analytical study shed light on the influence of customer spending behaviors on e-commerce revenue, providing strategic direction to enhance revenue streams. Consequently, the findings from this predictive inquiry are poised to equip e-commerce entities with the means to tailor strategies that enrich consumer-centric purchasing experiences. Moreover, the study delineates distinct customer segments that demonstrate a higher propensity for e-commerce transactions, offering a strategic advantage in targeting and customer engagement.

Conflicts of Interest

All the authors have declared that there has been no conflict of interest.

References

1. Karn, Arodh Lal, Rakshha Kumari Karna, Bhavana Raj Kondamudi, Girish Bagale, Denis A. Pustokhin, Irina V. Pustokhina, and Sudhakar Sengan. "Customer centric hybrid recommendation system for E-Commerce applications by integrating hybrid sentiment analysis." *Electronic Commerce Research* 23, no. 1 (2023): 279-314.
2. Sakalauskas, Virgilijus, and Dalia Kriksciuniene. "Personalized Advertising in E-Commerce: Using Clickstream Data to Target High-Value Customers." *Algorithms* 17, no. 1 (2024): 27.
3. Hai, Tao, Jincheng Zhou, Ye Lu, Dayang NA Jawawi, Anurag Sinha, Yash Bhatnagar, and Noble Anumbe. "Posterior probability and collaborative filtering based Heterogeneous Recommendations model for user/item Application in use case of IoVT." *Computers and Electrical Engineering* 105 (2023): 108532.
4. Obrenovic, Bojan, Jianguo Du, Danijela Godinic, Diana Tsoy, Muhammad Aamir Shafique Khan, and Ilimdorjon Jakhongirov. "Sustaining enterprise operations and productivity during the COVID-19 pandemic:"Enterprise Effectiveness and Sustainability Model"." *Sustainability* 12, no. 15 (2020): 5981.
5. Kumar, Nitendra, Priyanka Agarwal, Gauri Gupta, Sadhana Tiwari, and Padmesh Tripathi. "AI-Driven Financial Forecasting: The Power of Soft Computing." In *Intelligent Optimization Techniques for Business Analytics*, pp. 146-170. IGI Global, 2024.
6. Bharadiya, Jasmin Praful. "The role of machine learning in transforming business intelligence." *International Journal of Computing and Artificial Intelligence* 4, no. 1 (2023): 16-24.
7. Subasi, Abdulhamit. "Hospital readmission forecasting using artificial intelligence." In *Applications of Artificial Intelligence Healthcare and Biomedicine*, pp. 455-520. Academic Press, 2024.
8. Gupta, Kirti, Pravin Mane, Omprakash Sugdeo Rajankar, Mahua Bhowmik, Ranjana Jadhav, Sapna Yadav, Shitalkumar Rawandale, and Santoshkumar Vaman Chobe. "Harnessing AI for strategic decision-making and business performance optimization." *International Journal of Intelligent Systems and Applications in Engineering* 11, no. 10s (2023): 893-912.
9. Alrumiah, Sarah S., and Mohammed Hadwan. "Implementing big data analytics in e-commerce: Vendor and customer view." *Ieee Access* 9 (2021): 37281-37286.
10. Gupta, Shivam, Théo Justy, Shampy Kamboj, Ajay Kumar, and Eivind Kristoffersen. "Big data and firm marketing performance: Findings from knowledge-based view." *Technological Forecasting and Social Change* 171 (2021): 120986.
11. Fumey, Michael Provide, John Wiredu, and Agnes Nyamenaose Essuman. "Evaluating taxation's dual impact on business and social development: A case study of the Cape Coast metropolis in Ghana." *Financial Statistical Journal* 6, no. 2 (2024).
12. Tang, Huaizhi, Jiacheng Niu, Zibing Niu, Qi Liu, Yuanfang Huang, Wenju Yun, Chongyang Shen, and Zejun Huo. "System cognition and analytic technology of cultivated land quality from a data perspective." *Land* 12, no. 1 (2023): 237.
13. Fortino, Andres. "Data mining and predictive analytics for business decisions: a case study approach." (2023): 1-272.
14. Kleinstreuer, Nicole, and Thomas Hartung. "Artificial intelligence (AI)—it's the end of the tox as we know it (and I feel fine)." *Archives of Toxicology* (2024): 1-20.
15. Strielkowski, Wadim, Andrey Vlasov, Kirill Selivanov, Konstantin Muraviev, and Vadim Shakhnov. "Prospects and challenges of the machine learning and data-driven methods for the predictive analysis of power systems: A review." *Energies* 16, no. 10 (2023): 4025.
16. Poli, Naznin Sulata, and Abu Sayed Sikder. "Predictive Analysis of Sales Using the Apriori Algorithm: A Comprehensive Study on Sales Forecasting and Business Strategies in the Retail Industry.: Predictive Analysis of Sales Using the Apriori Algorithm." *International Journal of Imminent Science & Technology* 1, no. 1 (2023): 1-16.
17. De, Soumi, P. Prabu, and Joy Paulose. "Effective ML techniques to predict customer churn." In *2021 Third international conference on inventive research in computing applications (ICIRCA)*, pp. 895-902. IEEE, 2021.
18. Kumar, Akshi, and Arunima Jaiswal. "Scalable intelligent data-driven decision making for cognitive cities." *Energy Systems* 13, no. 3 (2022): 581-599.
19. Marchesi, Lodovica, Michele Marchesi, and Roberto Tonelli. "ABCDE—Agile block chain DApp engineering." *Blockchain: Research and Applications* 1, no. 1-2 (2020): 100002.

20. Leppla, Lynn, Sandra Hobelsberger, Dennis Rockstein, Viktor Werlitz, Stefan Pschenitza, Phillip Heidegger, Sabina De Geest, Sabine Valenta, Alexandra Teynor, and SMILe study team. "Implementation science meets software development to create eHealth components for an integrated care model for allogeneic stem cell transplantation facilitated by eHealth: the SMILe study as an example." *Journal of nursing scholarship* 53, no. 1 (2021): 35-45.
21. Rindell, Kalle, Jukka Ruohonen, Johannes Holvitie, Sami Hyrynsalmi, and Ville Leppänen. "Security in agile software development: A practitioner survey." *Information and Software Technology* 131 (2021): 106488.
22. Ahmed, Saadaldeen Rashid, Aso Kurdo Ahmed, and Swran Jawamir Jwmaa. "Analyzing The Employee Turnover by Using Decision Tree Algorithm." In *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1-4. IEEE, 2023.
23. Khajavi, Hamed, and Amir Rastgoo. "Predicting the carbon dioxide emission caused by road transport using a Random Forest (RF) model combined by Meta-Heuristic Algorithms." *Sustainable Cities and Society* 93 (2023): 104503.
24. Gohari, Kimiya, Anoshirvan Kazemnejad, Marjan Mohammadi, Farzad Eskandari, Samaneh Saberi, Maryam Esmaieli, and Ali Sheidaei. "A Bayesian latent class extension of naive Bayesian classifier and its application to the classification of gastric cancer patients." *BMC Medical Research Methodology* 23, no. 1 (2023): 190.
25. Yudhana, Anton, Andreyan Dwi Cahyo, Liya Yusrina Sabila, Arsyad Cahya Subrata, and Ilham Mufandi. "Spatial distribution of soil nutrient content for sustainable rice agriculture using geographic information system and Naïve Bayes classifier." *International Journal on Smart Sensing and Intelligent Systems* 16, no. 1 (2023).
26. Rofi'i, Yulianto Umar. "Analysis of E-Commerce Purchase Patterns Using Big Data: An Integrative Approach to Understanding Consumer Behavior." *International Journal Software Engineering and Computer Science (IJSECS)* 3, no. 3 (2023): 352-364.
27. Dol, Sunita M., and Pradip M. Jawandhiya. "Classification technique and its combination with clustering and association rule mining in educational data mining—A survey." *Engineering Applications of Artificial Intelligence* 122 (2023): 106071.
28. Chen, Kaile, Farhad Abtahi, Juan-Jesus Carrero, Carlos Fernandez-Llatas, and Fernando Seoane. "Process mining and data mining applications in the domain of chronic diseases: A systematic review." *Artificial Intelligence in Medicine* (2023): 102645.
29. Hu, Jianchang, and Silke Szymczak. "A review on longitudinal data analysis with random forest." *Briefings in Bioinformatics* 24, no. 2 (2023): bbad002.
30.