**Customer Shopping Trend Analysis with Machine Learning Approaches**

Contents

[Abstract: 2](#_Toc165558534)

[Keywords: 3](#_Toc165558535)

[Introduction: 3](#_Toc165558536)

[About the dataset : 4](#_Toc165558537)

[Experimental setup : 4](#_Toc165558538)

**[Method used :](#_Toc165558539)** [5](#_Toc165558539)

[Data Preprocessing 5](#_Toc165558540)

[Data cleaning 5](#_Toc165558541)

[Feature Engineering 5](#_Toc165558542)

[Splitting the data 5](#_Toc165558543)

[Scaling the Data 5](#_Toc165558544)

[Exploratory Data Analysis (EDA) - 6](#_Toc165558545)

[Statistical data analysis and visualisation 6](#_Toc165558546)

[Justification of the models 9](#_Toc165558547)

[Logistic Regression 9](#_Toc165558548)

[Model Evaluation 10](#_Toc165558549)

[Logistic Regression : 10](#_Toc165558550)

[RESULT & ANALYSIS 12](#_Toc165558551)

[Evaluation Metrics 12](#_Toc165558552)

[Results 12](#_Toc165558553)

[CONCLUSION 13](#_Toc165558554)

[References 13](#_Toc165558555)

Table of Figures

**[Figure 1: methods used 5](#_Toc165558406)**

**[Figure 2: Statistical data analysis and visualization 7](#_Toc165558407)**

**[Figure 3: Count chart for Items purchased and payment method 7](#_Toc165558408)**

**[Figure 4: Pie chart for gender distribution 8](#_Toc165558409)**

**[Figure 5: Heat map for correlation between features 8](#_Toc165558410)**

**[Figure 6: Count chart for review ratings and subscription status 9](#_Toc165558411)**

**[Figure 7: Correlation matrix of Logistic Regression 10](#_Toc165558412)**

**[Figure 8: PCA -1 and PCA-2 points 10](#_Toc165558413)**

**[Figure 9: Elbow Method 12](#_Toc165558414)**

**Customer Shopping Trend Analysis with Machine Learning Approaches**

# Abstract:

The customer shopping trend analysis is a vital aspect of the retail industry, as it helps businesses gain insights into the behavior of their customers. In recent times, customer behaviour models are typically based on data mining of customer data, and each model is designed to answer one question at one point in time. Predicting customer behaviour is an uncertain and difficult task. Thus, developing customer behaviour models requires the right technique and approach. This report presents a comprehensive analysis of customer shopping trends across various industries and market segments.

In this project, we aim to analyze customer shopping trends using historical data to identify patterns and predict future trends. Through extensive data collection and analysis of consumer purchase patterns, demographics this study identifies key drivers influencing shopping behaviors. It explores factors such as gender, item purchased, location, size, subscription status, shipping type that shape contemporary shopping habits.

Most customer behaviour models ignore so many pertinent factors that the predictions they generate are generally not very reliable. It examines the impact of season, frequency of purchases, payment method considerations on consumer choices.

By analyzing this data, we hope to provide businesses with valuable insights that can help them make informed decisions about them

marketing strategies, inventory management, and customer engagement tactics.

# Keywords:

Customer Shopping trend, Logistic Regression, Machine Learning, Ggradient boosting classifier, K-means clustering and data analysis.

# Introduction:

In recent years, the retail sector has witnessed a proliferation of data generated from various sources such as online transactions, e-commerce platforms, and customer interactions. This wealth of data presents both opportunities and challenges for retailers, as harnessing its potential requires advanced analytical techniques capable of distilling meaningful patterns from the noise.

The primary objective of this project is to leverage machine learning algorithms to analyze customer shopping trends effectively [6]. By harnessing the power of real-time data and sophisticated modeling techniques, we aim to provide retailers with actionable insights into consumer behavior, enabling them to make informed decisions regarding product offerings, pricing strategies, and marketing campaigns.

This project seeks to adapt and extend the machine learning methodologies applied in various sectors to the retail domain. By employing a diverse array of classification algorithms and feature selection methods, we aim to develop a robust framework for predicting customer shopping trends with high accuracy and reliability.

Key aspects of this report include a comprehensive analysis of various machine learning models, such as Logistic Regression, Gradient boosting classifier, and K-means clustering , to identify the most effective algorithm(s) for predicting customer behaviour [7]-[8]. Additionally, we will explore the importance of different variables in influencing shopping trends, providing valuable insights into the underlying drivers of consumer behavior.

To achieve our objectives, we will address the following critical questions:

1. Which machine learning models exhibit the highest efficacy in predicting customer shopping trends?
2. What are the key variables that significantly impact consumer behavior and influence shopping trends?

# About the dataset :

In this work, the customer shopping trend is used which consists of two databases. The dataset contains 3900 samples and 17 attributes, the target attribute represents that the customer has a subscription of the shopping website and constitutes yes or no value. The dataset includes various attributes related to the customer such as gender , age and size.

The dataset also includes details about shopping trends like categories of product , colour , purchase amount , review rating , subscription status , discount applied and mode of payment , season of purchase , frequency of purchase etc. The frequency of purchase has multiple values such as annually , bi-weekly , weekly , quarterly , monthly etc. In many studies, researchers used the dataset's features for different machine learning classification models to correctly identify trends in the shopping behaviour, exhibiting the dataset's utility. The dataset is available on : <https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset>

# Experimental setup :

## **Method used :**

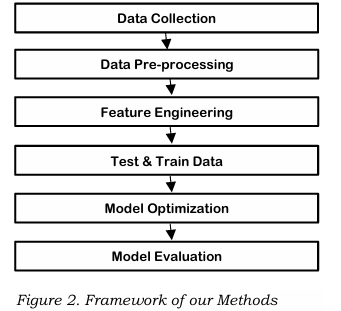


Figure 1: methods used

### Data Preprocessing

### Data cleaning

In this we have performed data analysis using manipulation libraries like pandas and numpy to find the missing values in the variables and cleaned them using the drop method [1]-[2]. The dataset doesn’t contain any duplicate value. Briefly described the data set by printing the size and shape of the dataset and by also displaying the , mean, median , mode , interquartile range , top five rows and the last five rows using describe, tail, and head functions to provide quick inspection into the structure of the dataset.

### Feature Engineering

There are many categorical features in the dataset, these features are converted to numerical features by using various methods. For every category in a categorical feature, additional features are generated by this method, where binary values are allocated to specify the presence or absence of the corresponding category,

Splitting the data **-** To ensure our model well generally on new data or unseen data, we split our data into training and test and dataset. We found two csv files from kaggle, we used the updated one for training the models and the older one for testing and validating the models.

Scaling the Data**:** Different features in our dataset have different scale, it is important to scale our data for the features to be on the same scale. We used StandardScaler to transform the features to have to have zero mean and the unit variance, this improved our machine learning algorithm performance and accuracy. We scaled the top 10 features selected before implementing our model to ensure all the features are on the same scale.

Encoding: Machine learning performs with numerical data. In this project, we adopted “label encoding” technique to transform our categorical variables into numerical features. The column with highest number of features has four variables and none of them are ordered or ranked, hence `change\_dtype` was used in the process. However, normalization technique was applied to the all the category which help return the frequency of the unique value for each category.

# Exploratory Data Analysis (EDA) -

Exploratory Data Analysis (EDA) provides understanding about the characteristics of the data and help us to understand trends and pattern. We imported necessary libraries such as NumPy, Pandas and Seaborn that gives opportunity to perform exploratory analyses such as value counts to obtain the frequency of the categorical variables [3].

## Statistical data analysis and visualisation

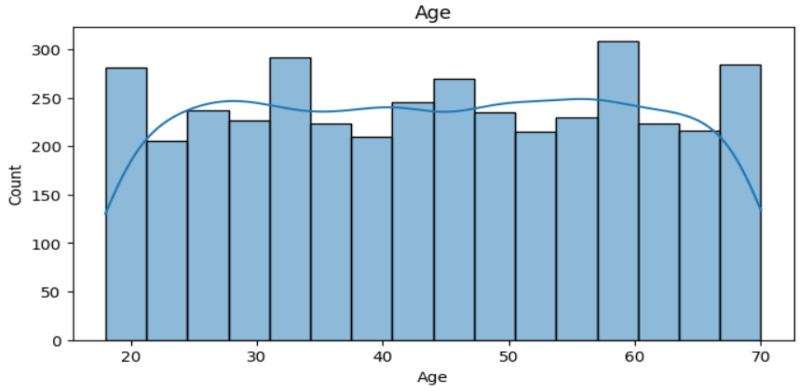


Figure 2: Statistical data analysis and visualization

Figure 1 shows the univariate analysis of the numerical attribute “Age” . This graph is displaying the distribution of the shopping trend among different age groups.

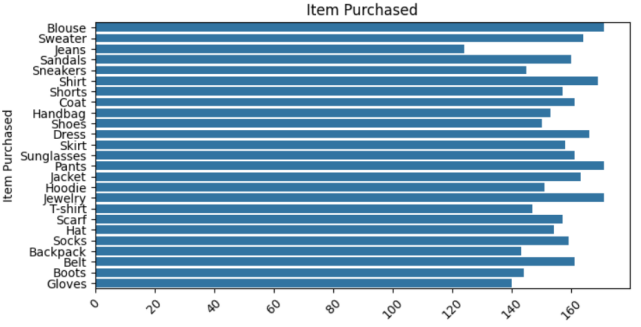
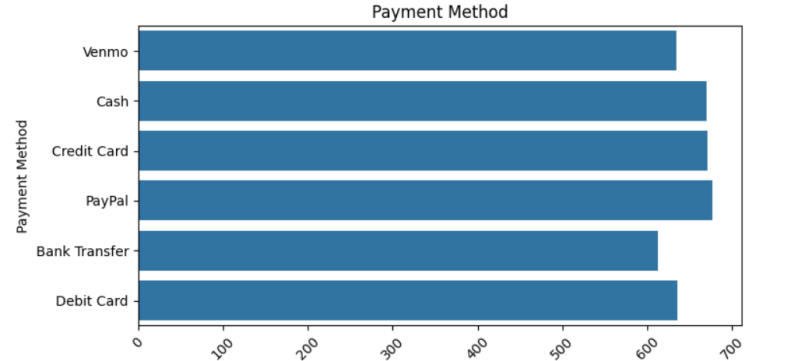
 

Figure 3: Count chart for Items purchased and payment method

Figure 2 depicts the univariate analysis of the categorical variables of the dataset “Item purchased” and “Payment method” through horizontal bar graphs for the counts distribution.

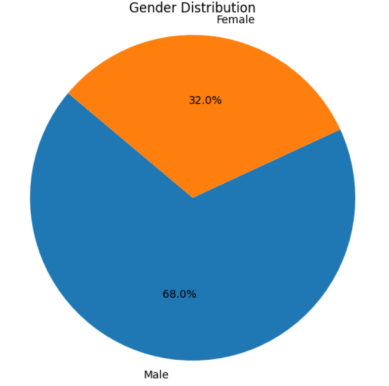


Figure 4: Pie chart for gender distribution

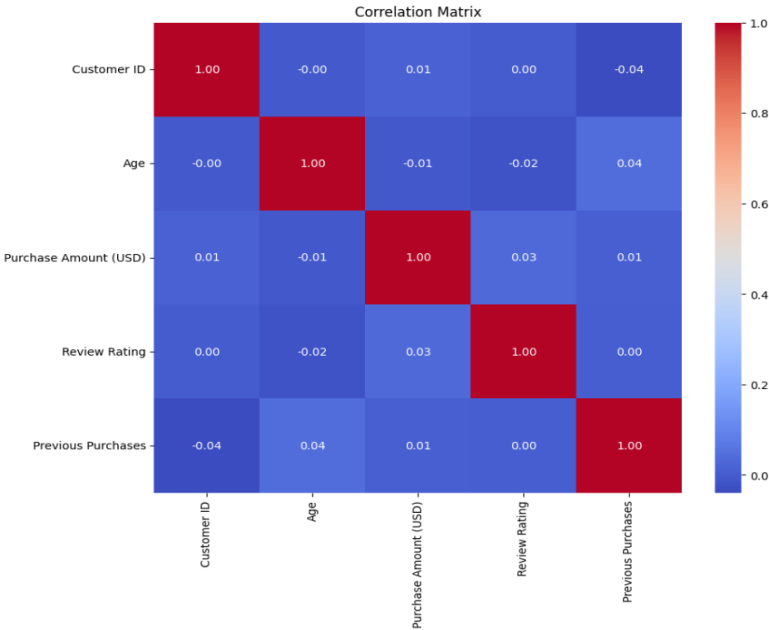


Figure 5: Heat map for correlation between features

Figure 4 is the correlation matrix used to show what form of correlation exist in between the attributes of the dataset and also shows that attributes are weakly

correlated with each other and highly correlated with themselves only. There is a weak correlation between age and previous purchases and also between purchase amount and review ratings.

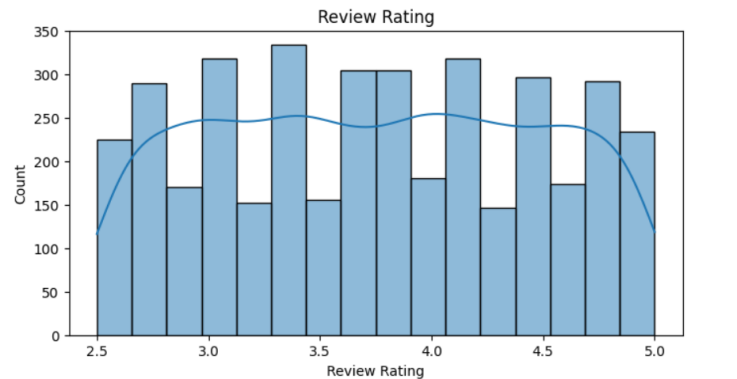
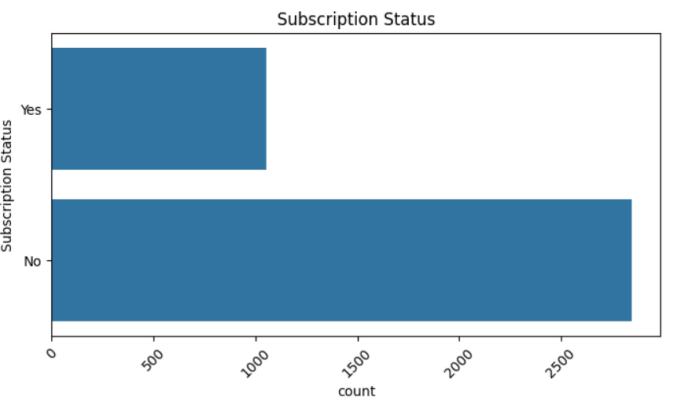
 

Figure 6: Count chart for review ratings and subscription status

Figure 5 depicts the count distribution of the different review ratings ranging from 0-5 given by the customers during their purchases. It clearly shows that mostly customers don’t prefer to have a subscription plan for shopping.

## Justification of the models

### Logistic Regression

Logistic regression (LR) is a statistical model used for binary classification tasks [4].

* It is best suited for a scenario where probabilities for dependent variable will fall into one of the two categories. The numerical values represent the model's predicted categories or classes . In our project we have used logistic regression to predict the preferred payment method by different genders.

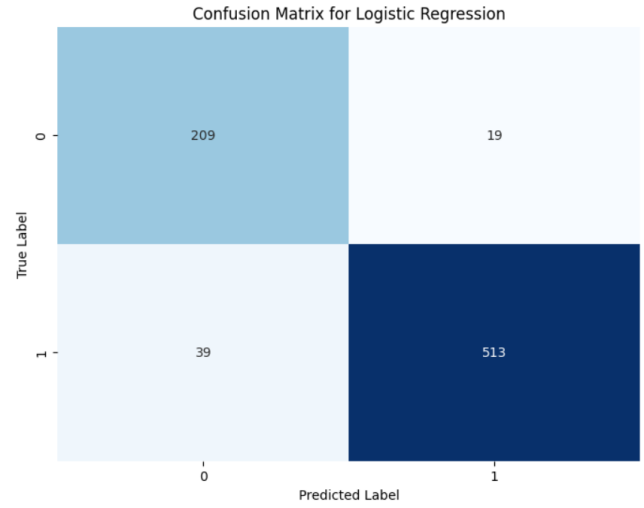


Figure 7: Correlation matrix of Logistic Regression

The performance of our models was further evaluated using confusion matrix based on these elements true positive, false positive, true negative and false negative for each class in the predicted data. The confusion matrix helps evaluate how well the model performs in respect to identification positive and negative instances of the previous purchase made.

**Gradient Boosting Machine(GBM)**

GBM is a powerful ensemble learning technique used for both regression and classification tasks. It builds multiple decision trees sequentially, where each tree corrects the errors of the previous one. GBM optimizes the model by minimizing a predefined loss function, enhancing its predictive accuracy with each iteratio. Here we have used GBM as a classification algorithm over the “Category” attribute to categories the products.

**K-Means Clustering**

1. means is an unsupervised machine learning algorithm for clustering data points into K distinct groups. It iteratively assigns data points to the nearest cluster centroid based on Euclidean distance. In our model , we have applied K-means clustering by creating 3 clusters.

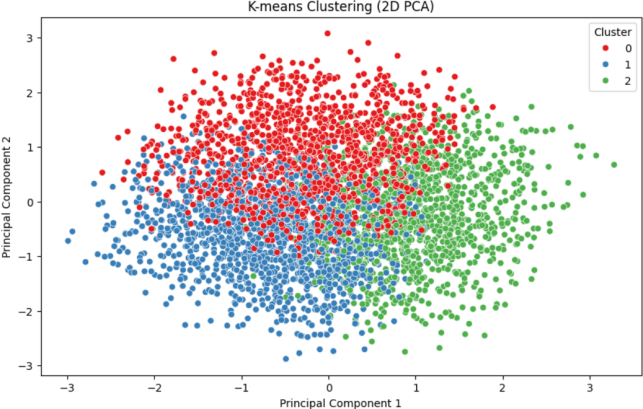


Figure 8: PCA -1 and PCA-2 points

Figure 12 displays a scatter plot where the first principal component (PC1) is displayed on the horizontal axis, and the second principal component (PC2) is displayed on the vertical axis. By visualising the data in this reduced feature space, insights into the distribution and degree of separation between the classes can be analysed. This plot shows how well the classes are separated in the reduced feature space.

## Model Evaluation

### Logistic Regression :

Following the selection of our machine learning algorithms, we evaluate the performance on these models on the split dataset based on different evaluation metrics which include accuracy, precision, recall, and F1-score. From these evaluation metrics, we understood how well the algorithms are able to predict the previous purchase of our the customer based on gender . The overall accuracy for the algorithms on both train and test dataset was visualized using bar chart for comparison.

**Table 1: Accuracy Table for Logistic Regression**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gender | Precision | Recall | F1-score | Support |
| 0 | 0.84 | 0.92 | 0.88 | 228 |
| 1 | 0.96 | 0.93 | 0.95 | 552 |

1. **Gradient Boosting Machine:**

GBM is a powerful ensemble learning technique used for both regression and classification tasks [5]. It builds multiple decision trees sequentially, where each tree corrects the errors of the previous one. At each iteration, t

the algorithm fits a new tree to the residuals (or gradients) of the previous predictions, hence the name "gradient boosting”. The final prediction is made by aggregating the predictions of all trees.

GBM is known for its high predictive accuracy and robustness to outliers.

The GBM accuracy Score is 1.0 suggests that the Gradient Boosting Machine (GBM) model achieved perfect accuracy on the test set. This means that the model made correct predictions for all instances in the test data.

1. **K-Means clustering**

**Table2: Unsupervised Learning Score Table**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Silhouette Score | Davis Bouldin Index | Calinski-Harabasz Index |
|  | 0.17525976417721367 | 1.2857011601443264 | 544.1324757989148 |

The summary of the performance was plotted in a table displaying all the four metrics, also

**Elbow Method -** Plotting the within-cluster sum of squares (inertia) for different values of K and looking for an "elbow" point where adding more clusters doesn't improve the model significantly.

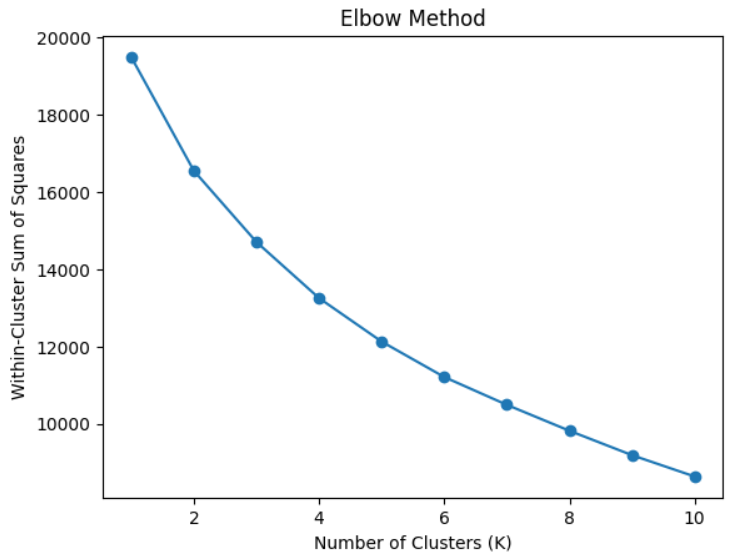


Figure 9: Elbow Method

# RESULT & ANALYSIS

## Evaluation Metrics

The accuracy of the classifier is calculated as,



The precision of the classifier is calculated as,



The recall of the classifier is calculated as,



The F1 score of the classifier is calculated as,



Here; TP = True Positive, FP = False Positive, and FN = False Negative

## Results

Various evaluation metrics, including accuracy score, precision, recall, and F1-score, are computed for the test dataset in order to assess the performance of the machine learning models.

The outcomes of these metrics for each of the four models trained in this work are shown in Table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| **Linear Regression** | 0.925641 | 0.93 | 0.93 | 0.93 |
| **Gradient Boosting Classifier** | 1.0 |  |  |  |

# CONCLUSION

For now , the model built gives considerable results and surely there is space for improvements by increasing the data length and the exploring more features which contribute to better analysis , insight gaining and predictions.

In our study and research we applied Logistic regression , Gradient Boosting machine , K-means clustering with labelencoding for categorical parameters resulting in acceptable results . In Gradient Boosting Machine we are getting 100% accuracy value for category of product purchased by customer.

Trying different ensemble learning to our current dataset would be part of future works that will worth trying to improve the performance of our low performing model.

On the whole conclusively , our aim through this research was to analyse the customer’s shopping trend using the various algorithm and analyse the performance to the selected algorithm. This is clear that machine learning algorithms are efficient for e-commerce and other businesses for understanding their customer potential , and take required actions to increase there sales and profit according to the factors. Further research is

required to validate the proposed model over more diverse and large datasets to investigate the customer shopping trend.

# 

# References

1. McKinney, W., 2012. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.".
2. McKinney, W. and Team, P.D., 2015. Pandas-Powerful python data analysis toolkit. *Pandas—Powerful Python Data Analysis Toolkit*, *1625*.
3. Waskom, M.L., 2021. Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), p.3021.
4. Hu, Z. and Lo, C.P., 2007. Modeling urban growth in Atlanta using logistic regression. *Computers, environment and urban systems*, *31*(6), pp.667-688.
5. Natekin, A. and Knoll, A., 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, p.21.
6. Orogun, Adebola , Onyekwelu, Bukola, “Predicting Consumer Behaviour in Digital Market: A Machine Learning Approach” in International Journal of Innovative Research in Science, Engineering and Technology Aug-2019.
7. Asante-Okyere, S., Shen, C., Ziggah, Y.Y., Rulegeya, M.M. and Zhu, X., 2020. A novel hybrid technique of integrating gradient-boosted machine and clustering algorithms for lithology classification. Natural Resources Research, 29, pp.2257-2273.
8. Zhou, S., Wang, S., Wu, Q., Azim, R. and Li, W., 2020. Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. Computational biology and chemistry, 85, p.107200.