SPRING 2023

CMPE 255 - DATA MINING PROJECT REPORT
On

“Black Friday Sales Prediction”

Instructor: TAEHEE JEONG

Submitted By:

Jesse Avanakshi[016664778]
Prashansa Evangeline Bonapalle[016678324]
Srikar Badugu[016697655]
Paavamaani Manchala[016626636]

Index

1. Introduction

The Black Friday sales event celebrated on the day following Thanksgiving Day in the United States, has become a highly significant period for retail stores and eCommerce businesses to generate profits by offering highly promoted sales. This event, which marks the beginning of the Christmas shopping season in the United States, has been a tradition since 1952, although the term "Black Friday" did not become widely used until more recent decades. In recent years, the popularity of digital shopping has increased significantly, with many customers preferring to shop online due to its convenience, wider selection of goods, easy price comparisons, and avoidance of crowds, among other benefits. Furthermore, the ongoing pandemic has further boosted the trend of online shopping, leading to a significant increase in online sales. As such, eCommerce businesses and retail stores must leverage historical sales data and machine learning approaches to predict and analyze customer behavior and preferences during the Black Friday sales event.

To this end, this project aims to determine the optimal product prices during the Black Friday sales event by utilizing machine learning regression models, such as linear regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regression. The Black Friday sales dataset, which is widely accepted by various e-commerce websites for training and prediction purposes, will be used in this project. We will analyze the relationship between various attributes, including city, age, marital status, and occupation, and their impact on customer shopping patterns. Through this project, we aim to develop insights into customer behavior and preferences during the Black Friday sales event and demonstrate how machine learning methods can enhance business growth and improve sales strategies. Ultimately, our project aims to help retail stores and eCommerce businesses determine optimal pricing strategies that maximize profits, improve customer satisfaction, and drive business growth during the Black Friday sales event.

2. Background and Need

Several research studies have been conducted to predict sales using various techniques. Several machine-learning approaches have been proposed and analyzed by researchers. In this section, we summarize some of these approaches. The study by "S. Yadav et al." evaluated and compared the effectiveness of hold-out validation and K-Fold cross-validation methods. The results showed that K-Fold cross-validation produced more precise findings, with accuracy values that were approximately 0.1-3% more accurate than hold-out validation for the same set of methods.

Another study conducted by "Aaditi Narkhede et al." implemented a machine learning algorithm to predict customer demand and manage inventory in shopping centres and large marts. These methods have shown potential for data shaping and decision-making, and provide opportunities for better-identifying customer demands and calculating marketing strategies to increase sales.

"Singh, K" et al. evaluated and graphically depicted sales data using complicated datasets, providing clarity about how it works, enabling business owners to assess and visualize sales data, resulting in appropriate choices and income generation. The data visualization is built around many factors and dimensions, allowing end-users to make better judgments, estimate future sales, boost production based on demand, and calculate regional sales.

"M.Sahaya Vennila et al." utilized machine learning approaches to estimate sales by analyzing, preprocessing, and training data. The Black Friday Sales Dataset from Kaggle was used for analysis and testing. The K-Fold technique was used to divide the dataset into training and testing datasets. The prediction model was built using Linear Regression, Decision Tree, Random Forest, Gradient boosting, and XGBoost, and the accuracy evaluation metrics included Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). After extensive testing, the Random Forest method performed well, with an accuracy of 75%, an RMSE of 2720, and an MAE of 2318.

Need

The Black Friday sales event is a critical period for businesses to boost their profits and increase customer loyalty by offering discounts on various product items. With the advent of the digital shopping trend and the ongoing pandemic, it has become more necessary for businesses to leverage machine learning methods to predict and analyze customer shopping behavior during the Black Friday sales event. The use of regression models such as linear regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regression can provide valuable insights into customer preferences and aid in optimizing sales strategies. Thus, the necessity of this project lies in its potential to demonstrate the benefits of machine learning techniques in enhancing business growth and profitability by developing a deeper understanding of customer behaviour and preferences during the Black Friday sales event.

3. Data Analysis & Preprocessing

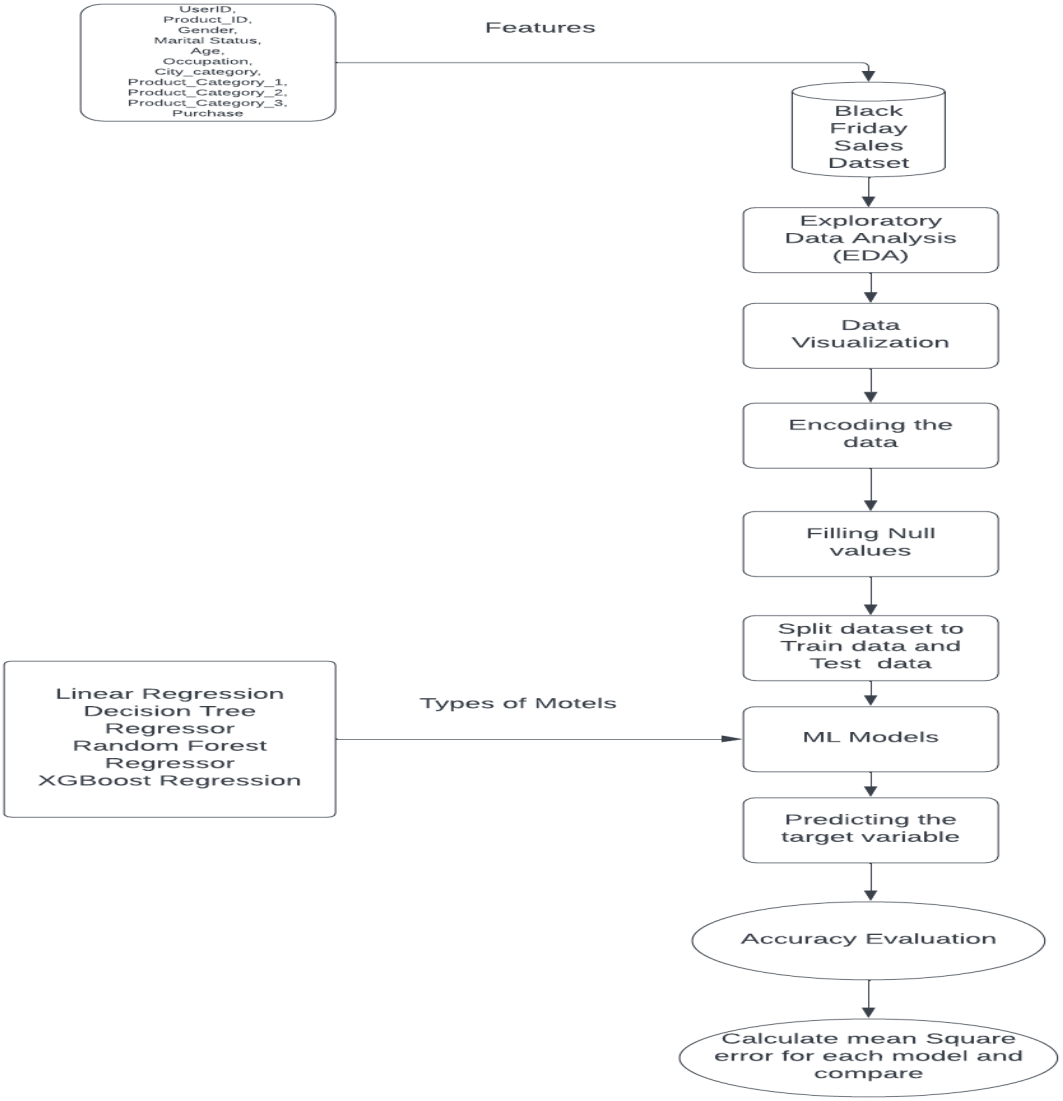
Dataset Description:

The dataset we have taken for this project is from Kaggle. The dataset contains 55068 records and 12 attributes. We have different attributes like age, gender, city, marital status, product category, and Purchase amount etc., Our target attribute is Purchase which is the total amount of an order. We are building the models to predict the purchase amount of the products.

Source: https://www.kaggle.com/code/midouazerty/black-friday-sales-prediction

Attributes	Non-null record for the attribute	Definition of Attributes
User_ID	550068	ID of the customer
Product_ID	550068	ID of the Product
Gender	550068	Customer Gender
Age	550068	Customer Age Group
Occupation	550068	Occupation of the customer
City_Category	550068	The current_ city where the customer lives in
Stay_In_Current_City_Years	550068	From how many years customer is staying the current city
Marital_Status	550068	Customer marital status whether married or unmarried
Product_Category_1	550068	Category of the product
Product_Category_2	376430	Category of the product
Product_Category_3	166821	Category of the product
Purchase	550068	Amount of the order placed by a Customer

4. Implementation & Work Flow



We have picked the dataset and analyzed the data using different plots. Our target variable is 'Purchase'. We perform Encoding the categorical values, filled null values. Removed User_ID and Product_ID as they wont be of much help in creating the Model because they are unique values.

We used 4 ML models Linear Regression, Decision Tree Regressor, Random Forest Regressor and XGBoost Regressor and compared the Mean Square Error.

Steps we followed:

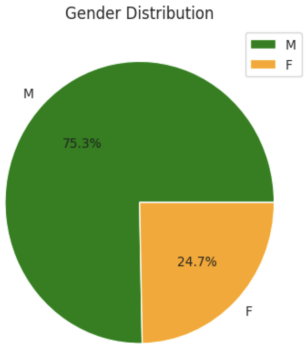
1. Exploratory Data Analysis
2. Data Preparation
3. Modelling
4. Results and Conclusion

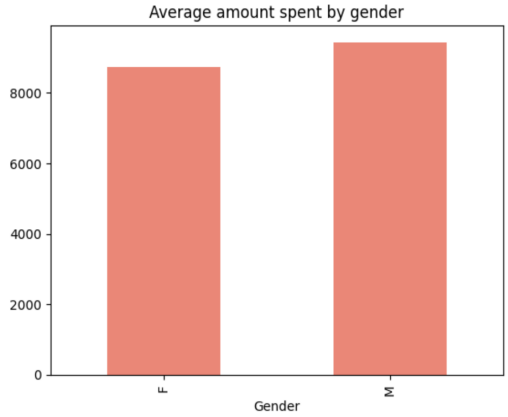
4.1 Exploratory Data Analysis:

Any data analyst or scientist seeking insight into their data should engage in exploratory data analysis (EDA). We can use EDA to understand the distribution, patterns, and relationship between the data, to identify outliers, and to identify any data quality problems. Using this process, we are able to generate hypotheses, perform data preprocessing, and improve the accuracy and reliability of our subsequent analysis or modeling. The EDA process involves the use of many graphical and statistical techniques, including histograms, pie plots, box plots, and correlation matrices. With the help of EDA, we were able to gain a deeper understanding of the data, allowing us to make better decisions and improve the performance of predictive models. In order to obtain the desired outcomes, it is crucial that EDA is performed in order to ensure that data analysis is of high quality and reliable.

4.2 Data Visualization

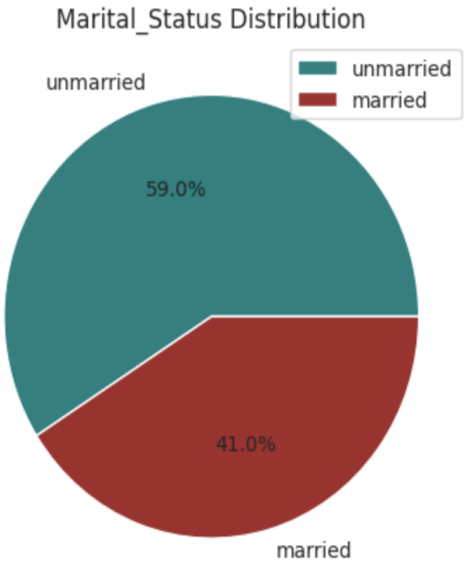
Gender

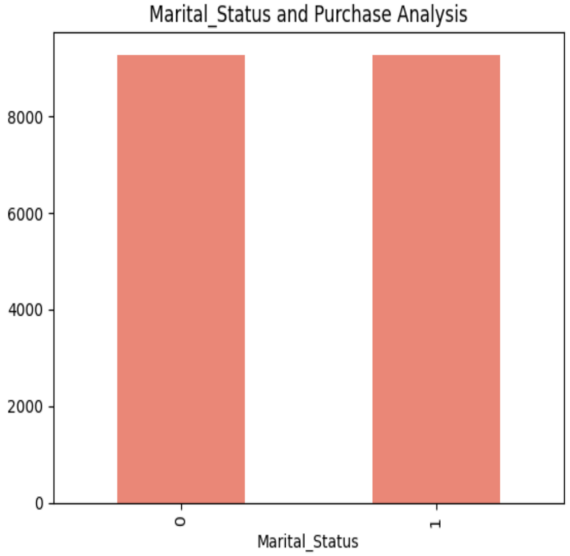




The above plots shows the distribution of gender where Male customers are more than Female customers and the average amount spent by male is more than the average amount spent by female.

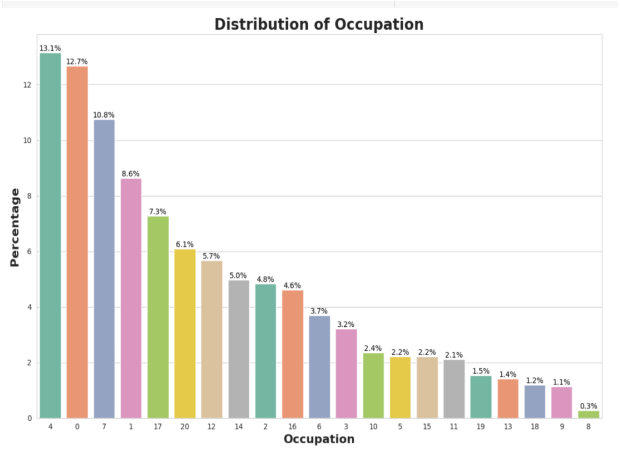
Marital status

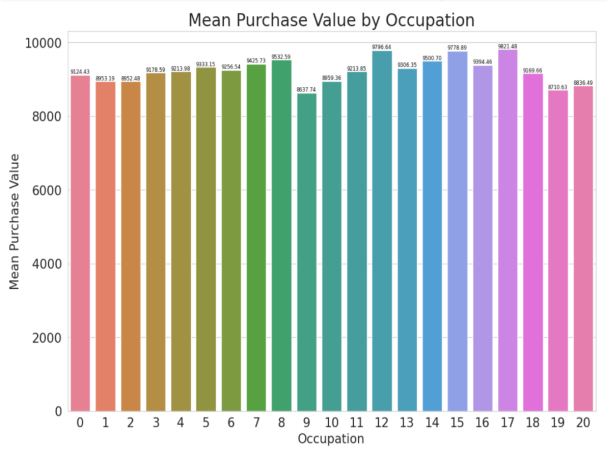




We have an interesting observation here, The amount spent by married and unmarried are same even though there are more unmarried customers than the married customers.

Occupation

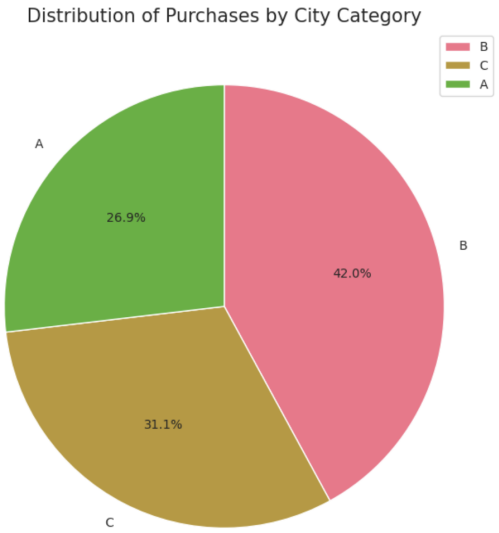


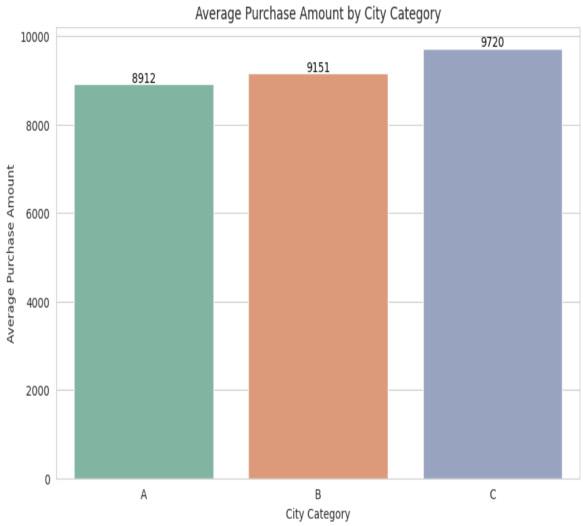


In the occupation variable, there are at least 20 different values, and we are unaware of which number corresponds to which occupation. As a result, analyzing the data is difficult. The number of occupation categories cannot be reduced in any feasible manner.

It is true that we have customers from a variety of occupational backgrounds, despite the lack of clarity regarding specific occupations. Intriguingly, the average purchase value is generally consistent across occupations, ranging between 8000 and ten thousand dollars. As a result of these initial observations, it appears that occupation may not be a significant determinant of purchase behavior. This hypothesis may need to be confirmed or refuted through further analysis.

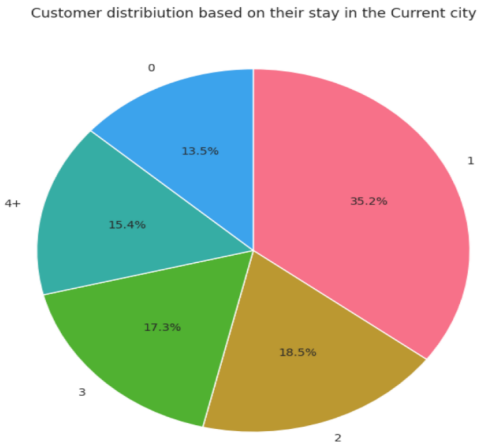
City_Category

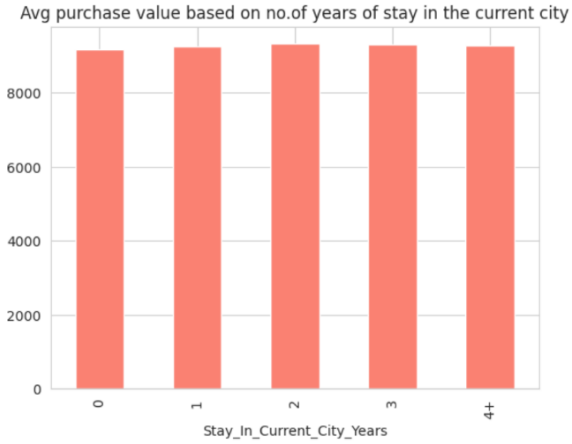




We observe that no.of purchases from City B is more than other cities A & B but the amount of purchases is more for City C rather than B.

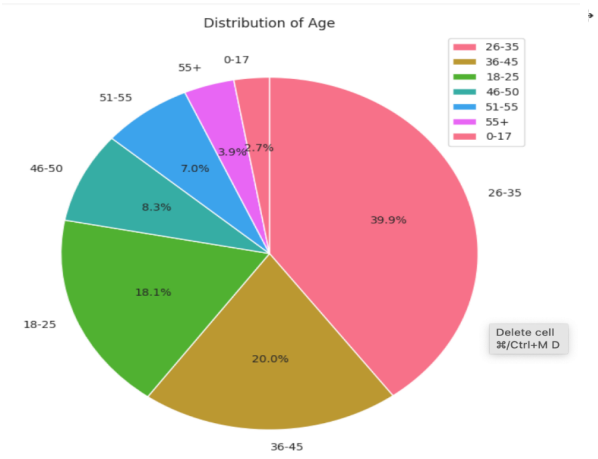
Stay_in_Current_City

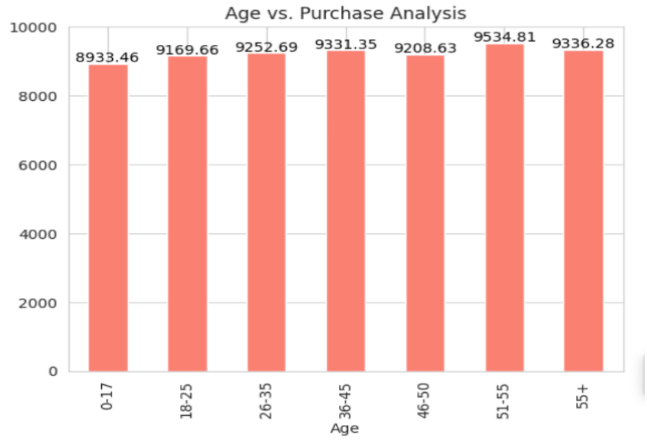




The distribution of the customers who are staying in the current city from years is varying and there are more no of purchases from the customers who are staying from past one year rather than 2 ,3, 4+ years. But still the average purchase value is the same for all different distributions.

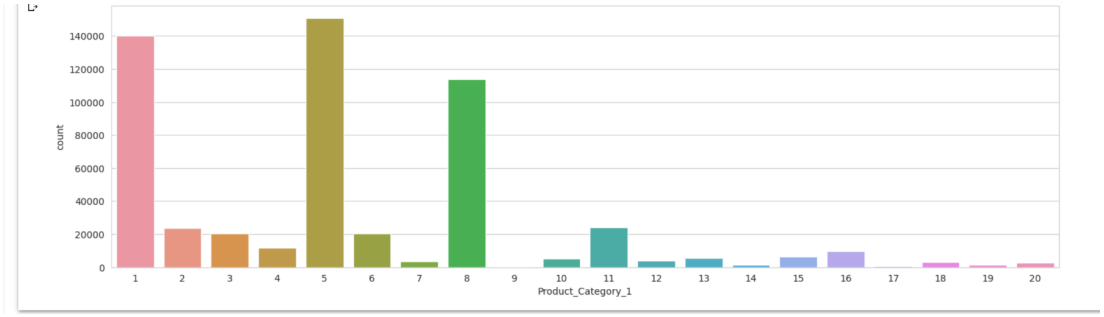
Age:

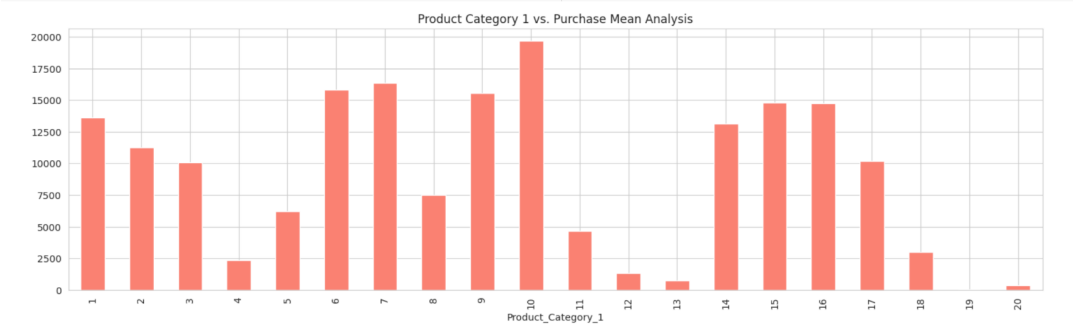




The above plot represents that customers are more from the age group 26-35, then 36-45 and then 18-25. But the average amount spent by each age group is more or less the same.

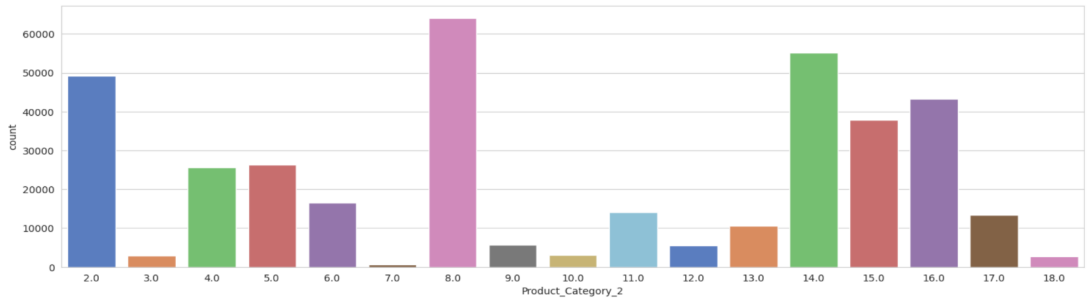
Product_Category_1





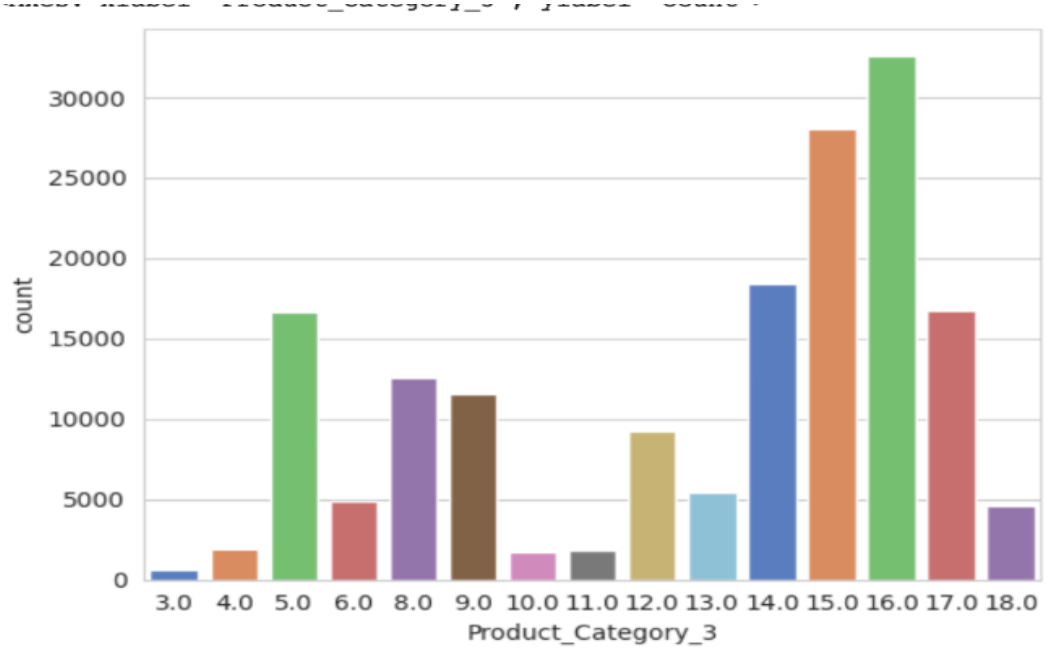
From the above plot, it is very clear that products 1, 5, 8 have more orders placed. But We could not say which product is that as it is not represented in the dataset.
As can be seen by the average amount spent for Product_Category_1, despite there being more products purchased in categories 1,5,8, the average amount spent on those three products isn't the highest. The fact that other categories had high purchase values despite having minimal impact on sales is interesting.

Product_Category_2



From the above plot, it is very clear that products 2, 8, 14 have more orders placed. But We could not say which product is that as it is not represented in the dataset.

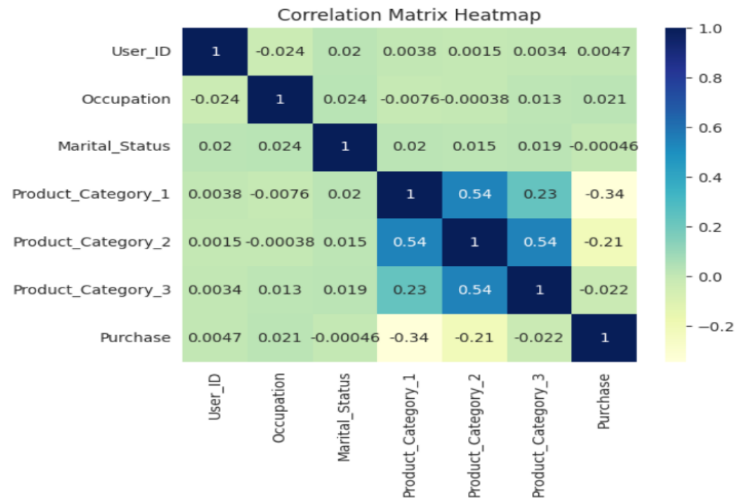
Product_Category_3



From the above plot, it is very clear that products 15, 16 have more orders placed. But We could not say which product is that as it is not represented in the dataset.

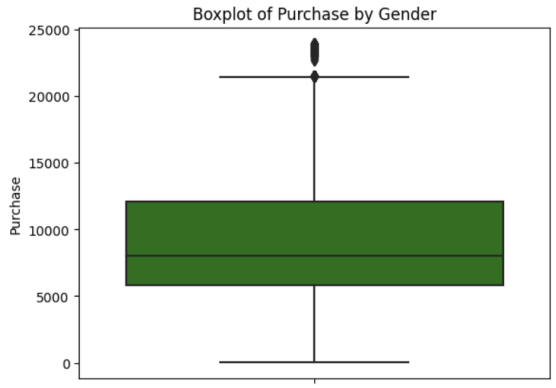
Heat Map:

A heatmap is a graphical representation that facilitates the identification and interpretation of relationships that exist among various attributes in a given dataset. This method employs a color scheme to depict the extent of association between multiple properties. Specifically, it encodes the degree of correlation between the target and attribute variables such that a stronger association corresponds to a more intense coloration.



We understand that there is a correlation between the product category groups and purchase

Box Plot for Target Variable



Skew : 0.6001400037087128
Kurtosis : -0.3383775655851702
IQR
User_ID -2962.0
Occupation -12.0
Marital_Status -1.0
Product_Category_1 -7.0
Product_Category_2 -10.0
Product_Category_3 -7.0
Purchase -6231.0

4.3 Data Preparation

Encoding the Categorical Values

When creating any Machine Learning model, it is important to ensure that the data is in numerical form. This means that categorical variables must be converted into numerical values. One way to do this is to use Label Encoder, which takes a categorical value and assigns it a numerical value based on its label in the dataset. This helps to reduce the complexity of the model, allowing it to make more accurate predictions. Label Encoder is a powerful tool for encoding categorical values and can be used in a variety of Machine Learning models. It is important to remember that the numerical values assigned to a label should be consistent across the dataset, otherwise the model could be misleading.

We found that there are categorical values in the dataset for the attributes gender, age and City. Gender is an important factor when considering data analysis. To facilitate the analysis, it can be transformed into numerical form with 0 representing female and 1 representing male. This process is known as Label Encoding. Other data categories, such as City Category and Age has also been transformed into numerical form using Label Encoding for more efficient and accurate analysis.

Filling the Missing Values.

Our analysis of the dataset revealed that some values in both the Product_category_2 and Product_category_3 attributes were null. To address this, we decided to fill these values with 0 as customers may not necessarily buy products from every category. This seemed like a reasonable solution as it allows customers to purchase items from one category in particular.

Dropping the irrelevant attributes

The attributes User ID and Product ID are unique identifiers that do not provide any meaningful insights for the prediction task. As they are unique to each user and product, they do not influence any patterns or relationships in the data. Accordingly, these attributes can be safely removed from the analysis without any significant information loss. This can help to reduce the dimensionality of the data and improve the performance of the prediction task. By removing these attributes, we can focus more on the relevant features that can provide more meaningful insights.

Splitting the dataset

In machine learning and data analysis, it is crucial to split the data into two or more sets, typically a training set and a testing set. It involves dividing the available dataset into two or more parts. Models are developed and trained using the training set, while models are evaluated based on the testing set.

A model that was trained on the entire dataset may perform well on the training set, but not necessarily on new, unseen data. The primary reason for splitting the data is to evaluate the generalization ability of the model. The model's performance can be evaluated in real-world scenarios more accurately when the data is split into a completely independent set.

The 70/30 or 80/20 rule is typically used to split up the data, where 70% or 80% of the data is used for training and 30% or 20% for testing. It is important to keep in mind that the optimal split depends on a number of factors, including the size and complexity of the dataset, as well as the specific domain of the problem.

We have taken a 70/30 split for our dataset. It is essential to split the data in order to evaluate a model's performance and ensure that it can generalize well to new data.

Modeling

Implementing multiple supervised models is a common approach in machine learning, and it can be a great way to get the most accurate predictions out of your data. Linear Regressor, Decision Tree Regressor, and Random Forest Regressor are three popular models that can be used in this type of approach.

Linear Regressor is a model that uses linear equations to fit data points, and it is best suited for data that is linearly separable. Decision Tree Regressor is a model that uses decision trees to make predictions, and it is best used with data that has a lot of features. Finally, Random Forest Regressor is an ensemble model that uses multiple decision tree regressors to make predictions, and is best used when the data is complex and there are multiple features.

These models use different algorithms and techniques to learn from the data and make predictions. Linear Regressor uses linear equations to fit the data, Decision Tree Regressor uses decision trees to make predictions, and Random Forest Regressor uses a combination of decision tree regressors to make predictions. Each model has its own strengths and weaknesses, and it is important to understand the data before selecting the best model for your problem.

Implementing multiple supervised models such as Linear Regressor, Decision Tree Regressor, and Random Forest Regressor is a common approach in machine learning. These models use different algorithms and techniques to learn from the data and make predictions, and selecting the best model for the problem depends on understanding the data.

We have taken our measure as Root Mean Square Error. Root Mean Square Error (RMSE) is a standard measure of the error of a model in predicting quantitative data. It is the square root of the average of squared differences between predictions and actual observations. RMSE quantifies the differences between values predicted by a model and values actually observed from the environment. It provides a single measure of the overall prediction error, which makes it an effective tool for comparing different models on a common basis. The lower the RMSE, the better a model is at predicting the data. RMSE is often used in regression analysis, where it measures the difference between observed and predicted values of a dependent variable. It is calculated by taking the square root of the mean squared error, which is the average of the squared differences between the predicted and actual values. RMSE is an important metric to consider when evaluating a model's performance, as it provides a clear indication of how much error is present in the model.

4.4 Modeling

Linear Regression:

Linear regression is a supervised machine learning algorithm used for predicting continuous target variables. It works by predicting the target variable Y based on a linear combination of the independent variable or variables X.

This can be expressed mathematically as

$Y = AX + B$

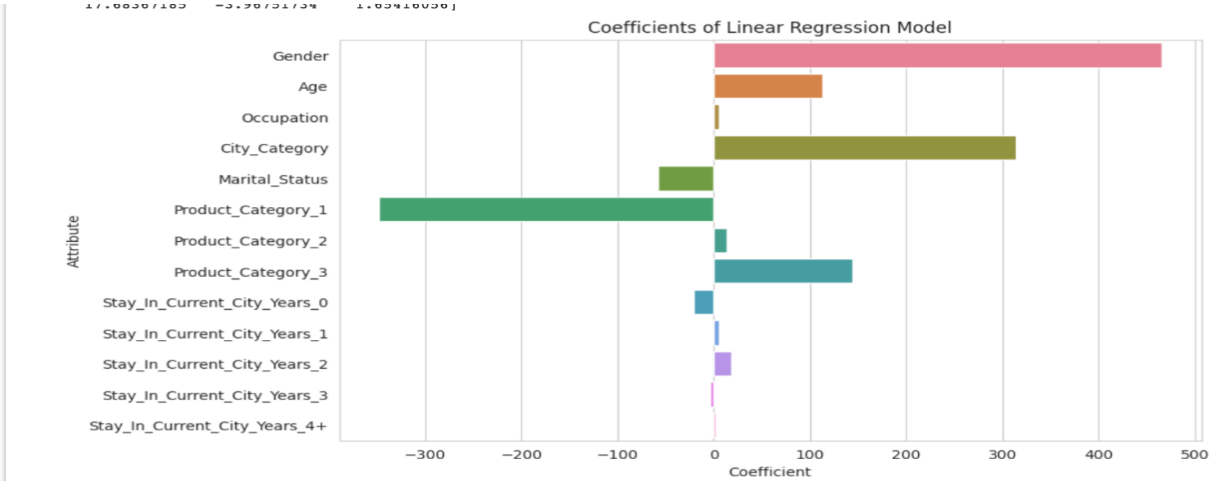
where Y is the target variable, X is the independent variable or variables, B is the intercept, and A is the coefficient of X.

Linear regression is useful for a variety of tasks, including predicting prices of stocks, predicting future sales, predicting the weather, and more. It is also used for classification tasks, where the output is a discrete variable instead of a continuous variable.

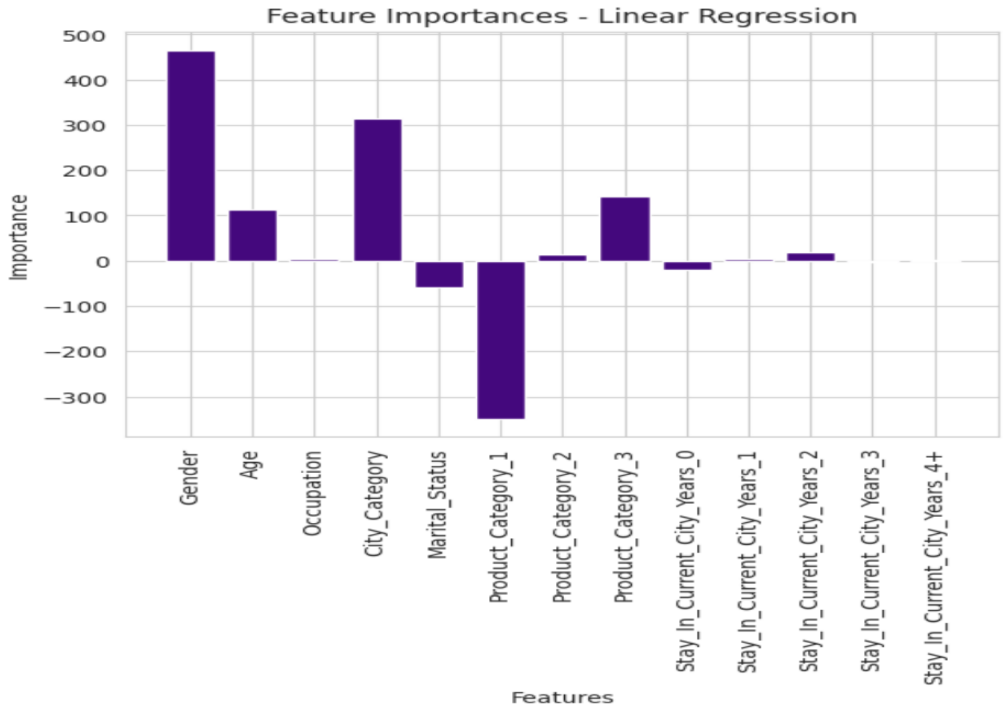
When fitting a linear regression model, the goal is to find the coefficients A and B that minimize the error between the predicted and actual values of Y. This can be done by using optimization algorithms such as the gradient descent algorithm.

Once the model is fit, it can be used to predict the value of Y given new values of X. The accuracy of the predictions depends on the quality of the fit, which is determined by the amount of variance in the data and the size of the dataset.

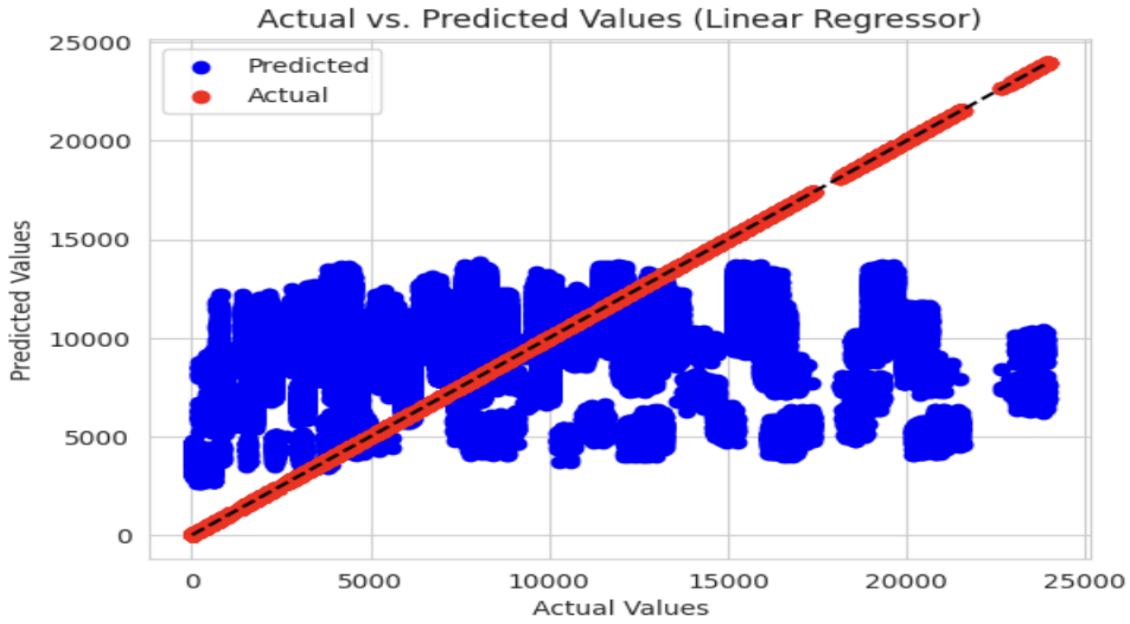
Here is the plot of Coefficients for Linear Regression model.



Here is the plot of Feature Importance in the Linear Regression model.



The below plot represents the Actual vs Predicted values of the Target variable “Purchase” using Linear regression Model



The values we got for the dataset

Mean Absolute Error: 3532.069226165843
Mean Squared Error: 21397853.26940751
R2 Score: 0.15192944521481688
Root Mean Squared Error: 4625.78

Decision Tree Regression

Data analysis is extremely commonly performed using decision trees, which provide both classification and regression models with a tree-like structure. Attribute choice plays an important role in the selection of the root node. The model's underlying process is that it breaks up the main dataset into smaller data sets (subsets). The major terms in a decision tree are

Information Gain:

After a dataset is split by attribute, information gain is measured by the reduction of entropy or uncertainty.

$$Info(D) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Gain ratio:

It is based on the number of branches and instances distributed to these branches that results from a split, which is a form of information gain.

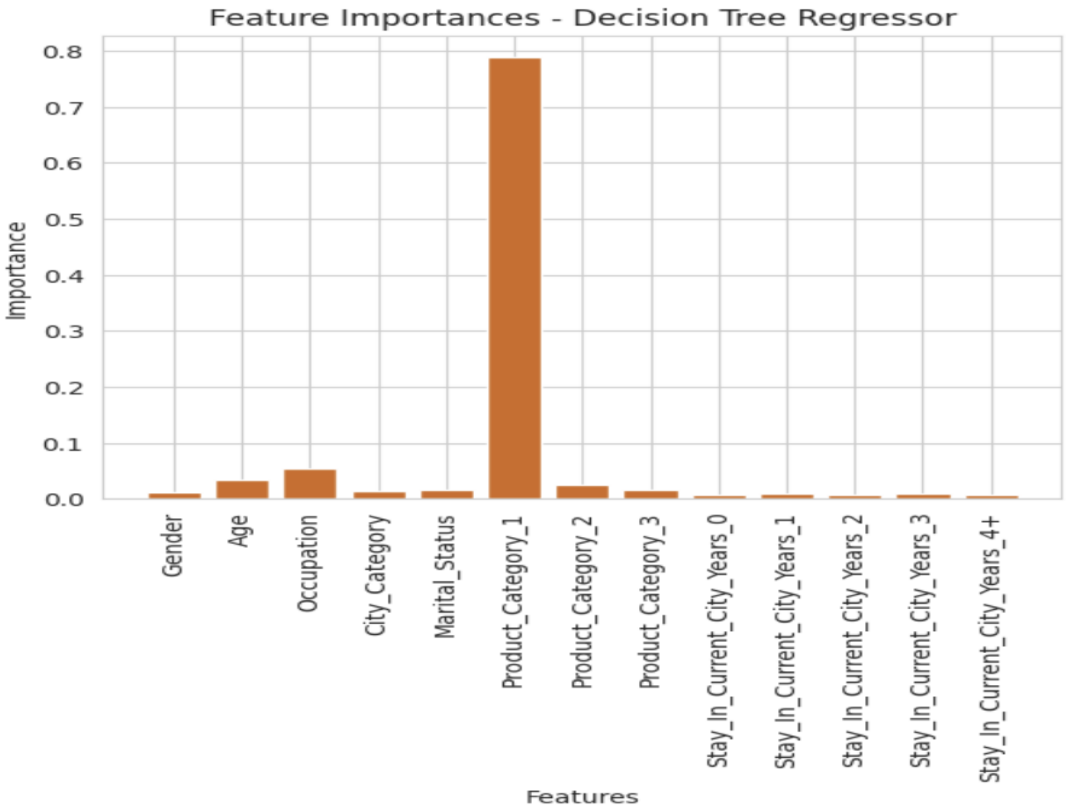
$$\text{Gain Ratio} = \frac{\text{Information Gain}}{\text{Split Info (or) Intrinsic info}}$$

Gini Index:

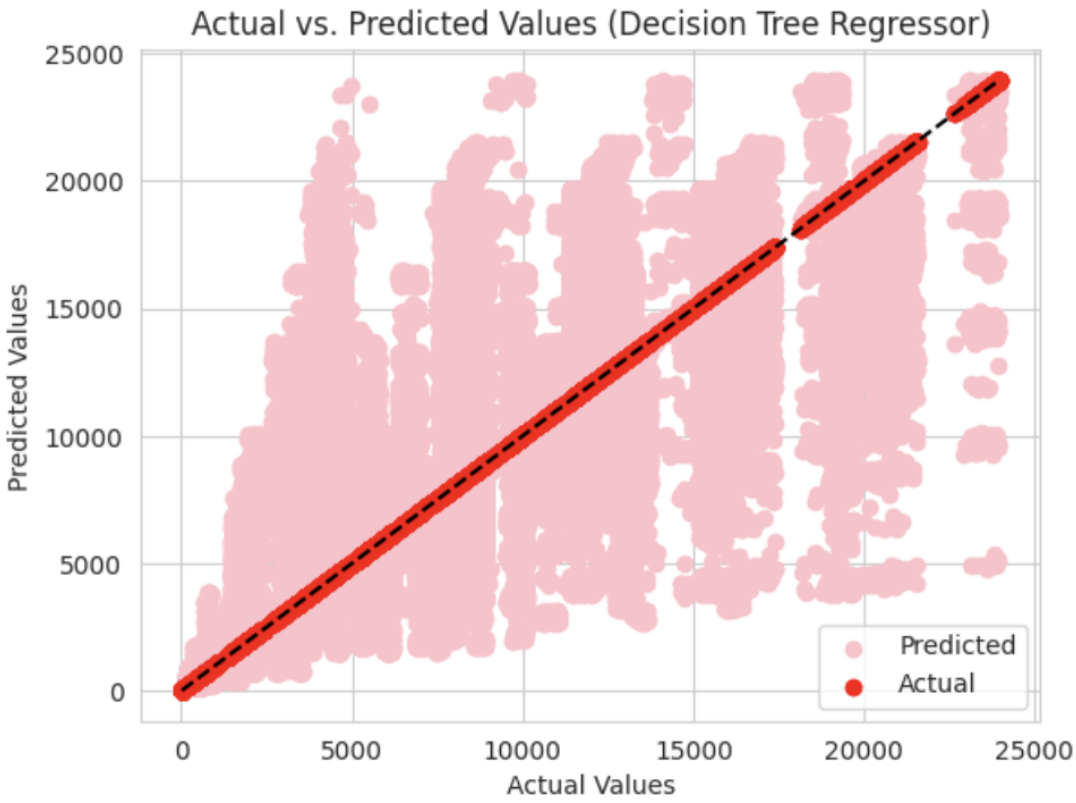
An element of a dataset is assigned a Gini index because of its level of impurity, or its probability of misclassification.

$$Gini = 1 - \sum_{j} p^2 j$$

Here is the plot of Feature Importance in the Decission Tree Regressor.



The below plot represents the Actual vs Predicted values of the Target variable “Purchase” using Linear regression Model



The values we got for the dataset

Mean Absolute Error: 2372.0357559134654

Mean Squared Error: 11300579.466797074

R-squared Score: 0.5521191505924365

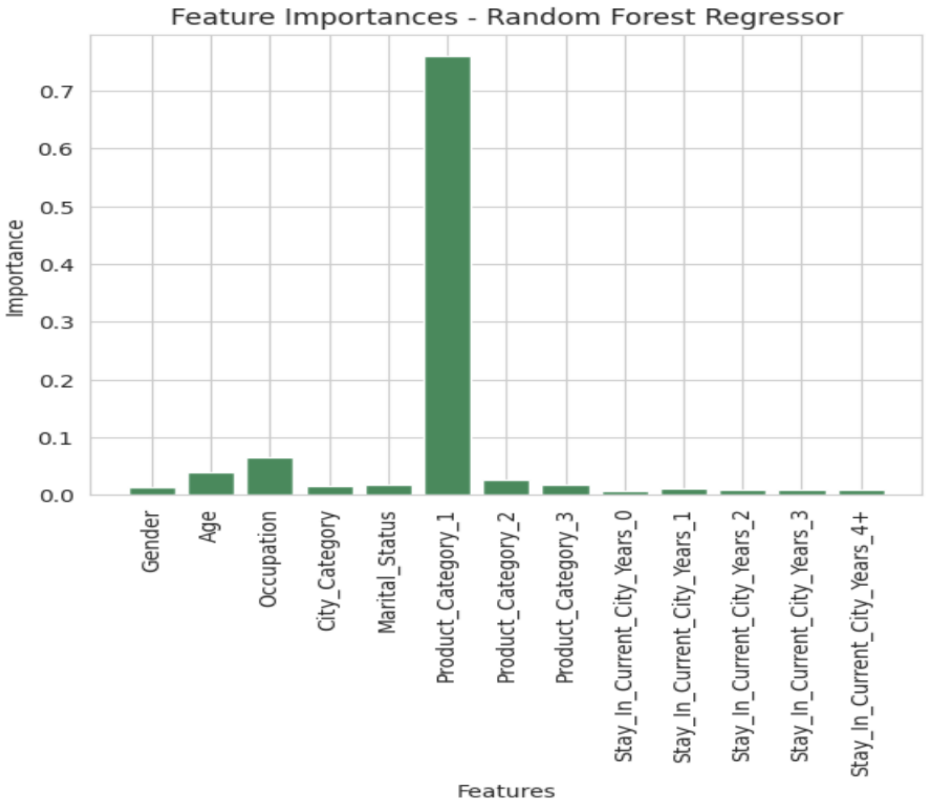
Root Mean Squared Error: 3361.633452177241

Random Forest Regressor

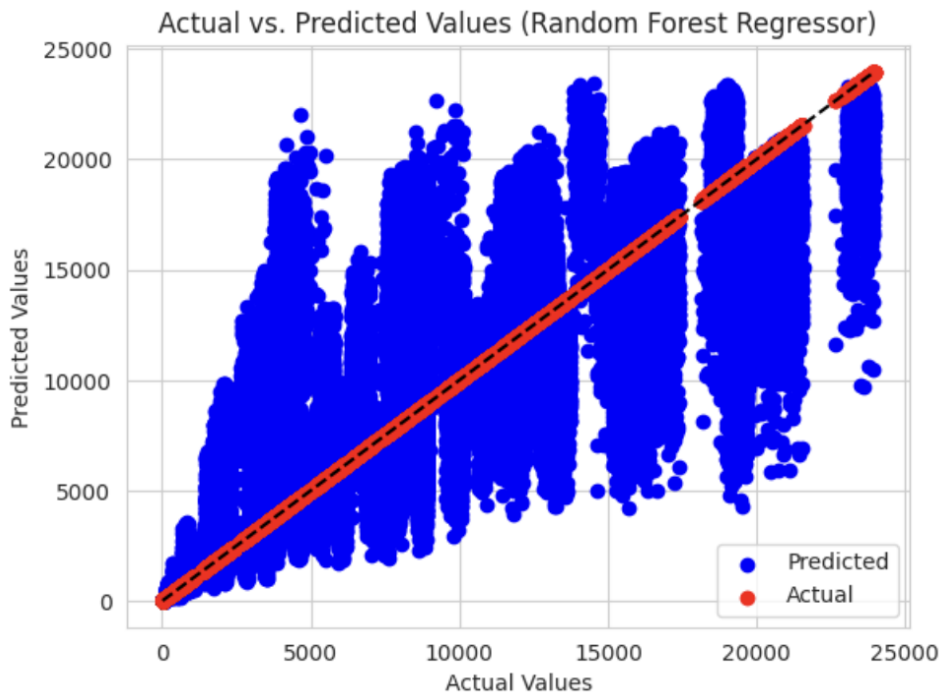
Based on a forest of decision trees, Random Forest is a method for classifying and predicting problems by using a random subset of features and sample data.

The sklearn.ensemble module contains the Random Forest Regressor class, which takes a number of arguments, including n_estimators and random_state (for reproducibility). The trained model is then used to predict target values based on the training data. As a final step, we measure the model's performance based on metrics such as mean absolute error, mean squared error, R2 score, and root mean squared error.

Here is the plot of Feature Importance in the Random Forest Regressor.



The below plot represents the Actual vs Predicted values of the Target variable “Purchase” using Random Forest Regressor



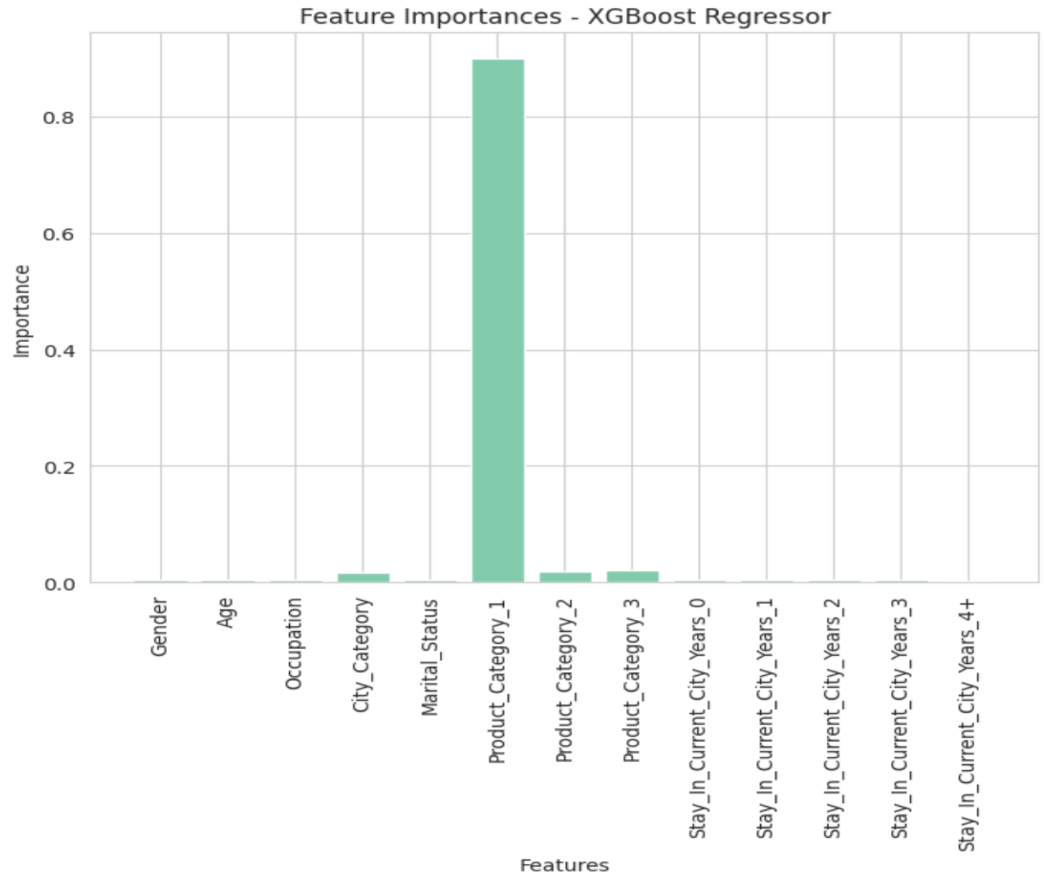
The values we got for the dataset

Mean Absolute Error : 2222.049109204734
Mean Squared Error : 9310769.87311957
R2 Score : 0.6309821516972987
RMSE of Linear Regression Model is 3051.35541573242

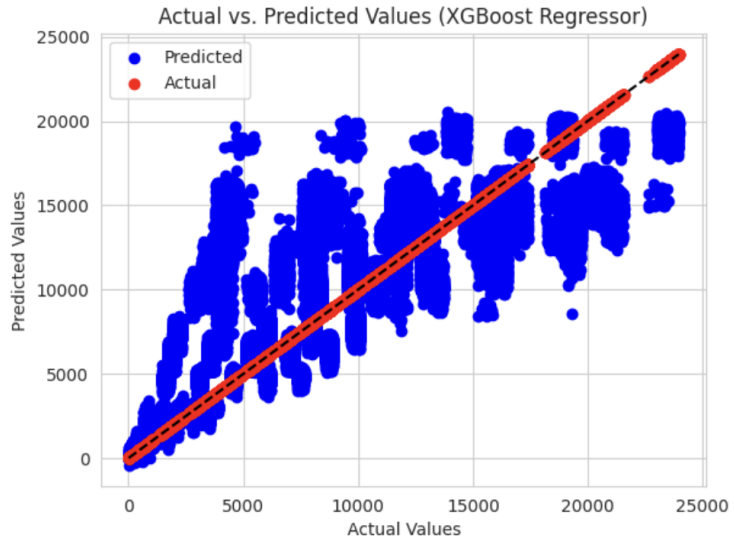
XGB booster Regression

XGBoost regressor (eXtreme Gradient Boosting) is a powerful machine-learning algorithm that can be used to solve both classification and regression problems. This algorithm is based on decision trees and works on large-sized datasets, which makes it a great choice for big data analysis. It is optimized to run efficiently and quickly even on datasets with millions of data points and scales its training method to avoid overfitting. With its powerful model, the XGBoost regressor is capable of producing accurate results with minimal effort and training time. It is a highly efficient algorithm that can be used for a variety of tasks and can be used in combination with other algorithms for even greater accuracy and performance.

Here is the plot of Feature Importance in the XGBooster Regressor.



The below plot represents the Actual vs Predicted values of the Target variable “Purchase” using XGBooster Regressor

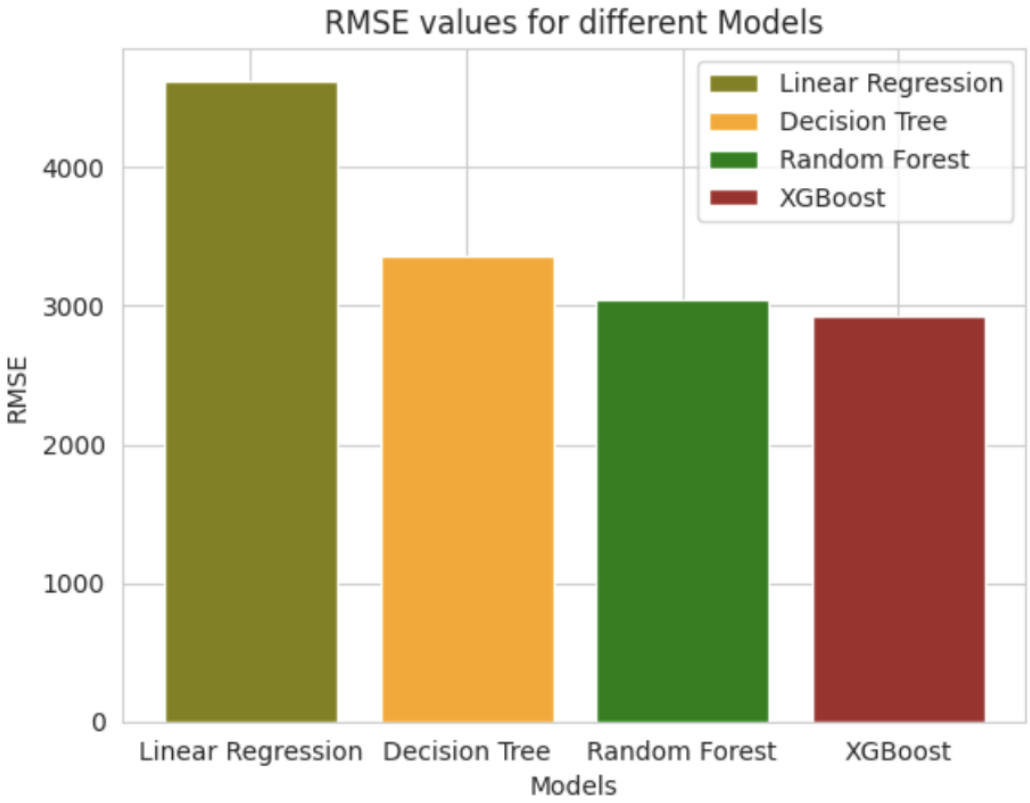


The values we got for the dataset using XGBoost Regression

Mean Squared Error (MSE): 8553036.108839238
Mean Absolute Error (MAE): 2202.242240341369
R-squared (R2): 0.6610137481271811
Root Mean Squared Error (RMSE): 2924.5574210193304

5. Results

ML Model	RMSE Value
Linear Regression	4625.78
Decision Tree Regression	3361.633452177241
Random Forest Regression	3051.35541573242
XGBoost Regression	2924.5574210193304



The RMSE for XGBoost is less and therefore it is efficient compared with others.

6. CONCLUSION

In order to generate income and drive business growth, traditional business techniques have been found insufficient. By integrating machine learning methodologies, businesses can develop effective strategies that take consumer buying patterns into account.

Analyzing sales data for in-demand products based on factors such as previous year's sales is one such strategy.

On the Black Friday sales dataset from Kaggle, we used four regression models to accomplish this – Linear Regression, Decision Tree Regression, Random Forest Regression, and XGBoost Regression. As a standard measure of error, RMSE (root mean squared error) was used to evaluate these models' performance.

Our evaluation result indicates that XGBoost Regressor is the best algorithm for forecasting sales based on the given data. Businesses can use this model to forecast Black Friday customer purchases and tailor discounts based on customer preferences, resulting in increased profits for both the business and the customer.

Contribution to the Society

Retail sales prediction can help the social community in several ways, some of which include:

1. Better resource allocation: With accurate sales predictions, retailers can allocate resources such as inventory, personnel, and marketing efforts more effectively, leading to better overall business performance and increased customer satisfaction.
2. Improved economic growth: Retail sales make up a significant portion of a country's economy, and accurate sales predictions can help to improve economic growth by allowing businesses to make informed decisions about investment and expansion.
3. Job creation: Accurate sales predictions can help retailers to make informed decisions about staffing levels, leading to job creation and improved employment opportunities in the community.
4. Increased tax revenue: Improved retail sales can lead to increased tax revenue for local and national governments, which can then be used to support social services and community initiatives.
5. Better planning for events and promotions: Retailers can use sales predictions to better plan for events and promotions, ensuring that they have adequate supplies and staffing in place to meet demand.

Overall, retail sales predictions can help to improve the economic health of a community by providing businesses with valuable information that can inform their decision-making and lead to increased job opportunities, improved economic growth, and increased tax revenue.

REFERENCES

1. S.Yadavi and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, 2016, pp. 78-83, doi: 10.1109/IACC.2016.25.

2. GeeksforGeeks. Linear Regression Python Implementation. [Online]. Available: https://www.geeksforgeeks.org/linear-regression-python-implementation/. [Accessed: May 1, 2023].

3. N. Aaditi, A. Mitali, G. Suvarna and Prof. M. Amrapal, " Big Mart Sales Prediction Using Machine Learning Techniques," *International Journal of Scientific Research and Engineering Development (IJSRED)* Vol3-Issue 4 | 693-697.

4.GeeksforGeeks. Linear Regression Python Implementation. [Online]. Available: https://www.geeksforgeeks.org/linear-regression-python-implementation/. [Accessed: April 30th, 2023].

5. C. Liew. (2020, January 16). Decision Tree Algorithm, Explained. [Online]. Available: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html. [Accessed: May 1, 2023].

6. N.Garg, (2019, June 24). Random Forest Regression. [Online]. Available: https://towardsdatascience.com/random-forest-regression-5f605132d19d. [Accessed: May 1, 2023].

7.GeeksforGeek. XGBoost for Regression. [Online]. Available: https://www.geeksforgeeks.org/xgboost-for-regression/ [Accessed: May 1, 2023].