# Black Friday Sales Prediction

| | | |
|---|---|---|
| Name | : | Lokesh Nalamasa |
| Date | : | 20/02/2024 |

# Abstract

This study focuses on the development of a predictive model for forecasting Black Friday sales, a crucial event in the retail industry characterized by heightened consumer activity. Leveraging historical sales data, demographic variables, promotional strategies, and potentially external factors such as economic indicators and weather conditions, the aim is to construct a robust predictive tool. By employing advanced data analysis and machine learning techniques, this model seeks to provide accurate insights into anticipated sales figures, thereby enabling retailers to make informed decisions regarding inventory management, pricing strategies, and marketing campaigns. The dataset Black Friday Sales Dataset available on Kaggle has been used for analysis and prediction purposes. The models used for prediction are linear regression, lasso regression, ridge regression, Decision Tree Regressor, and Random Forest Regressor. Mean Squared Error (MSE) is used as a performance evaluation measure. Random Forest Regressor outperforms the other models with the least MSE score.

**Keywords** - Regression, Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, Random Forest Regressor, Mean Squared Error, Data Analysis

# 1.Problem statement

In the realm of retail, Black Friday stands as a pivotal event, heralding a surge in consumer spending and shaping the annual financial landscape for businesses. However, accurately forecasting sales during this frenzied period remains a challenging . The problem at hand is to develop a robust predictive model that can anticipate Black Friday sales with precision. The goal is to create a predictive tool that not only provides insights into anticipated sales figures but also informs strategic decision-making for retailers, enabling them to optimize inventory management, pricing strategies, and marketing efforts to capitalize on the Black Friday frenzy effectively. Through this analysis, the aim is to empower businesses to navigate the Black Friday landscape with foresight and agility, maximizing profitability and customer satisfaction amidst the whirlwind of consumer activity

# 2.Market/Customer

Black Friday, the annual shopping extravaganza following Thanksgiving in the United States, presents a critical juncture for retailers to assess market trends and customer needs. Retailers utilize advanced analytics techniques such as machine learning algorithms and data mining to forecast sales volumes, identify popular products, and optimize pricing strategies. Customer need analysis involves delving into consumer preferences, spending patterns, and demographic shifts to tailor marketing strategies and product offerings. By leveraging data-driven insights, retailers can anticipate market fluctuations, personalize promotions, and optimize inventory management to capitalize on the lucrative Black Friday sales period effectively. Thus, a comprehensive approach to Black Friday sales prediction analysis encompassing market trends and customer needs enables retailers to strategize effectively and drive business growth in the ever-evolving retail landscape

# 3.Target specification and characterization

Our primary aim is to forecast the sales figures for the upcoming Black Friday event accurately. The target specification involves predicting the total revenue generated during the Black Friday sale period across various product categories and regions. This prediction will be based on historical sales data, demographic information, promotional strategies, and other relevant factors. Additionally, the characterization of the target involves understanding the underlying patterns, trends, and drivers influencing Black Friday sales, such as consumer behaviour, economic indicators, seasonal trends, and competitor activities. By comprehensively characterizing the target, we can develop robust predictive models that effectively capture the complex dynamics of Black Friday sales and provide valuable insights for optimizing marketing efforts and inventory management.

# 4.Business Model

**1.Subscription-Based Model**: Offer subscription plans to businesses and retailers, allowing access to the predictive analytics platform for Black Friday sales forecasting. Different subscription tiers can be offered based on the scale of usage, features provided, and the number of users.

**2.Pay-Per-Use Model**: Implement a pay-per-use model where businesses pay based on the volume of data processed or the number of predictions generated. This model offers flexibility for businesses with varying needs and budgets.

**3.Consulting Services**: Offer consulting services to assist businesses in implementing and optimizing the predictive analytics solution for Black Friday sales forecasting. Consulting services can include data analysis, model development, implementation support, and training.

**4.Premium Support**: Offer premium support services, including priority technical support, dedicated account managers, and access to exclusive resources and training materials. Premium support can be provided as an add-on to subscription plans or as a standalone service.

**5.Revenue Sharing**: Explore revenue-sharing partnerships with retailers and e-commerce platforms, where a percentage of the additional revenue generated through improved sales forecasting is shared with the predictive analytics provider.

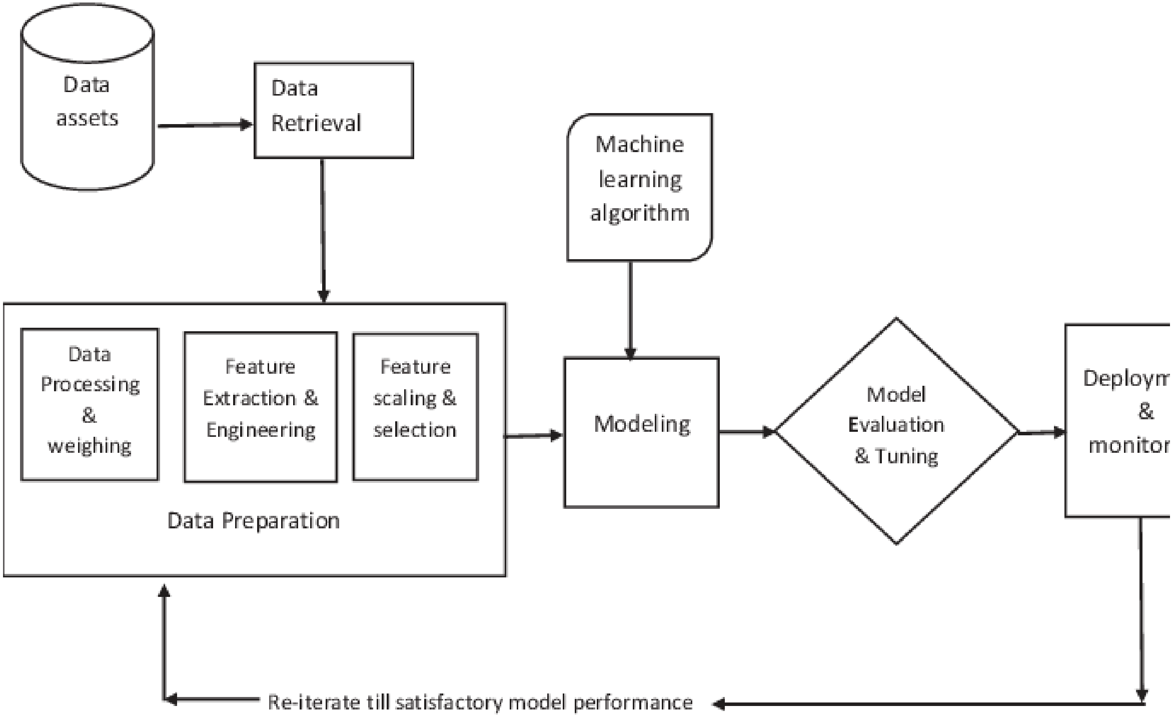
**6.White-Label Solutions**: Offer white-label solutions for businesses that wish to brand the predictive analytics platform as their own. This allows businesses to leverage the technology while maintaining their brand identity and customer relationships.

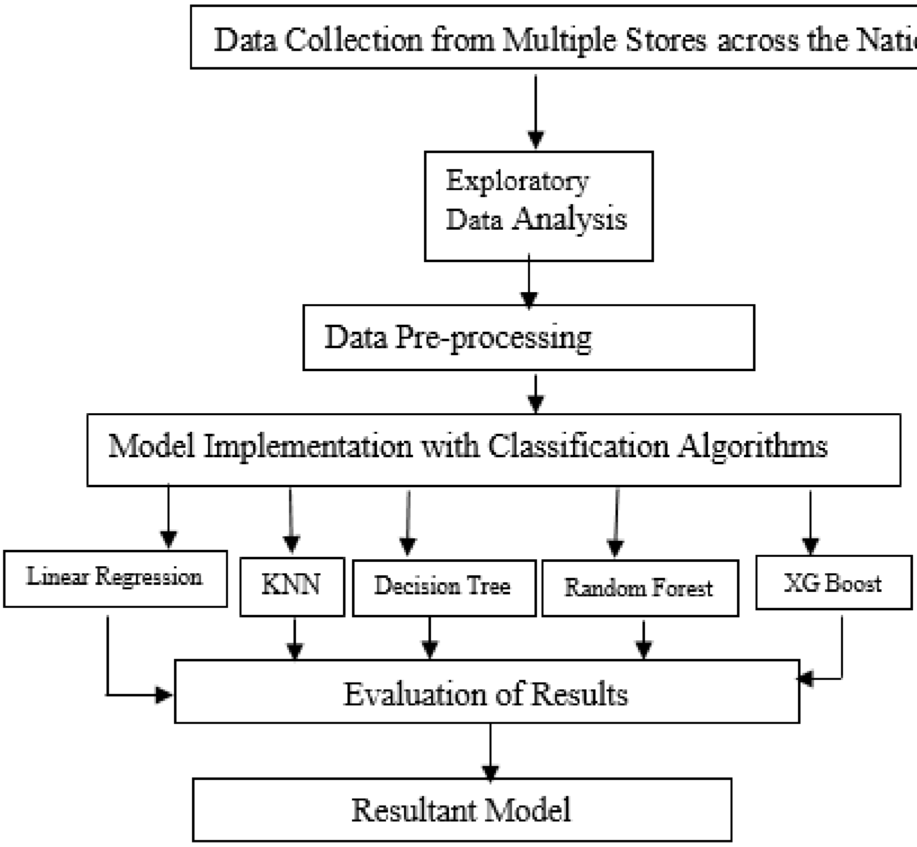
**7.Freemium Model**: Adopt a freemium model where a basic version of the predictive analytics platform is offered for free, with limited features and usage. Premium features and advanced functionality are available through paid subscription plans.

By adopting a combination of these monetization strategies, the Black Friday Sales Prediction Analysis service can generate revenue while providing value to businesses in optimizing their sales strategies and maximizing revenue during the Black Friday sales event.

# 5.System Architecture/Final Product Prototype

Machine learning is defined as the learning of different algorithms that can develop and improve their own performance by means of previous experience & old data. A study of such types of models is called as Machine Learning, where models improve their performance by learning from their previous errors. The steps followed to complete this study make up the system architecture. The steps include collection of Data, Exploratory Data Analysis, data pre-processing, comparison of different techniques of classification and regression, evaluation of results and finally model implementation as shown in figure


Data assets
Data Retrieval
Machine learning algorithm
Data Processing & weighing
Feature Extraction & Engineering
Feature scaling & selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deploym & monitor
Re-iterate till satisfactory model performance


Data Collection from Multiple Stores across the Nation
Exploratory Data Analysis
Data Pre-processing
Model Implementation with Classification Algorithms
Linear Regression
KNN
Decision Tree
Random Forest
XG Boost
Evaluation of Results
Resultant Model

The different models used for analysis and comparison include linear regression, KNN, Decision Tree, Random Forest and XG Boost (extreme gradient boost), the working of each of these models has been explained in this section. Regression and Classification are basic machine learning techniques for Supervised Learning. Both the algorithms, after being trained on the training data set, are used for prediction of the test data set. These algorithms of machine learning work fine with labelled dataset. While classification is used to classify or segregate data sets in one or other category, regression techniques are effectively utilized for future sales prediction. Another difference between regression and classification techniques in machine learning is that regression works very well to predict continuous variables like cost, expenses, salary, age etc., whereas classification reflects good results for discrete variables like whether certain attributes classify the customer fit for loan granting or not, whether it will rain or not.

# DATASET USED

**Dataset:**
**https://www.kaggle.com/code/vishnu691999/black-friday-sales-prediction-analytics**

The Black Friday Discounts dataset is used to train a variety of machine learning models as well as to forecast the amount of people who will make purchases during Black Friday sales. Retailers will be able to study and tailor offers for more customers' favourite products using the purchase prediction provided. The dataset mentioned is the Analytics Vidhya Black Friday

Sales Dataset. They have machine learning models like XGBoost, Linear Regression, MLK classifier, Decision Tree, Decision Tree with bagging, and Deep Learning model using Keras. The models utilized are assessed using the performance evaluation metric Root Mean Squared Error (RMSE). The dataset used is the Black Friday sales dataset, it is a popular open-source dataset that can be found on Kaggle as well. The dataset contains 2 CSV files namely, test.csv and train.csv.

- The training data has 474330 entries and 12 columns.
- The test data has 233599 rows and 11 columns (the purchase column is missing).

Before proceeding with the stated dataset in discrete models of machine learning, an analysis is done for the removal of redundant or trivial data. This is needed to maintain uniformity in the dataset. Besides this removal of redundant or trivial data reduces system complexity.

# DATASET VISUALIZATION

Heatmap is used for determining the correlation between dataset attributes. The data of a given dataset can be easily represented graphically by using a Heatmap. It uses a colour system to represent the correlation among different attributes. It is a data visualization library (Seaborn) element.

Heatmap colour encoded matrix can be described as lower the intensity of the colour of an attribute related to the target variable, higher is the dependency of target and attribute variables.

Based on the Black Friday Sales Dataset the heatmap obtained gives output as Figure 1. The observation based on the heatmap is the attributes age and marital status, product_category_3 and purchase have a correlation

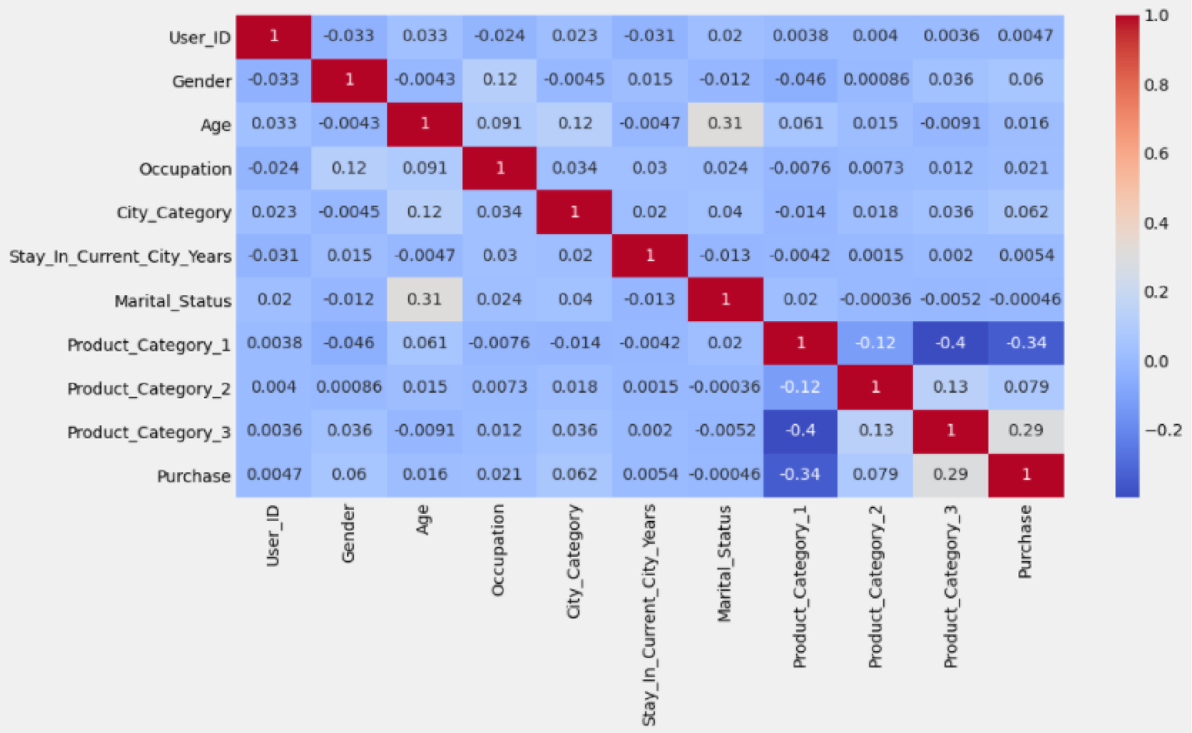


Figure:1

The count plots for different attributes are visualized as different figures given below. The count plot for gender attributes is as Figure 2 Based on the count plot for gender attribute it is observed that feature M (Male) has the maximum count. The count for F features is less

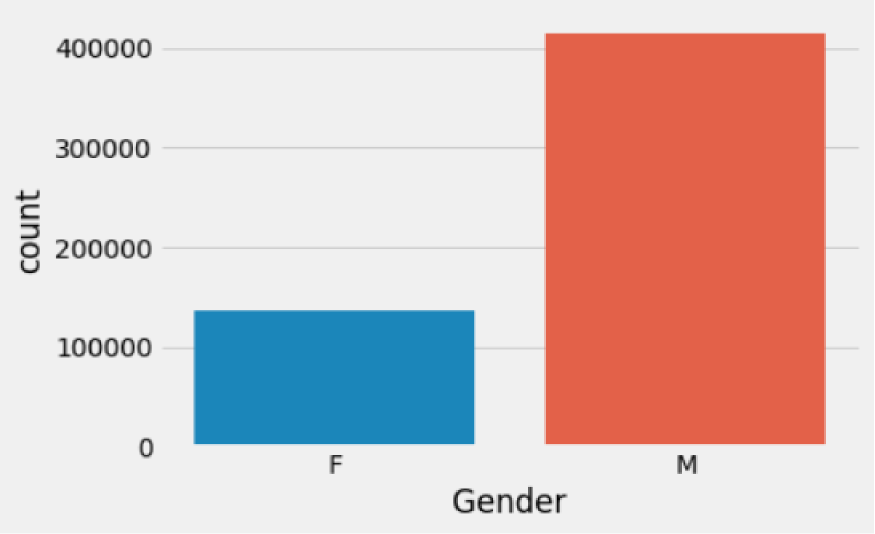


Figure: 2

The count plot for the age attribute is as Figure 3. Based on the count plot the observations noted are the age group 26-35 has a maximum count. The second maximum count observed is for the age group 36-45. The third maximum count observed is for the age group 18-25.

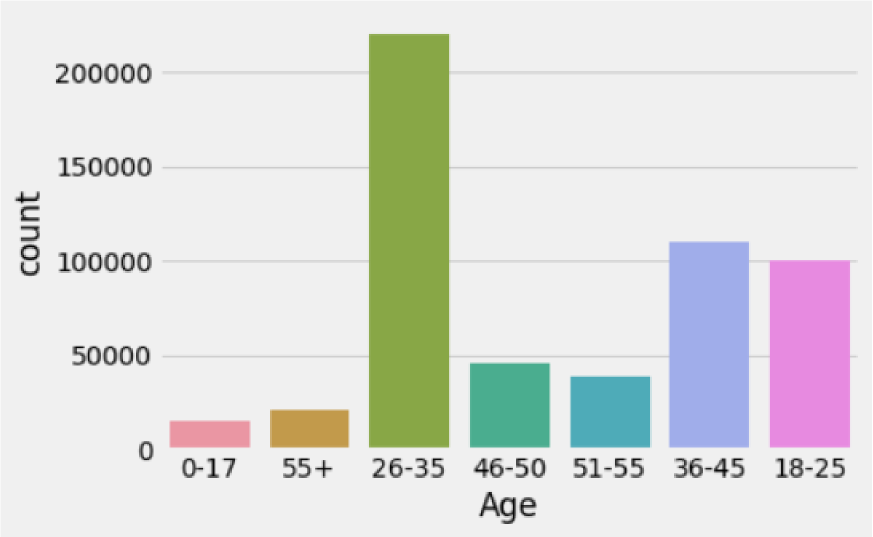


Figure:3

The count plot for the occupation attribute is as Figure 4. The observation based on the count plot is that the masked occupation 4 has maximum count. The second maximum based on the count plot is occupation 0.

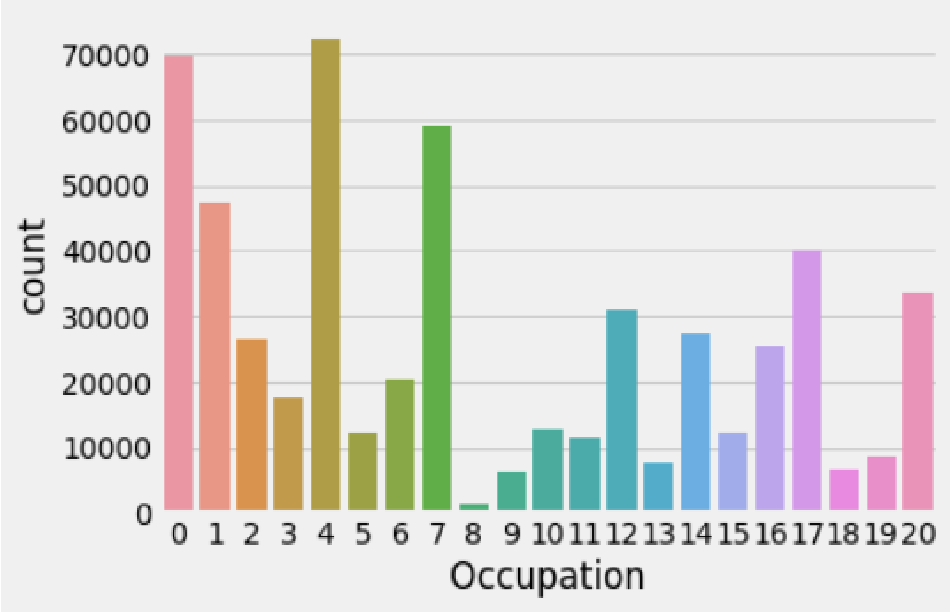


Figure:4

The count plot for city_category is as given in Figure 5. The count plot depicts the maximum count for category B. The second maximum count is for category C. The minimum count is for category A.

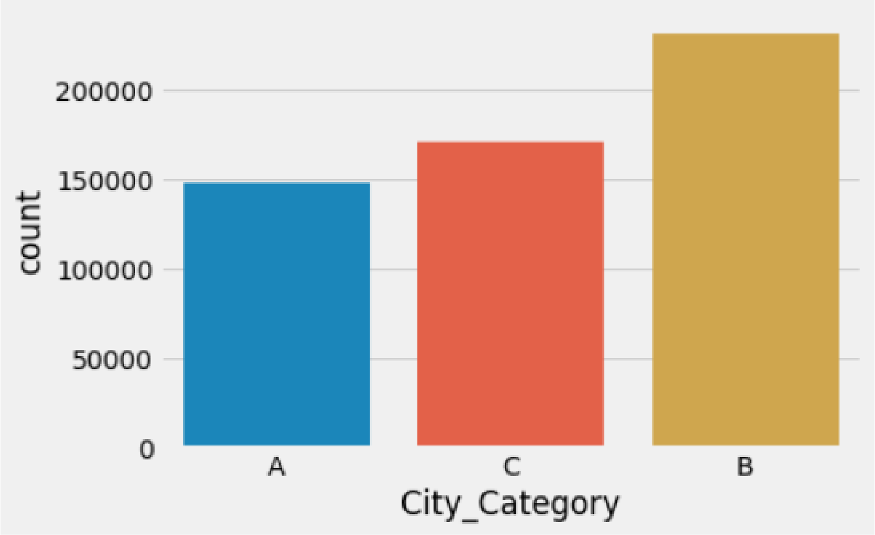
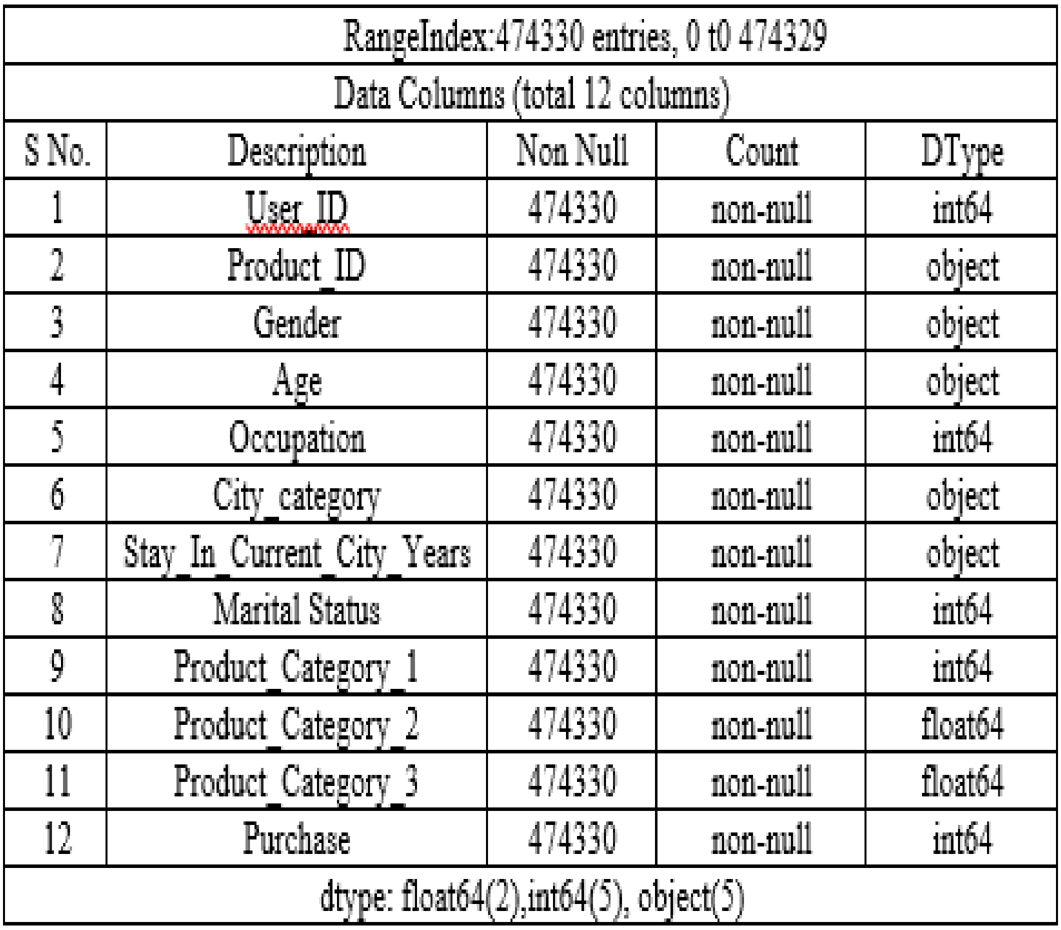


Figure:5

# EXPLORATORY DATA ANALYSIS

For summarizing the primary characteristics of a dataset, for analysing and investigating datasets exploratory data analytics is needed. This is basically done by means of visualization methods. Exploratory data analytics helps to manipulate or make modifications in data sets to get the requisite results. The first step is to go through the data and understand the datatypes (Dtype) in the dataset as shown in Table



| | | | | |
|---|---|---|---|---|
| RangeIndex:474330 entries, 0 t0 474329 | | | | |
| Data Columns (total 12 columns) | | | | |
| S No. | Description | Non Null | Count | DType |
| 1 | User_ID | 474330 | non-null | int64 |
| 2 | Product_ID | 474330 | non-null | object |
| 3 | Gender | 474330 | non-null | object |
| 4 | Age | 474330 | non-null | object |
| 5 | Occupation | 474330 | non-null | int64 |
| 6 | City_category | 474330 | non-null | object |
| 7 | Stay_In_Current_City_Years | 474330 | non-null | object |
| 8 | Marital Status | 474330 | non-null | int64 |
| 9 | Product_Category_1 | 474330 | non-null | int64 |
| 10 | Product_Category_2 | 474330 | non-null | float64 |
| 11 | Product_Category_3 | 474330 | non-null | float64 |
| 12 | Purchase | 474330 | non-null | int64 |
| dtype: float64(2),int64(5), object(5) | | | | |

The above table gives information about the data types of the columns in the training dataset. To further explore the dataset, finding null values and unique values in the data are common approaches. The observations made from these two operations are:

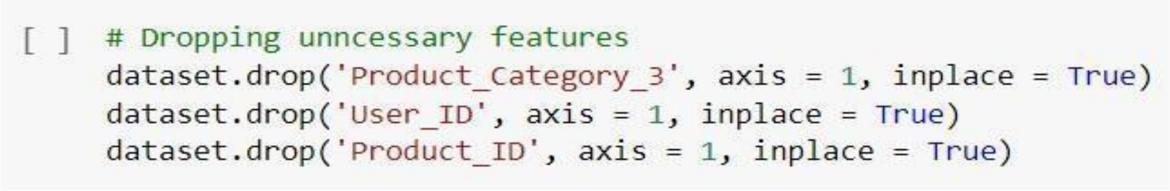
● Product_Category_2 comprises 31.57% null values which can be dropped before moving ahead.

● Product_Category_3 comprises 69.67% null values so this feature can be dropped. .

# DATA PROCESSING

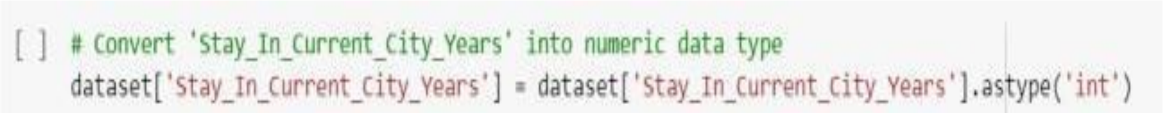
Data processing mainly involved getting the data ready for training. The changes to the data had to be made based on the observations made in the exploratory data analysis, this included, the code snippets for some important steps are added below the point itself:

● Dropping unnecessary features: User_ID and Product_ID, these features are better for database management. Product Category 3 as most of the values are null.



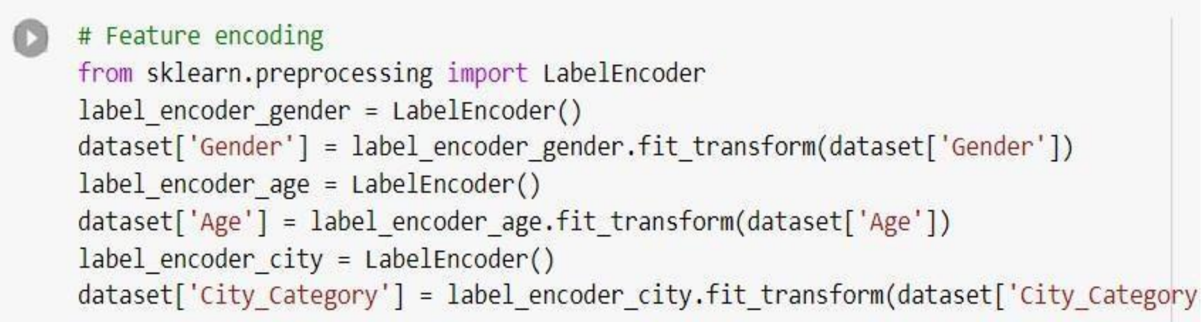
```
[ ]  # Dropping unncessary features
     dataset.drop('Product_Category_3', axis = 1, inplace = True)
     dataset.drop('User_ID', axis = 1, inplace = True)
     dataset.drop('Product_ID', axis = 1, inplace = True)
```

- Fixing null values: based on observations in the exploratory data analysis.
- Replacing values
- Merging train and test data
- Converting ‘Stay in current years’ into numeric data type



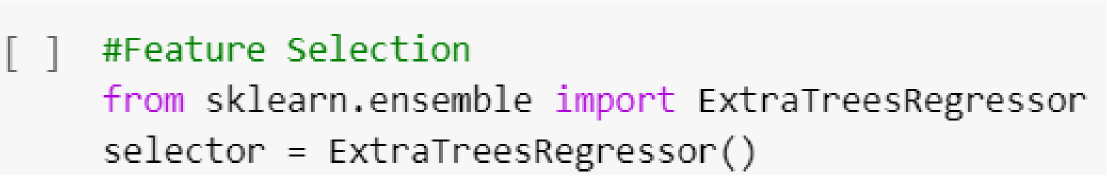
```
# Convert 'Stay_In_Current_City_Years' into numeric data type
dataset['Stay_In_Current_City_Years'] = dataset['Stay_In_Current_City_Years'].astype('int')
```

- Feature encoding: to convert the remaining data types into numerical data type for model training.



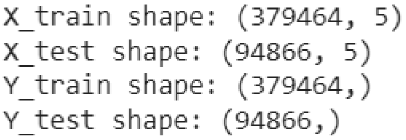
```
# Feature encoding
from sklearn.preprocessing import LabelEncoder
label_encoder_gender = LabelEncoder()
dataset['Gender'] = label_encoder_gender.fit_transform(dataset['Gender'])
label_encoder_age = LabelEncoder()
dataset['Age'] = label_encoder_age.fit_transform(dataset['Age'])
label_encoder_city = LabelEncoder()
dataset['City_Category'] = label_encoder_city.fit_transform(dataset['City_Category
```

- Feature scaling: in order to standardise all independent features in given range.
- Separating the data back in train and test
- Feature selection: ExtraTreeRegressor from the scikit learn library.



```
#Feature Selection
from sklearn.ensemble import ExtraTreesRegressor
selector = ExtraTreesRegressor()
```

- Splitting the data into test and train sets: the shape of each subset can be viewed in figure 13.



```
X_train shape: (379464, 5)
X_test shape: (94866, 5)
Y_train shape: (379464,)
Y_test shape: (94866,)
```

# ALGORITHMS

## I. Linear Regression

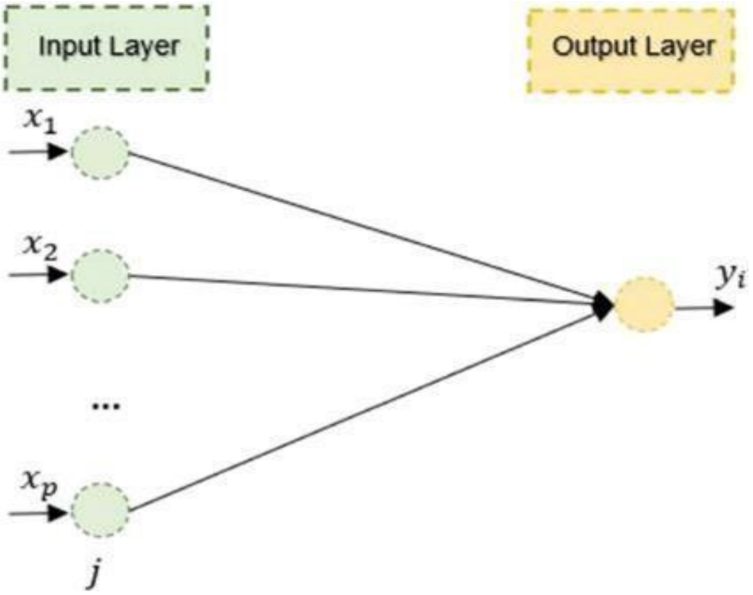


Figure: 6

Figure 6 shows the working of Linear regression. Since regression searches for correlation between dependent and independent attributes, the task of the Regression models is to find the mapping function that maps the input variable(independent) to the continuous output variable(dependent).

**Example**: Suppose we want to predict sales in a particular season, so for this, we can very well utilize the Regression algorithm. Firstly, the model needs to be trained from the past data, and after completion of training the model can be
deployed for future prediction. Regression models are finding wide usage in weather forecast prediction, stock market predictions, game outcome predictions, polls predictions etc. Besides this, many other realistic domains find wide usage of regression models. Variations of regression are logistic regression and root mean square regression which finds major applications.

# II KNN

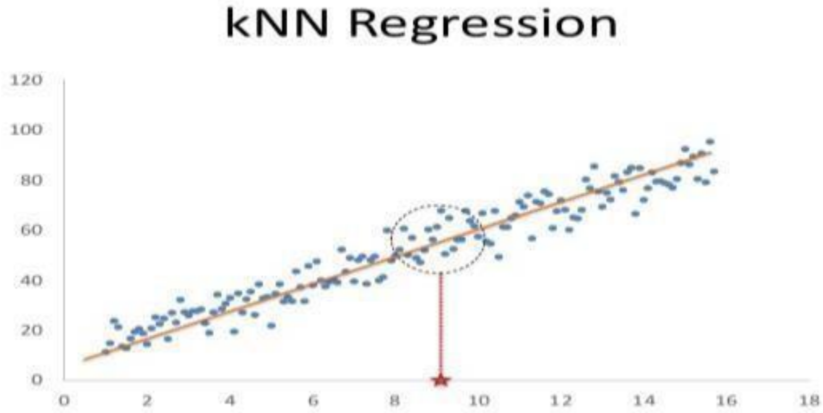


Fig. 7. KNN

Figure 7 shows the plot drawn between dependent and independent variable in case of KNN. Another supervised machine learning model which is widely used for classification and regression is K-Nearest Neighbour. The logic behind classifying data in K NN is the similarity of features in test data with the available data set.

K-NN algorithm is a lazy learner parametric algorithm because its didn't initiate training upon arrival of training data set, instead it stores the data set and when classification of test data is needed, it performs the action of training the data set used for same. It can be used both for classification as well as for regression, but it mostly finds its applications in classification problems.

**Example**: Suppose, we have a picture of a fruit that resembles in features with that of an apple and pear, but we need to recognize the fruit as an apple or pear. One of the algorithms which can be used for this is KNN algorithm, since its working is dependent on similarity measure. KNN model works by correlating features of the new data set to that of apples and pears images and based on the most frequent attributes it will classify the test data in one of the categories

# III. Decision Tree

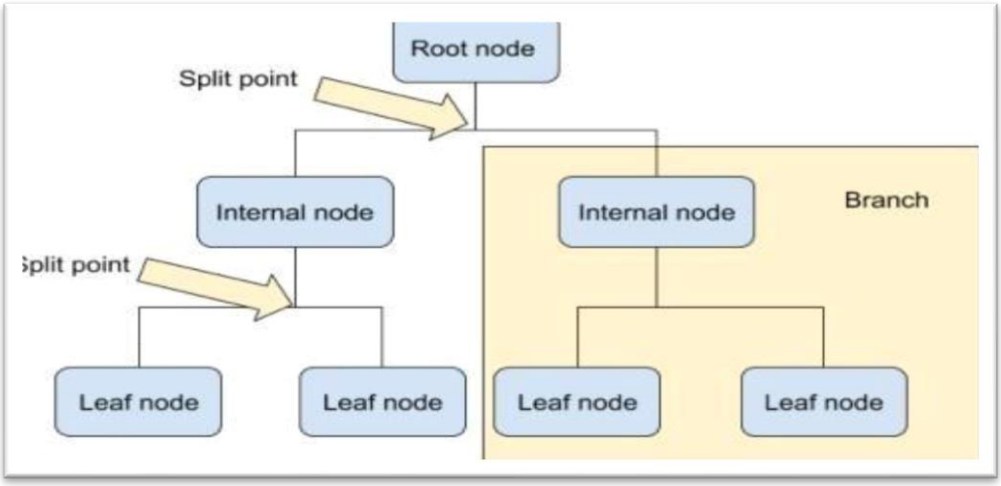


Figure:8

Figure describes how a tree is drawn depending on the results obtained at points of splitting in a node. A decision Tree is again a Supervised machine learning technique that finds its applications for both classification and Regression problems, but generally, it is chosen for solving Classification problems. The structure of the classifier developed in the decision tree is that of a tree as the name suggests. Internal nodes of the trees represent different features in the dataset, decision rules are represented as branches of the decision tree, and the resulting outcome of the rules are represented as leaf nodes in the tree. Decision nodes represented as branches of the tree can take multiple values but the outcome or result of the decision represented by Leaf didn't further continue with any more branches. The decision for moving to a particular branch or Leaf is dependent on features present in the dataset.

It uses two methodologies Iterative Dichotomise (ID3) the CART algorithm, which stands for Classification and Regression Tree algorithm.ID3 basically calculates information gain of features for decision while CART calculates Gini Index of features present in data set for decision. The answer to the question being asked while constructing a decision tree results in either yes or No. The decision whether to split the tree further or not is based on the binary answer received.

# IV. Random Forest

Random Forest is another famous widely used supervised machine learning algorithm. It finds its usage for both Classification and Regression techniques in ML. Ensemble learning is the principle upon which Random Forest works. Ensemble learning solves complex classification problems by mixing multiple classifiers. This further improves the efficiency of the overall model built. Below Figure reflect that Random forest is combination of multiple decision trees.

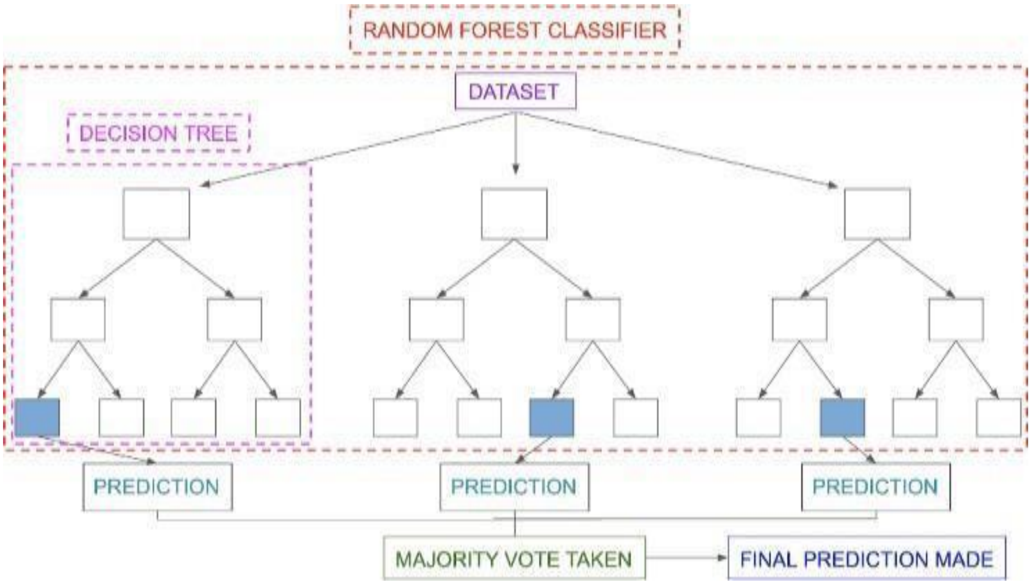


Figure :9

Random Forest is a unique type of classifier that consists of a combination of multiple decision trees designed on discrete subsets of the original dataset. To improve predictive accuracy of the test dataset, it takes help of the average of results of each decision tree

# V. XG Boost

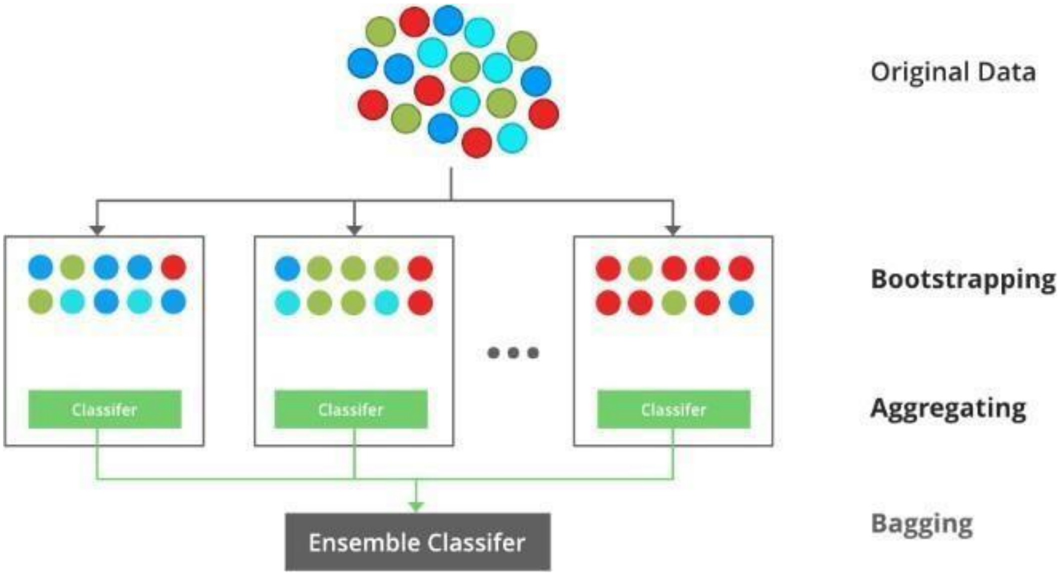
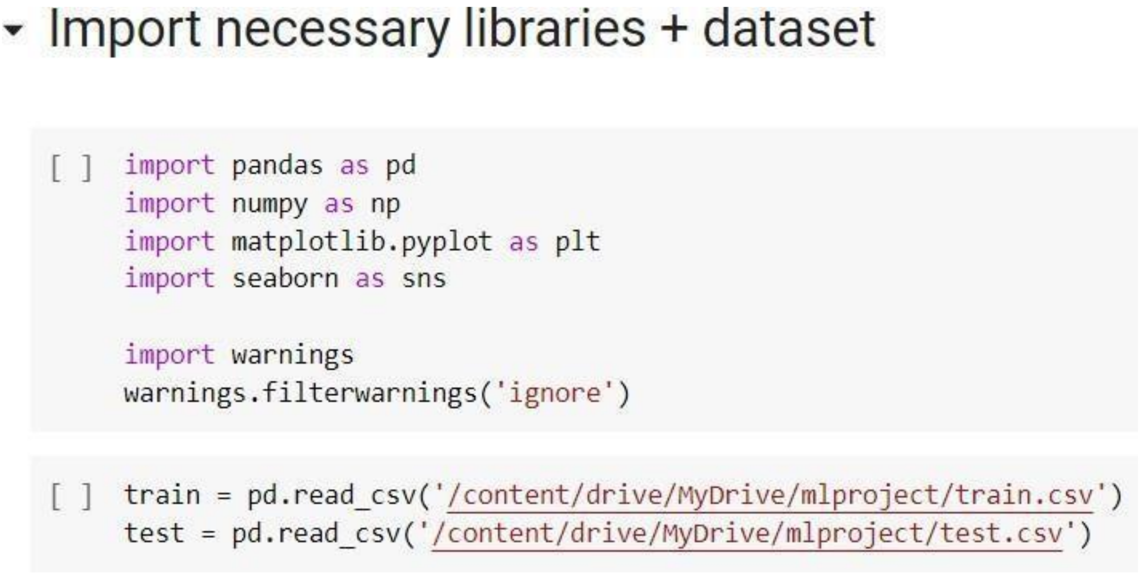


Figure:10

Above Figure shows XG Boost as a combination of Bagging, Aggregating and Bootstrapping. After having discussed different machine learning models and algorithms that we will be utilizing in our data set. A comparative study of the results obtained by these models will be done before selecting the final model for the dataset used in this work. Among all the models discussed, the supervised machine learning whose performance is at par with its peers will be selected for final design.

# MODEL IMPLEMENTATION

One of the most important aspects of model implementations is figuring out its hardware specifications and the environment for training. The models trained as a part of this implementation were trained on an intel i7 10th gen laptop with 16 GB RAM. The environment chosen for training is the Google Colab, the freely available platform for machine learning and data science. The GPU hardware accelerator was picked for better training, the Tesla K80 with 12 GB Ram.



Import necessary libraries + dataset

```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns

    import warnings
    warnings.filterwarnings('ignore')
```

```
[ ] train = pd.read_csv('/content/drive/MyDrive/mlproject/train.csv')
    test = pd.read_csv('/content/drive/MyDrive/mlproject/test.csv')
```

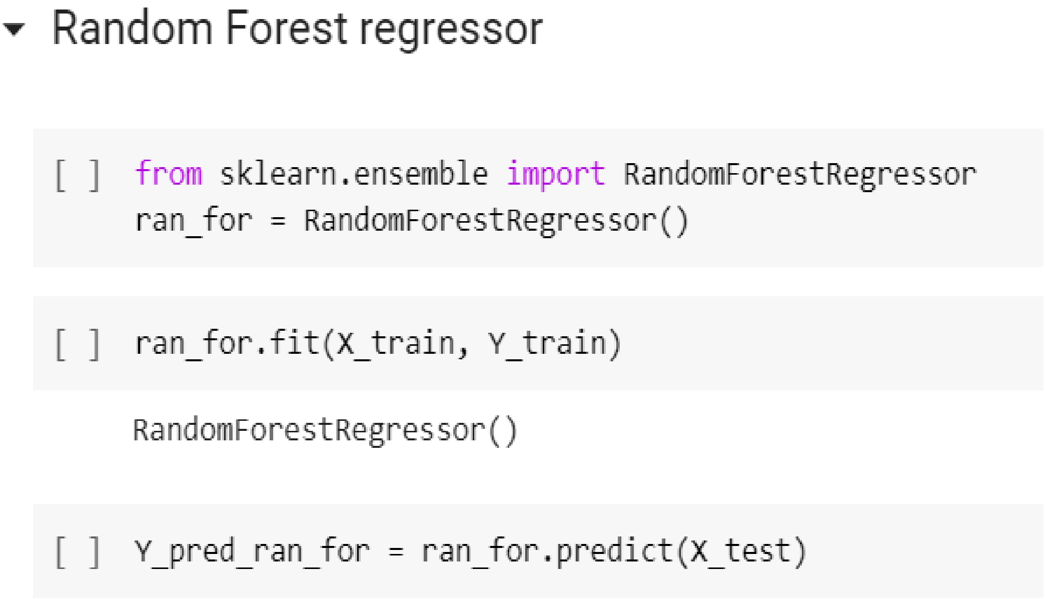
The models implemented are:

- Linear regression
- KNN
- Decision tree
- Random forest
- XGboost

The procedure for model training is using the same protocol and syntax.

- First, import the needed model
- Fit the model on X_train and Y_train dataset
- Use the predict function on X_test and save the result in Y_pred

Figure 16 correctly describes the process used for model implementation.



Random Forest regressor

```
[ ]  from sklearn.ensemble import RandomForestRegressor
     ran_for = RandomForestRegressor()
```

```
[ ]  ran_for.fit(X_train, Y_train)

     RandomForestRegressor()
```

```
[ ]  Y_pred_ran_for = ran_for.predict(X_test)
```

The metrics used for evaluating the model performance are RMSE (root mean squared error) and *R2* score.

## A. Root mean square error:

The standard deviation of the residuals is defined as the Root Mean Square Error (RMSE). The residuals measure the distance between the data points and the regression line, and the RMSE measures the spread of these residuals

## B. R2 Score

The r2 score is between 0% and 100%. It is closely connected to the mean square error even if they are not the same. The proportion of the dependent variable's variation that can be predicted using the independent variable is known as r2.

If the correlation coefficient is 100%, the two variables are perfectly associated, that is, there is absolutely no variation. Though not necessarily, a low number would suggest a poor correlation, which would suggest that the regression model is flawed.

Table3. Tabular Representation of Regression Model's Performance **Results**

| ***Model*** | ***RMSE*** | ***R2*** |
|---|---|---|
| Linear regression | 4722.69 | 0.1029 |
| KNN regression | 3295.77 | 0.5631 |
| Decision tree regression | 3085.12 | 0.6171 |
| Random forest regression | 3047.12 | 0.6265 |
| XGB regression | 3023.49 | 0.9323 |

Table shows a comparative study of different regression techniques based on RMSE and R2 results obtained. Root mean square error RMSE, denotes the error. This should be as low as possible for any machine learning models. When different models are compared, the model which is having lesser RMSE is selected for deployment, since the lesser the error, the greater the model efficacy. Another parameter considered while evaluating model performance is R2 Score, which denotes the correlation between variables in the model. It lies between 0 and 1. The greater its value, the higher correlation between different attributes used while designing model. So, a value near 1 is the desirable R2 value in a good machine learning model.

# Conclusion and Future Work

After analysing the scores obtained by different classification algorithms of machine learning based on RMSE and R2 it is found that the machine learning model best suitable for dataset used in this study is XGB regression XGB regression which is an ensemble model in machine learning and theoretically outperforms the other models. This is verified in the obtained results as well.

The work can be carried forward as a customized business solution in wise numeric values can be turned to view the form of the predicted price. We have considered only 3 product applications in the study

results obtained can be further tested for more no of products. Also, an analysis of customer behaviour or shopping pattern can be done based on time of the year and currents needs of the customer as well. In machine learning, more models can be trained and visualized using different graphs. Other metrics can be used for comparison of performance amongst models and specific metrics for models (for e.g., information gain for decision tree) can be used to evaluate individual performance of machine learning models