REPORT FOR DATA ANALYSIS ON BLACK FRIDAY SALES

ABSTRACT

Black Friday marks the beginning of the Christmas shopping festival across the US. On Black Friday big shopping giants like Amazon, Flipkart, etc. lure customers by offering discounts and deals on different product categories. The product categories range from electronic items, Clothing, kitchen appliances, Décor.

Research has been carried out to predict sales by various researchers. The analysis of this data serves as a basis to provide discounts on various product items. With the purpose of analysing and predicting the sales, we have used three models.

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.

CONTENT

FIGURES

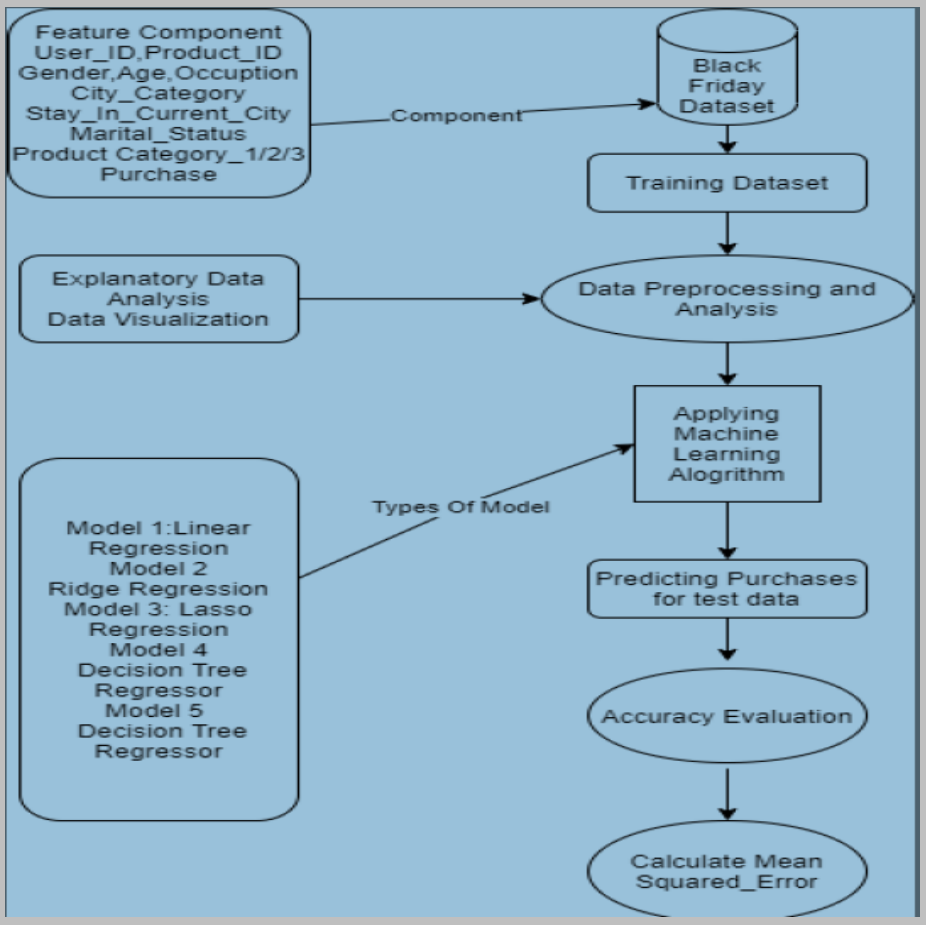

Fig-1: Flowchart of Proposed System

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-2. The observation based on the heatmap is the attributes age and marital_status, product_category_3 and purchase have a correlation.

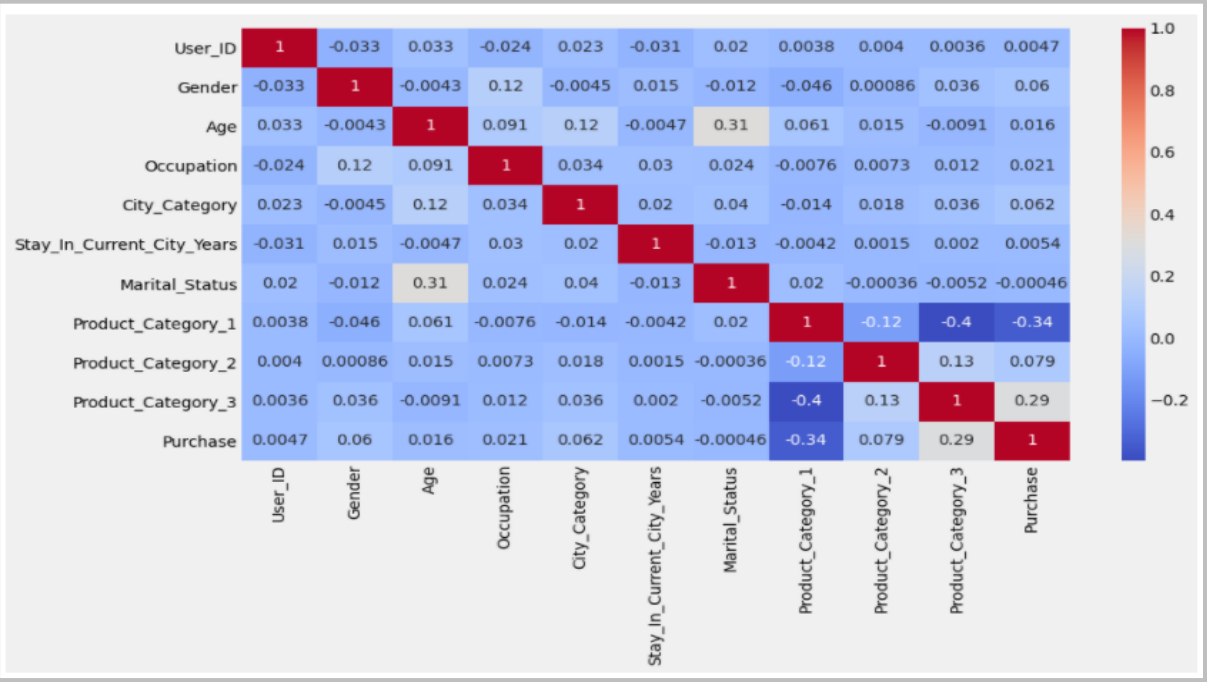

Fig-2: Heatmap for correlation between attributes

The count plots for different attributes are visualized as different figures given below. The count plot for gender attributes is as Figure 3. 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.

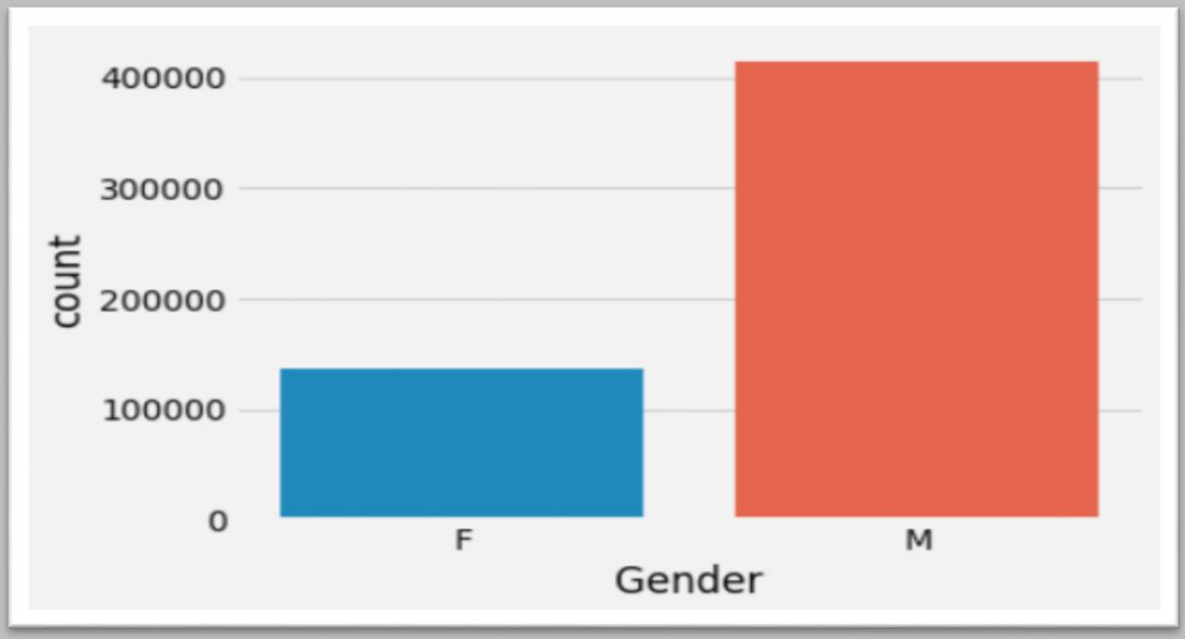

Fig-3: Count Plot for Gender

The count plot for the age attribute is as Figure 4. 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.

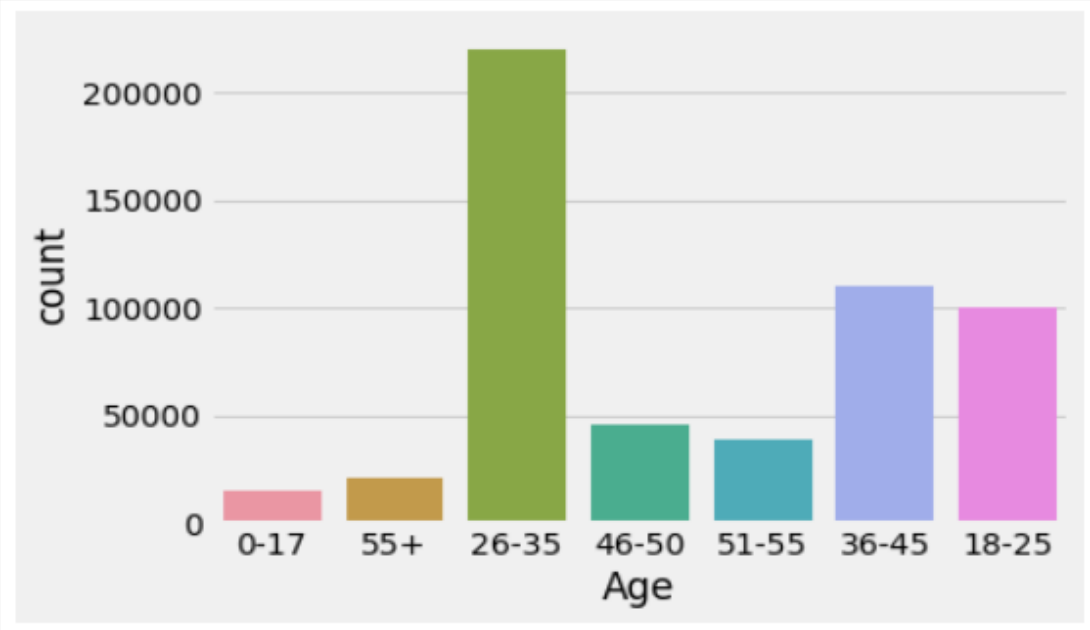

Fig-4: Count Plot for Age

The count plot for the occupation attribute is as Figure-5. 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.

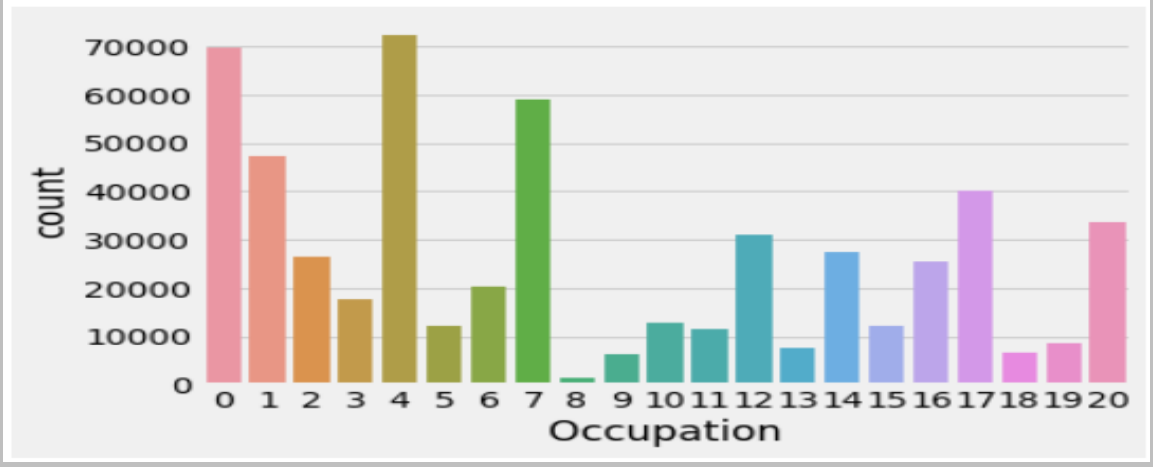

Fig-5: Count Plot for Occupation

The count plot for city_category is as given in Figure-6. 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.

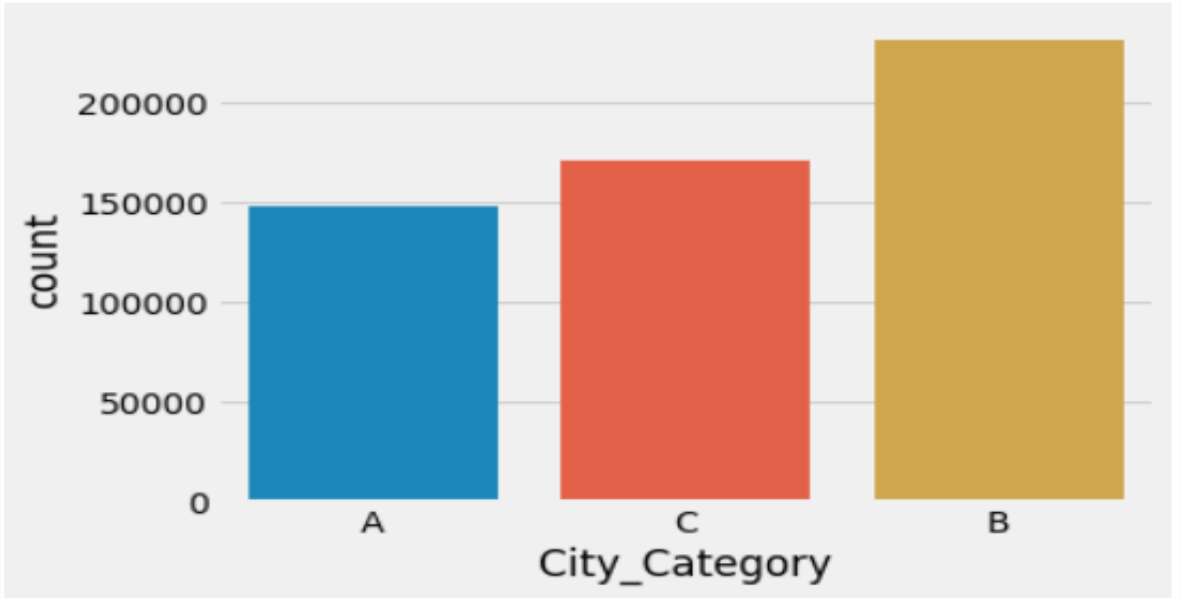

Fig-6: Count Plot for City_Category

The count plot for Stay_In_Current_City is as given in Figure-7. The observations based on the count plot can be stated as the maximum count is for 1 year. The minimum count is for 0 years.

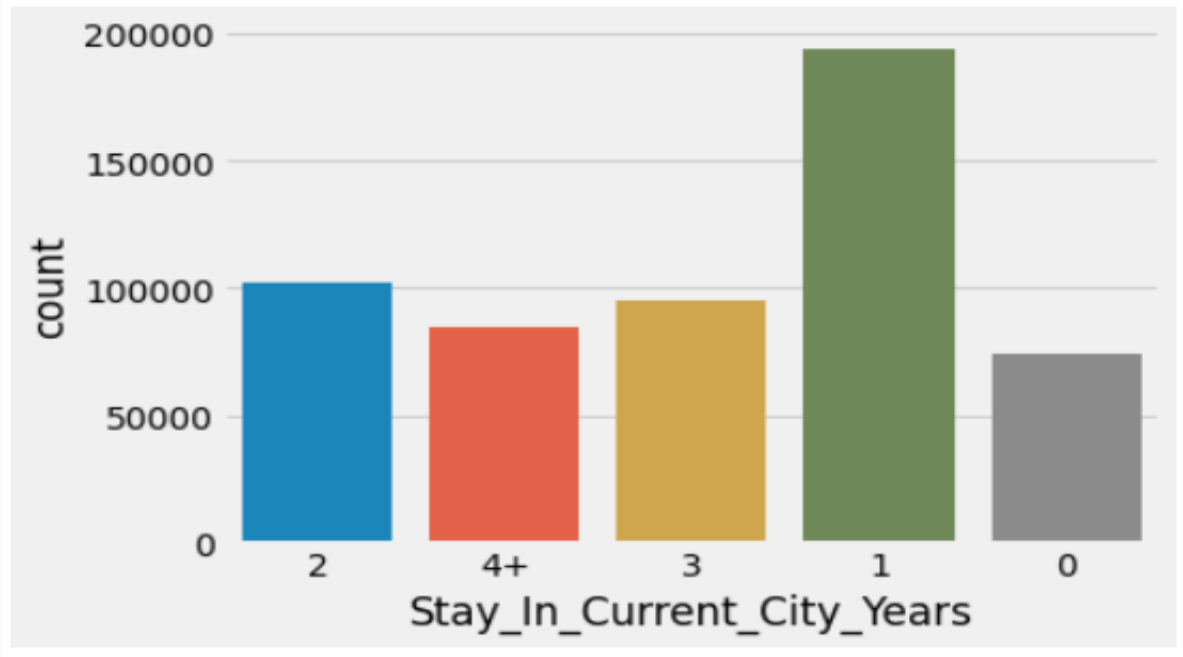

Fig-7: Count Plot for Stay_In_Current_City_Years

TABLES

TABLE-1: DATASET DEFINITION

SR NO	VARIABLE	DEFINITION	MASKED
1	USER_ID	UNIQUE ID OF CUSTOMER	FALSE
2	PRODUCT_ID	UNIQUE PRODUCT ID	FALSE
3	GENDER	SEX OF CUSTOMER	FALSE
4	AGE	CUSTOMER AGE	FALSE
5	OCCUPATION	OCCUPATION OF CUSTOMER	TRUE
6	CITY_CATEGORY	CITY CATEGORY OF CUSTOMER	TRUE
7	STAY_IN_CURRENT_CITY	NUMBER OF YEARS CUSTOMER STAYS IN CITY	FALSE
8	MARITIAL_STATUS	CUSTOMER MARITAL STATUS	FALSE
9	PRODUCT_CATEGORY_1	PRODUCT CATEGORY	TRUE
10	PRODUCT_CATEGORY_2	PRODUCT CATEGORY	TRUE
11	PRODUCT_CATEGORY_3	PRODUCT CATEGORY	TRUE
12	PURCHASE	AMOUNT OF CUSTOMER PURCHASE	FALSE

TABLE-2: COMPARATIVE ANALYSIS

Model	MSE
Linear Regression	4638.09
Ridge Regression	4846.28
Lasso Regression	4638.09
Decision Tree Regressor	3788.33
Random Forest Regressor	**2748.16**

INTRODUCTION

The shopping sector has greatly evolved due to the Internet revolution. Most of the population takes into consideration online shopping more than the traditional method of shopping. The biggest perks of online shopping are convenience, better prices, more variety, easy price comparisons, no crowds, etc. The pandemic has boosted online shopping. Though online shopping keeps growing every year, the total sales for the year 2021 are expected to be much higher.

Black Friday originated in the USA and is also referred to as Thanksgiving Day. This sale is celebrated on the fourth Thursday of November once every year. This day is marked as the busiest day in terms of shopping. The purpose of organizing this sale is to promote customers to buy more products online to boost the online shopping sector.

The prediction model built will help to analyze the relationship among various attributes. Black Friday Sales Dataset is used for training and prediction. Black Friday Sales Dataset is the online biggest dataset and the dataset is also accepted by various e-commerce websites.

The prediction model built will provide a prediction based on the age of the customer, city category, occupation, etc. The prediction model is implemented based on models like linear regression, ridge regression, lasso regression, Decision Tree Regressor, Random Forest Regressor.

The paper further walks through various sections. It gives an introduction to the problem, illustrates the prior research done in this field, provides the data set description and presents the proposed model, with the conclusion in the last section.

LITERATURE SURVEY

Ample research is carried out on the analysis and prediction of sales using various techniques. There are many methods proposed to do so by various researchers. In this section, we will summarize a few of the machine learning approaches.

First I have proposed a prediction model to analyze the customer's past spending and predict the future spending of the customer. The dataset referred is Black Friday Sales Dataset from Emerging India Analytics. They have machine learning models such as Linear Regression, MLK classifier, Deep learning model using Keras, Decision Tree, and Decision Tree with bagging, and XGBoost. The performance evaluation measure Root Mean Squared Error (RMSE) is used to evaluate the models used. Simple problems like regression can be solved by the use of simple models like linear regression instead of complex neural network models.

Then, I have proposed a sales forecasting model. The machine learning models used for implementation are K-Nearest Neighbour, Random Forest, and Gradient Boosting. The dataset used for the experimentation is provided by Data Science Nigeria, as a part of competitions based on Machine Learning. The performance evaluation measures used are Mean Absolute Error (MAE). Random Forest outperformed the other algorithms with a MAE rate of 0.409178.

Then, I have analysed and visually represented the sales data provided in the complex dataset from which we ample clarity about how it works, which helps the investors and owners of an organization to analyze and visualize the sales data, which will outcome in the form of a proper decision and generate revenue. The data visualization is based on different

parameters and dimensions. The result of which will enable the end-user to make better decisions, ability to predict future sales, increase the production dependencies on the demand, and also regional sales can be calculated.

Then,I have analysed and compared the performance of K-Fold cross-validation and hold-out validation method. The result of the experimentations where k-fold cross-validation gives more accurate results. The accuracy results of K - Fold cross-validation were around 0.1 - 3% more accurate as compared to hold-out validation for the same set of algorithms.

Then, I have performed sales prediction based on a dataset collected from a grocery store. The algorithms used for experimentations are Linear Regression, K-Nearest Neighbours algorithm, XGBoost, and Random Forest. The result precision is based on Root Mean Squared Error (RMSE), Variance Score, Training, and Testing Accuracies. The Random Forest algorithm outperforms the other three algorithms with an accuracy of 93.53%.

Then, I have applied machine learning algorithms to predict sales. The dataset for the experimentation purpose is taken, named as Black Friday Sales Dataset. The algorithms used for the implementation of the system are linear regression, Ridge Regression, XGBoost, Decision Tree, Random Forest, and Rule-Based Decision Tree. Root Mean Squared Error is used as the performance evaluation measure. As per RMSE lower the RMSE value better the prediction. As a result, based on the RMSE rate Rule-Based DT outperforms other machine learning techniques with a RMSE rate of 2291.

Then, I have applied machine learning algorithm in tracking sales at places like shopping center big mart to

anticipate the demand of customers and handle the management of inventory accordingly the methods presented here are an effective method for data shaping and decision-making. New ways that can better identify consumer needs and calculate marketing plans which will improve sales.

Lastly I have analyzed, preprocessed, and applied machine learning techniques to predict sales. The dataset used for the analysis and experimentation purpose is Black Friday Sales Dataset. The dataset is preprocessed. K - Fold method is used for the purpose of splitting the dataset into training and testing datasets. The prediction model is implemented using Linear Regression, Decision Tree, Random Forest, Gradient Boost, and XGBoost. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used as the accuracy evaluation measures. As a result of experimentation, the Random Forest performed significantly with an accuracy of 77%, with an RMSE value of 2730 and MAE value of 2349.

DATA DESCRIPTION

The study uses Black Friday Sales Dataset publicly available on Kaggle. The dataset consists of sales transaction data. The dataset consists of 5, 50,069 rows.

The dataset consists of attributes such as user_id, product_id, martial_status, city_category, occupation, etc. The dataset definition is mentioned in Table-1.

The Black Friday Sales dataset is used for training various machine learning models and also for predicting the purchase amount of customers on black friday sales. The purchase prediction made will provide an insight to retailers to analyze and personalize offers for more customer's preferred products.

The Purchase Variable will be the predictor variable. The Purchase Variable will predict the amount of purchase made by a customer on the occasion of black friday sales.

As mentioned in the introduction, the proposed approach tries to implement the machine learning models such as Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, and Random Forest.

Regressor to forecast sales. Fig-1 depicts the flow of data through the proposed model.

Exploratory Data Analysis has been performed on the dataset [5]. The tools used for the data analysis are python, pandas, matplotlib, NumPy, array, seaborn and jupyter notebook.

The Black Friday Sales Dataset is the input dataset. Data visualization of the various attributes of this dataset is performed.

Data pre-processing which mainly includes filling missing values is performed. The categorical values are label encoded

to numeric form. The categories such as Gender where F represents female and M represents Male is converted to numerical form as 0 and 1 also other categorical values such as City_Category, Stay_In-Current_City, Age are converted to numerical form by applying Label Encoding.

The attributes such as User_id and Product_id are removed to train the model with no bias based on user_id or product_id and to achieve better performance.

The algorithms used for implementing the system are linear regression, Ridge Regression, Lasso Regression, Decision Tree Regression, and Random Forest Regressor. The models are trained using 5 fold cross-validation . The performance evaluation measure used is Mean Squared Error (MSE).

Random Forest Regressor performs better than the other algorithms with a MSE score of 3062.719.

MODELLING

A. Linear Regression.

Linear Regression is one of the supervised machine learning algorithms. A regression problem can be stated as a case when the output variable is continuous. Linear regression predicts a dependent variable (y) based on a given independent variable (x). The model depicts a linear relation among the variables. Function for linear regression is:

$$Y = \Theta 1 + \Theta 2 .x$$

Here, the input variable is x, the output value is y and Θ1 represents intercept and Θ2 represents the coefficient of x. This algorithm aims to calculate and find the best fit line to target variable and independent variable.

The features which majorly affect the linear regression model are depicted in Figure 8.

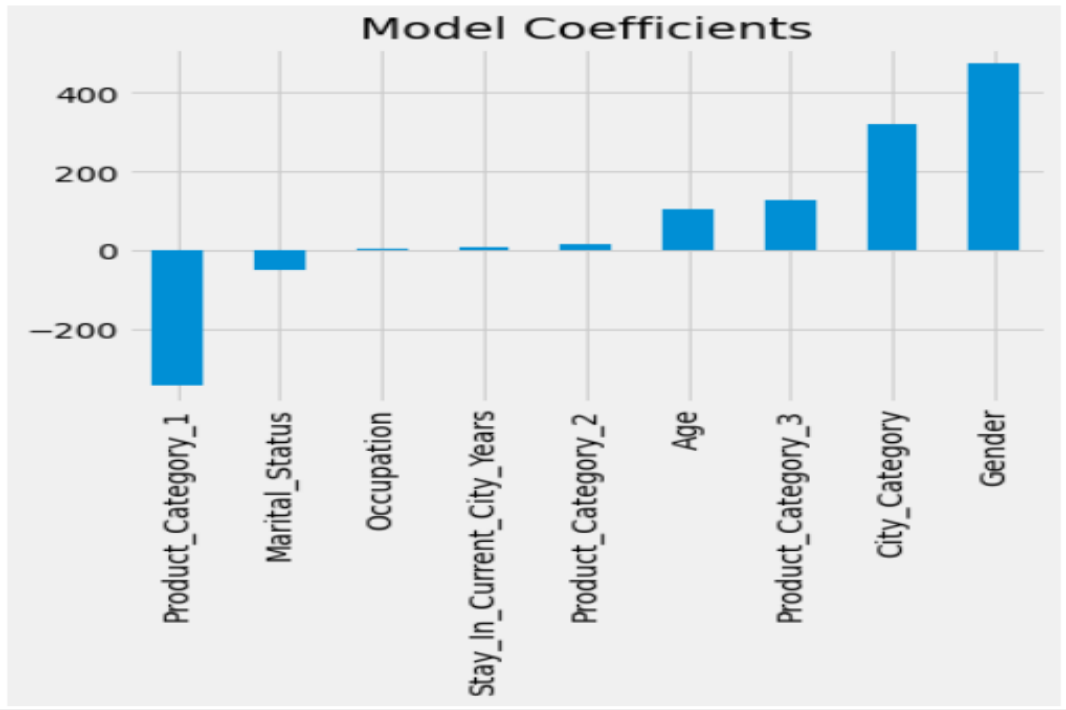

FIG-8: Attributes affecting Linear Regression

B. Ridge Regression.

Multiple regression data can be analyzed using Ridge Regression. Least Square estimates are unbiased when multicollinearity occurs. Based on the degree of bias it reduces the standard errors that is added to the regression estimates. The formula for the ridge regression is [18]:

$$\beta = (X^T X + \lambda * I)^{-1} X^T y$$

The attributes affecting the ridge model for the given dataset are depicted in Figure 9.

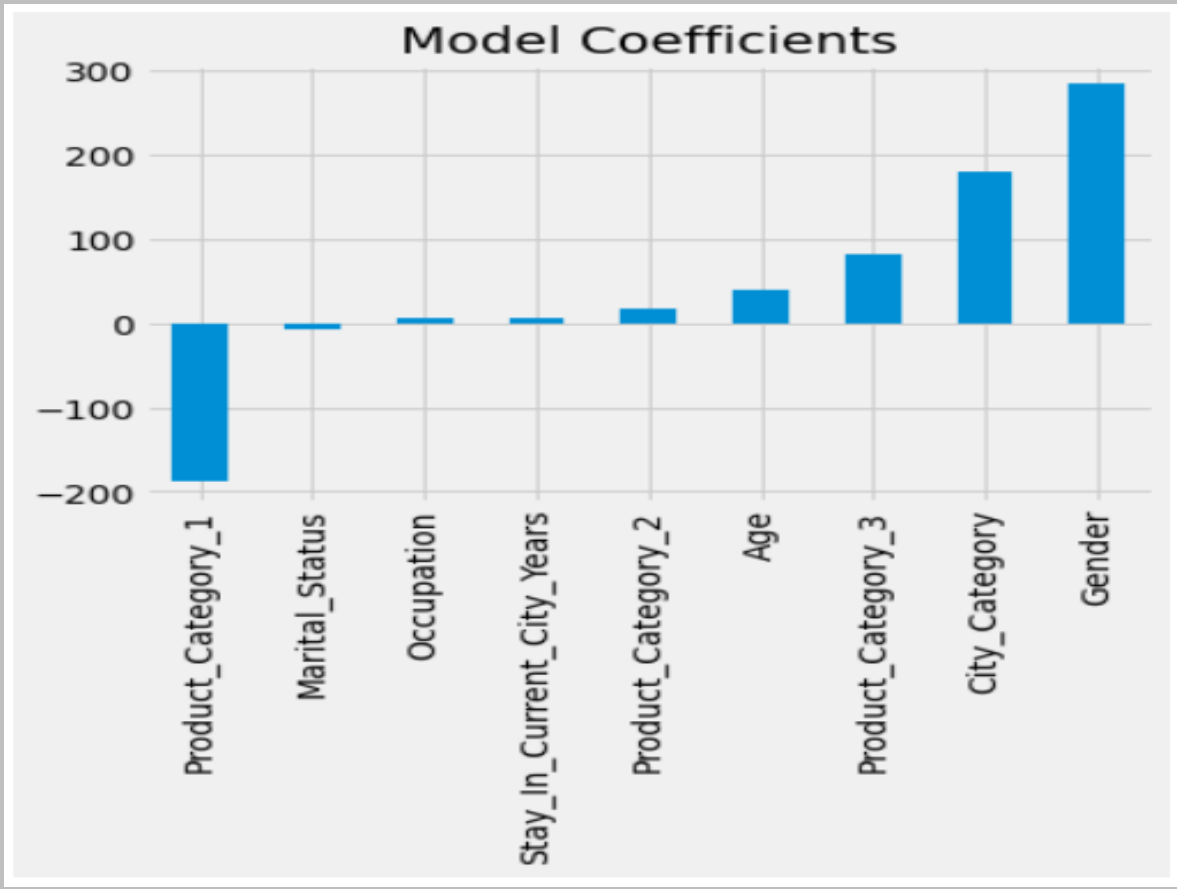

Fig-9: Attribute affecting Ridge Regression

C. Lasso Regression.

Lasso Regression provides both variable selection and regularization. It makes use of soft thresholding. Only a subset of the covariates provided is select for use in the final model in the case of Lasso regression. It can be denoted as :

$$N^{-1} \sum_{i=1}^{N} f(x_i, y_I, \alpha, \beta)$$

The attributes affecting lasso regression for the given dataset are as shown in Figure 10.

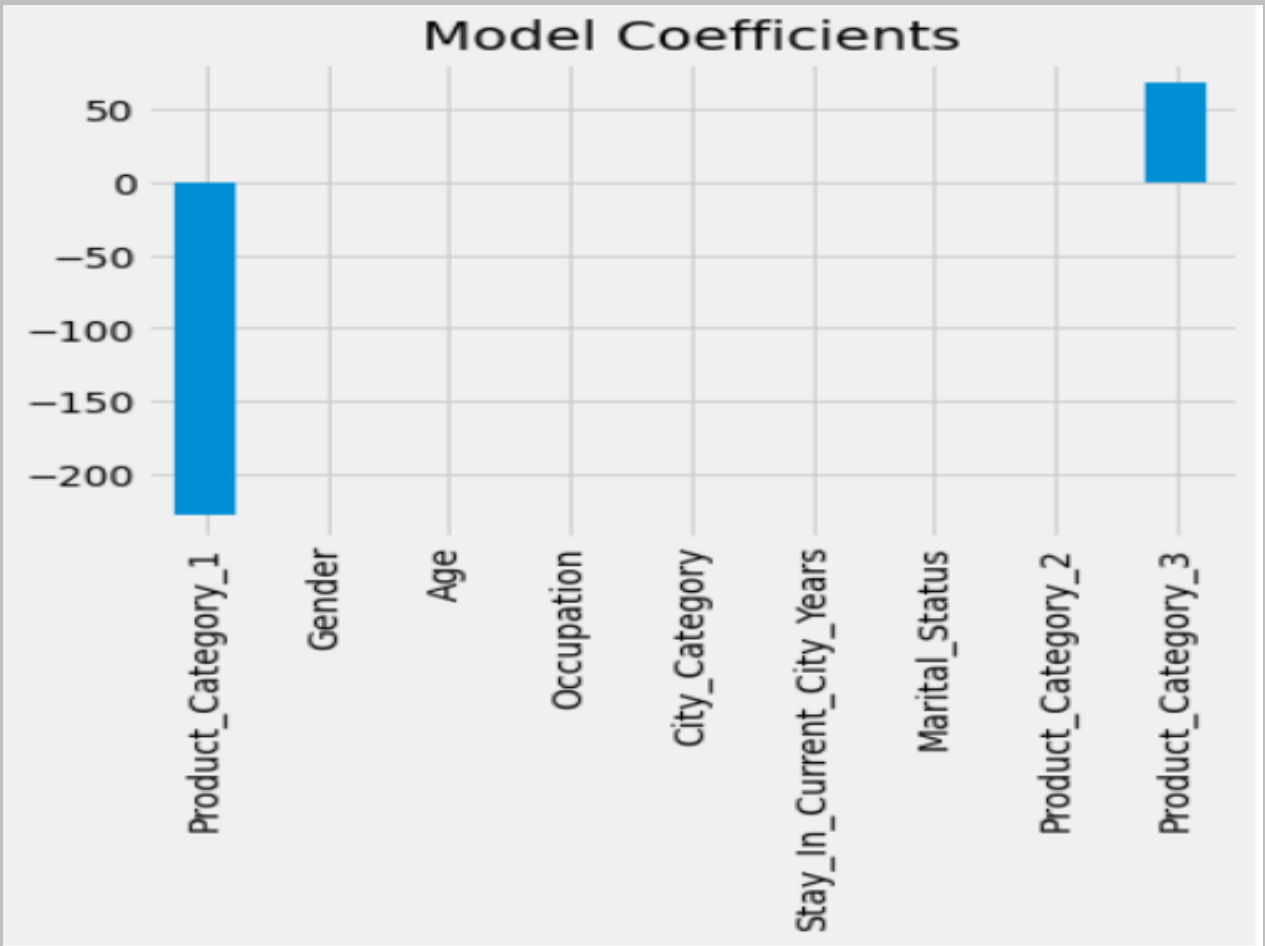

Fig-10: Attribute affecting Lasso Regression

D. Decision Tree Regressor.

The Decision Tree model builds a tree-like structure for regression or classification models. The dataset is simply broken down into smaller subsets. In a DT the control statements or values are a basis for branching, and the splitting node contains data points on either side**,** depending on the value of a specific attribute. The attribute selection measure plays an important role in root node selection.

1. Information Gain: The splitting attribute is calculated based on the amount of information required to describe the tree. The formula for the same is :

$$Info(D) = -\sum_{i}^{m} p_i \log_2(p_i),$$

2. Gain Ratio: Attributes that have a large number of values are selected by this attribute selection measure.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}.$$

3. Gini Index: The formula for calculating the Gini index is as given below :

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2,$$

Where the probability that a tuple in D belongs to class Ci is defined by pi and is estimated by |Ci,D|/|D|. The sum is computed over m classes.

The attributes that majorly affect the decision tree models are shown in Figure 11.

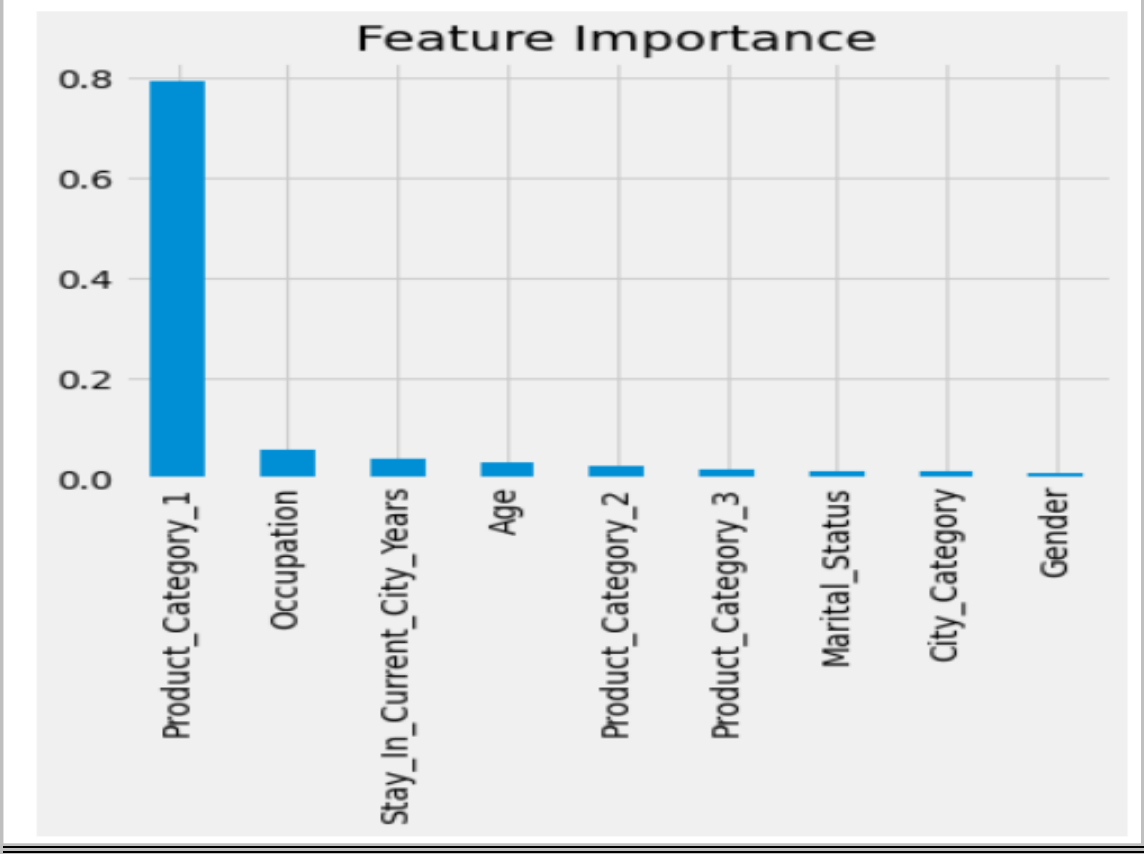

Fig-11: Attribute affecting Decision Tree Regressor

E. Random Forest Regressor

Random Forest being an ensemble technique is capable of both tasks namely regression and classification. The Random Forest is based on the ideology of combining multiple DT rather than single DT dependency.

The attributes that affect the Random Forest Regressor are as given in Figure 12.

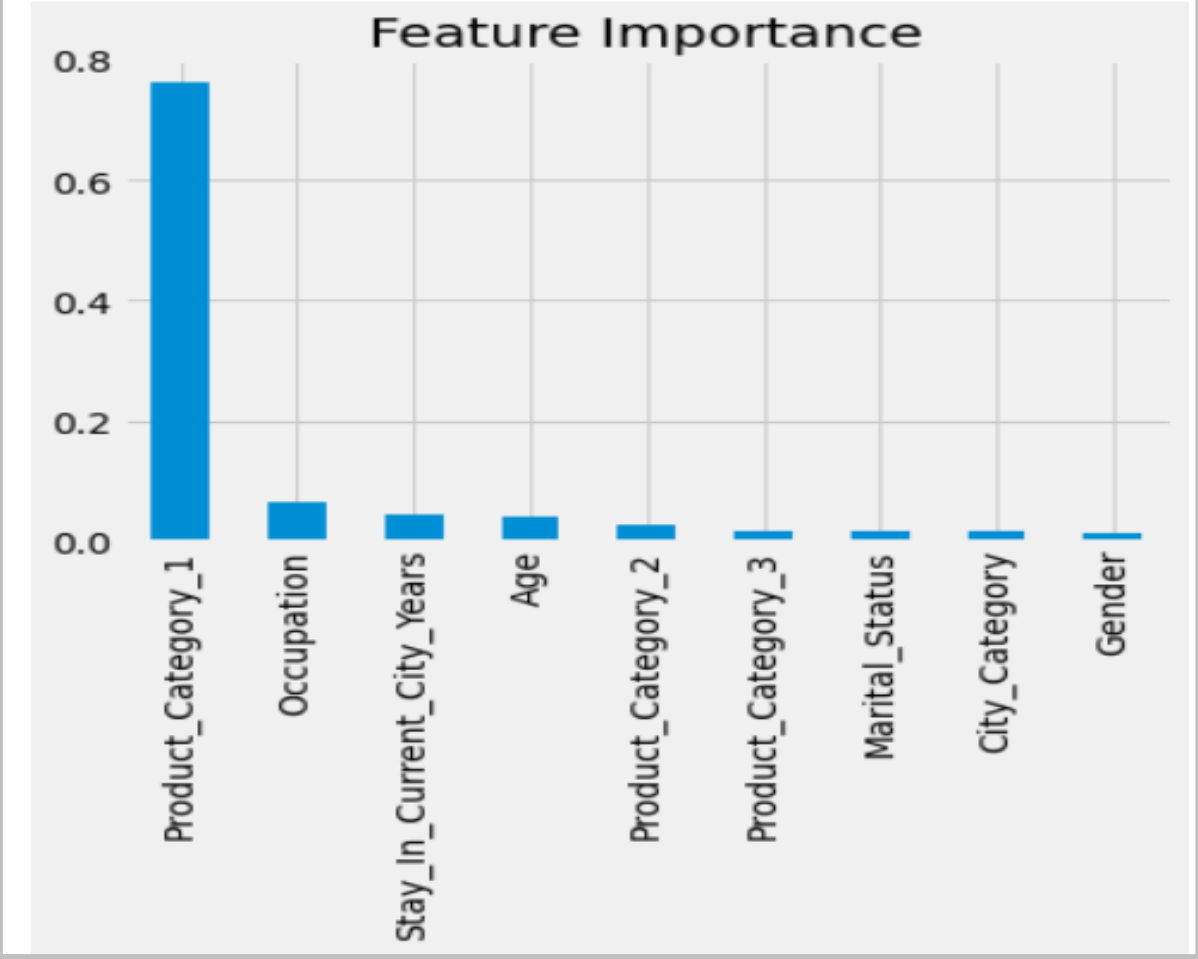

Fig. 12. Training and Validation Accuracy of Random Tree Regressor

DISCUSSION

The comparison between the MSE rates of all algorithms is depicted in Table 2 in the tables.

Based on Table 2 it can be observed that Random Forest Regressor gives better performance with comparison to other machine learning models namely linear regression and Decision tree regressor.

The MSE rate of Random Forest Regressor is 3062.72 and hence it is more suitable for the prediction model to be implemented.

CONCLUSION

With traditional methods not being of much help to business growth in terms of revenue, the use of Machine learning approaches proves to be an important point for the shaping of the business plan taking into consideration the shopping pattern of consumers.

Projection of sales concerning several factors including the sale of last year helps businesses take on suitable strategies for increasing the sales of goods that are in demand.

Thus the dataset is used for the experimentation, Black Friday Sales Dataset from Emerging India Analytics. The models used are Linear Regression, Lasso Regression, Ridge Regression, Decision Tree Regressor, and Random Forest Regressor. The evaluation measure used is Mean Squared Error (MSE). Based on Table II Random Forest Regressor is best suitable for the prediction of sales based on a given dataset.

Thus the proposed model will predict the customer purchase on Black Friday and give the retailer insight into customer choice of products. This will result in a discount based on customer-centric choices thus increasing the profit to the retailer as well as the customer.