ANALYSIS

R Aravind	AM.EN.U4AIE21151
Vishnu Shaji	AM.EN.U4AIE21167
Karthik G Nair	AM.EN.U4AIE21138
Akhilesh	AM.EN.U4AIE21108
Mamidi Sowji Krishna	AM.EN.U4AIE21181

ABSTRACTION

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, etc. 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. For the purpose of analyzing and predicting the sales, we have used three models. 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.

INTRODUCTION

The shopping sector has greatly evolved due to the Internet revolution. Nowadays, most people use the Online Shopping method as it is easier than the traditional method of shopping. The biggest advantages of online shopping are convenience, better prices, more variety, easy price comparisons, no crowds, etc. The covid pandemic has boosted online shopping. As we know that in today's time online shopping is growing, also if we see the total sales of the year 2021 it will go much higher in 2022.

Basically, Black Friday originated in the USA. This day is referred to as Thanksgiving Day. This sale is celebrated on the fourth Thursday of November once every year. The purpose of this day is to organize this sale to promote customers to buy more products online so that they can boost the online sector of shopping.

Dataset is used for training and prediction. The Dataset of Black Friday Sales is the online biggest dataset, and the dataset is also accepted by various e-commerce websites.

The prediction model which we 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.

DATASET

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 customers' preferred products.

TABLE I. 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

The Purchase Variable will be the predictor variable. The Purchase Variable will predict the amount of purchases made by a customer on Black Friday sales.

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.

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 color system to represent the correlation among different attributes. It is a data visualization library (Seaborn) element. Heatmap color encoded matrix can be described as the lower the intensity of the color of an attribute related to the target variable, the higher 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. From correlation matrix we take age, gender, and occupation as they have high correlation to purchase

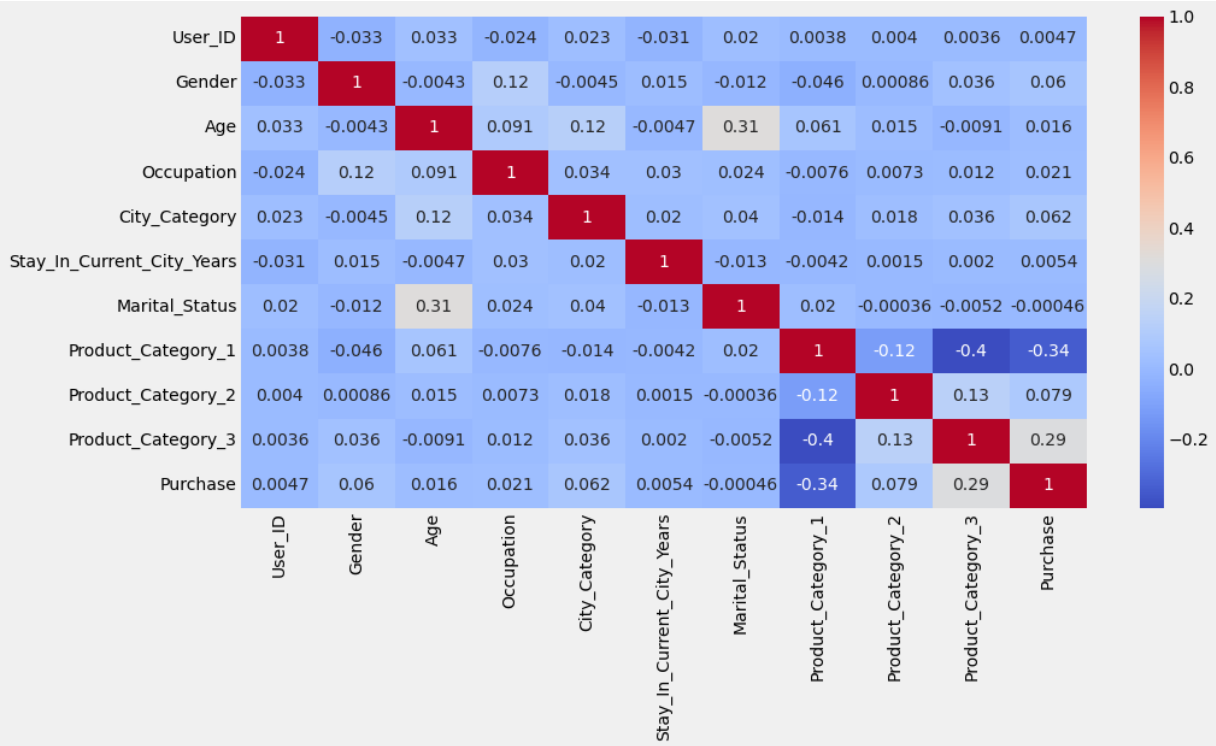

Fig 1.

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.

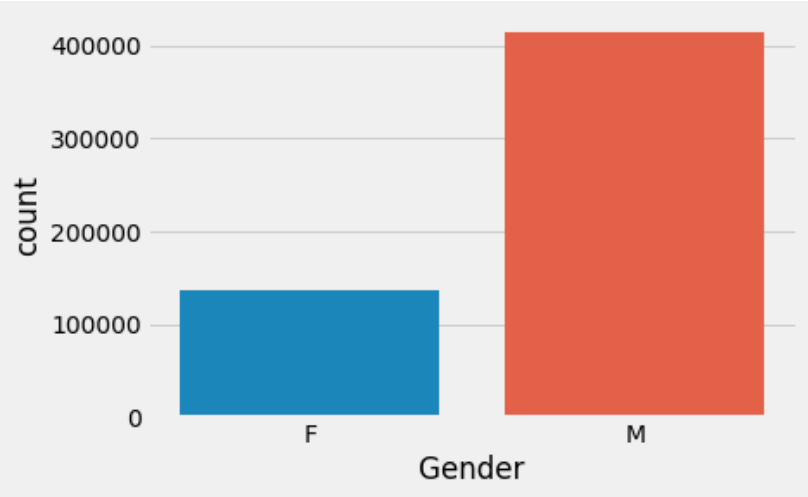

Fig 2

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.

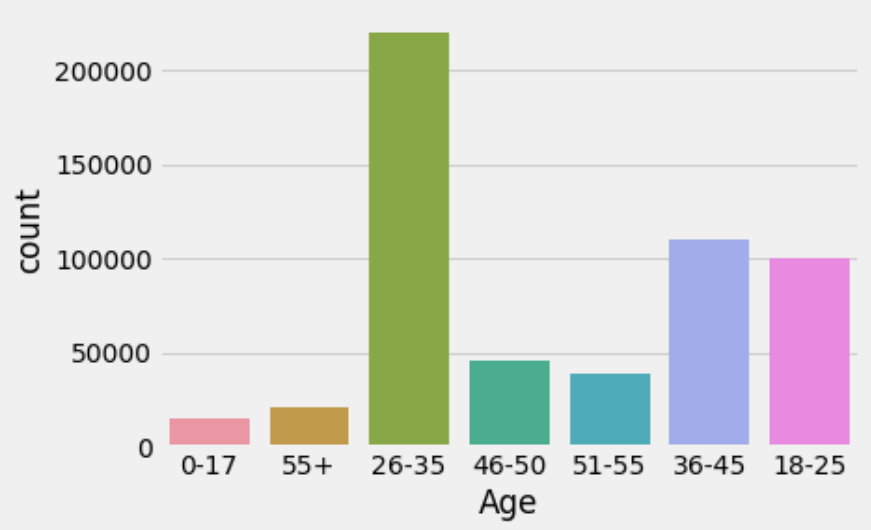

Fig 3

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 the maximum count. The second maximum based on the count plot is occupation 0.

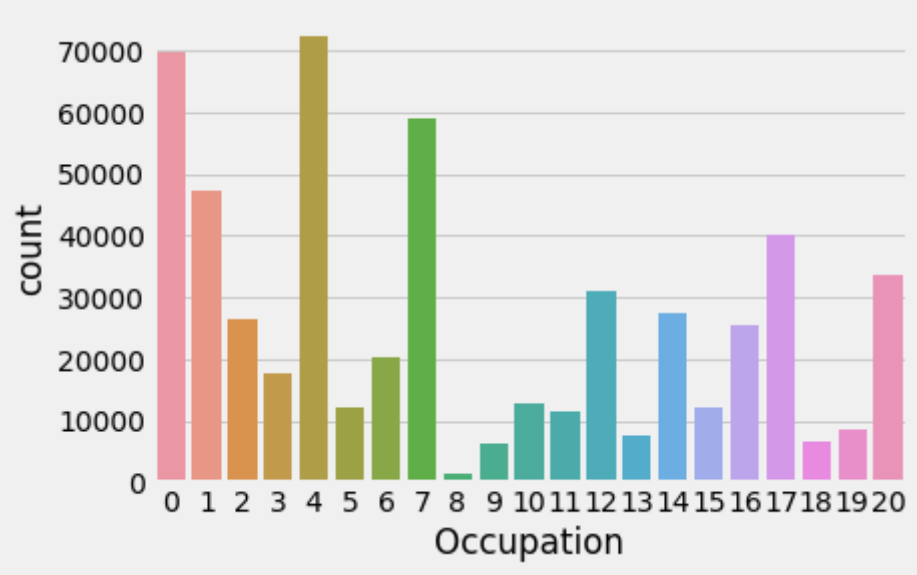

Fig 4

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.

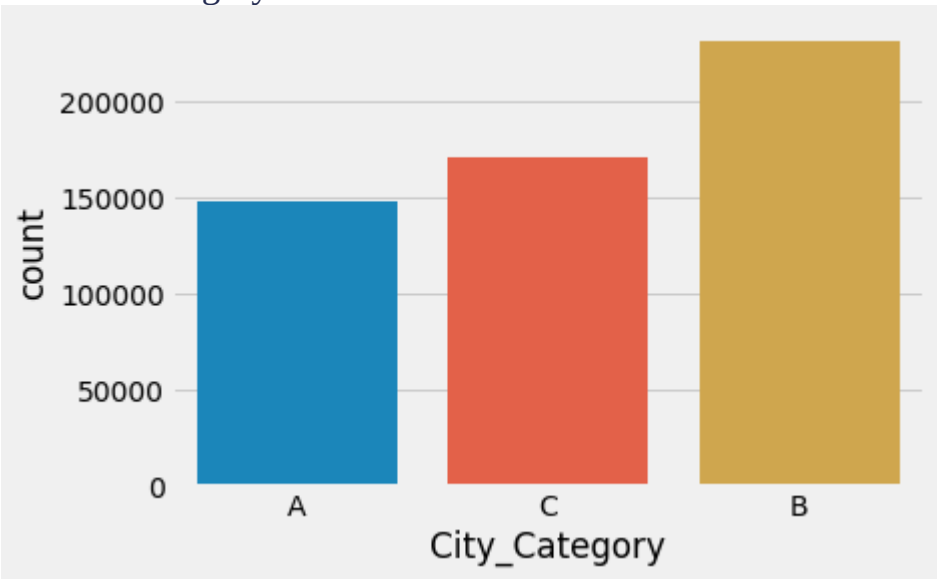

Fig 5

The count plot for Stay_In_Current_City is as given in Figure 6. The observations based on the count plot can be stated as the maximum count for 1 year. The minimum count is for 0 years.

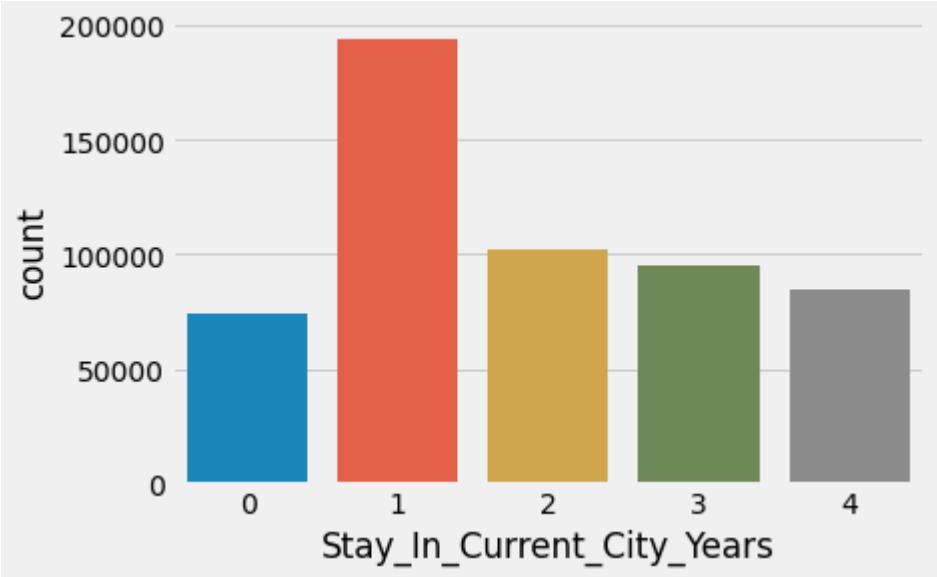

Fig 6

The density of purchase is as given in figure 7.

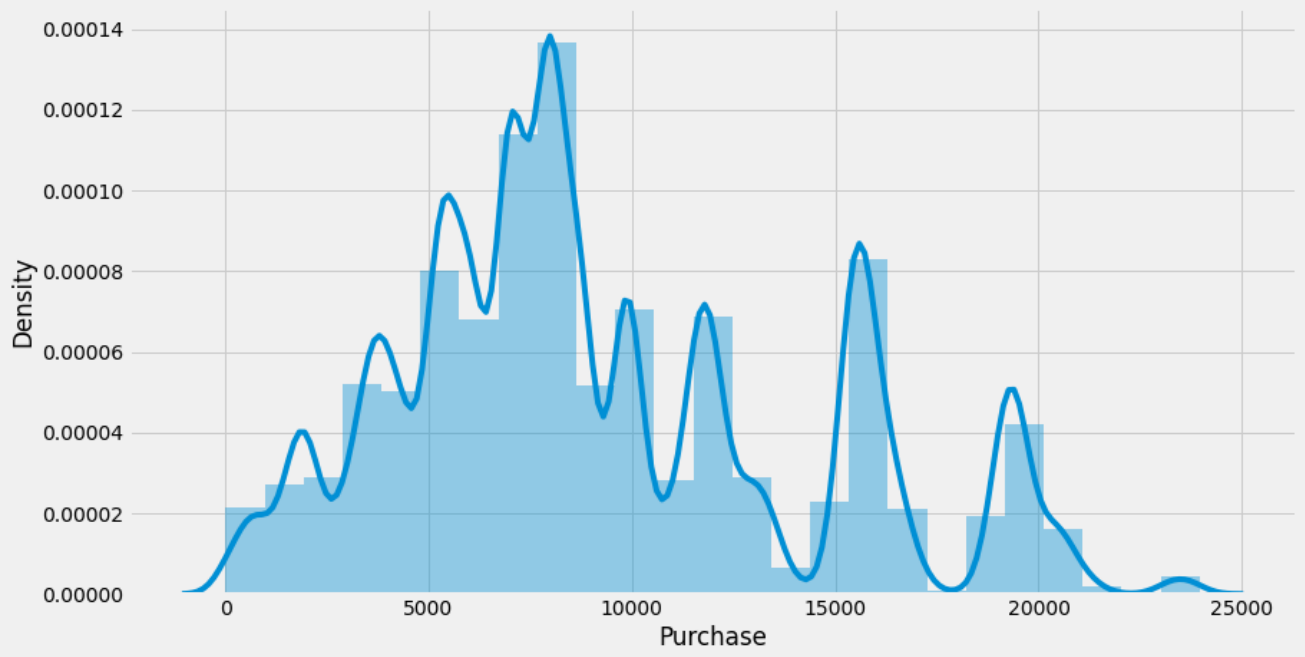

Fig 7

Bivariate analysis between occupation and purchase is shown in figure 8, between age and purchase is shown in figure 9, between gender and purchase is shown in figure 10.

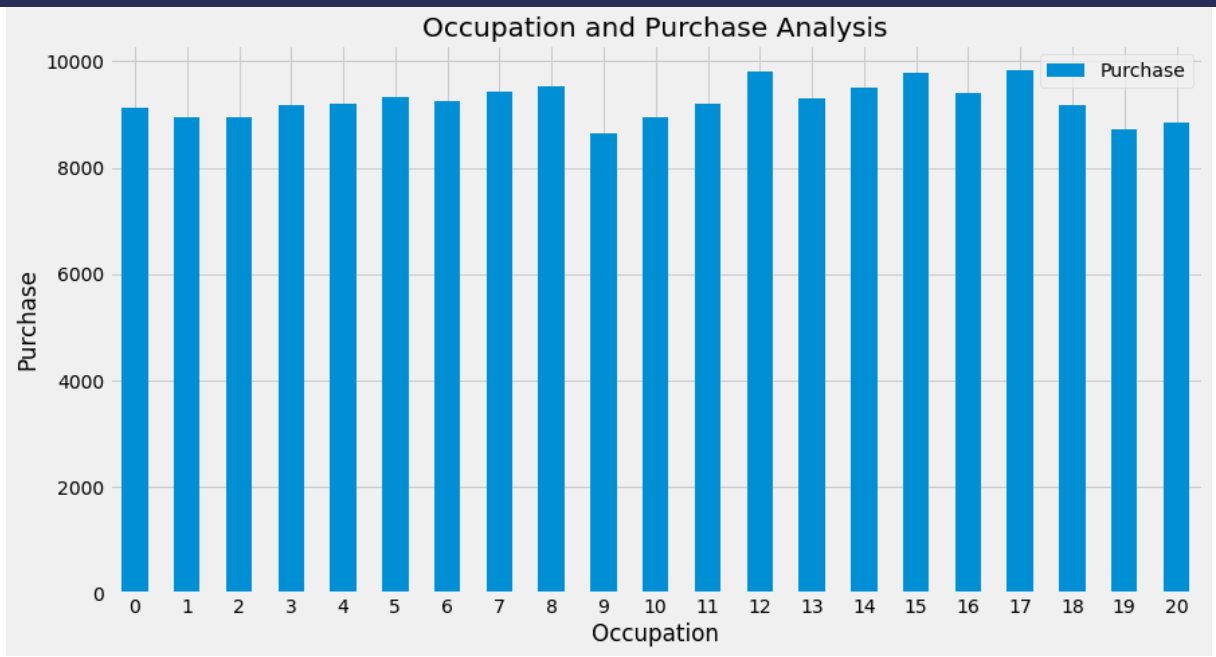
Occupation and Purchase Analysis
Purchase
10000
8000
6000
4000
2000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Occupation

Fig 8

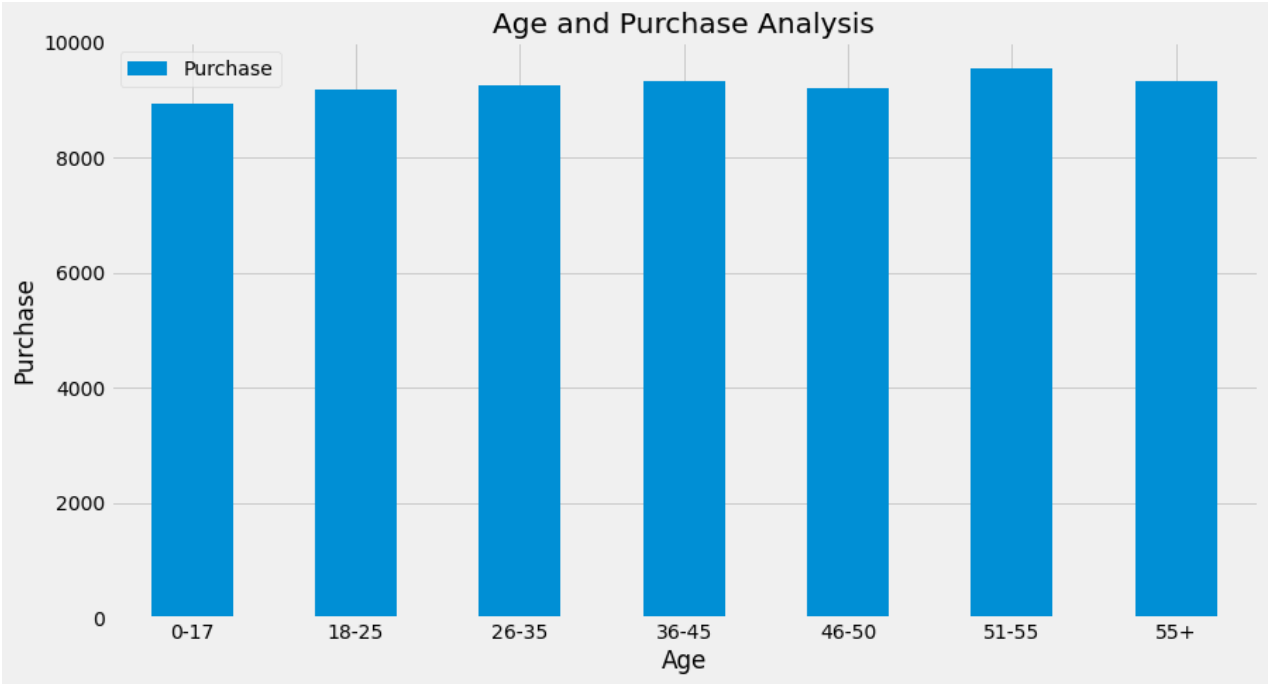
Age and Purchase Analysis
Purchase
10000
8000
6000
4000
2000
0
0-17
18-25
26-35
36-45
46-50
51-55
55+
Age

Fig 9

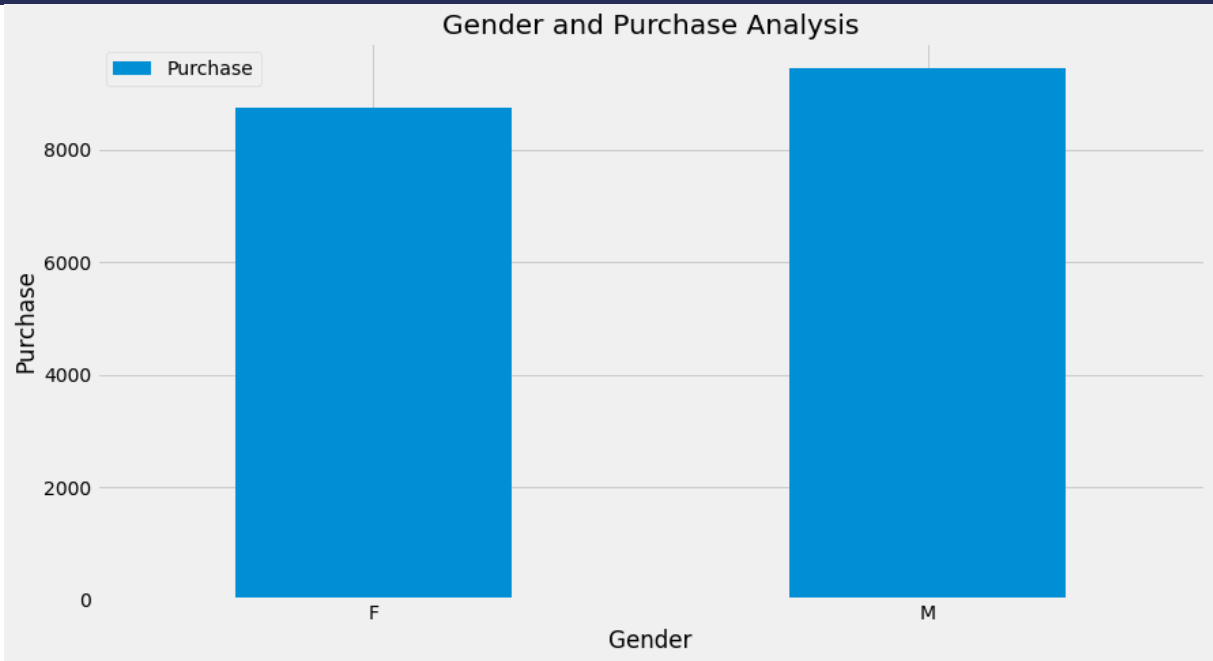
Gender and Purchase Analysis
Purchase
8000
6000
4000
2000
0
Purchase
F
M
Gender

Fig 10

METHODOLOGY

Checking how different parameter affects the purchase field

.

1) Polynomial regression: Polynomial features are those features created by raising existing features to an exponent. Here we are using Polynomial Features with a degree as 3.

a) Age:

```
Results
MAE :  4064.9071463952996
MSE :  5018.5251009811045
r2score :  0.0003968401162388613
```

b) Gender:

```
Results
MAE :  4049.0281304607497
MSE :  5009.581797776315
r2score :  0.003956367420788531
```

c) Occupation:

```
Results
MAE :  4057.7072494538434
MSE :  5014.465811529174
r2score :  0.0020132662304295224
```

2) 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).

a) Age:

```
Results
MAE :  4060.1805567155034
MSE :  5014.609255171635
r2score :  0.00030796409289413074
```

b) Gender:

```
Results
MAE :  4045.498726965293
MSE :  5006.510625204182
r2score :  0.0035343762856595573
```

c) Occupation:

```
Results
MAE :  4059.4018306562616
MSE :  5014.35528545359
r2score :  0.0004092222616890462
```

3) Linear regression on multiple parameters: also known simply as multiple regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

Parameters (User_ID', 'Product_ID', 'Purchase')

```
Results
MAE :  3523.3557657213196
MSE :  4617.994034201719
r2score :  0.1521895377687933
```

4) 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.

```
Results
MAE :  2376.1949528899095
MSE :  3366.698893523624
r2score :  0.5493902452705477
```

5) Random Forest Regressor: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

```
Results
MAE :  2228.166266312957
MSE :  3064.9665564520924
r2score :  0.6265404955293731
```

6) Extra tree regressor model: This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

```
Results
MAE :  2286.592730511008
MSE :  3194.682177146094
r2score :  0.5942604396643059
```

RESULTS

The comparison between the MSE rates of all algorithms is depicted in the table II.

MODEL	MSE
Polynomial (Age)	5018.5251009811045
Polynomial (Gender)	5009.581797776315
Polynomial (Occupation)	5009.581797776315
Linear Regression (Age)	5014.609255171635
Linear Regression (Gender)	5014.609255171635
Linear Regression (Occupation)	5014.609255171635
Linear regression on multiple parameters	4617.994034201719
Decision Tree Regressor	3366.698893523624
Random Forest Regressor	**3064.9665564520924**
Extra Trees Regressor	3194.682177146094

TABLE II

Based on the Table II , 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 3064.64 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 Kaggle. The models used are Linear Regression, Polynomial Regression, Linear regression on multiple parameters, Decision Tree Regressor, Extra Trees 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

As future research, we can perform hyperparameter tuning and apply different machine learning algorithms.

SOURCE CODE & DATASET

Source Code: https://www.kaggle.com/code/souvi11/maths-project-sem-2
Dataset: https://www.kaggle.com/datasets/souvi11/black-friday-sales