Regression Analysis
on
Black Friday Sales Prediction

BY

GROUP-I

Kiran Kumar Katamneni

1002115997

Abhinay Akula

1002135976

Bhargav Sai Cherukuri

1002126215

Abstract: The nationwide Christmas shopping carnival gets underway on Black Friday. Large online retailers like Amazon, Flipkart, and others entice buyers by providing discounts and offers across a variety of product categories on Black Friday. The product categories include clothing, kitchen appliances, electronics, and home decor. Numerous academics have conducted studies to forecast sales. Discounts are offered on a variety of product items based on the examination of this data.

We conducted a regression analysis of the data to examine and forecast sales. Black Friday Sales Dataset, a dataset that is accessible on Kaggle, has been utilized for analysis and forecasting. Regression models that are used for prediction are linear. Mean Squared Error (MSE) and R-squared is a metric used to assess performance.

Introduction:

The Internet revolution has significantly changed the retail industry. Most people perceive internet buying more favorably than conventional purchasing. Convenience, better costs, greater choice, simple price comparisons, absence of crowds, etc. are some of the greatest advantages of Internet shopping. The epidemic has increased internet sales. Although internet purchasing continues to expand this year, 2021 is predicted to have greater overall sales.

Thanksgiving Day is another name for Black Friday, which has its roots in the United States. Every year on the fourth Thursday in November, this sale is held. In terms of shopping, this day is designated as being the busiest. The goal of holding this deal is to encourage people to purchase more items online to grow the online shopping industry.

The built-in prediction model will assist in examining how various qualities relate to one another. For training and prediction, Black Friday Sales Dataset is utilized. The largest dataset available online is called the Black Friday Sales Dataset, and several e-commerce businesses have agreed to use it.

Based on the customer's age, city category, employment, etc. The prediction model will offer a prediction. The implementation of the prediction model is based on concepts like linear regression.

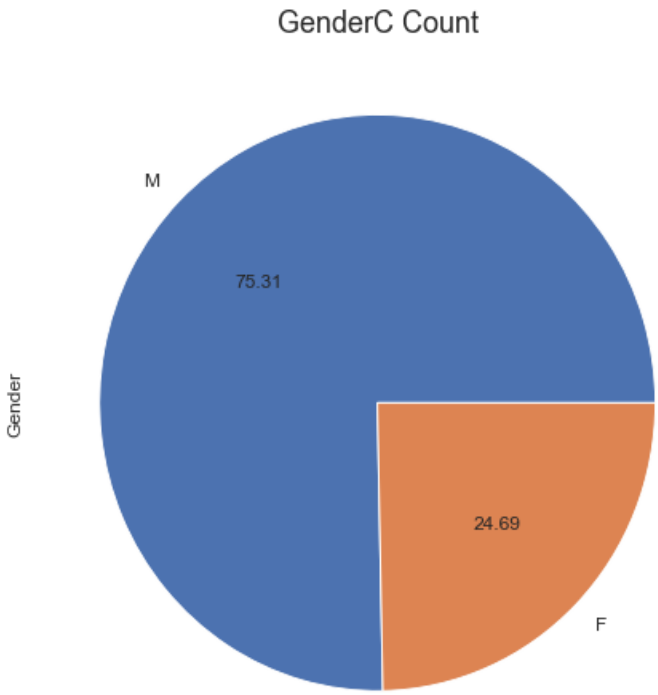
Implementation:
The study makes use of the Black Friday Sales Dataset, a free resource on Kaggle. (https://www.kaggle.com/datasets/cerolacia/black-friday-sales-prediction) The dataset contains information on sales transactions. The collection has 5,50,068 records. User_id, Product_id, Marital Status, City Category, Occupation, etc. are some of the attributes in the dataset.

A summary of the dataset is provided below.
The dataset for Black Friday sales is used to train a machine learning model and predict how much each customer will spend during the Black Friday sales. The purchase forecast offered will allow retailers to research and provide offers for more consumers' preferred goods.

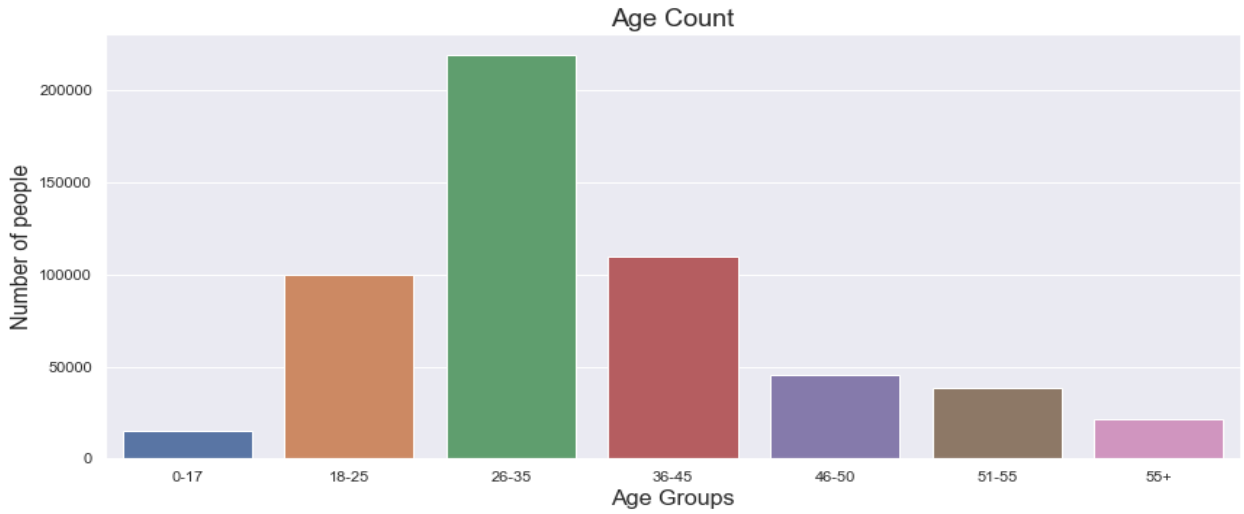
Variable	**Definition**
User_ID	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in bins
Occupation	Occupation (Masked)
City_Category	Category of the City (A, B, C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status
Product_Category_1	Product Category (Masked)
Product_Category_2	Product may belong to another category also (Masked)
Product_Category_3	Product may belong to another category also (Masked)
Purchase	Purchase Amount (Target Variable)

The predictor variable is going to be the purchase variable. The Purchase Variable will forecast how much a person will spend during the Black Friday discounts.
The suggested method attempts to use the machine learning model Linear Regression to anticipate sales, as was described in the introduction.
Exploratory data analysis has been performed on the data. Python, pandas, matplotlib, NumPy, array, seaborn, and jupyter notebook are the tools utilized for data analysis.

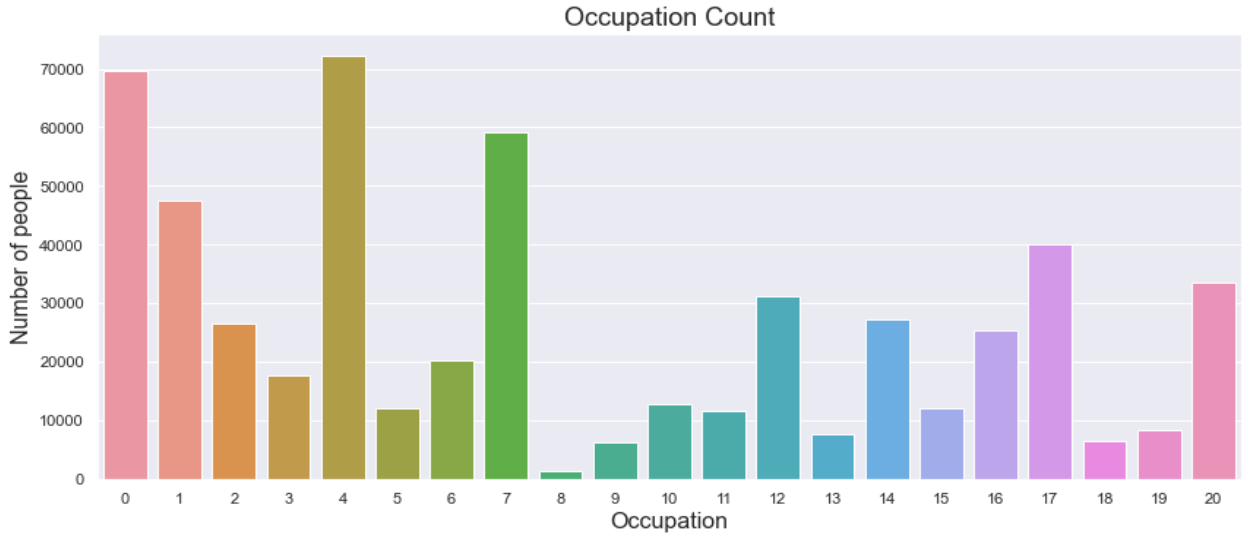
Exploratory Data Analysis:



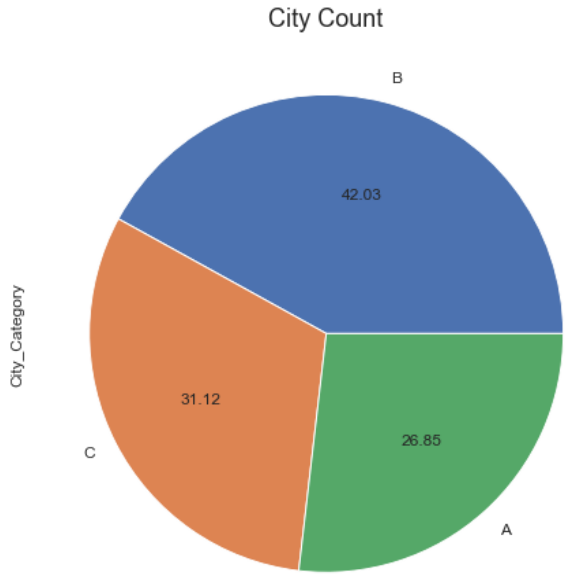
The count plots for different attributes are visualized as different figures given below. The count plot for gender attributes is as above. Based on the count plot for gender attribute it is observed that feature M (Male) has the maximum count. The count for F features is less.



Above plot depicts the count plot for the age property. According to the count plot's findings, the age group 26–35 has the highest count. The second-highest total recorded belongs to the 36-to-45 age group. The age group of 18 to 25 years old has the third highest observed count.



The count plot depicts the count plot for the occupation attribute. Masked occupation 4 has the highest count, according to the count plot. Based on the count plot, occupation 0 is the second maximum.



The count plot for city_category is 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.

From the above heatmap, we can say the target variable is not highly correlated to any of the variables. Also, all the product categories are having negative correlation with purchase. However, they have a positive correlation with each other.

There are 3 categorical variables in the predicting variables present in the provided data which are:

- Gender
- City Category
- Stay in Current City Years

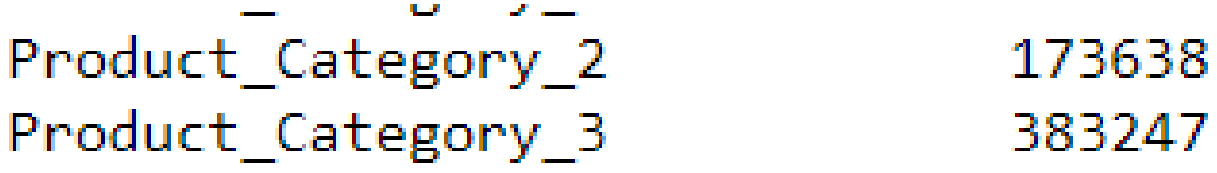
The Label encoder class has been used to convert these categorical features into numerical values. The label encoder transforms the variables with values in between 0 and n_classes 1. The function used to transform the variable is.

le=LabelEncoder ()
le.fit_transform

The Age variable which is divided into interval classes is replaced with the average values of the classes.

0-17	17
18-25	20
26-35	30
36-45	40
46-50	47
51-55	52
55+	56

There are a large number of null values in the Product_Category_2 and Product_Category_3 labels which are dropped from the data set as they do not seem to help in predicting the response variable due to insufficient data.



LINEAR REGRESSION

Linear regression model has been built using the data set after the EDA and data preprocessing steps have been performed. The target variable**(Y)** is the purchase label of the data set.

The dataset is divided using the **train_test_split** feature from the **sklearn.model** module into a training set which contains 70% of records and a test set which contains 30% of records of the data. The training data set has been used to train the model and the test data is used to verify the accuracy of the model.

The model is built using the **Original Least Squares (OLS)** method which minimizes the distance between observed responses in the data set and the responses predicted by the linear approximation.

The model performance is verified based on the following metrics:

- Mean_Absolute_Error – 3602.206
- Mean_Squared_Error – 22099033.78
- Coefficient of Determination(R^2) – 0.1241

OLS Regression Results

Dep. Variable:	Purchase	R-squared:	0.124
Model:	OLS	Adj. R-squared:	0.124
Method:	Least Squares	F-statistic:	1.117e+04
Date:	Sat, 06 May 2023	Prob (F-statistic):	0.00
Time:	16:14:12	Log-Likelihood:	-5.4315e+06
No. Observations:	550068	AIC:	1.086e+07
Df Residuals:	550060	BIC:	1.086e+07
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.033e+04	27.905	370.089	0.000	1.03e+04	1.04e+04
Gender	514.1218	14.818	34.696	0.000	485.079	543.165
Age	14.8222	0.636	23.293	0.000	13.575	16.069
Occupation	6.2726	0.984	6.374	0.000	4.344	8.202
City_Category	352.4752	8.402	41.950	0.000	336.007	368.943
Stay_In_Current_City_Years	8.1018	4.919	1.647	0.100	-1.539	17.743
Marital_Status	-52.8036	13.561	-3.894	0.000	-79.383	-26.224
Product_Category_1	-437.2734	1.615	-270.735	0.000	-440.439	-434.108

Omnibus:	61928.125	Durbin-Watson:	1.706
Prob(Omnibus):	0.000	Jarque-Bera (JB):	87083.095
Skew:	0.886	Prob(JB):	0.00
Kurtosis:	3.812	Cond. No.	167.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Observations from Model-I:

1. The significance level selected for the model is 0.05.
2. The R^2 value of the model-1 is 0.124 i. e 12.4% of the variance of the data is explained by the model, meaning they are not good predictors of the purchase target variance.
3. The F-stat is used to judge whether the model is a good fit which is extremely high and probability of obtaining the F-stat is near zero which is an indicator of a statistically significant model.
4. The significance of all the predicting variables is judged based on the probability of their respective t-values. The values of all the variables are lower than the significance value except for the Stay_In_Current_City_Year which can be removed as it is to be statistically insignificant.

MODEL-II

OLS Regression Results

Dep. Variable:	Purchase	R-squared:	0.122
Model:	OLS	Adj. R-squared:	0.122
Method:	Least Squares	F-statistic:	1.523e+04
Date:	Sat, 06 May 2023	Prob (F-statistic):	0.00
Time:	16:14:14	Log-Likelihood:	-5.4324e+06
No. Observations:	550068	AIC:	1.086e+07
Df Residuals:	550062	BIC:	1.086e+07
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.061e+04	25.722	412.593	0.000	1.06e+04	1.07e+04
Gender	509.7860	14.840	34.351	0.000	480.699	538.873
Age	17.7749	0.633	28.061	0.000	16.533	19.016
Occupation	7.2819	0.985	7.393	0.000	5.351	9.213
Marital_Status	-51.3445	13.582	-3.780	0.000	-77.965	-24.724
Product_Category_1	-438.7600	1.617	-271.287	0.000	-441.930	-435.590

Omnibus:	61803.071	Durbin-Watson:	1.701
Prob(Omnibus):	0.000	Jarque-Bera (JB):	86757.420
Skew:	0.886	Prob(JB):	0.00
Kurtosis:	3.803	Cond. No.	155.

Notes:

Observations from Model-II:

1. The significance level selected for the model is 0.05. All the variables are significant in this model.
2. The R^2 value of the model-1 is 0.122 i. e 12.2% of the variance of the data is explained by the model, meaning they are not good predictors of the purchase target variance.
3. The F-stat is extremely high and probability of obtaining the F-stat is near zero which is an indicator of a statistically significant model.

Comparing both the models:

The ANOVA table is used to compare the two models modeled with all the variables and the one with reduced number of variables using the hypothesis.

H_0: Model-II better than Model-I

H_a: Model-I is a better fit than Model-II

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	550062.0	1.219114e+13	0.0	NaN	NaN	NaN
1	550060.0	1.215212e+13	2.0	3.901456e+10	882.987576	0.0

From the anova table the p-value for the F-test is 0, which is less than the typical significance level of 0.05. Therefore, we reject the null hypothesis and conclude that the reduced model-II is a better fit for the data than the previous model-I.

Multicollinearity:

The Multicollinearity of the models can be checked used by the **Variation Inflation Factors (VIF).** The VIF of the two models:

```
Full model VIF: [23.469371621285664, 1.0186529087539893, 1.1232526873286566, 1.0255263339496978, 1.1074089389597352, 1.0012349384773567, 1.1404432265136697, 1.138449734694281]
Reduced model VIF: [21.502416242097876, 1.018536839532248, 1.123244699979447, 1.0246704001578417, 1.1072528250624811, 1.1404397856952033, 1.1384370112159916]
```

Based on the VIF values, there does not seem to be high multicollinearity in the full model, as several variables do not have VIF values greater than 10, except one. In the reduced model, the VIF values are lower than the full

model, but the reduced model appears to have less multicollinearity compared to the full model.

Conclusion: In conclusion, this project utilized a regression analysis on the Black Friday Sales Dataset to forecast sales during the nationwide Christmas shopping carnival, which starts on Black Friday. Linear regression models were used for prediction, and the performance was assessed using the Mean Squared Error (MSE) and R-squared metrics. Also, the anova depicts that the model after removing non-significant variables performs better. Moreover, there is not much multicollinearity between the independent variables in both the models.

Future Work: As future research, we can perform hyperparameter tuning and apply different machine learning models like Decision Tree, Random Forest, etc.

References:

[1] C. M. Wu, P. Patil and S. Gunaseelan, "Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 16-20, doi: 10.1109/ICSESS.2018.8663760.

[2] Odegua, Rising. (2020). Applied Machine Learning for Supermarket Sales Prediction.

[3] K. Singh and R. Wajgi, "Data analysis and visualization of sales data," 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), 2016, pp. 1-6, doi: 10.1109/STARTUP.2016.7583967

[4] S. Yadav 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.2016.2