Project Report

Sales Prediction on Black Friday using ML

Project Report submitted on the fulfilment of the requirements of Post graduate Diploma in Big data Analytics

Authors:

Anjali Maria V – 220960925001

Sagar Chandre – 220960925002

Bhairav Chaudhari – 220960925003

Darshan Darekar – 220960925004

Apoorva Eadke – 220960925005

Co-ordinator:

Ms. Roopa Panicker (Course Coordinator)

Ms. Soorya M. (CDAC)

Ms. Divya Das (CDAC)

STDC

CDAC Thiruvananthapuram

Trivandrum, Kerala 695581

Abstract:

A retail company “ABC Pvt. Ltd” wants to understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for a selected high-volume product from last month. We are building a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

Introduction:

Black Friday is an informal name for the Friday following Thanksgiving Day in the United States, which is celebrated on the fourth Thursday of November. The day after Thanksgiving has been regarded as the beginning of the United States Christmas shopping season since 1952, although the term "Black Friday" did not become widely used until more recent decades.

Many stores offer highly promoted sales on Black Friday and open very early, such as at midnight, or may even start their sales at some time on Thanksgiving. The major challenge for a Retail store or eCommerce business is to choose product price such that they get maximum profit at the end of the sales.

Our project deals with determining the product prices based on the historical retail store sales data. After generating the predictions, our model will help the retail store to decide the price of the products to earn more profits.

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.

C. M. Wu et al. [1] 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 Kaggle. 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.

Odegua, Rising [2] 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.

- **Proposed System:**

a) Dataset pre-processing:

Converting Gender to binary

Gender
0
0
0
0
1

Converting City_Category to binary

City_Category
1
1
1
1
3

Converting Stay_In_Current_City_Years to binary

Stay_In_Current_City_Years
2
2
2
2
4

Converting Age to numeric values

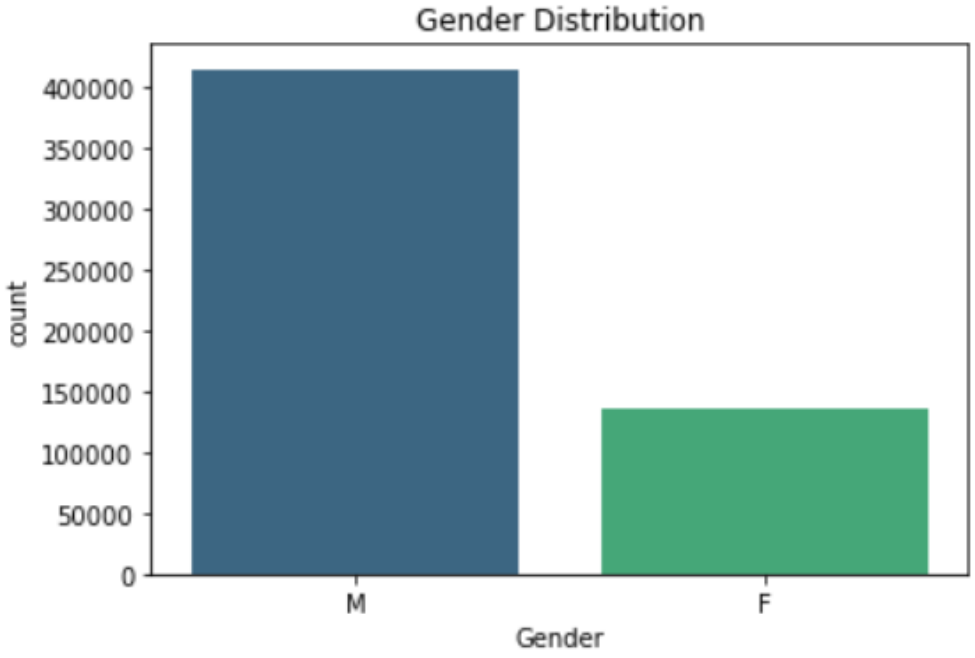
```
one_hot=pd.get_dummies(data=train_df['Age_Groups'],prefix='Age_Group',drop_first=True)
one_hot
```

	Age_Group_18-25	Age_Group_26-35	Age_Group_36-45	Age_Group_46-50	Age_Group_51-55	Age_Group_55+
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	1
...	...	...	...	...	...	...
550063	0	0	0	0	1	0
550064	0	1	0	0	0	0
550065	0	1	0	0	0	0
550066	0	0	0	0	0	1
550067	0	0	0	1	0	0

550068 rows × 6 columns

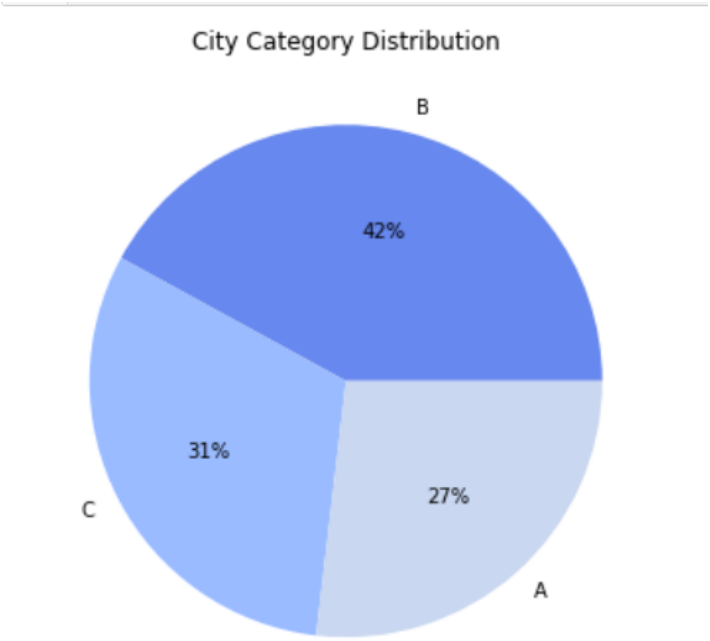
b) Feature Extraction:

- **Univariate Analysis:**



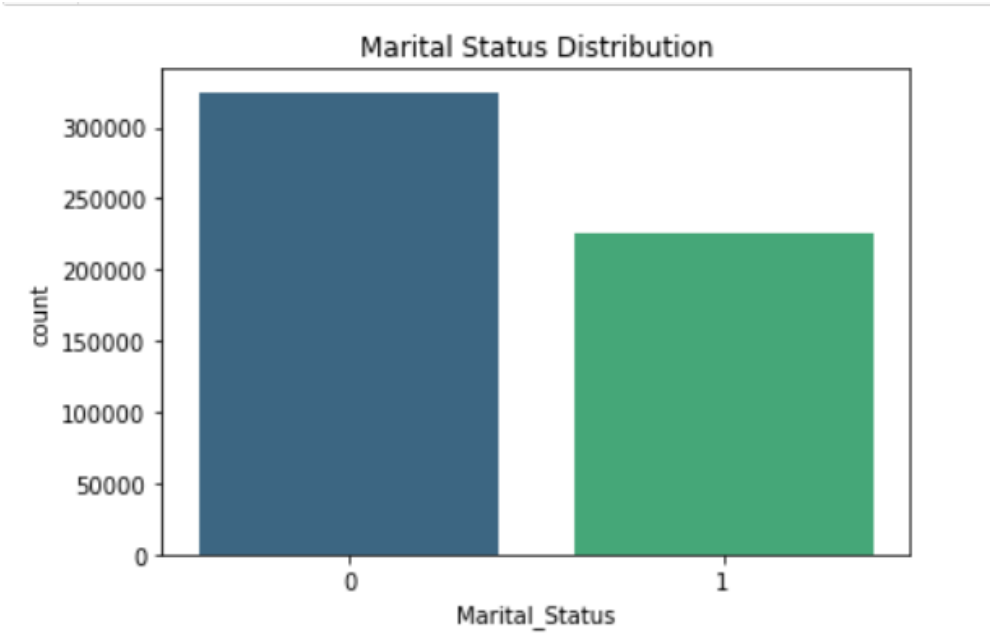
Insights:

1. The Gender feature has high data imbalance. The ratio of count of female customers is very less compared to male customers.
2. Need to handle this class imbalance using SMOTE/Oversampling techniques.



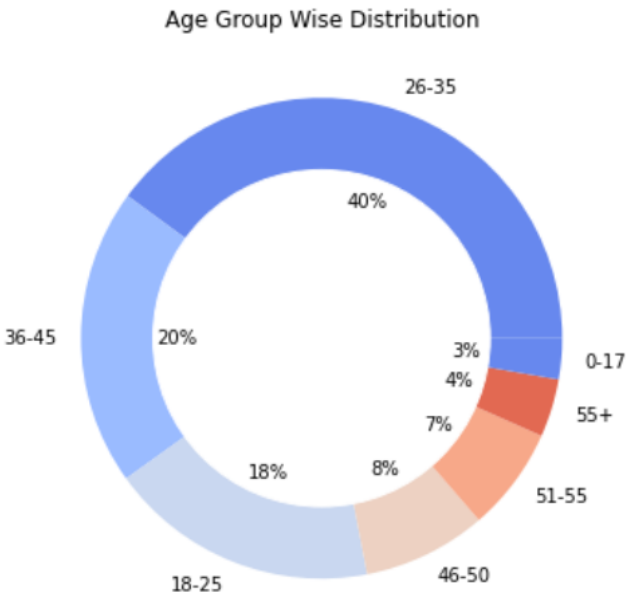
Insight:

1. City category B has observed to be having highest percentage of customers who have purchased in Black Friday Sale i.e. 42% compared to A and C.



Insight:

1. Data shows unmarried customers have spent on Black Friday sale more than married customers.



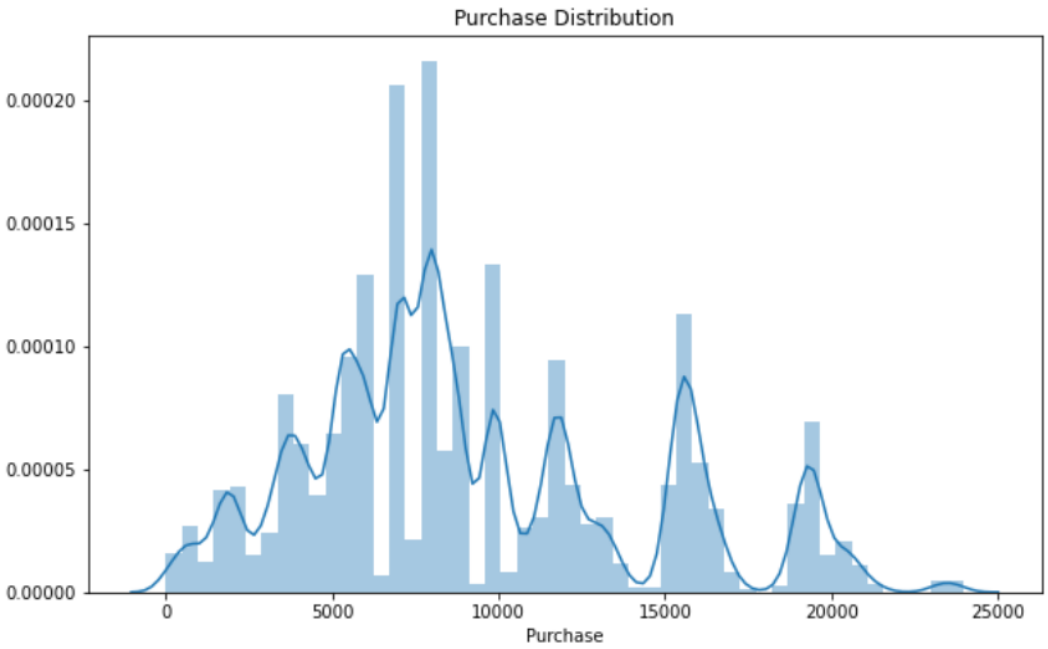
Insights:

1. Age group 26-35 years has highest percentage of customers who have spent on Black Friday sale i.e almost around 40%.

2. Age group 0-17 years has lowest percentage of customers who have spent (3%). And it is reasonable as teenage customers are less probable to have income.

3. Age group 18-25 & 36-45 years has average percentage of customers who spent on sale (around 20%).

4. Customers with Age above 45 years has observed to have decreasing percentage of customers as trend.

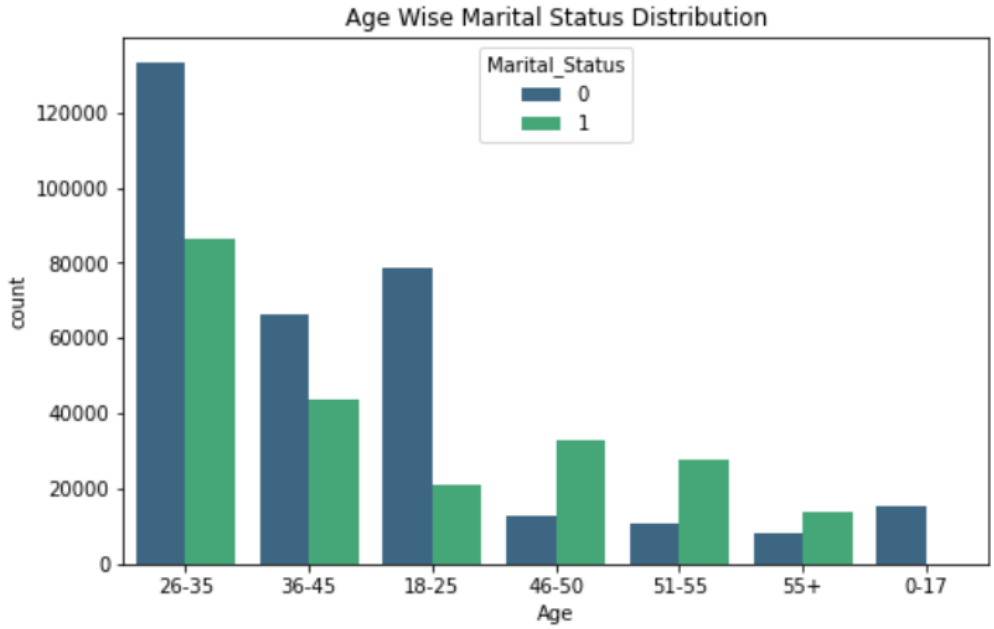


Insights:

1. The important observation form the above visualisation can be made that there are some outliers present in the dependent/target feature "Purchase".

2. According to distplot data is nearly normally distributed.

- ## Bivariate Analysis:



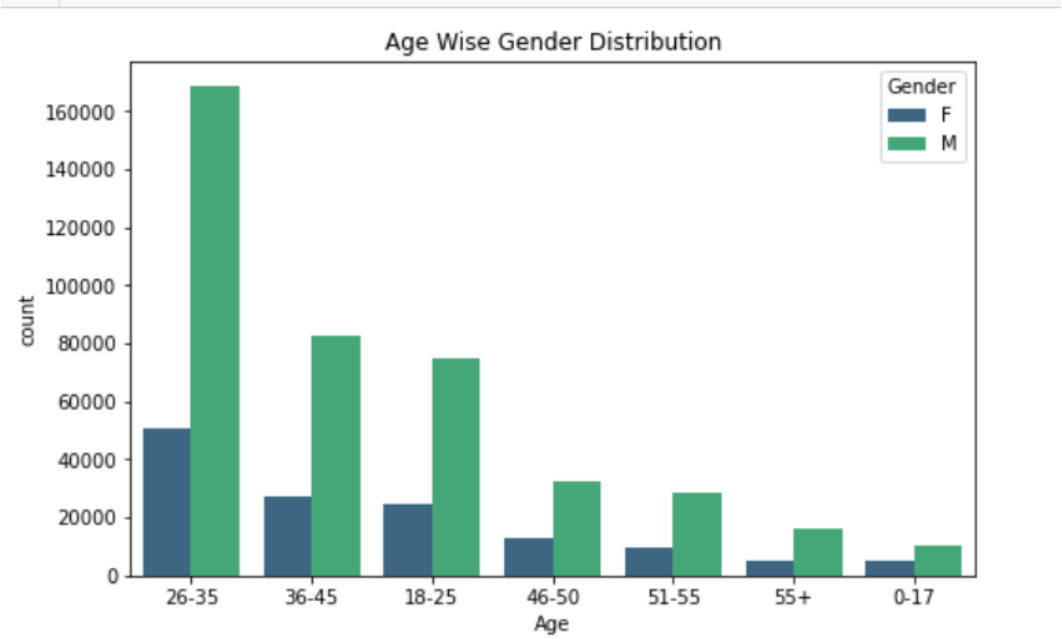
Insights:

1. Age group 0-17 years has all the single customers.

2. Age group 18-25 & 36-45 years has high single ratio than married.

3. Age group 26-35 years has highest ratio of both being single and married customers.

4. As age group is getting increased the ratio of being single is reduced. For example, 46-50, 51-55 & 55+, etc.



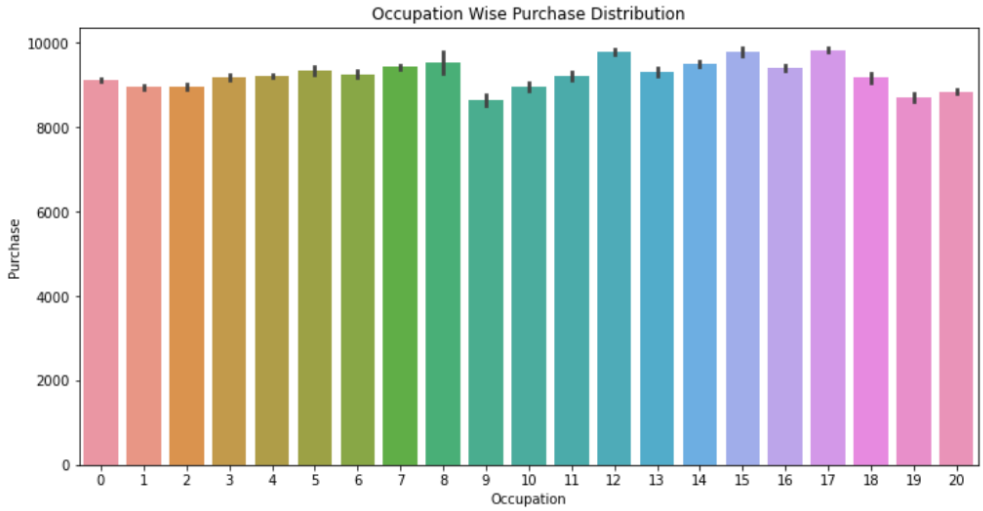
Insights:

1. In each Age group Male customers are dominating in spending/purchase in Black Friday Sale.

2. Age group 26-35 years has highest number of customers, whereas group 18-25 & 36-45 years has average number of customers.

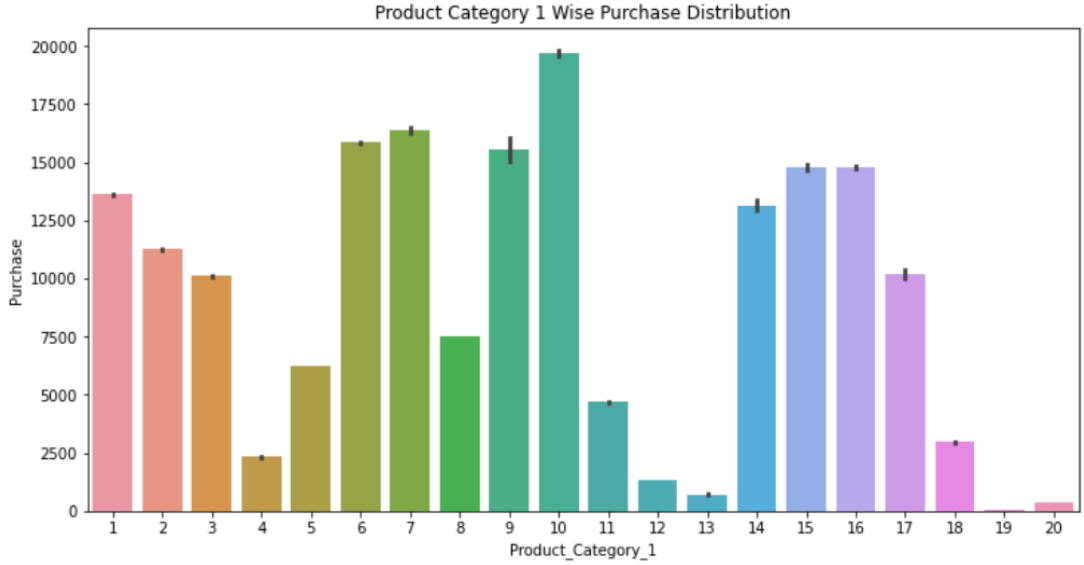
3. Age groups 0-17 & 55+ years has lowest number of customers.

4. Less number of customers are witnessed in age groups 46-50 & 51-55 years to purchase in Sale.



Insight:

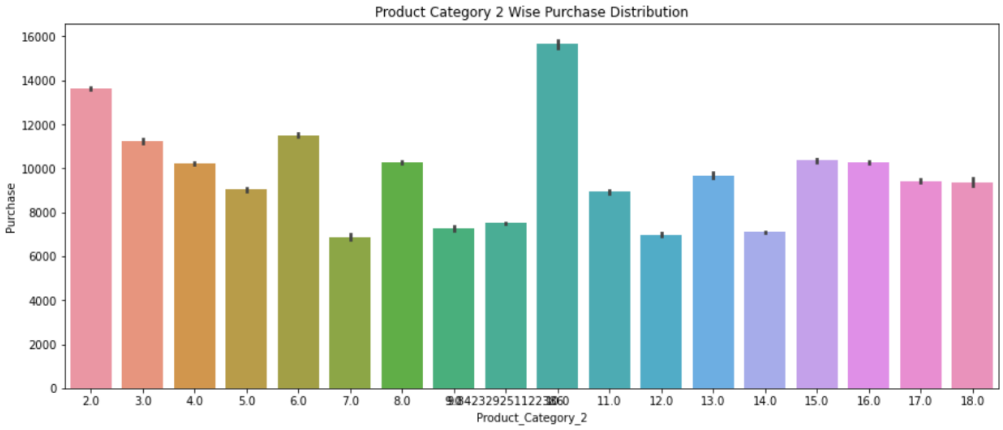
1. Almost all of the Occupation categories have spent 8000-10000 in Black Friday sale.



Insights:

1. Product category 10 has highest Purchase happened in Black Friday Sale.

2. Other Categories has a dispersed sale.



Insights:

1. Product category 10 has highest Purchase happened in Black Friday Sale.

2. Other Categories has a dispersed sale.

3. We can't make any solid statement from above visualizations.

- ## Multivariate Analysis:

Marital_Status	0							1						
Age	0-17	18-25	26-35	36-45	46-50	51-55	55+	18-25	26-35	36-45	46-50	51-55	55+
Gender														
F	0.924068	3.337224	5.468051	3.026717	0.575565	0.650829	0.346866	1.140041	3.758444	1.912673	1.823956	1.147858	0.577201
M	1.821411	10.941738	18.764589	9.040337	1.731422	1.319655	1.086229	2.698757	11.928889	6.020165	4.177302	3.880975	1.899038

Insights:

1. In all unmarried customers, Age groups 18-25 & 26-35 years has highest percentage of customers in both genders Male & Female.

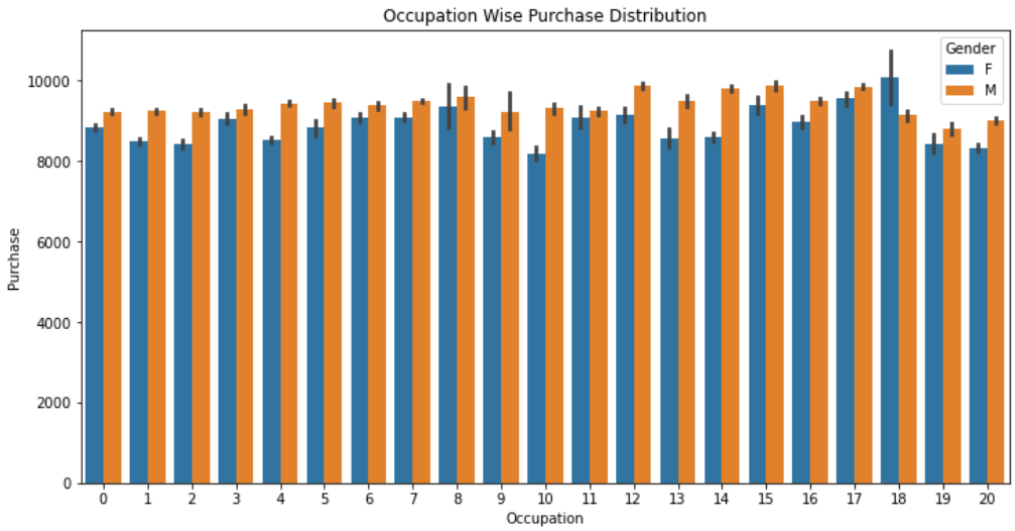
2. In all Male married customers, Age groups 18-25 & 26-35 years has highest percentage of customers.

3. In all Female married customers, Age groups 26-35 & 36-45 years has highest purchase customers.

	Gender	F											
	Age	0-17	18-25	26-35	36-45	46-50	51-55	55+	0-17	18-25	26-35	36-45	46-50	51-55
City_Category	Marital_Status													
A	0	0.263058	0.861348	1.816866	0.853894	0.048358	0.085989	0.038541	0.199430	3.100526	6.613728	1.822684	0.394497	0.237243
	1	0.000000	0.278329	1.362922	0.437764	0.178887	0.237243	0.027633	0.000000	0.765542	3.613008	1.724514	0.761179	0.548296
B	0	0.284510	1.557989	2.550594	1.209305	0.269421	0.327778	0.077263	0.703549	4.561800	7.664689	4.268200	0.612833	0.539751
	1	0.000000	0.566475	1.346561	0.810445	0.894798	0.443582	0.168343	0.000000	1.175855	5.087735	2.365162	1.932670	1.914127
C	0	0.376499	0.917887	1.100591	0.963517	0.257786	0.237062	0.231062	0.918432	3.279413	4.486173	2.949454	0.724092	0.542660
	1	0.000000	0.295236	1.048961	0.664463	0.750271	0.467033	0.381226	0.000000	0.757361	3.228146	1.930489	1.483453	1.418552

Insights:

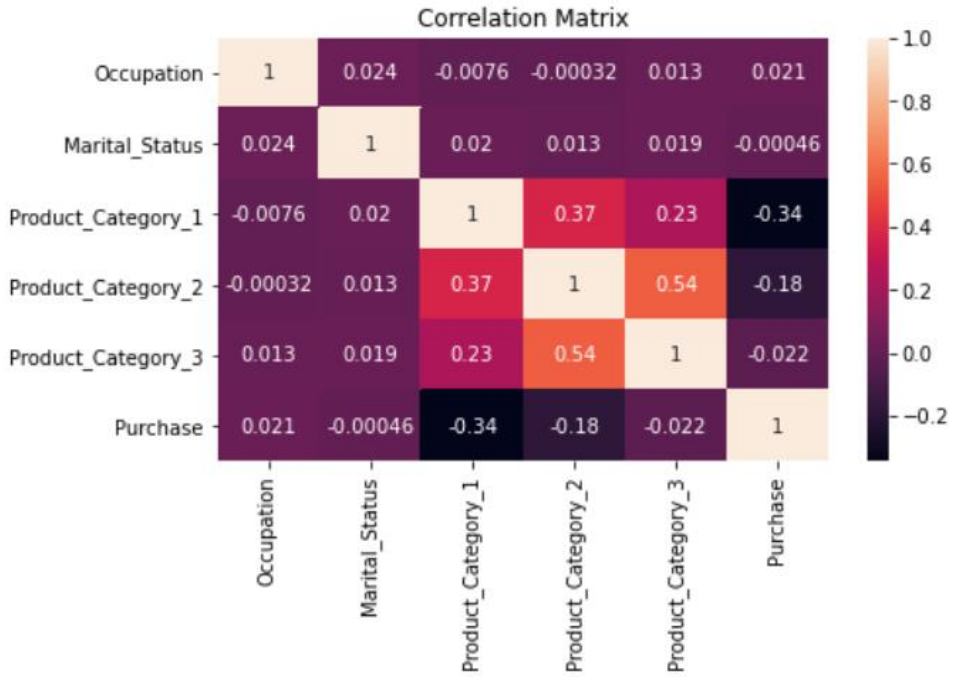
1. City B has highest percentage of Female customers in Age group 26-35 years in both Married and Unmarried marital status.
2. City B has highest percentage of Male customers in Age group 26-35 years in both Married and Unmarried marital status.
3. The above tabular representation shows that, in all the city categories A, B & C, the customers who have purchased/spent on Black Friday Sale always have high percentage of Unmarried customers irrespective of their Gender, Age groups.
4. Hence, we can say that Unmarried customers are more tend to spend in Black Friday sale.



Insights:

1. In the Occupation level 18, Female customers has high purchase amount compared to Male customers.

2. The purchase amount of all the Occupation levels is almost in the range of 8000-10000, which indicates Occupation levels has no greater impact on the purchase in Black Friday.



Insights

1. Occupation levels has no impact/correlation on independent as well as dependent features. Occupation feature can be dropped based on the various performance comparisons of the Model.
2. Product Category 3 has no correlation with target feature (Purchase).Also it has 70% of missing values, Hence, Product Category 3 feature can be dropped.
3. Product Category 1 & Product Category 2 has strong negative correlation with target feature(Purchase). Which means the change of value in one feature varies with change of value in other feature. This is called as Inverse correlation. In other words, If the value of Purchase Category 1 increases, that will result in reduction in Purchase Value.

c) Model Used:

Since I'll be making many models, instead of repeating the codes again and again, I would like to define a generic function which takes the algorithm and data as input and makes the model, performs cross-validation and generates submission.

Linear Regression Model

Regression is to examine two things: (1) does a set of predictor variables do a good job of predicting an outcome (dependent) variable? (2) Which variables, in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The

simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Random Forest Model

A random forest is an estimator that fits a number of classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True. The idea behind this technique is to decorrelate the several trees. It generates on the different bootstrapped samples(i.e. self-generated samples) from training Data. And then we reduce the Variance in the Trees by averaging them. Hence, in this approach, it creates a large number of decision trees in python or R. The random forest model is very good at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. Random forests have the ability to capture the non-linear interaction between the features and the target.

XGBoost Model

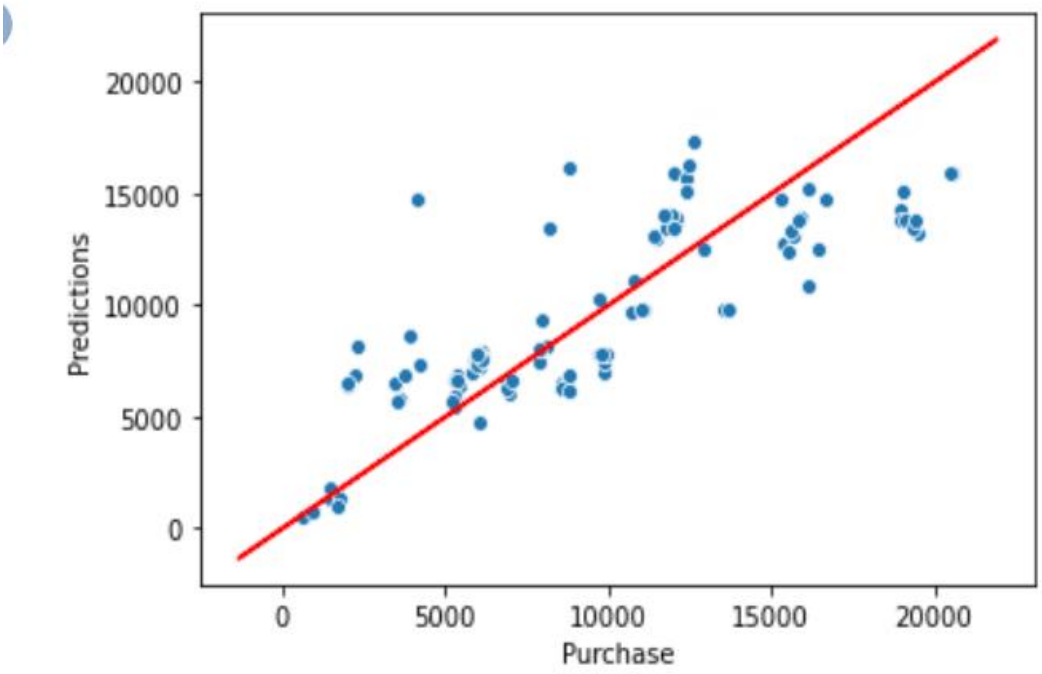
Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Boosting can be interpreted as an optimization algorithm on a suitable cost function. The latter two papers introduced the view of boosting algorithms as iterative functional gradient descent algorithms. That is, algorithms that optimizes a cost function over function space by iteratively choosing a function (weak hypothesis) that points in the negative gradient direction. This functional gradient view of boosting has led to the development of boosting algorithms in many areas of machine learning and statistics beyond regression and classification. Gradient boosting combines weak "learners" into a single strong learner in an iterative fashion. It is easiest to explain in the least-squares regression setting, where the goal is to "teach" a model F to predict values of the form y ^ = F (x) by minimizing the mean squared error.

- **Discussion and Results:**

a) Output obtained:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase	Predictions
2	1000001	P00087842	F	0-17	10	A	2	0	12	9.842329	NaN	1422	1348.173950
12	1000005	P00031342	M	26-35	20	A	1	1	8	9.842329	NaN	6073	7207.742676
14	1000006	P00231342	F	51-55	9	A	1	0	5	8.000000	14.0	5378	6909.993164
22	1000008	P00213742	M	26-35	12	C	4+	1	8	9.842329	NaN	9743	7862.944824
23	1000008	P00214442	M	26-35	12	C	4+	1	8	9.842329	NaN	5982	7862.944824
...	...	...	...	...	...	...	...	...	...	...	...	...	...
550044	1006004	P00370853	F	26-35	15	C	2	0	19	9.842329	NaN	62	245.385040
550047	1006009	P00372445	F	26-35	12	C	3	0	20	9.842329	NaN	244	381.417999
550048	1006010	P00371644	M	36-45	0	C	1	0	20	9.842329	NaN	591	420.123199
550051	1006013	P00375436	F	26-35	20	C	3	0	20	9.842329	NaN	489	381.417999
550064	1006035	P00375436	F	26-35	1	C	3	0	20	9.842329	NaN	371	381.417999

110014 rows × 13 columns



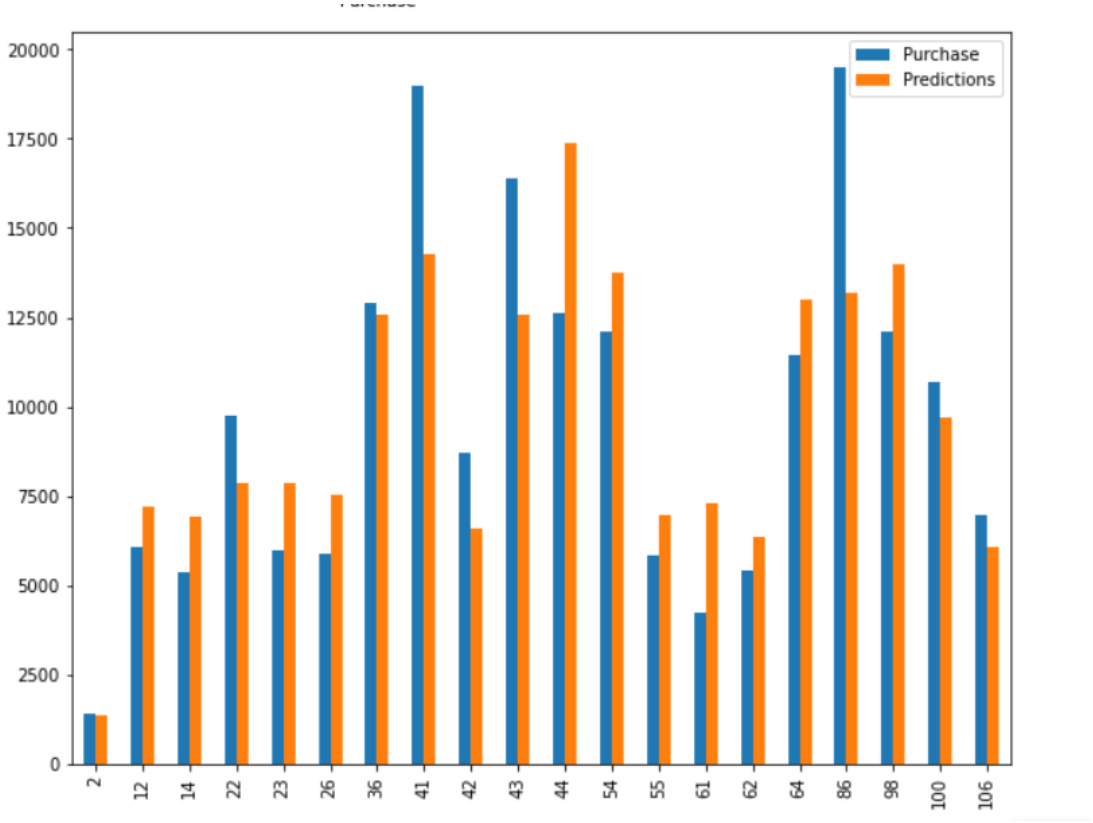


Fig. Purchase Vs Predicted

The Blue bar represents Actual purchase value and orange bar represents Predicted purchase value. We can see that most of the Actual and predicted values are close. And model is Performing well and not overfitting.

b) Evaluation measures used:

Performance Metrics Comparison:

RMSE: 4694.56

MAPE: 1.11

R2 Score: 12.63

RMSE: 3015.32

MAPE: 0.34

R2 Score: 63.96

RMSE: 2972.35

MAPE: 0.35

R2 Score: 64.98

By above Performance Metrics, Linear Regression gives R2 Score as 12.63, Random Forest Regressor gives 63.96 and XGBoost Regressor gives R2 Score as 64.98. Hence, we conclude that XGBoost Regressor is giving the highest accuracy among other algorithms. We are selecting XGBoost Regressor as best suited for our dataset.

Best model selected for prediction – XGBoost Regressor**.**

Conclusion:

The ML algorithm that performs the best was XGBoost Regressor Model with R2 Score 64.98, MAPE is 0.35 and RMSE 2972.35, compared with other models.
We have implemented an algorithm that allows us to predict the purchase of the basket of a consumer knowing his gender, marital status, age, occupation, and number of years spent in the current city of residence. This algorithm is based on easily identifiable low-level indicators and can be used to estimate the purchasing potential of a consumer. In the visualization part, we were also able to draw the profile of the perfect consumer to target in priority marketing campaigns including (Black Friday or others). A targeted or retrospective analysis that takes into account all the indicators of the dataset used could complement this study and determine in particular the products generating the most sales.

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. https://www.ijert.org/data-analysis-and-price-prediction-of-black-friday-sales-using-machine-learning-techniques
4. http://www.ijirset.com/upload/2022/june/297_Black_NC.pdf
5. Barbaro, M. (2006, November 25). Attention, holiday shoppers: We have fisticuffs in aisle 2 [Electronic version]. The New York Times Late Edition.
6. Retrieved June 11, 2007 from LexisNexis Academic database. · Bellizzi, J. & Hite, R. (1992). Environment color, customer feelings, and purchase likelihood, Psychology and Marketing, 9, 347-363.
7. https://www.datacamp.com/community/tutorials/ml-black-Friday-dataset