KOÇ
ÜNİVERSİTESİ
GRADUATE SCHOOL OF
SCIENCES AND ENGINEERING

DASC591

Non-Thesis Master of Science in Data Science

Project Report:

E-Commerce Sales Prediction Based on Black Friday Customer Data

Ezgi Aydın Bozkürk

0078635

23/01/2023

TABLE OF CONTENTS

ABSTRACT

Sales prediction is an important field in the e-commerce industry, especially for the events such as Black Friday which gives a great opportunity for consumers to purchase items with discounted prices as well as creating an environment for retailers to increase their sales. In order to better understand the customer, and benefit from such opportunities, retailers should utilize sales prediction methods.

The purpose of this project was to build machine learning algorithms that analyze consumer purchase behavior based on an e-commerce sales dataset. Black Friday sales data which is a public dataset gathered from Kaggle, consists of nearly 550K observation with 11 features which are both qualitative and quantitative. Since the predictor label (purchase) is continuous, regression models were suited for the purpose. Six different regression models were trained: Linear regression, Decision tree regression, Random Forest regression, K-Nearest Neighbors regression, Extra tree regression and Extreme Gradient Boosting regression.

For data pre-processing, data cleaning, dimensionality reduction, feature encoding and scaling was performed. Root Mean Squared Error (RMSE) and R^2 (R-Squared) were used as model performance evaluation metrics. For the hyperparameter tuning Randomized Search cross validation was performed. The hyperparameter tuned Extreme Gradient Boosting regressor performed a minimum score of 2907 for RMSE and 0.66 R-Squared score, which means that for 66% of the time the model predicts accurately, therefore the hyperparameter tuned Extreme Gradient Boosting regressor outperformed Extreme Gradient Boosting regressor which resulted scores of 2916 RMSE and 0.66 R-Squared. As a result, even though both tuned and the original version of Extreme Gradient Boosting regressor performed similarly, the hyperparameter tuned Extreme Gradient Boosting regressor performed better than the other algorithms.

1.INTRODUCTION

Originated in Philadelphia around the year 1961, Black Friday is a huge event for especially e-commerce industry, in which online retailers run special offers, deals, and discounts on countless products. While the sales number are becoming massive, it becomes almost impossible to track, reach out and analyze manually. In such cases, big data analytics and sales prediction comes in aid with benefits. The process of sales prediction refers to estimation of future sales by the prediction of purchase amount based on the customer data in a fixed time period. Sales prediction is an important field in the e-commerce industry, especially for the events such as Black Friday which gives a great opportunity for consumers to purchase items with discounted prices as well as creating an environment for retailers to increase their sales. This creates large number of sales transactions and in order to better understand the customer, to be prepared and benefit from such opportunities, retailers should utilize sales prediction methods.

This project aims to build a prediction model that can predict purchases based on various attributes of the customers. For the project purposes, Black Friday Sales dataset which is a public dataset extracted from Kaggle, containing around 550K observation with 11 features which are both qualitative and quantitative, was used for training and testing. The dataset provides demographic information of customers and product information such as category information. Using this dataset, several machine learning algorithms were trained in order to create a machine learning model that can predict the purchase amount. The process consisted of descriptive and exploratory data analysis, pre-processing, data modeling, model evaluation, hyperparameter tuning and finally the analysis of the results. The regression models used were Linear regression, Decision tree regression, Random Forest regression, K-Nearest Neighbors regression, Extra tree regression and Extreme Gradient Boosting regression. Root mean squared error and R-squared values were used to evaluate the performance of the models. Based on these performance metrics, tuned Extreme Gradient Boosting regressor outperformed the other models which resulted scores of 2907 RMSE and 0.66 R-Squared.

The following sections of this project report will cover related work, details of approach and implementation, evaluation, results and discussion. First, some insights from related work in the field will be shown in related work section, then Approach will be provided in detail, including data description, pre-processing, feature extraction, learning approach and structure, and hyperparameter selection. Next section will cover implementation details such as environment used, and following implementation, evaluation, results and discussion will be covered. Lastly, summary of the results, strength and weaknesses of the approach will be covered under conclusions section.

2.RELATED WORK

E-commerce platforms have more data than ever before, which creates ton of instances to work on. Additionally, ecommerce is a field where machine learning is practiced often, in order to optimize customer segmentation, targeted ads, personalized recommendations, churn prediction, search optimization and many more. Therefore, e-commerce datasets attract much attention from researchers, due to their vast opportunities. In this section related works of researchers will be covered.

In Data Analysis and Price Prediction of Black Friday Sales using Machine Learning Techniques paper published by International Journal of Engineering Research & Technology, Amruta Aher et al [1] used linear, lasso, ridge, Decision Tree, and Random Forest Regressor for analysis and prediction on Black Friday Sales Dataset. For performance evaluation measure, Mean Squared Error was used and based on the MSE score Random Forest Regressor provided the best performance with a MSE score of 3062.72. As a result, Random Forest Regressor is proposed to be the model that predicts customer purchase. Lastly, the author argues hyperparameter tuning and training alternative algorithms as a future work.

Singh, K et al [2] provided detailed analysis and visualization of a similar sales dataset gathered from Kaggle to help decision makers via providing user friendly visuals of the complex data. The data visualization is based on different parameters and dimensions. The authors provide a framework for analysis of real time sales data to forecast, visualize, extract insights or patterns for future sales. Facebook Prophet Algorithm for Sales conjecture was used for insights, Plotly Express visualizing tool used for Sales visualizations. Kaggle sales dataset is used to train and model building with Prophet Package for prediction considering the components of the database constant at the time of prediction. Authors also argue, usage of larger dataset and multiple forecasting tools for future work.

Ramasubbareddy,S et al [3] also worked on Black Friday dataset from Kaggle, focusing on building a prediction model based on an exact and proficient algorithm to analyze the clients purchases before and yield the future spending of the client with similar features. Regressors, NN and classifiers were used for the model, and evaluation was performed on six different algorithms, based on exhibition and exactness of predicting. The algorithms used in this paper was linear regression, Ridge Regression, XGBoost, Decision Tree, Random Forest, and Rule-Based Decision Tree, and the main performance metric is RMSE. Author concludes the best performing algorithm as Rule-Based Decision Tree which yielded the lowest RMSE score of 2291.

Another widely used public dataset Walmart Sales Data was used in the paper Sales Prediction Using Machine Learning Algorithms by Purvika Bajaj et al. [4]. In this paper, authors performed sales prediction of a grocery store data including features like item weight, item fat content, item visibility, etc. Several algorithms were used: Linear Regression, K-Nearest Neighbors algorithm, XGBoost, and Random Forest.RMSE, variance score, training, and testing accuracies were used as main performance measures and with a 93% accuracy score Random Forest algorithm outperformed the other algorithms.

C. M. Wu et al. [5], trained a prediction model to predict future spendings based on past spendings of the customers. The algorithms used in this work are Linear Regression, MLK classifier, Deep learning model using Keras, Decision Tree, and Decision Tree with bagging, and XGBoost. Usage of keras, MLK classifier and deep learning approach differentiates this particular work from current Project. Performance of the algorithms were evaluated based on RMSE and the author further argues the data pre-processing, visualization techniques and usage of complex neural network models.

Ramachandra, H et al.[6], mention retail industry as the most common application domain for machine learning. The authors mention understanding the purchase behavior of customers against different product categories and demographics, would benefit the retailers in areas such as inventory management, financial planning, advertising and marketing. Based on the RMSE score and accuracy rate of 83.6%, this paper proposes Random Forest regressor as the winner algorithm for Black Friday dataset.

3.APPROACH

The aim was to build a model that can predict future purchases based on customer historical data. For this, a public dataset is first gathered from Kaggle website. Then preliminary descriptive analysis was performed. Following that, exploratory data analysis was performed. Lastly, Preprocessing was done before model training.

Black Friday sales data is a public sales transaction dataset gathered from Kaggle website. The dataset is provided as train and test set. Respectively consisting of 412551 instances and 11 features for train, and 137517 instances and 11 features for test set. The features are 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 and target variable/predictor label is Purchase. The dataset contains both categorical and numerical features. Overview of Data sets are shown below.

TRAIN SET

Column1	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
0	1000001	P00069042	F	0-17	10	A	2	0	3	NaN	NaN	8370
1	1000001	P00248942	F	0-17	10	A	2	0	1	6.0	14.0	15200
2	1000001	P00087842	F	0-17	10	A	2	0	12	NaN	NaN	1422
3	1000001	P00085442	F	0-17	10	A	2	0	12	14.0	NaN	1057
4	1000002	P00285442	M	55+	16	C	4+	0	8	NaN	NaN	7969

After importing the data set to a python environment, shape, columns, info and description of the dataset was observed. Statistical overview was conducted in order to further understand the numerical features. Inspecting the null values and their percentage values are important while handling a dataset. So, null values and their volumes were observed. As a result of this observation, it is seen that Product_category_3 has almost 70% null values, which indicates that it will be performing poorly while training the model since it will not be giving many information.

	count	mean	std	min	25%	50%	75%	max
User_ID	412551.0	1.003028e+06	1727.412174	1000001.0	1001510.5	1003074.0	1004478.0	1006040.0
Occupation	412551.0	8.077668e+00	6.523470	0.0	2.0	7.0	14.0	20.0
Marital_Status	412551.0	4.098596e-01	0.491808	0.0	0.0	0.0	1.0	1.0
Product_Category_1	412551.0	5.402185e+00	3.929854	1.0	1.0	5.0	8.0	20.0
Product_Category_2	282287.0	9.853628e+00	5.085291	2.0	5.0	9.0	15.0	18.0
Product_Category_3	124956.0	1.266545e+01	4.128713	3.0	9.0	14.0	16.0	18.0
Purchase	412551.0	9.262248e+03	5019.582812	12.0	5824.0	8046.0	12052.0	23961.0

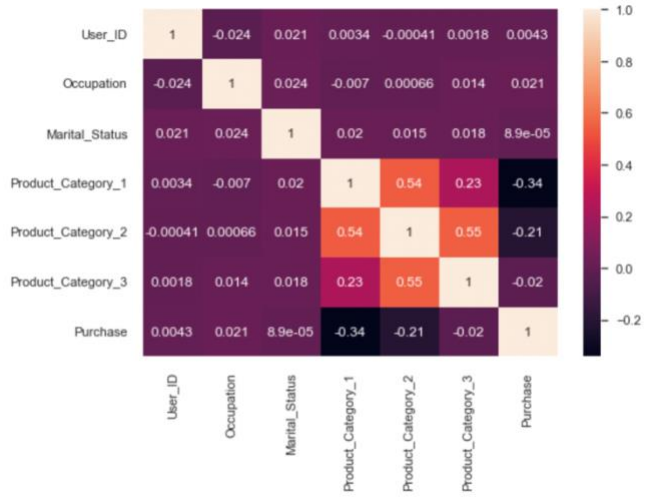
NULL VALUES COUNT - PERCENTAGE

User_ID	0	0.0 %
Product_ID	0	0.0 %
Gender	0	0.0 %
Age	0	0.0 %
Occupation	0	0.0 %
City_Category	0	0.0 %
Stay_In_Current_City_Years	0	0.0 %
Marital_Status	0	0.0 %
Product_Category_1	0	0.0 %
Product_Category_2	130264	31.58 %
Product_Category_3	287595	69.71 %
Purchase	0	0.0 %

dtype: int64

Unique values and value counts of the columns were observed in order to understand the distribution of the dataset. As a results, it is observed that majority of the Age group is within 26-35 of ages, for occupation majority of percentage is taken by 4, city category B has the maximum percentage with 42% of the column. The number of males shopping during the Black Friday is much higher than females in this dataset. Marital status is evenly distributed, while 35% of the city in the current year column is 1 year. Product categories distribute greatly but the majority groups are 5,1 and 8. Lastly, purchase appears to be normally distributed with a right/positive skew. Distribution of features table is provided in Appendix 1.

Numerical analysis was followed with plotting in order to better understand the distributions of the features while performing the univariate exploratory data analysis. Plots and visual representations of the features are represented under Appendix 2.



After completing the univariate analysis, bivariate observations were made to understand the relationship and distribution of features with purchase. Results have shown that, purchase amount across occupation, occupation groups 12,15,17 are in the lead. Age wise purchase amount is slightly higher for the 51-55 age group, and gender wise male's purchase amount is higher than the females. As for the city category, people from city C made the largest amount of purchases and the products that are most purchased are from the product group 10. Amount wise, marital status did not reveal any observable difference. Additionally multivariate analysis was conducted via creating pair plots for all features, and a correlation matrix was illustrated with heatmap. To observe any correlation within the features. Based on the correlation scores on the heatmap, it appears to be there is a 0.55 correlation between Product categories 1 and 2. However other than that, correlation scores appear to be low. Heatmap is shown above. Bivariate plots are provided under Appendix 3.

At the pre-processing step, first the values were organized by removing '+' signs from Age and Stay in Current City years columns to be able to treat these columns numerically. As previously mentioned, Product Category 3 column includes almost 70% null values, therefore would be a poor feature, in order to avoid poor learning rate Product Category 3 column were dropped along with the product ID and User ID columns which were also irrelevant for the purposes of this model. Null values on Product Category 2 column were filled with the median values, and data type of Stay in current city years column were changed to integer. In order to normalize the data and transform non-numerical labels such as occupation, product category and city category, one hot encoding/get dummies was used. Gender was turned into 1 and 0 values. Since the age groups were ordinal, guided encoding was performed. After completing the above-mentioned steps, purchase column is assigned as the y value which is our target value while rest of the columns left assigned to X.

Feature selection is performed based on domain knowledge and previously performed exploratory data analysis. Encoded columns were kept while the original categorical columns were dropped. Product Category 3 column was dropped. Lastly with the remaining features, a feature scaling was performed with StandarScaler which is a widely used scaling method, in order to avoid bias towards features with higher values/magnitudes. Scaling provides features to distribute in the same range enabling model to train on features better. After the scaling step, the above-mentioned manipulations were also performed on the test set.

Learning Approach and the System Structure:

i) Performance Analysis:

Performance of a regression algorithm is computed via the results of training process. It gives an indicator of how did the model learned and performed on the dataset. This performance can be measured by Root Mean Squared Error (RMSE) and R-Squared (R^2) scores. RMSE is one of the most commonly used measure for evaluation. As the math behind it shown in the figure below, it calculates the square root of mean difference between prediction and the actual point in order to show how far predictions fall from the true values based on Euclidian distance. Therefore, a model with a lower RMSE score is considered to be better. Second metric R^2 refers to the closeness of the data points to the fitted regression, this score ranges between 0 and 1 and as the score gets higher the model is considered to be better fitted. In this Project, RMSE and R^2 scores are used to evaluate the performance of models and lower RMSE and higher R^2 scores are targeted.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

$$\text{R2 Squared} = 1 - \frac{SSr}{SSm}$$

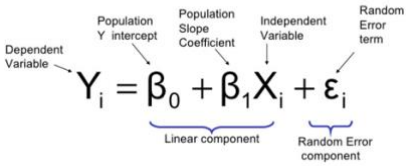
SSr = Squared sum error of regression line

SSm = Squared sum error of mean line

ii) Machine Learning Algorithms Used:

Linear Regression

In machine learning Linear Regression is a widely used algorithm. It is used to establish a linear relationship between dependent and independent variables. The algorithm can be stated as shown below where y stands for dependent variable, x is for independent variable B0 stands for intercept B1 as slope coefficient and E for error. The algorithm works on finding the best fit line between dependent and independent variables of the dataset by optimizing the coefficients with minimum error.



Decision Tree Regression

Decision tree algorithm is another popularly used algorithm for supervised learning. It works with a tree structure via breaking down the dataset in subsets, creating a tree with decision and leaf nodes. Sum of squared error or variance is used to evaluate impurity of the model in regression whereas Gini index is used in classification problems. It could take parameters such as max/min depth, features, leaf nodes, impurity, samples split in order to optimize the algorithm as fit.

Random Forest Regression

Based on Ensemble learning, Random Forest is an algorithm that acts as a collection of decision trees. This collection method, provides the usage of multiple classifiers in order to improve performance, so as the number of trees included increases accuracy improves and

provides a better chance to avoid overfitting. Algorithm takes similar parameters as decision tree algorithm like min max depth, leaf, impurity, and features. Additionally takes parameters like n estimators, max samples, parameter in order to control the sub-samples. The hyperparameter for the algorithm is number of decision trees to be considered.

KNN Regression

K nearest neighbors algorithm is a supervised learning algorithm that is considered to be an effective algorithm for large datasets. It is a non-parametric and nonlinear algorithm. It works based on distances in which it classifies instances based on proximities under neighbor classes. Since it is non parametric, while training it is not possible to fine tune the parameters. The algorithm takes parameter k which refers to the number of neighbors. A low number of k can create noise and represent the effects of outliers largely which results in poor prediction; therefore, a larger k value would yield better prediction and avoid over-fitting. The most commonly used distance metric for the algorithm is Euclidian distance:

$$d(x,\ y)\ =\ \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Extra Tree Regression

Extra tree algorithm, which is short for extremely randomized trees is ensemble supervised learning algorithm, that is similar to random forest but considered to be faster in process. It creates a meta-estimator fitting a group of randomized decision trees that are sampled randomly, uses averaging to enhance accuracy and avoid over-fitting. The difference of the model comes from its randomness while selecting number of features and splitting points, which creates unique and uncorrelated trees. The algorithm takes similar parameters as random forest such as n estimator, max depth, min splits and leaf.

XGB Regression

Extreme gradient boosting algorithm is considered a powerful supervised learning model. It is based on decision tree with improved qualities, and considered to be working well with large datasets. The cost function of XGB regressor is Mean Square Error. Gradient boosting is an ensemble machine learning algorithm, in which trees are added and fit to correct the predictive errors of prior models. The gradient descent optimization algorithm and loss function are used while model fitting, and as the gradient loss is minimized model performance maximizes. The algorithm takes general parameters, boosting parameter and learning task parameters which could be used to tune hyperparameters. Max depth, eta which refers to learning rate, subsample, colsample bytree, bylevel, bynode, gamma are some of the parameters that the algorithm takes. For this project XGB regressor first trained with the default parameters and 150 n_estimators, afterwards several attempts of hyperparameter tuning were performed with varying learning rates, subsamples, n_estimators and max dept values. Hyperparameter tuning was performed with RandomizedSearchCv. Best parameters were as follows: {'min_child_weight': 61, 'max_depth': 17, 'learning_rate': '0.5','gamma': 0.0,'colsample_bytree': 0.4}

In this project linear regression algorithm is taken as a baseline which resulted in 2999 RMSE and 0.64 R2 score. Additional to linear regression a simpler approach was also taken for baseline with dummy regressor method which scored 5033 RMSE and -1.8691 R^2.

4.IMPLEMENTATION DETAILS

Black Friday Sales dataset was gathered from Kaggle website. Dataset consisted of 412551 instances and 11 features for train, and 137517 instances and 11 features for test set. The features were preprocessed with above mentioned step. For descriptive statistics and exploratory analysis NumPy and Pandas libraries were used. For visualization Seaborn library and Matplotlib module were used. The proposed model was implemented using Python 3.9.7 and SKlearn (Scikit-learn) 0.24.2. Six different regression models were trained: Linear regression, Decision tree regression, Random Forest regression, K-Nearest Neighbors regression, Extra tree regression and Extreme Gradient Boosting regression. Feature encoding was performed with Label Encoder from sklearn preprocessing. Feature importances were calculated via calling several model feature importances. Feature scaling was conducted using StandartScaler from sklearn preprocessing. Models were evaluated based on RMSE and R^2 scores. For hyperparameter tuning Randomized Search Cross validation was used on the highest scorer XGBoost Regression algorithm.

5. EVALUATION & RESULTS & DISCUSSION

	RMSE	R^2
Linear Regression	2999,58	0,64
Decision Tree Regression	3141,53	0,61
Random Forest Regression	3573,09	0,49
K-Nearest Neigbours	3022,45	0,63
Extra tree regressor	3087,68	0,62
Extreme Gradient Boosting	2916,68	0,66
Tuned Extreme Gradient Boosting	2907,28	0,66

For this project, a model using various algorithms were trained in order to get best possible predictions. Since the predictor label (purchase) is continuous, regression models were suited for the purpose. Algorithm evaluations were made based on Root Mean Squared Error and R-Squared scores. As shown on the above table, Extreme Gradient Boosting has the lowest RMSE score combined with the highest R^2 score, therefore outperformed the other models.

Linear regression was the worst performer on the dataset based on its RMSE and R^2 scores. Tuned Extreme Gradient Boosting has score performance measures to original XGB model, however it failed to outperform after hyperparameter tuning.

The dataset was split randomly as 0.75 for train 0.25 for test. Several steps of preprocessing were performed including feature selection. Changes in the methods of preprocessing steps or approaching the feature selection/reduction differently might have provided different results. Additionally, for obtaining the best parameter and hyperparameter tuning Randomized Search cross validation method was used. For future work, research on alternative applications of cross validation techniques is needed. Also, testing different scaling methods would be beneficial.

For the purposes of the project, many different algorithms were trained. As a discussion smaller number of models but with multiple approaches could have been trained for several times in order to optimize parameter and yield better performing model.

6.CONCLUSION

Black Friday sales prediction model was trained and evaluated in this project. Based on the RMSE and R^2 scores Extreme Gradient Boosting algorithm yielded the best results with the minimum score of 2907 RMSE and maximum score of 0.66 R^2. We tried to further optimize the model via hyperparameter tuning. To increase the scores even further, alternative tuning methods can be used to fine tune. Additionally, for future work, to obtain better results with such large datasets neural networks would be beneficial to use in order to build a predictive model. Additionally, to neural networks, alternative hyperparameter tuning methods can be applied in order to avoid failing to outperform the previous models.

7.REFERENCES

1. Amruta Aher, Rajeswari Kannan, Sushma Vispute, 2021, Data Analysis and Price Prediction of Black Friday Sales using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 07 (July 2021).

2. 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.

3. Ramasubbareddy, S., Srinivas, T., K, G. & Swetha, E. (2021). Sales analysis on back friday using machine learning techniques. Advances in Intelligent Systems and Computing, 1171:313–319. doi: 10.1007/978-981-15-5400-1_32

4. Purvika Bajaj1, Renesa Ray2, Shivani Shedge3, Shravani Vidhate4, Prof. Dr. Nikhilkumar Shardoor5,SALES PREDICTION USING MACHINE LEARNING ALGORITHMS",International Research Journal of Engineering and Technology (IRJET) ,Vol 7 Issue 6,2020,e-ISSN: 2395-0056 | p-ISSN: 2395-0072

5. 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.

6. Ramachandra, H.V., Balaraju, G., Rajashekar, A., & Patil, H. (2021). Machine Learning Application for Black Friday Sales Prediction Framework. 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 57-61.

7. Black Friday Sales Dataset Kaggle https://www.kaggle.com/kkartik93/black-friday-salesprediction?select=train.csv

8. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

9. Poornima, p., Jennifer Joyce, B., 2022, Black Friday Sales Prediction Analysis using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 11, Issue 06 (June 2022).
10. Karandeep Singh *et al* 2020 *J. Phys.: Conf. Ser.* **1712** 012042

8.APPENDIX

Appendix 1: Distributions of Features

Distribution of Features

Unique_values

```
train_df.nunique()

User_ID                       5891
Product_ID                    3588
Gender                           2
Age                              7
Occupation                      21
City_Category                    3
Stay_In_Current_City_Years       5
Marital_Status                   2
Product_Category_1              20
Product_Category_2              17
Product_Category_3              15
Purchase                     17535
dtype: int64
```

Age

```
26-35    39.93 %
36-45    20.03 %
18-25    18.09 %
46-50     8.29 %
51-55     7.02 %
55+       3.91 %
0-17      2.74 %
Name: Age, dtype: object
```

City Category

```
B    42.06 %
C    31.04 %
A    26.89 %
Name: City_Category, dtype: object
```

Marital_Status

```
0    59.01 %
1    40.99 %
Name: Marital_Status, dtype: object
```

Stay_In_Current_City_Years

```
1     35.23 %
2     18.54 %
3     17.33 %
4+    15.38 %
0     13.53 %
Name: Stay_In_Current_City_Years, dtype: object
```

Occupation

```
4     13.13 %
0     12.68 %
7     10.77 %
1      8.61 %
17     7.31 %
20     6.09 %
12     5.67 %
14     4.93 %
2      4.84 %
16     4.65 %
6      3.69 %
3      3.22 %
10     2.33 %
5       2.2 %
15      2.2 %
11     2.11 %
19     1.54 %
13     1.42 %
18     1.19 %
9      1.15 %
8      0.28 %
Name: Occupation, dtype: object
```

Gender

```
M    75.36 %
F    24.64 %
Name: Gender, dtype: object
```

Product_category_1

```
5     27.48 %
1      25.5 %
8     20.76 %
11     4.42 %
2      4.35 %
6      3.72 %
3      3.66 %
4      2.12 %
16     1.78 %
15     1.13 %
13      1.0 %
10     0.93 %
12     0.72 %
7      0.68 %
18     0.56 %
20     0.46 %
19     0.29 %
14     0.28 %
17     0.11 %
9      0.07 %
Name: Product_Category_1, dtype: object
```

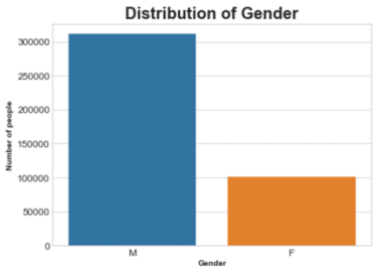
Product_category_2

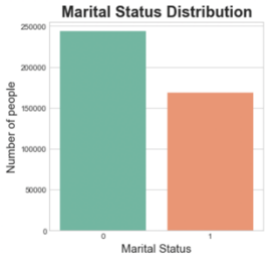
```
8.0     17.0 %
14.0    14.69 %
2.0     13.03 %
16.0    11.47 %
15.0    10.12 %
5.0      6.95 %
4.0      6.83 %
6.0      4.36 %
11.0     3.76 %
17.0     3.55 %
13.0     2.78 %
9.0      1.51 %
12.0     1.48 %
10.0     0.81 %
3.0      0.75 %
18.0     0.74 %
7.0      0.17 %
Name: Product_Category_2, dtype: object
```

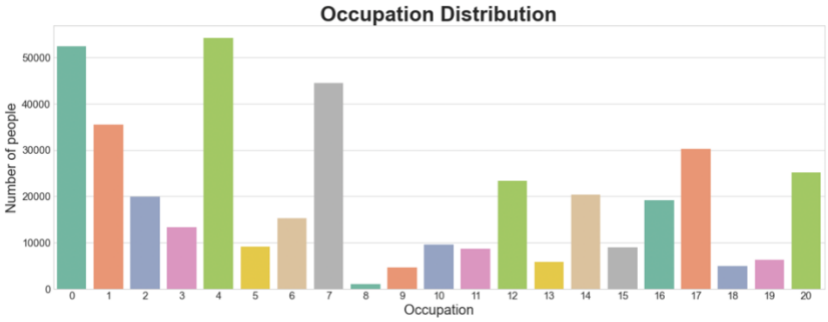
Product_category_3

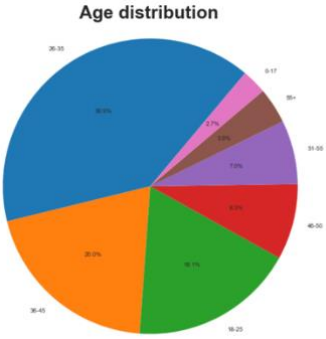
```
16.0    19.6 %
15.0    16.75 %
14.0    11.01 %
5.0     10.04 %
17.0    10.02 %
8.0      7.56 %
9.0      6.98 %
12.0     5.51 %
13.0     3.26 %
6.0      2.91 %
18.0     2.79 %
4.0      1.12 %
11.0     1.08 %
10.0     1.02 %
3.0      0.36 %
Name: Product_Category_3, dtype: object
```

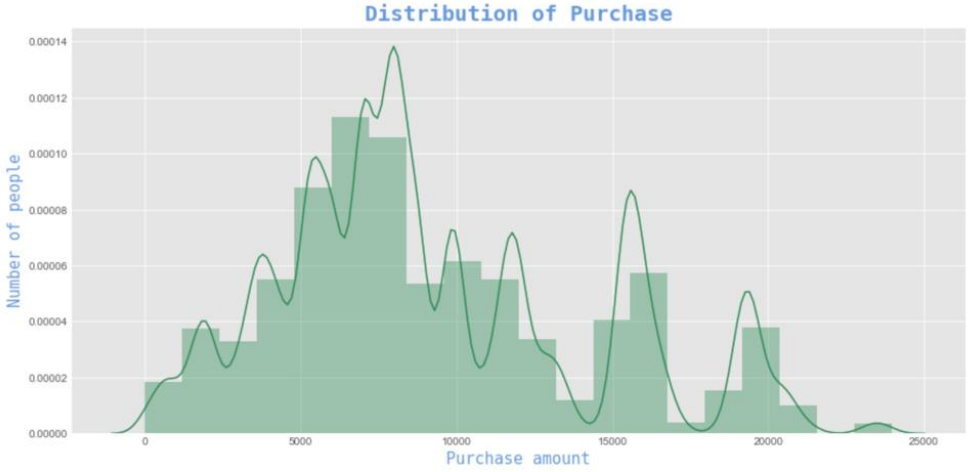
Appendix 2: Feature Graphs



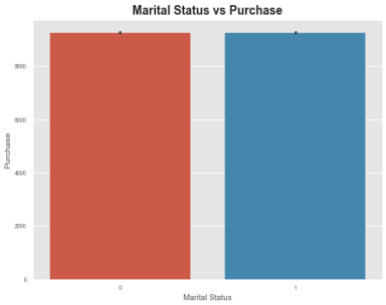


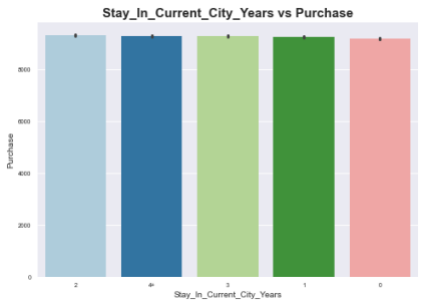


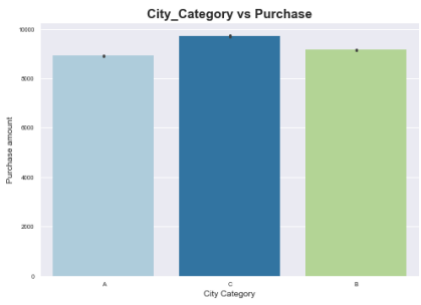


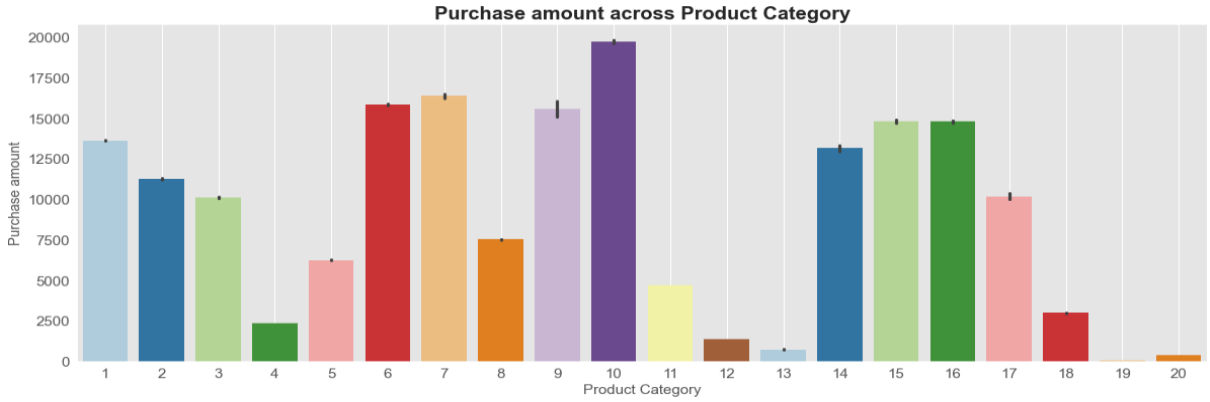


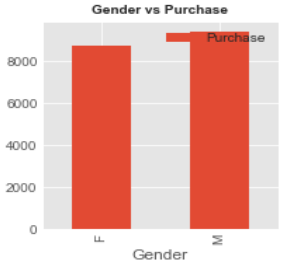
Appendix 3: Bivariate Plots

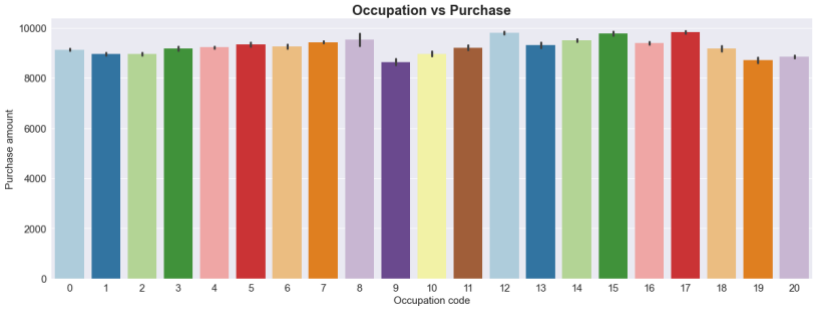


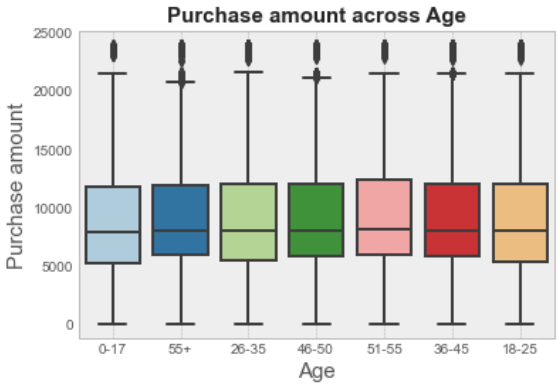












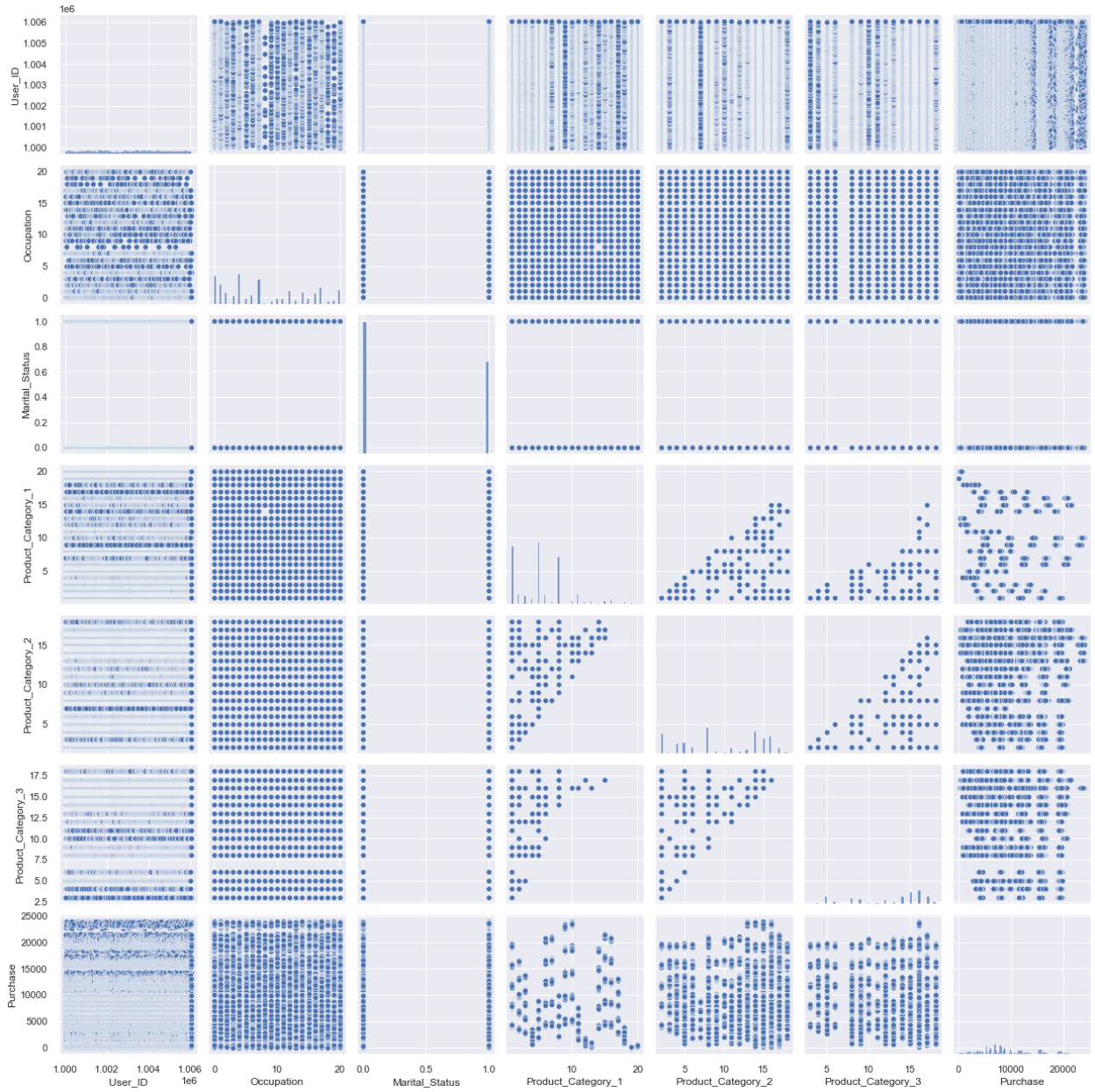


table. Multivariate pair plot of Black Friday Dataset