Black Friday Sale Predictor

UML 501 Machine Learning Project Report

Submitted by:

(102103590) Anraobir Singh

(102103310) Nitish Gaba

BE Third Year, COE

Group No: 3CO21

Submitted to: Suchita Sharma

Computer Science and Engineering Department

TIET, Patiala

November 2023

1. Introduction

1.1 Background

In the dynamic landscape of retail, Black Friday stands as a pivotal day, driving a substantial surge in sales. Recognizing the critical importance of making informed decisions, the Black Friday Sale Predictor project is designed to harness the power of advanced machine learning techniques. Its goal is to accurately forecast individual customer sales amounts on this significant shopping day.

1.2 Dataset

The foundation of this project lies in the Black Friday Sale dataset, a comprehensive resource available on Kaggle (link). This dataset encapsulates a wealth of information, including customer demographics, intricate product details, and insightful purchase information. This rich repository serves as a robust basis for developing predictive models that can uncover valuable insights.

2. Problem Statement

The primary objective of this endeavor is to construct a highly precise predictive model capable of estimating the sales amount an individual customer is likely to generate on Black Friday. This predictive model is poised to be a strategic asset for retailers, empowering them to tailor promotions and optimize sales strategies to cater to the diverse preferences of various customer segments.

3. Data Exploration and Preprocessing

3.1 Data Overview

Within the dataset, a myriad of features such as User_ID, Product_ID, Age, Occupation, City_Category, and Product_Category provide a comprehensive and nuanced view of customer behavior and preferences.

3.2 Data Cleaning

A meticulous data cleaning process is undertaken to ensure the integrity of the dataset. This involves addressing missing values, outliers, and duplicates. Categorical variables are encoded, and any inconsistencies within the data are carefully rectified.

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a crucial step in uncovering hidden patterns within the dataset. It involves delving into feature distributions and visualizing relationships, providing valuable insights for subsequent modeling.

4. Feature Engineering

4.1 Creation of New Features

To augment the predictive power of the model, new features are engineered. These additions, along with scaling and transforming existing features, contribute to an optimized and refined model.

5. Model Selection and Training

5.1 Model Selection

A diverse set of regression models, including Linear Regression, Decision Trees, Random Forests, and XGBoostRegressor, are considered for their ability to predict continuous outcomes.

5.2 Model Training

The dataset is strategically divided into training and testing sets. Models are trained on the former, and a meticulous process of hyperparameter tuning is executed to optimize their performance.

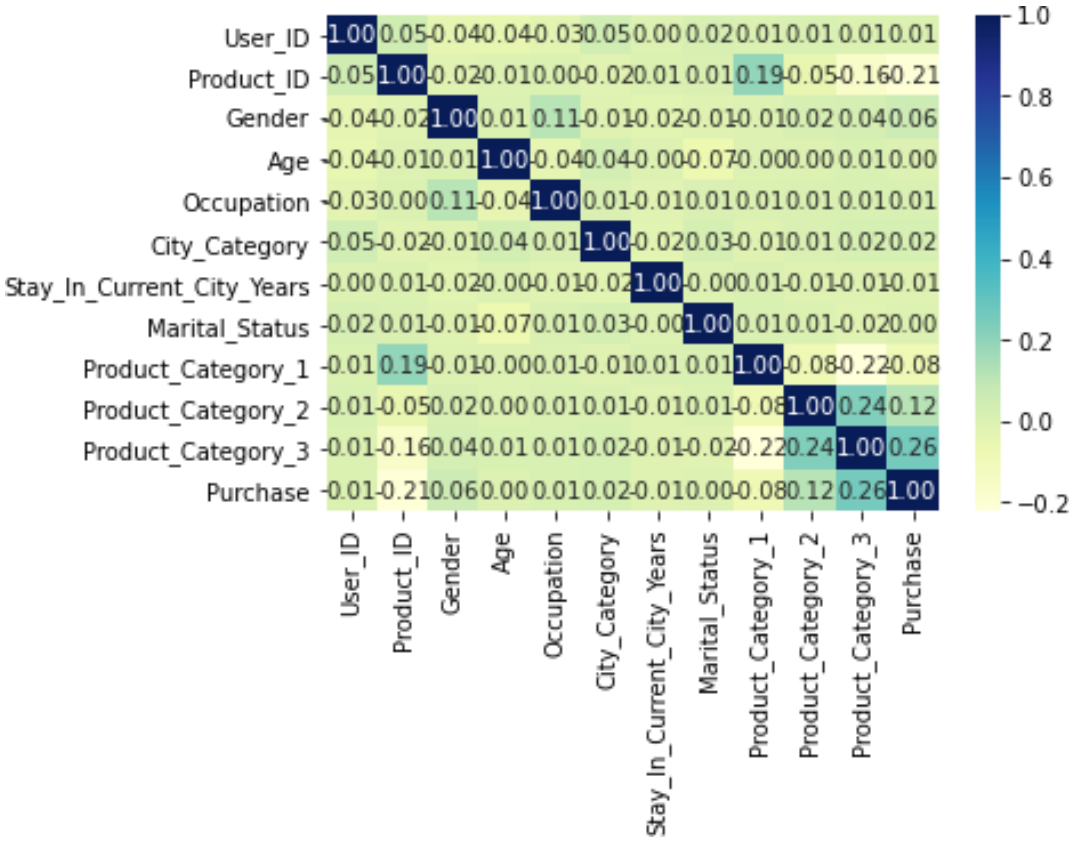
6. Model Evaluation

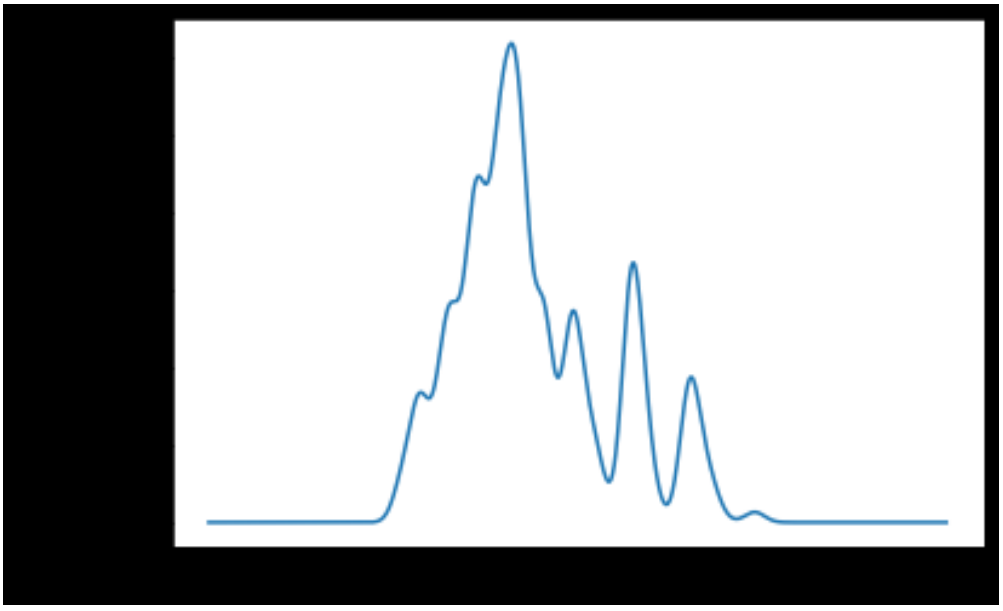
6.1 Performance Metrics

The evaluation of model performance is quantified through essential metrics such as Mean Squared Error (MSE) and R-squared, providing a comprehensive understanding of predictive accuracy and goodness-of-fit.

6.2 Visualization of Results

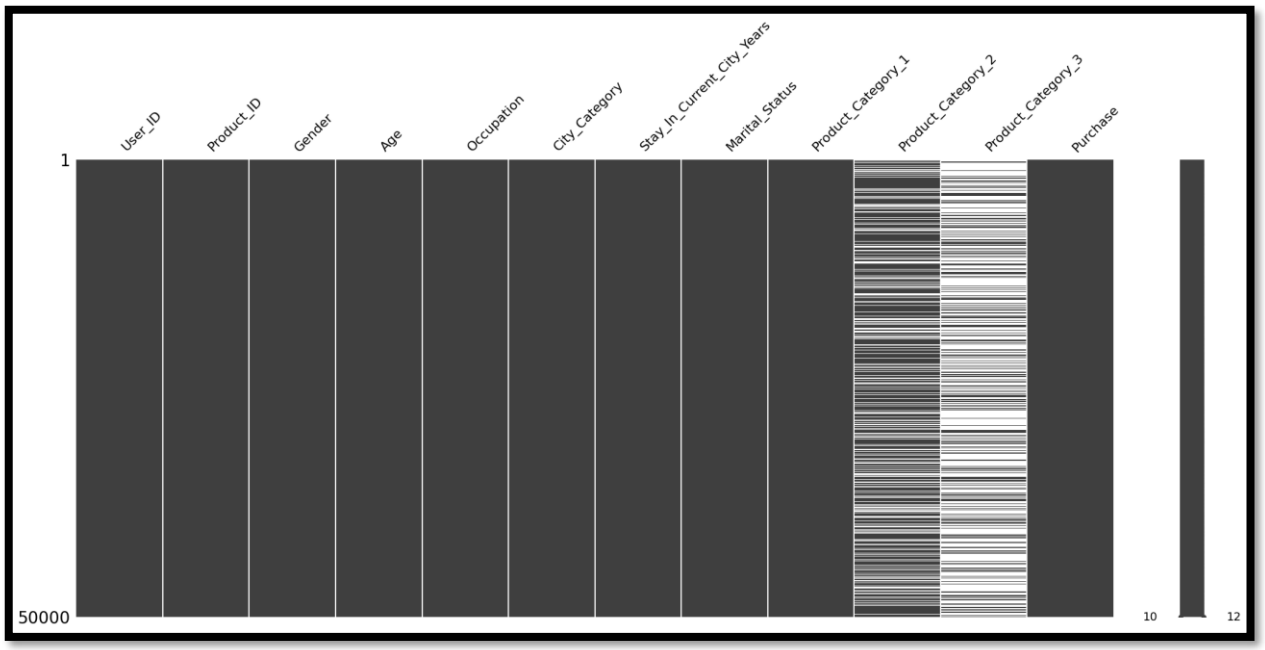
Results are not just numerically presented but also visually communicated. The comparison of predicted sales amounts against actual values is showcased through visualizations, offering a clear assessment of the model's accuracy and efficacy.

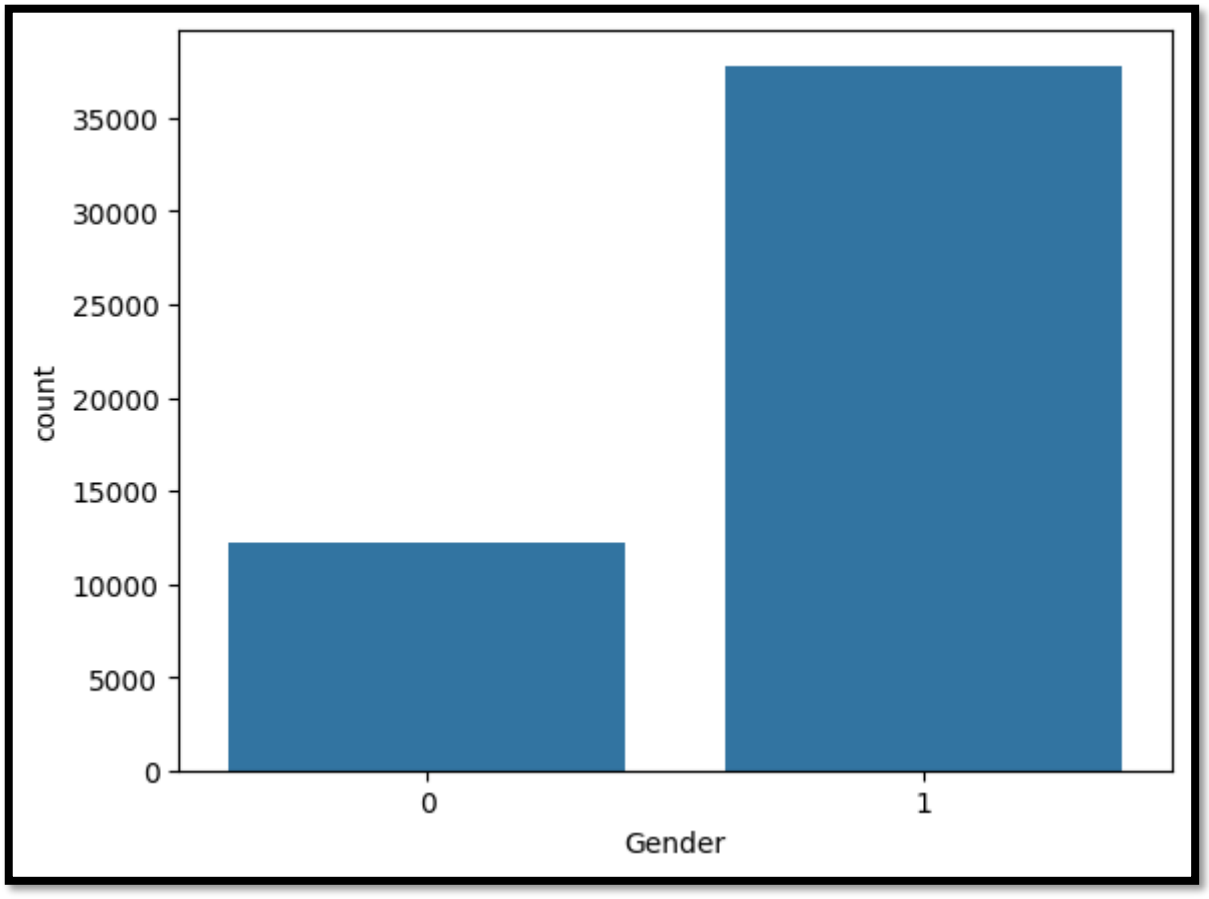

User_ID
Product_ID
Gender
Age
Occupation
City_Category
Stay_In_Current_City_Years
Marital_Status
Product_Category_1
Product_Category_2
Product_Category_3
Purchase
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

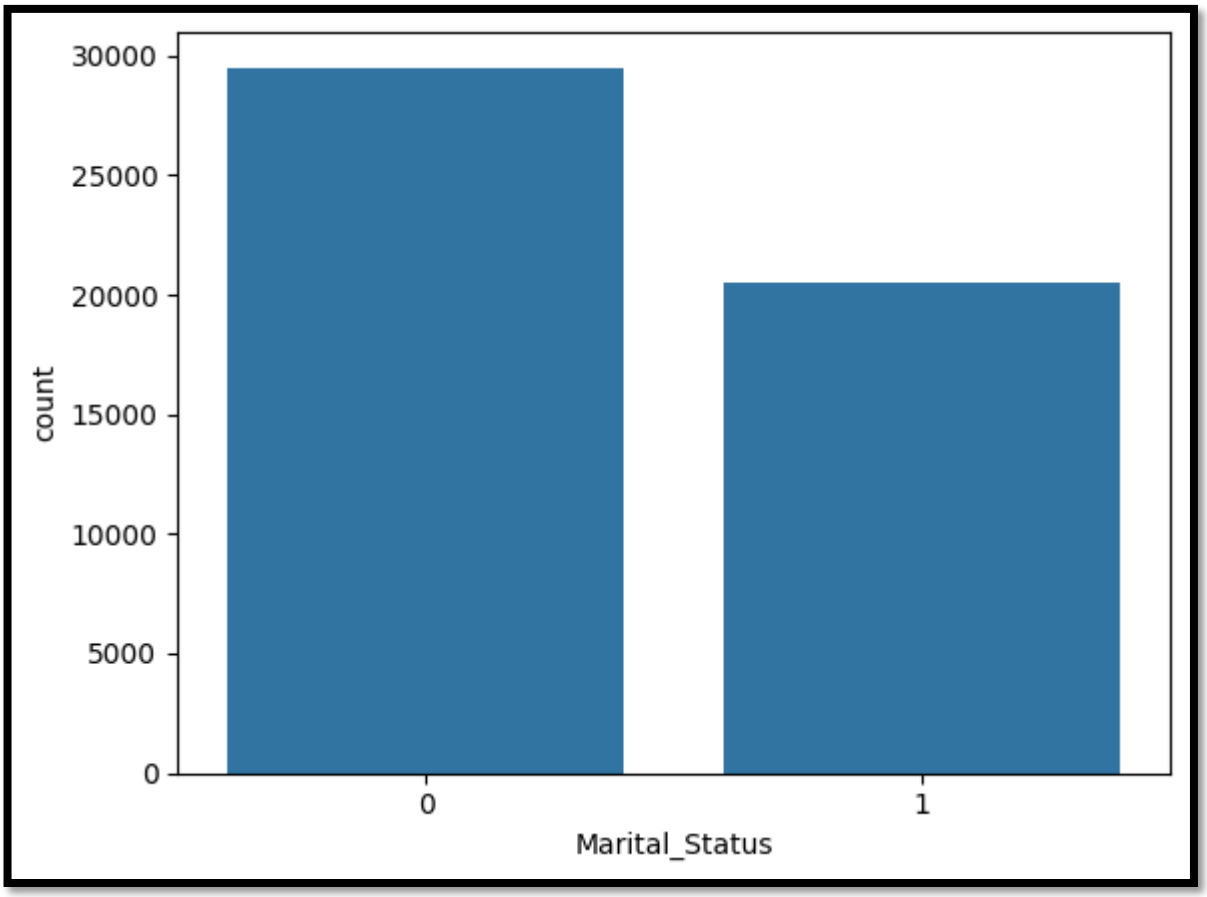


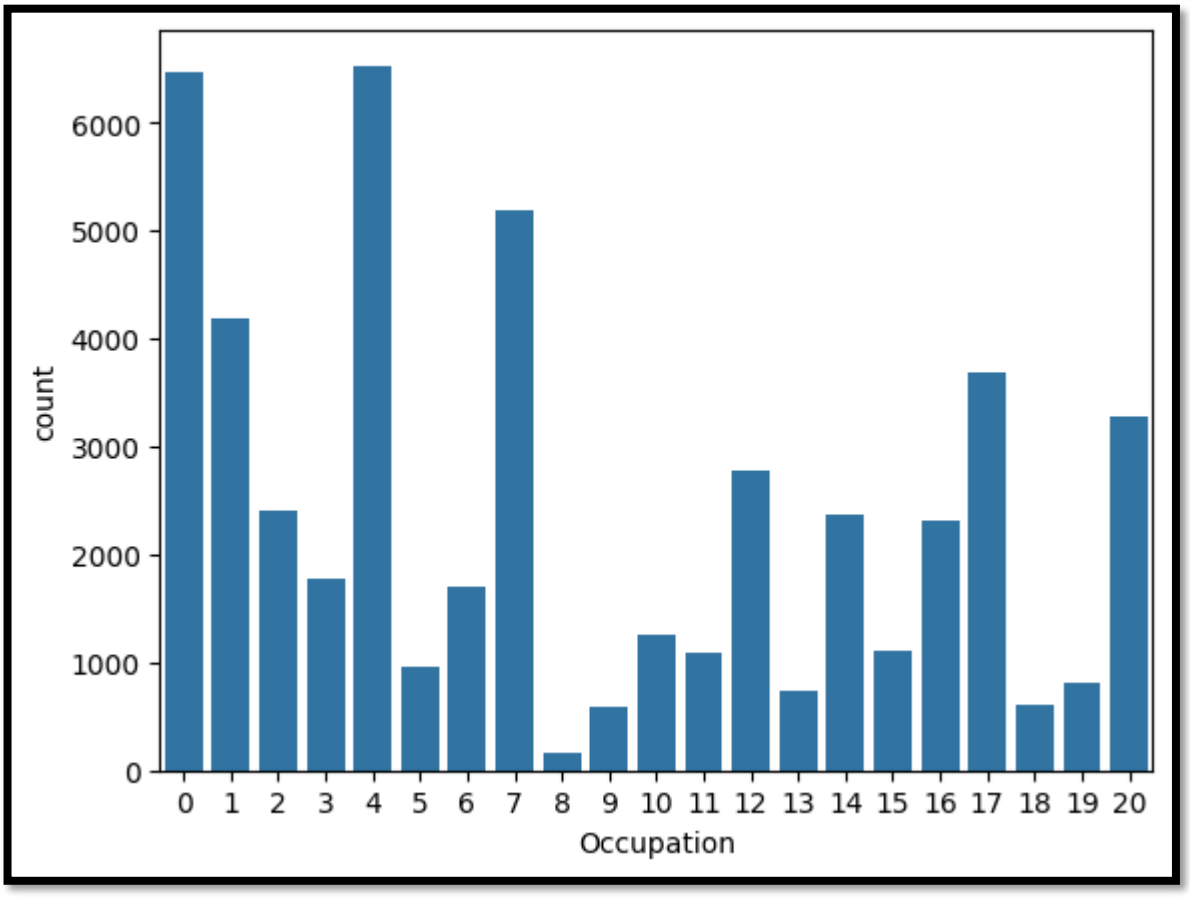
- Decision Tree Regressor - 60.622781412422945
- Linear Regression - 9.735664501653485
- Random forest Regressor - 66.78581833217073
- XGB Regressor - 68.74816218771024

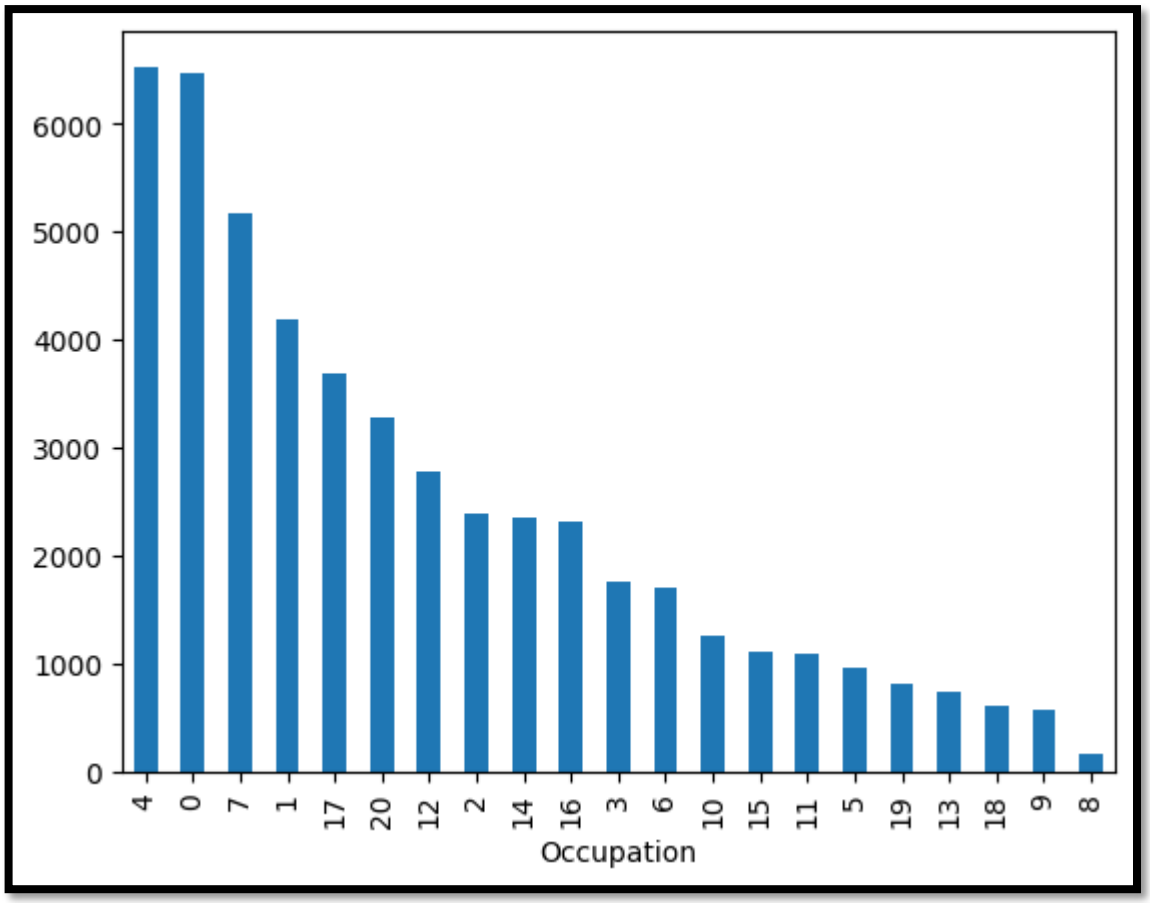
Plots:

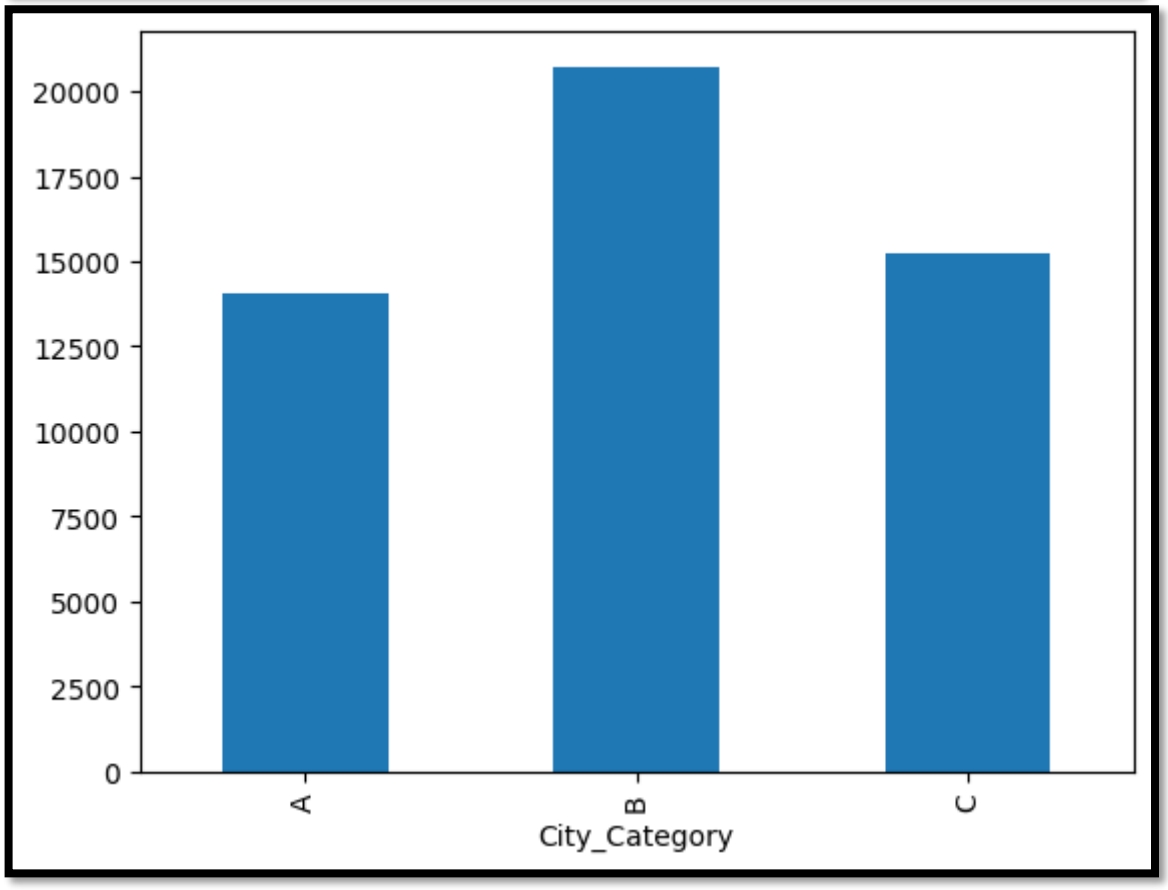


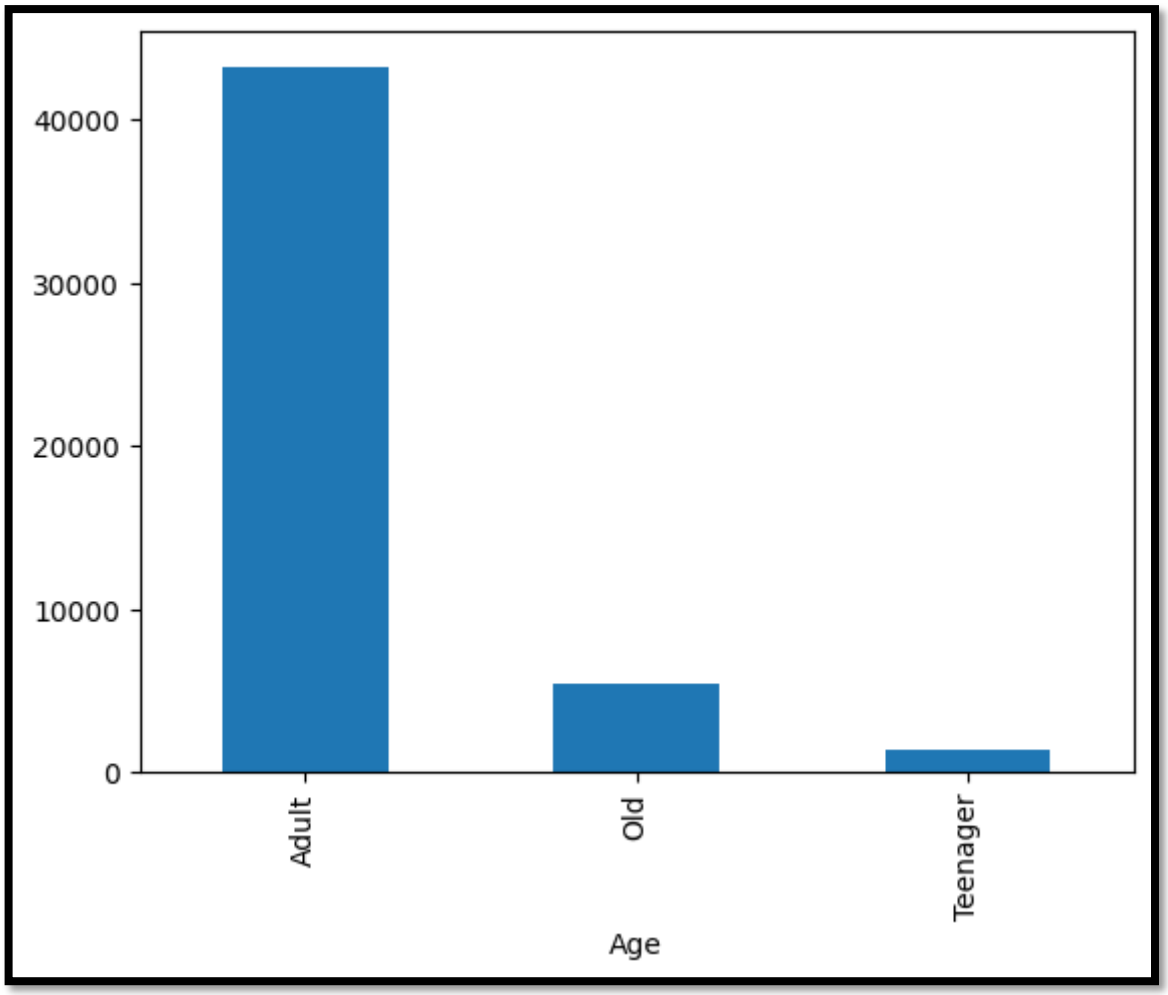


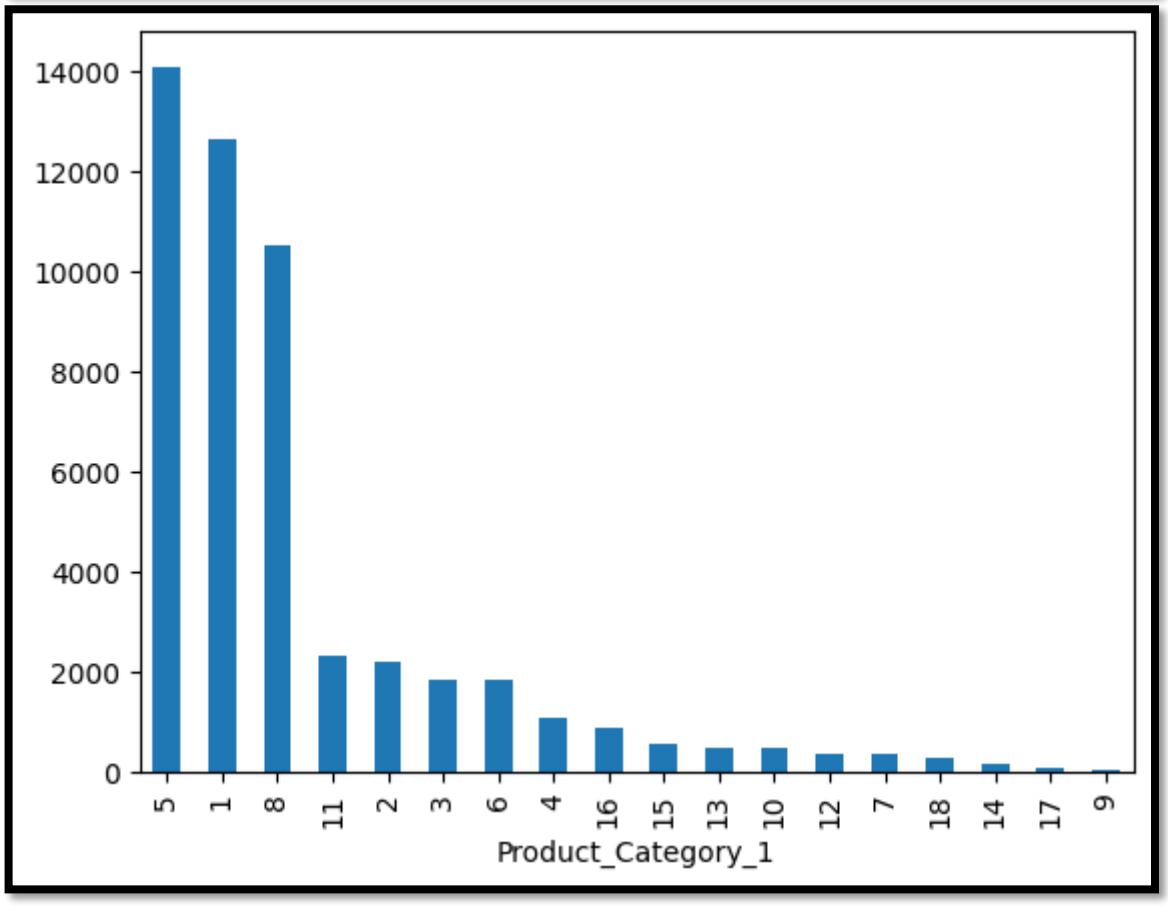

30000
25000
20000
15000
10000
5000
0
count
0
1
Marital_Status

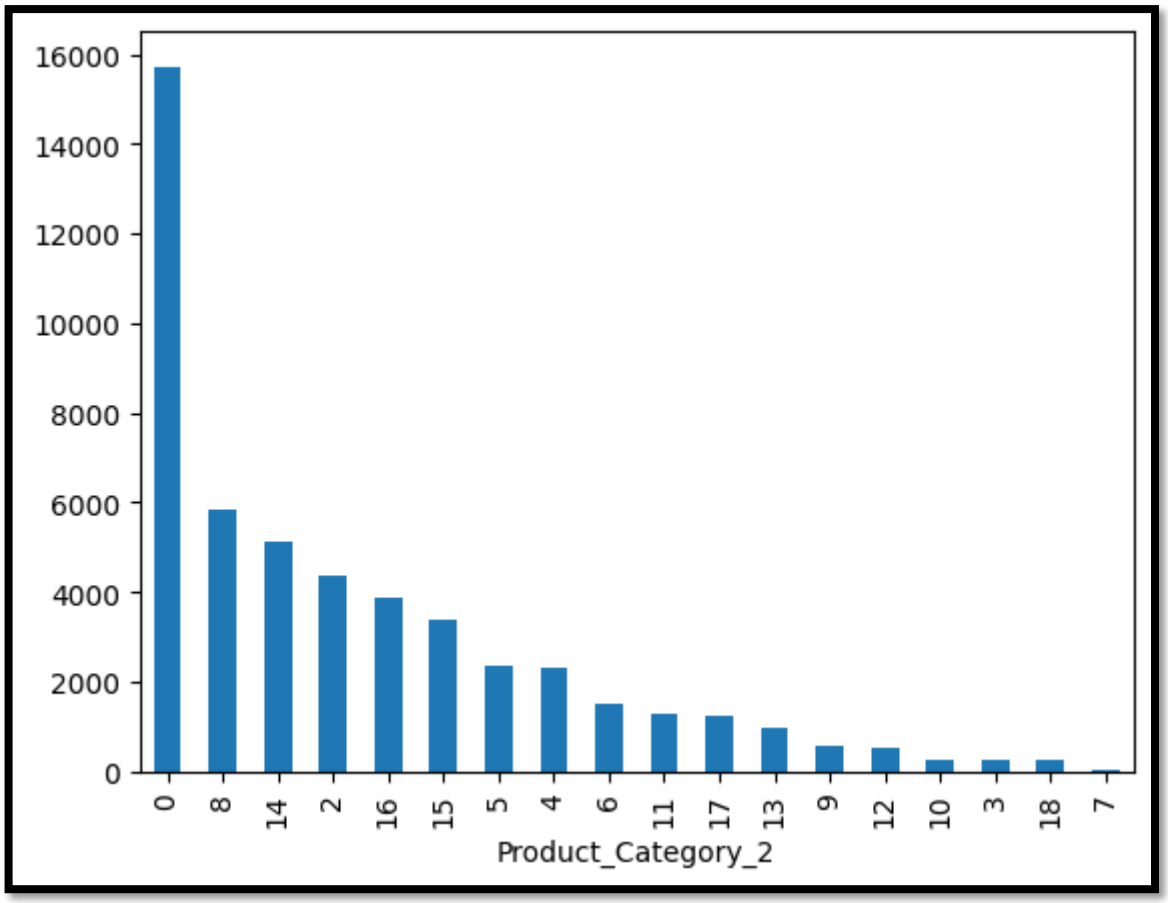

6000
5000
4000
3000
2000
1000
0
count
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Occupation

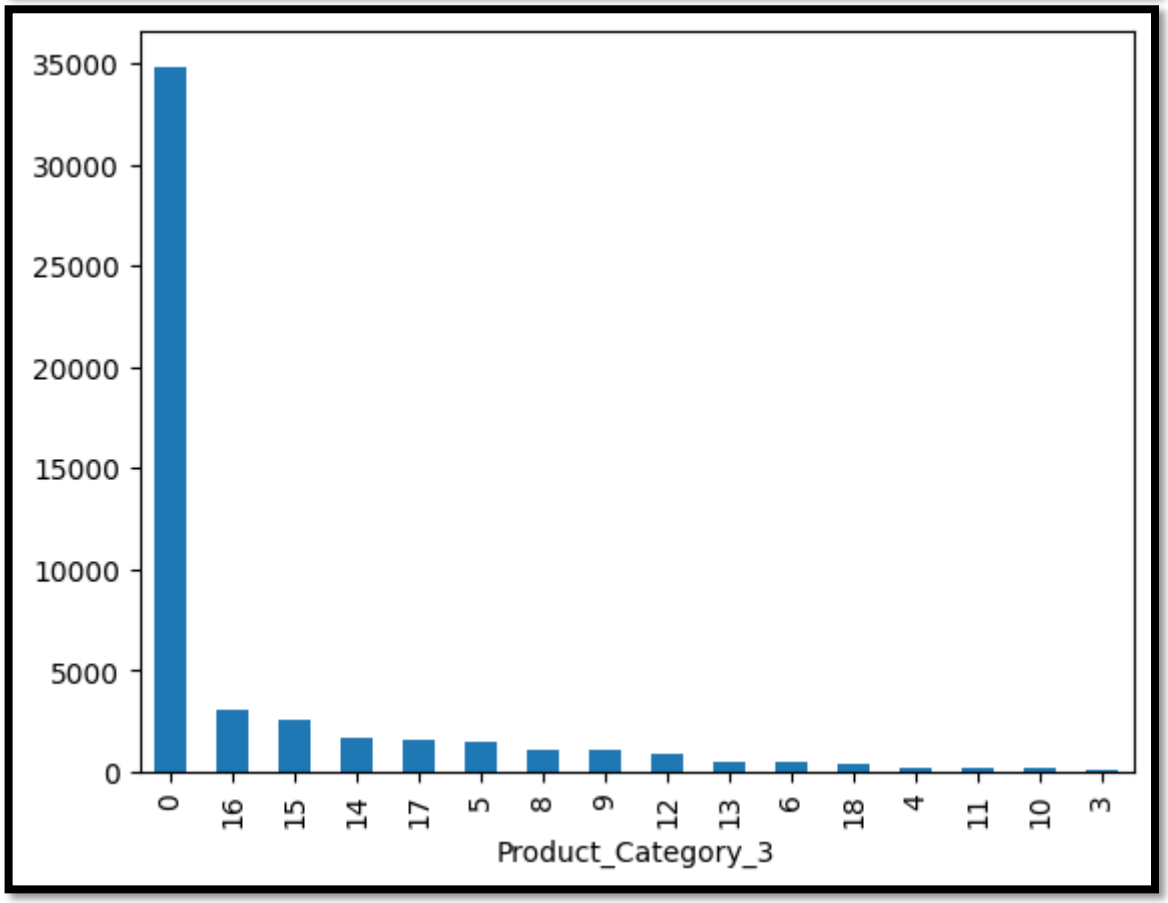

6000
5000
4000
3000
2000
1000
0
4
0
7
1
17
20
12
2
14
16
3
6
10
15
11
5
19
13
18
9
8
Occupation


20000
17500
15000
12500
10000
7500
5000
2500
0
A
B
C
City_Category


40000
30000
20000
10000
0
Adult
Old
Teenager
Age


14000
12000
10000
8000
6000
4000
2000
0
5
1
8
11
2
3
6
4
16
15
13
10
12
7
18
14
17
9
Product_Category_1


16000
14000
12000
10000
8000
6000
4000
2000
0
0
8
14
2
16
15
5
4
6
11
17
13
9
12
10
3
18
7
Product_Category_2


35000
30000
25000
20000
15000
10000
5000
0
0
16
15
14
17
5
8
9
12
13
6
18
4
11
10
3
Product_Category_3

7. Conclusion

7.1 Summary of Findings

The comprehensive analysis conducted in this project has uncovered valuable insights into customer behavior and preferences. These insights are poised to become crucial drivers for data-driven decision-making within the retail sector. Through the exploration of the dataset, we have identified both the strengths and limitations of the predictive model, shedding light on areas of excellence and potential refinement.

7.2 Recommendations

Based on the predictive insights derived from our model, we offer strategic recommendations for retailers. These recommendations are grounded in the data-driven understanding of customer segments and their purchasing patterns. Additionally, we provide suggestions for refining the existing model, emphasizing potential improvements for more accurate predictions in future iterations.

8. Future Work

8.1 Model Enhancement

Looking ahead, there is ample room for enhancing the predictive model. Future work involves delving into advanced machine learning techniques and exploring ensemble methods to elevate predictive accuracy. Additionally, considering the dynamic nature of data, the integration of external data sources is considered to provide a more comprehensive and nuanced analysis. This ongoing commitment to improvement ensures the model remains adaptive and continues to meet the evolving needs of the retail landscape.

9. References

Black Friday Sale | Kaggle