IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication 2023,12,08, date of current version 1

# Unlocking Consumer Insights: A Comprehensive Analysis of Black Friday Sales Data

# (December 2023)

**D.Wassef**
Humber college institute of technology & advanced learning

**ABSTRACT** This comprehensive report undertakes a thorough exploration of Black Friday sales data, employing a nuanced approach to develop predictive models aimed at forecasting customer purchases. Employing sophisticated tools like Python and Jupyter Notebooks for meticulous data analysis, coupled with the efficiency of Visual Studio Code for streamlined code editing, this project seamlessly integrates data cleaning, feature engineering, and advanced machine learning techniques. The dataset under, sourced from Kaggle, encapsulates a rich tapestry of customer transactions during the Black Friday sales event. This multifaceted dataset spans crucial details, including but not limited to, gender, age, occupation, city, marital status, product categories, and the corresponding purchase amounts. The analysis seeks to unravel the intricate patterns and correlations within this dataset, shedding light on the underlying factors influencing customer purchasing behavior during the Black Friday shopping extravaganza. The exploration is not merely confined to the raw data but delves deeper into the realms of data cleaning and feature engineering. The transformative journey involves converting intricate strings into meaningful numerical representations, handling missing values judiciously, and crafting new features that amplify the predictive power of machine learning models. This meticulous approach is quintessential in preparing the dataset for the ensuing machine learning models, ensuring they can discern meaningful patterns and generate accurate predictions.

## I. INTRODUCTION

The central goal of this analysis is to gain insights into Black Friday sales data and develop predictive models to forecast customer purchases. By exploring and cleaning the dataset, implementing feature engineering, and leveraging machine learning techniques, we aim to understand the factors influencing purchase behavior and build models capable of accurate predictions. This report will detail the data cleaning process, the choice of machine learning models, their hyperparameters, evaluation results, and conclusions drawn from the analysis.

### *Data Collection*

The dataset for this project is sourced from Kaggle and pertains to Black Friday sales. The dataset contains information about customer transactions during Black Friday sales, including details such as gender, age, occupation, city, marital status, product categories, and purchase amounts.

### *Problem Statement*

Black Friday represents a crucial sales event for retailers, providing an opportunity to understand customer preferences and tailor marketing strategies. The primary problem addressed in this analysis is predicting the purchase amount a customer is likely to make on Black Friday. This predictive capability is valuable for retailers seeking to optimize inventory, tailor promotions, and enhance the overall shopping experience.

## II. Methodology

The project harnessed the power of Python, a versatile programming language, and Jupyter Notebooks, an interactive and collaborative coding environment, to conduct data analysis and model development. Visual Studio Code (VS Code) was employed for efficient code editing and version control, enhancing the overall development workflow.

For data manipulation and exploration, the project relied on key libraries such as pandas and numpy, enabling seamless handling of datasets and efficient numerical computations. Statsmodels and scipy were instrumental for statistical analyses and hypothesis testing, providing valuable insights into the underlying patterns in the data. Scikit-learn emerged as a cornerstone for machine learning tasks, offering a diverse set of tools and algorithms. Functions like Standard Scaler aided in preprocessing, while train_test_split facilitated dataset splitting for training and testing. Cross_val_score provided a robust mechanism for cross-validation, ensuring reliable model performance evaluation.

The project explored different regression models, leveraging scikit-learn's Linear Regression, RFE (Recursive Feature Elimination), Random Forest Regressor, and Gradient Boosting Regressor. These models played a pivotal role in predicting the target variable based on the selected features.

Data visualization was achieved using Seaborn and Matplotlib, offering a comprehensive view of the dataset's characteristics and model performance. Additionally, Plotly and iplot were utilized to create interactive plots, enhancing the interpretability and engagement of the visualizations.

## III. Data Cleaning Process

**Initial Data Exploration:** Upon loading the dataset, an initial exploration revealed various columns such as 'Gender,' 'Age,' 'Occupation,' 'City_Category,' 'Stay_In_Current_City_Years,' and 'Marital_Status.' The dataset also included columns 'User_ID' and 'Product_ID,' which were determined to be irrelevant for predicting purchases.

*Visualizing raw data*

The raw data, extracted from the Black Friday Sales dataset, presented an unfiltered and unaltered snapshot of customer transactions. This section will showcase the unprocessed data, providing insights into its structure, distributions, and inherent characteristics, setting the stage for the subsequent data cleaning and transformation processes.

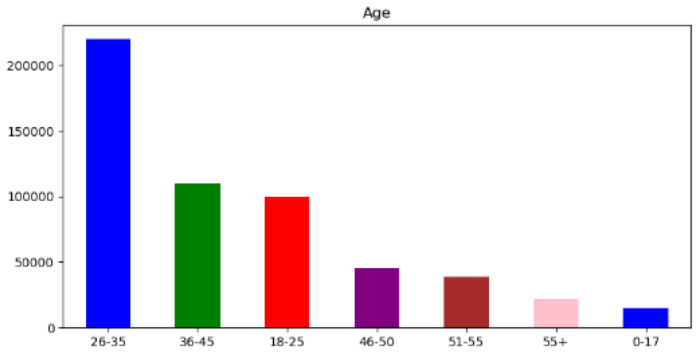


**Figure 1**. The frequency of age distribution of customers

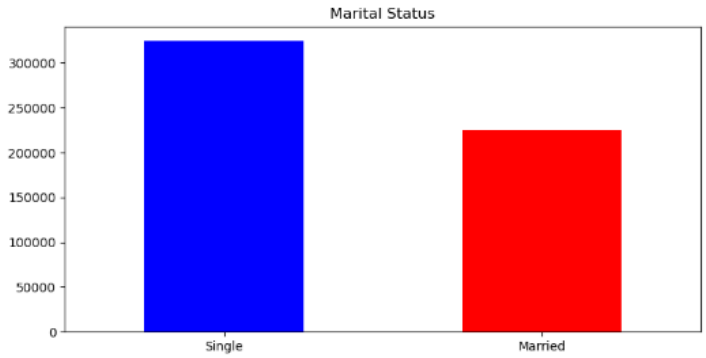


**Figure 2**. Frequency of single and married customers

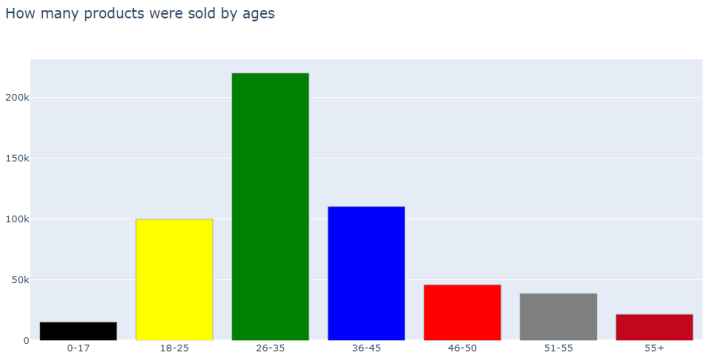


**Figure 3**. Frequency of products sold by age

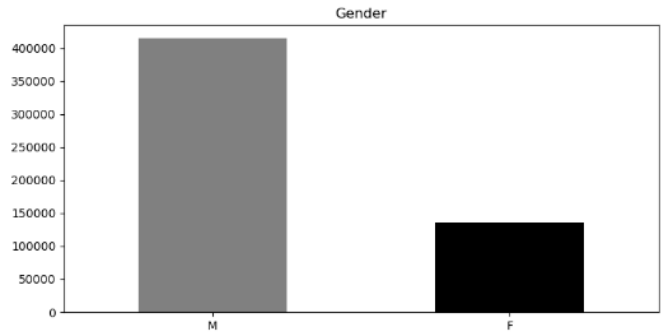


**Figure 4.** Frequency of male and female customers

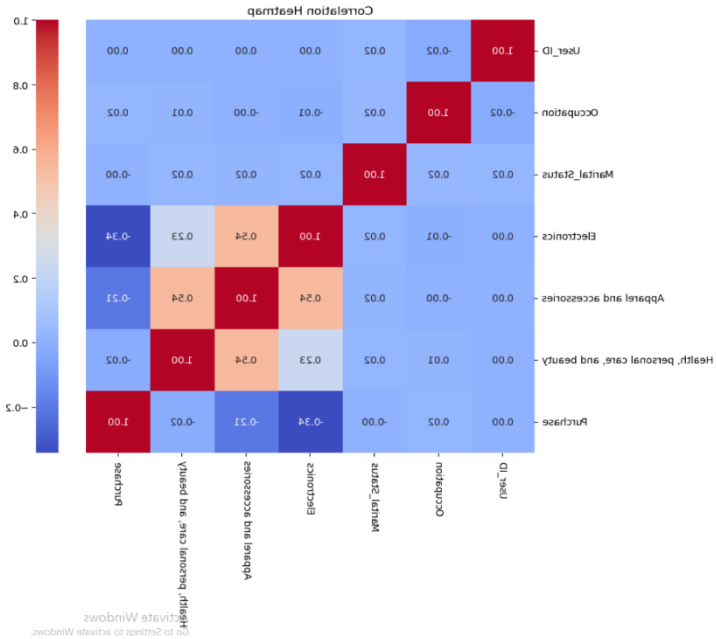


**Figure 5**. Heatmap showings the correlation matrix of numerical features.

*Handling Categorical Variables*

Categorical variables, such as 'City_Category' and 'Age,' were mapped to more meaningful categories. 'City_Category' was mapped to actual city names ('Toronto,' 'Brampton,' 'Scarborough'), and 'Age' was split into dummy variables for different age categories.

*Outlier Removal*

Outliers in the 'Purchase' column were identified using the Z-score method, and rows containing outliers were removed to ensure model stability.

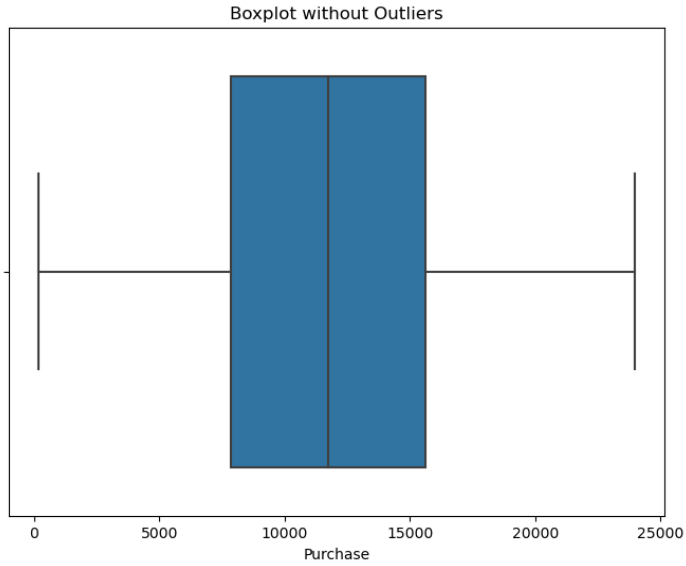


**Figure 6**. Visualization the range of data which highlights a lack of anomalies

## IV. Feature Engineering

New features were created, including dummy variables for 'City_Category' and 'Age' categories, which were crucial for subsequent machine learning modeling. 'Gender' was converted to binary (0 for male, 1 for female).

*Scaling*

The 'Purchase' column was scaled using StandardScaler to bring features to a similar scale for better model performance.

## V. Machine Learning Models

Feature Selection:

A correlation matrix was analyzed to select features relevant to predicting 'Purchase.' The top five features were identified: 'City_Scarborough,' 'Middle Age,' 'Senior Adult,' 'Senior,' and 'Electronics.'

```
Top 5 Selected Features:
Index(['City_Scarborough', 'Mature Adult', 'Middle Age', 'Senior Adult',
       'Senior', 'Electronics'],
      dtype='object')
                            OLS Regression Results
==============================================================================
Dep. Variable:               Purchase   R-squared:                       0.163
Model:                            OLS   Adj. R-squared:                  0.163
Method:                 Least Squares   F-statistic:                     5422.
Date:                Fri, 08 Dec 2023   Prob (F-statistic):               0.00
Time:                        10:03:33   Log-Likelihood:            -2.2185e+05
No. Observations:              166821   AIC:                         4.437e+05
Df Residuals:                  166814   BIC:                         4.438e+05
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3531      0.004     89.067      0.000       0.345       0.361
City_Scarborough  0.1367      0.005     28.574      0.000       0.127       0.146
Mature Adult   0.0431      0.006      7.470      0.000       0.032       0.054
Middle Age     0.0463      0.008      5.492      0.000       0.030       0.063
Senior Adult   0.1224      0.009     13.408      0.000       0.105       0.140
Senior         0.0930      0.012      7.520      0.000       0.069       0.117
Electronics   -0.1542      0.001   -176.859      0.000      -0.156      -0.152
==============================================================================
Omnibus:                     1999.872   Durbin-Watson:                   1.829
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2138.728
Skew:                           0.248   Prob(JB):                         0.00
Kurtosis:                       3.250   Cond. No.                         21.5
==============================================================================
```

**Figure 7**. Summary of the model using Recursive Feature Elimination

The selected features for predicting the 'Purchase' category were determined through an extensive feature selection process, resulting in a subset of six impactful variables. The chosen features include 'City_Scarborough', representing a specific city category; 'Mature Adult', 'Middle Age', 'Senior Adult', and 'Senior', capturing various age groups; and 'Electronics', denoting the product category related to electronics. The Ordinary Least Squares (OLS) regression results reveal that the overall model explains approximately 16.3% of the variance in the 'Purchase' variable, as indicated by the R-squared value. The coefficients associated with each feature provide insights into their individual impact on the purchase behavior. Notably, 'City_Scarborough' and 'Electronics' exhibit statistically significant coefficients, suggesting a considerable influence on the predicted purchase amounts. The statistical significance, combined with the high F-statistic and low p-values, reinforces the reliability of the selected features in the predictive model. The coefficients' signs and magnitudes further aid in interpreting the direction and strength of the relationships between the features and the target variable. Overall, this comprehensive analysis equips businesses with valuable information for targeted marketing strategies and future sales predictions.

*Model Selection and Hyperparameters*

Three machine learning models were employed: Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor. Hyperparameters were chosen based on default values provided by the scikit-learn library.

**Linear Regression Model:**
Linear regression is a fundamental statistical model used to establish a linear relationship between the independent variables (features) and the dependent variable (target).
**Strengths:** Simple and interpretable; provides a baseline understanding of the data.
**Limitations:** Assumes a linear relationship, may not capture complex patterns in the data.
**Random Forest Regressor Model:**
Random Forest is an ensemble learning method that builds a multitude of decision trees to enhance predictive performance.
**Strengths:** Handles non-linear relationships, reduces overfitting, and is less sensitive to outliers.
**Limitations:** May be computationally expensive and challenging to interpret compared to linear models.
**Gradient Boosting Regressor Model:**
Gradient Boosting is another ensemble learning technique that builds trees sequentially, with each tree correcting the errors of the previous one.
**Strengths:** Excellent predictive performance and handles non-linear relationships well and is less prone to overfitting.
**Limitations:** May require fine-tuning of hyperparameters and longer training times.

## VI. Model Evaluation and Insights

**RMSE (Root Mean Squared Error):** Measures the average magnitude of the errors between predicted and actual values. Lower RMSE indicates better model performance.

- yi is the actual value for the n-ith observation.
- ŷi is the predicted value for the n-ith observation.
- N is the number of observations.

**R2 Score:** Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R2 score indicates a better fit.
**Overall Consideration:**
The choice of models involves a trade-off between simplicity and predictive accuracy. Linear Regression provides simplicity and interpretability, while Random Forest and Gradient Boosting enhance predictive power by capturing more complex relationships.

| | RMSE | $R^2$ Score |
|---|---|---|
| **Linear Regression** | 0.9129 | 0.1728 |
| **Random Forest Regressor** | 0.7354 | 0.4633 |
| **Gradient Boosting Regressor** | 0.7056 | 0.5058 |

**Table 1**. Evaluation results of the three trained models against the dataset

*Visualizations post train and test*

Various visualizations were created to understand data distributions, feature importance, and the impact of different variables on purchase predictions.

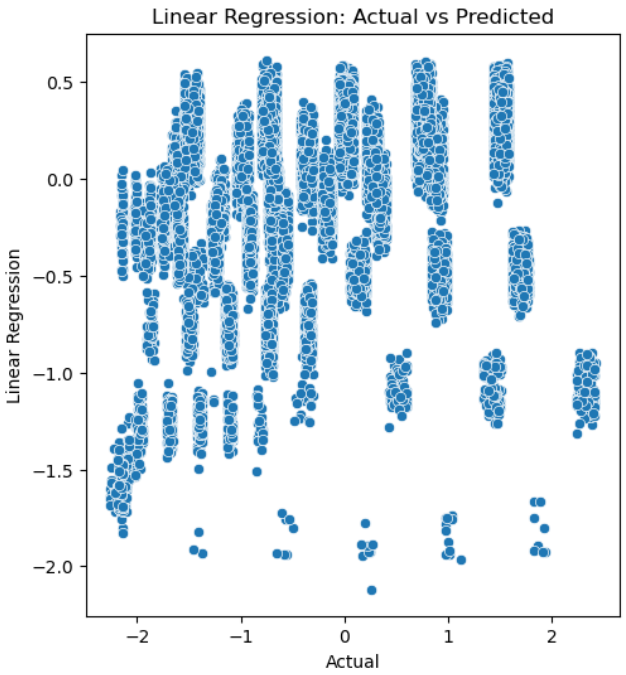


**Figure 8**. Scatter plot of actual vs predicted values for Linear Regression

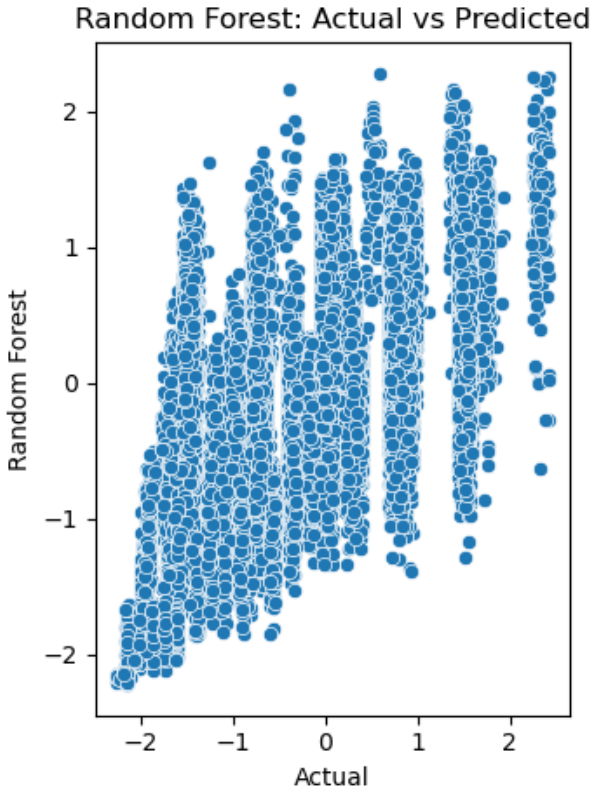


**Figure 9**. Scatter plot of actual vs predicted values for Random Forest

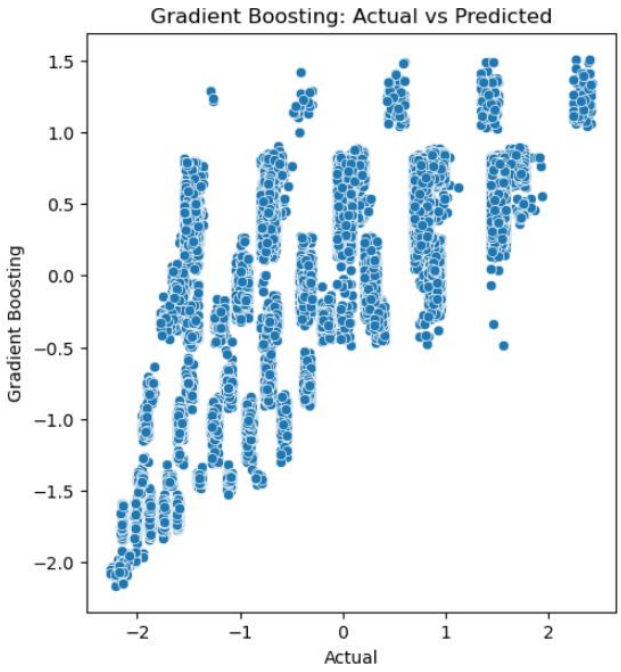


**Figure 10**. Scatter plot of actual vs predicted values for gradient boosting

X-axis (Actual): This represents the actual values of the target variable (dependent variable) in your dataset. Each point on the X-axis corresponds to the true values of the target variable.

Y-axis (Linear Regression): This represents the predicted values of the target variable generated by your Linear Regression model. Each point on the Y-axis corresponds to the predicted values of the target variable.

Points on the Plot: Each point on the plot represents a data point in your dataset. The position of the point is determined by its actual value on the X-axis and its predicted value on the Y-axis.

**Interpretation**:
Diagonal Line (45-degree line): Ideally, in a perfect model, all points would lie on the 45-degree line (the line where actual equals predicted). Deviations from this line indicate the extent to which the model's predictions differ from the actual values.

Scattering of Points: The spread or scattering of points around the diagonal line provides insights into the model's accuracy. If points are tightly clustered around the diagonal, it suggests that the model is making accurate predictions. If there is a significant spread, it indicates that the model has difficulty accurately predicting certain instances.
Pattern or Lack Thereof: Patterns in the scatter plot, such as a systematic bias (all predictions consistently too high or too low), can reveal issues with the model that may need attention.

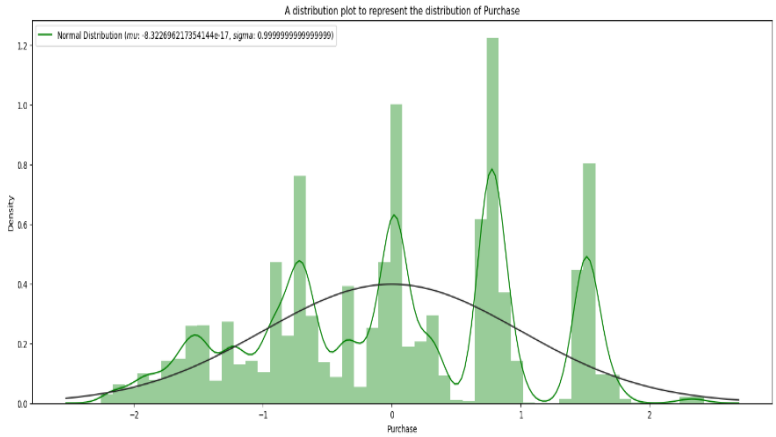


**Figure 11**. distribution plot for the target variable (Purchase)

Understanding the Distribution:
The primary purpose of this visualization is to understand the distribution of the 'Purchase'.
Visualizing Data Spread: The histogram (green bars) represents the distribution of the 'Purchase' values. It gives you an idea of how the values are spread across different ranges.

Fitting a Normal Distribution Curve: The green curve overlaid on the histogram is a fitted normal distribution curve. It is based on the mean (mu) and standard deviation (sigma) calculated from the actual data. This curve helps you assess how well the 'Purchase' variable follows a normal distribution. A normal distribution is characterized by a bell-shaped curve. By fitting a normal distribution curve to the data, you can visually inspect how closely the 'Purchase' variable resembles a normal distribution.

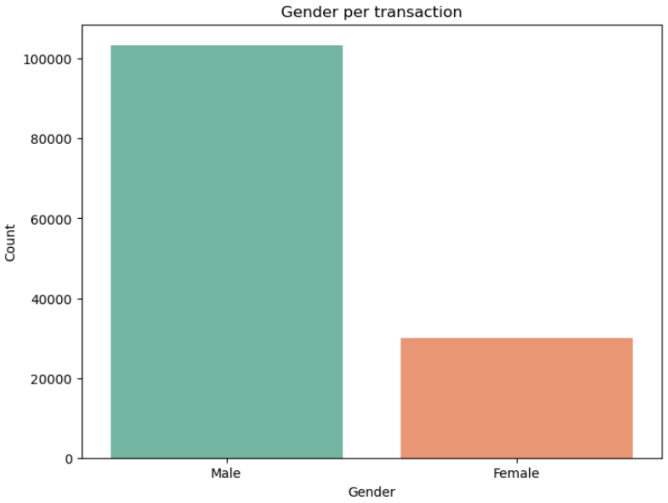


**Figure 12**. Frequency of transactions completed by customer gender

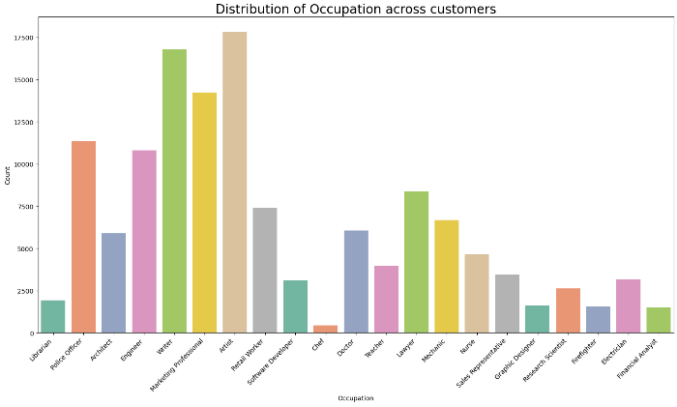


**Figure 13**. Distribution of occupation among customers

## VII. Conclusions

The comprehensive analysis undertaken to predict customer purchases on Black Friday demonstrated that the Gradient Boosting Regressor emerged as the most effective model, exhibiting the lowest RMSE and the highest $R^2$ score among the tested models. This outcome underemphasizes the model's robust predictive capabilities, making it a valuable tool for businesses seeking accurate forecasts for Black Friday sales.

The identification of top features influencing purchase decisions adds a layer of strategic insight, enabling businesses to tailor marketing efforts and promotions more effectively. By understanding the pivotal factors contributing to customer purchases, companies can optimize their product offerings, target specific demographics, and design promotions that resonate with their audience.

There are several promising avenues for future work. Refining the model further, exploring additional feature engineering, and incorporating external data sources could potentially enhance predictive accuracy. Additionally, delving into customer segmentation and behavior analysis could unveil nuanced patterns, allowing for more personalized and targeted marketing strategies. Furthermore, as the dataset represents a single year's Black Friday, extending the analysis to multiple years could capture seasonal variations and trends, providing a more comprehensive understanding of customer behavior during this high-impact shopping event.

## Appendix

Shmueli, G., Bruce, P. C., Patel, N. R., & Gedeck, P. (2020). *Data mining for Business Analytics: Concepts, techniques, and applications in Python*. John Wiley & Sons.