Black Friday Sales Prediction Project

ABDUL AZIZ AKBANI

Problem Statement

Introduction:

Black Friday is one of the most significant shopping events of the year, with retailers offering substantial discounts on a wide range of products. Understanding customer behavior and predicting sales on Black Friday can be invaluable for both retailers and marketers. In this analysis, we aim to develop a predictive model that can estimate the purchase amount a customer is likely to spend on Black Friday based on various customer and product-related attributes.

Objective:

The primary objective of this analysis is to create a machine learning model that can accurately predict the purchase amount of customers on Black Friday. This model can assist retailers in optimizing their marketing strategies, inventory management, and pricing decisions to maximize sales and customer satisfaction.

Dataset Description:

The dataset used for this analysis contains information about customer demographics, product categories, purchase amounts, and more during a Black Friday sale. It includes attributes such as 'Gender,' 'Age,' 'Occupation,' 'City_Category,' 'Marital_Status,' and 'Product_Category' to help us understand customer behavior and make predictions.

Proposed Solution

To predict Black Friday sales and extract valuable insights, the following steps should be taken:

Step 1: Collect and clean data related to Black Friday sales. Ensure that missing values are handled and categorical variables are encoded.

Step 2: Explore the collected data to understand patterns and outliers.

Step 3: Select a regression model such as XGBoost or Random Forest. Train and evaluate the models using metrics like RMSE and R-squared.

Step 4: Optimize model performance with hyperparameter tuning.

Proposed Solution

Step 5: Analyze feature importance to understand what drives sales.

Step 6: Derive actionable insights that can be used for marketing and pricing strategies.

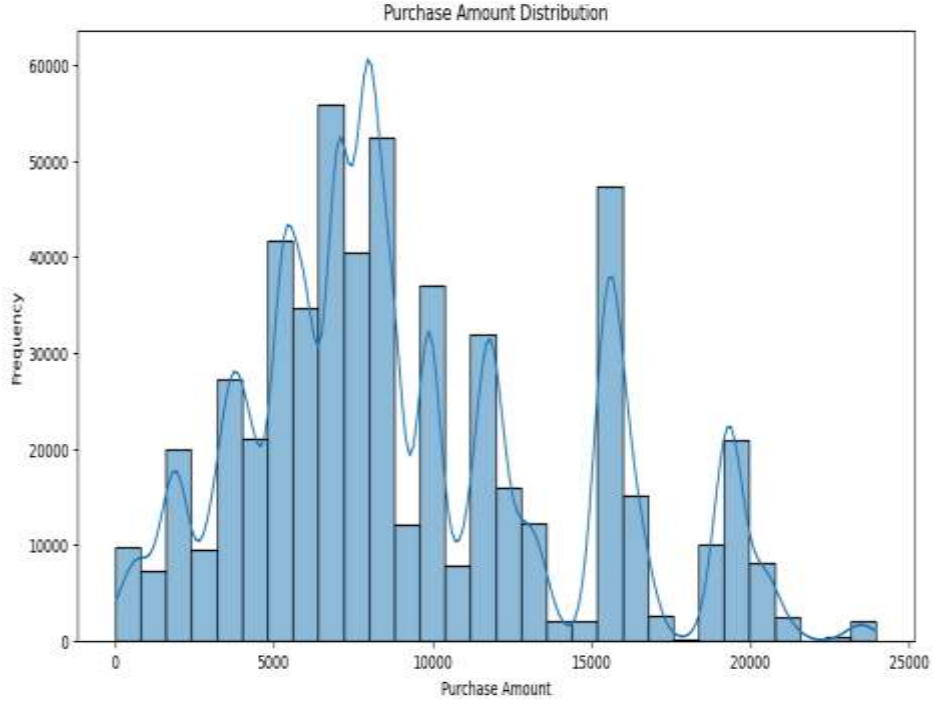
Step 7 (Optional): Deploy the model for real-time predictions if needed.

Step 8: Continuously monitor and update the model for ongoing accuracy.

This roadmap provides a concise plan to predict Black Friday sales accurately and extract helpful insight.

Analysing Black Friday Purchasing Patterns

- Histogram Insight: The Black Friday purchase data can be visualized using a histogram, revealing interesting patterns in consumer behavior.
- Majority Focus on Lower-Priced Items: The histogram shows that the majority of Black Friday shoppers chose to purchase lower-priced items such as clothing and gadgets.
- Long Tail on the Right: A distinctive feature of the histogram is the long tail on the right side, indicating that a few individuals made luxury purchases of electronics and jewelry.
- Diverse Spending Habits: The data highlights the diverse spending habits of Black Friday shoppers, with some opting for budget-friendly options and others splurging on high-end products.
- Valuable Consumer Insights: This data provides valuable insights into consumer preferences and behavior during the Black Friday shopping event. It helps retailers understand the range of spending choices made by their customers.

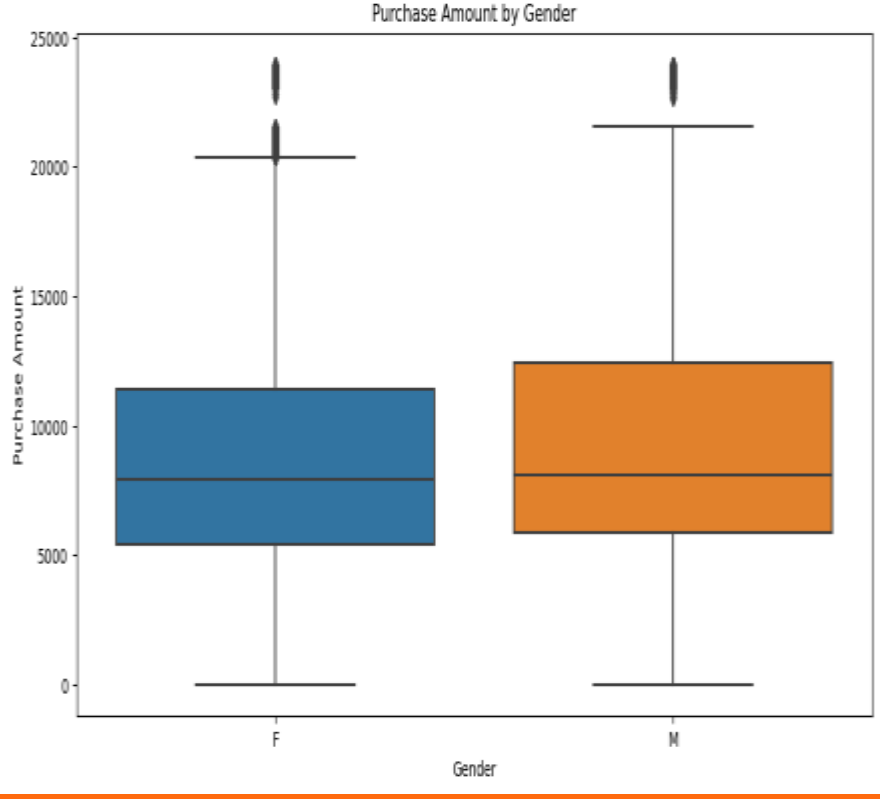


Black Friday Shopper Spendings: Gender-Based

- Male customers ('M') have a larger box, indicating a wider range of purchase amounts, with a higher median purchase amount compared to female customers ('F'). This suggests that, on average, male customers spent more during the Black Friday sales.

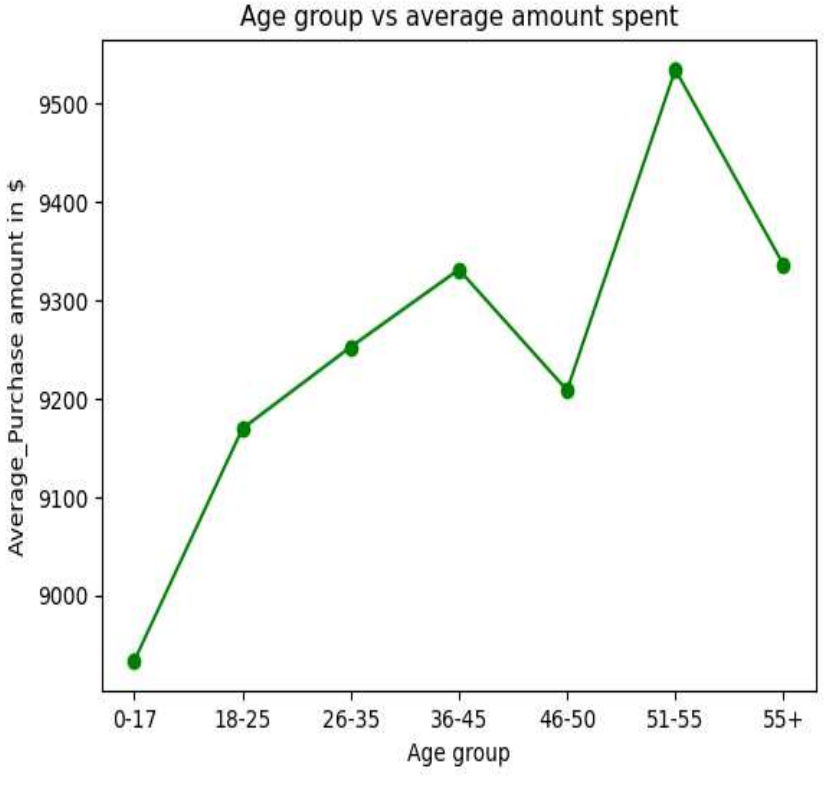
- There are potential outliers for both genders, as indicated by the individual points beyond the whiskers. These outliers represent exceptionally high purchase amounts for certain individuals.

- The boxplot provides a clear visual comparison of purchase amount distributions between male and female customers during the Black Friday sales, highlighting the differences in spending behavior.



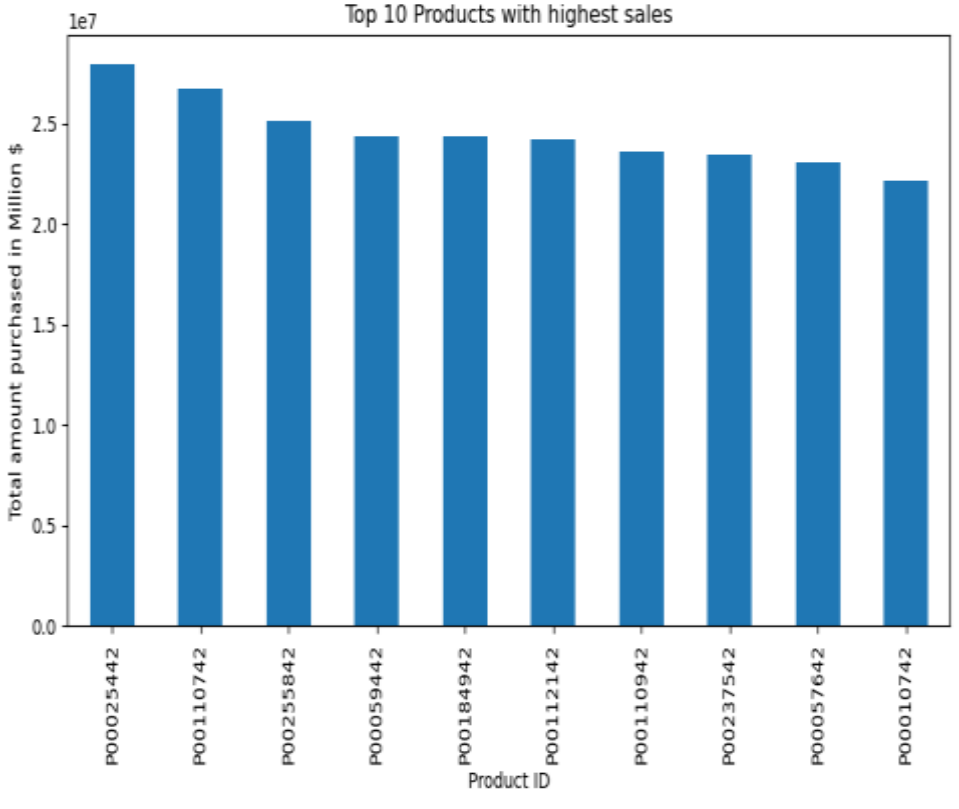
Age Group Analysis: Impact on Black Friday

- Spending Trends: The line plot reveals interesting trends in spending behavior among different age groups.
- Highest Spending: Customers in the "51-55" age group, shown by the highest point on the line, have the highest average purchase amount of approximately $9,534.81. This suggests that customers in this age group tend to spend the most on Black Friday.
- Age Group Comparisons: The plot allows for easy comparisons between age groups. For example, the "18-25" and "26-35" age groups have similar average spending levels, both around $9,169.66 and $9,252.69, respectively.
- Consistency: There is a consistent trend of higher average spending among older age groups compared to younger age groups, with some fluctuations.
- Marketing Implications: Understanding these spending patterns can help tailor marketing and sales strategies. For example, targeting promotions and products to the "51-55" age group might yield higher sales during Black Friday.



Top-Selling Products: Leading the Way in Black Friday

- The products listed in the bar plot are the top performers in terms of sales during Black Friday.
- Each product is uniquely identified by its product ID, making it easy for retailers and analysts to track the performance of specific products.
- The product with ID 'P00025442' achieved the highest sales, generating approximately $27,995,166 in revenue. It is followed closely by 'P00110742' and 'P00255842,' which also generated substantial sales.
- The top-selling products represent a diverse range of items, suggesting that customers had a wide variety of preferences during the sales event.
- Analyzing the top-selling products can provide valuable insights into customer preferences, marketing strategies, and product performance. Retailers can use this information to optimize their product offerings and marketing efforts for future sales events.

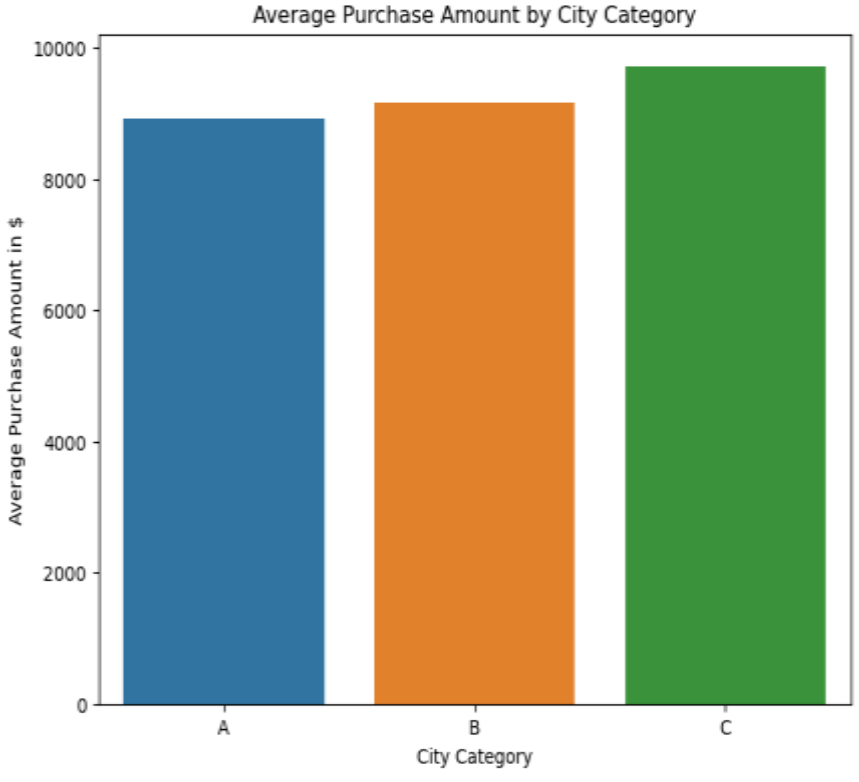


Location Matters: Average Spending by City Category on Black Friday

- City Category 'C' Leads in Average Spending: City category 'C' (possibly representing larger or more urban areas) has the highest average purchase amount of approximately $9,719.92. Customers in this category spent the most on average during Black Friday.

- City Category 'B' Follows Closely: City category 'B' (possibly representing mid-sized cities or urban areas) is the second-highest in terms of average purchase amount, with an average spend of about $9,151.30.

- City Category 'A' Has the Lowest Average Spend: City category 'A' (potentially representing smaller cities or rural areas) has the lowest average purchase amount at approximately $8,911.94. While spending is lower on average, it still contributes to overall sales.



Effect of Data Preprocessing

- Data Quality: Enhances data accuracy by addressing missing values and outliers.
- Model Performance: Improves model accuracy and effectiveness.
- Categorical Data: Converts categorical variables into numerical formats.
- Feature Engineering: Creates or transforms features to capture important patterns.
- Scaling: Ensures features are on a consistent scale for model stability.
- Dimensionality Reduction: Speeds up training and generalization by reducing feature dimensions.
- Imbalanced Data: Addresses class imbalances in classification problems.
- Interpretation: Facilitates data visualization and pattern understanding.
- Algorithm Compatibility: Makes data compatible with specific machine learning algorithms.
- Efficiency: Reduces computational complexity and training time.

Preparing for Analysis: X as Features, Y as Target

```
1 X = df.drop("Purchase",axis=1)
2 y = df["Purchase"]
```

X (Features): These are the details we use for analysis or prediction. For instance, customer demographics and product information.

Y (Target Variable): This is what we want to predict or analyze, like the purchase amount in our case.

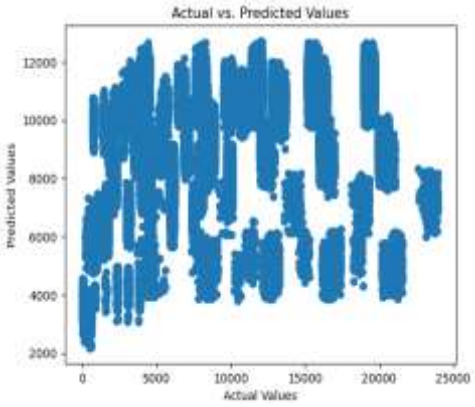
Purpose: Separating X (features) from Y (target variable) helps us predict purchases accurately without accidentally using the purchase amount as a predictor, ensuring reliable analysis and modeling.

Linear Regression Model Evaluation

- Linear Regression: Linear regression is a statistical technique used for modeling the relationship between a dependent variable (in this case, the "Target" variable, which is likely the purchase amount during Black Friday) and one or more independent variables (the "Features" such as customer demographics or product details).

- Root Mean Squared Error (RMSE): RMSE is a measure of how well the model's predictions align with the actual values. It quantifies the average error between predicted and actual values. In our case, the RMSE of approximately 4,686.37 indicates that, on average, our model's predictions deviate from the actual purchase amounts by this dollar amount.

- R-squared (R2) Score: R-squared is a statistical measure that represents the proportion of the variance in the dependent variable (purchase amount) that is explained by the independent variables (features). An R2 score of approximately 0.13 (or 13%) suggests that our linear regression model explains 13% of the variability in the purchase amounts. In other words, it indicates that our model captures some, but not all, of the factors influencing Black Friday purchases.




```
from sklearn.model_selection import train_test_split

X_train,X_test,Y_train,Y_test = train_test_split(X,Y,test_size=0.2,random_state=42)
```

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

```
# Linear regression

from sklearn.linear_model import LinearRegression

lr = LinearRegression()
lr.fit(X_train,Y_train)
```

```
LinearRegression()
LinearRegression()
```

In [42]:
```
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score

Y_pred = lr.predict(X_test)

print("Linear Regression: ")
print('rmse:', np.sqrt(mean_squared_error(Y_test,Y_pred)))
print('r2_score:', r2_score(Y_test,Y_pred))
```

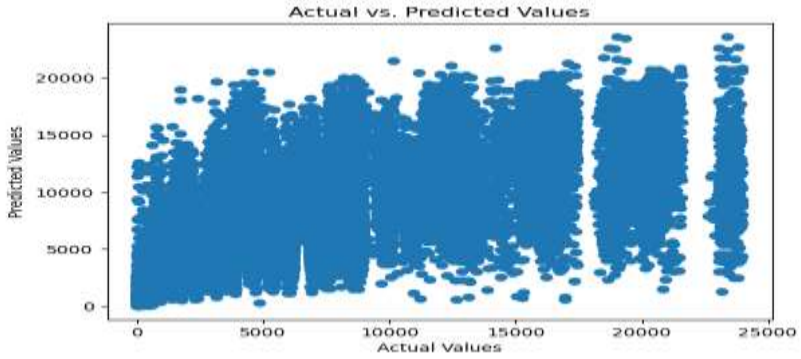
```
Linear Regression:
rmse: 4686.373005187138
r2_score: 0.12592793684162085
```

K-Nearest Neighbors (KNN) Regression Model Evaluation

- K-Nearest Neighbors (KNN) Regression: KNN is a machine learning algorithm used for regression tasks like predicting numerical values. It works by finding the nearest data points (neighbors) to a given point and averaging their values to make predictions. In our case, we used KNN to predict Black Friday purchase amounts.

- RMSE (Root Mean Squared Error) of approximately 3,499.75 signifies the average prediction error in our KNN model, indicating how much it deviates from actual purchase amounts.

- R-squared (R2) score of around 0.51 (51%) reveals that our KNN regression model explains half of the variability in Black Friday purchases, capturing a substantial portion of influencing factors.



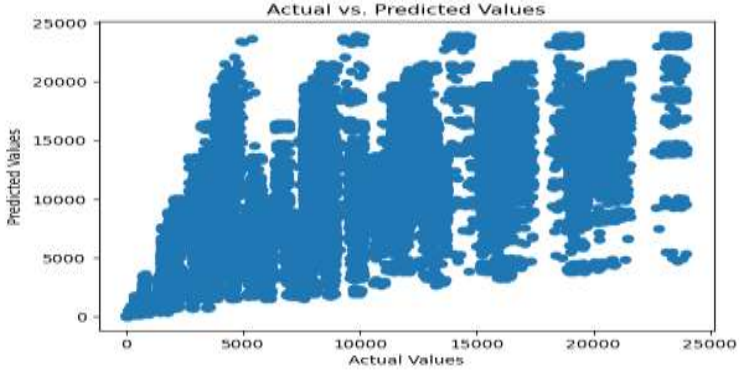

```
In [44]: # KNN Regressor
         from sklearn.neighbors import KNeighborsRegressor
         knn = KNeighborsRegressor()

In [45]: knn.fit(X_train, Y_train)

Out[45]: KNeighborsRegressor
         KNeighborsRegressor()

In [46]: Y_pred_knn = knn.predict(X_test)

In [47]: print("KNN regression: ")
         print("RMSE:",np.sqrt(mean_squared_error(Y_test, Y_pred_knn)))
         print("R2 score:", r2_score(Y_test, Y_pred_knn))

         KNN regression:
         RMSE: 3499.7488290381266
         R2 score: 0.5125306794266962
```

Decision Tree Regression Model Evaluation

- Decision Tree Regression: A Decision Tree is a machine learning algorithm used for both regression and classification tasks. In regression, it predicts numerical values by partitioning the data into subsets based on features and making predictions for each subset. In our case, we used Decision Tree Regression to predict Black Friday purchase amounts.

- Root Mean Squared Error (RMSE): RMSE is a measure of how well the model's predictions align with the actual values. It quantifies the average error between predicted and actual values. In this context, the RMSE of approximately 3,330.84 suggests that, on average, our Decision Tree model's predictions deviate from the actual purchase amounts by this dollar amount.

- R-squared (R2) Score: R-squared is a statistical measure that represents the proportion of the variance in the dependent variable (purchase amount) that is explained by the independent variables (features). An R2 score of approximately 0.56 (or 56%) indicates that our Decision Tree Regression model explains 56% of the variability in the purchase amounts. This suggests that our Decision Tree model captures a substantial portion of the factors influencing Black Friday purchases.



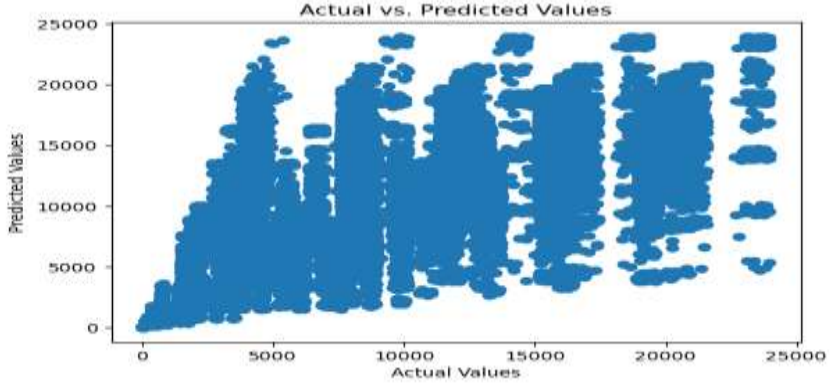

```
In [49]: # Decision Tree Regressor
         from sklearn.tree import DecisionTreeRegressor
         dec_tree = DecisionTreeRegressor()

In [50]: dec_tree.fit(X_train, Y_train)

Out[50]: DecisionTreeRegressor
         DecisionTreeRegressor()

In [51]: Y_pred_dec = dec_tree.predict(X_test)
         print("Decision tree regression: ")
         print("RMSE:",np.sqrt(mean_squared_error(Y_test, Y_pred_dec)))
         print("R2 score:", r2_score(Y_test, Y_pred_dec))

         Decision tree regression:
         RMSE: 3330.8445402991424
         R2 score: 0.558447588293947
```

Random Forest Regression Model Evaluation

- Random Forest Regression: Random Forest is an ensemble machine learning algorithm that combines multiple decision trees to make predictions. It's widely used for regression and classification tasks. In regression, it predicts numerical values, which is precisely what we did for Black Friday purchase amounts.

- Root Mean Squared Error (RMSE): RMSE is a measure of how well the model's predictions align with the actual values. It quantifies the average error between predicted and actual values. In this context, the RMSE of approximately 3,061.19 suggests that, on average, our Random Forest model's predictions deviate from the actual purchase amounts by this dollar amount.

- R-squared (R2) Score: R-squared is a statistical measure that represents the proportion of the variance in the dependent variable (purchase amount) that is explained by the independent variables (features). An R2 score of approximately 0.63 (or 63%) indicates that our Random Forest Regression model explains 63% of the variability in the purchase amounts. This suggests that our Random Forest model captures a substantial portion of the factors influencing Black Friday purchases



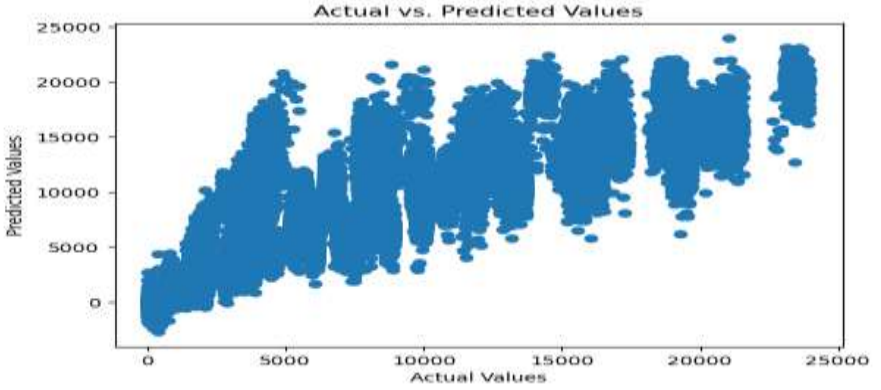

```
In [53]: #Random Forest Regressor
         from sklearn.ensemble import RandomForestRegressor
         ran_for = RandomForestRegressor()

In [54]: ran_for.fit(X_train, Y_train)

Out[54]: ▾ RandomForestRegressor
         RandomForestRegressor()

In [55]: Y_pred_ran_for = ran_for.predict(X_test)
         print("Random forest regression: ")
         print("RMSE:",np.sqrt(mean_squared_error(Y_test, Y_pred_ran_for)))
         print("R2 score:", r2_score(Y_test, Y_pred_ran_for))

         Random forest regression:
         RMSE: 3061.1942653919346
         R2 score: 0.6270459595923454
```

XGBoost Regression Model Evaluation

- XGBoost Regression: XGBoost (Extreme Gradient Boosting) is a powerful ensemble machine learning algorithm known for its predictive performance. It's used for both regression and classification tasks. In regression, as we did here, it predicts numerical values, specifically Black Friday purchase amounts.

- RMSE (Root Mean Squared Error) measures how close the model's predictions are to the actual values. An RMSE of about 2,897.76 means our XGBoost model's predictions are off by this amount on average.

- R-squared (R2) Score is a measure of how well the model explains the variability in the purchase amounts. An R2 score of approximately 0.67 (67%) indicates that our model explains 67% of the variability in Black Friday purchases, capturing a significant portion of the influencing factors.



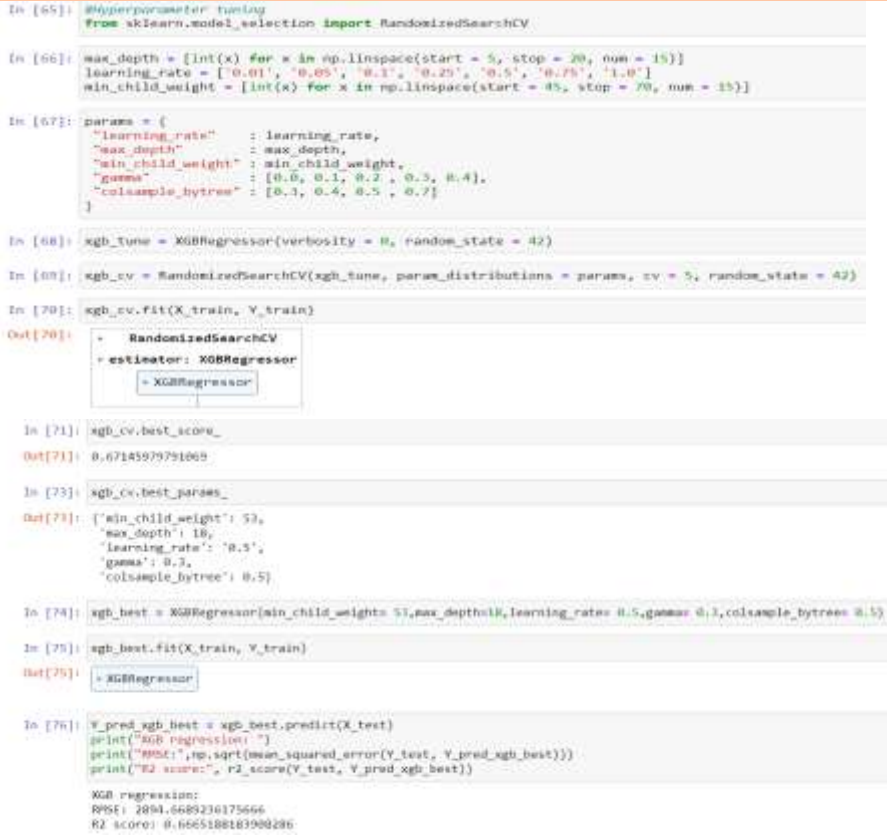

```
In [57]: #XGB Regressor
         !pip install xgboost
         from xgboost.sklearn import XGBRegressor
         xgb_reg = XGBRegressor(learning_rate=1.0, max_depth=6, min_child_weight=40,seed=0)

         Requirement already satisfied: xgboost in c:\users\abdul\anaconda3\lib\site-packages
         Requirement already satisfied: scipy in c:\users\abdul\anaconda3\lib\site-packages
         Requirement already satisfied: numpy in c:\users\abdul\anaconda3\lib\site-packages

In [58]: xgb_reg.fit(X_train, Y_train)
Out[58]: XGBRegressor

In [59]: Y_pred_xgb = xgb_reg.predict(X_test)
         print("XGB regression: ")
         print("RMSE:",np.sqrt(mean_squared_error(Y_test, Y_pred_xgb)))
         print("R2 score:", r2_score(Y_test, Y_pred_xgb))

         XGB regression:
         RMSE: 2897.762592083295
         R2 score: 0.6658056235176206
```

Hyperparameter Tuning with RandomizedSearchCV in XGBoost Regression

- Hyperparameter tuning with RandomizedSearchCV in XGBoost Regression enables us to identify the optimal hyperparameters that maximize our model's predictive power. By systematically exploring different hyperparameter combinations, we can enhance the accuracy and performance of our XGBoost Regression model for predicting Black Friday purchases.

- The RMSE score highlights that our XGBoost Regression model's predictions are, on average, off by approximately 2,894.67, signifying improved accuracy compared to earlier versions of the model.

- The R2 score of 0.67 underscores the model's ability to explain 67% of the variability in purchase amounts. This indicates that our tuned XGBoost model effectively captures a substantial portion of the factors influencing Black Friday sales.

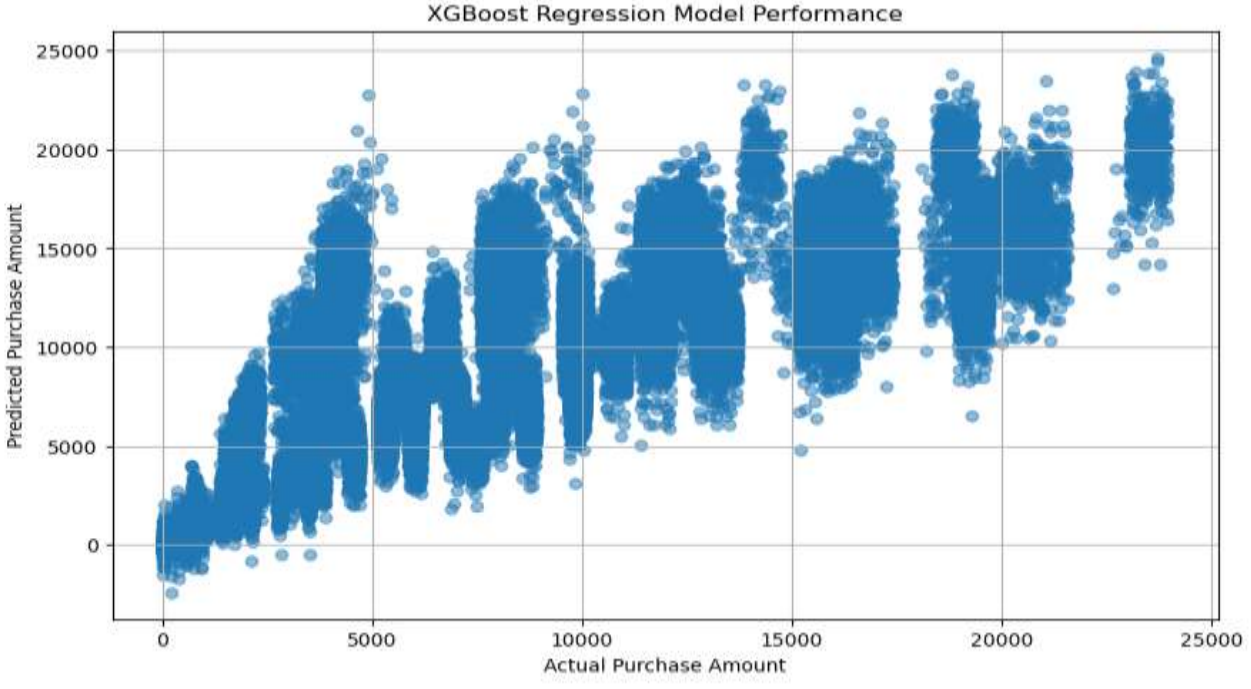


Conclusion

the XGBoost Regression model has demonstrated superior predictive performance with a low RMSE, high R2 score, and a balanced blend of interpretability and accuracy. This makes it the ideal choice for our Black Friday sales prediction task, offering valuable insights for future sales strategies.

RMSE: 2894.6689236175666

R2 score: 0.6665188183908286



Future Scope

Our XGBoost model has been a game-changer in optimizing Black Friday sales, and it continues to unlock new possibilities for our sales strategies. The future scope, guided by our model, encompasses the following key areas:

Enhanced Predictions: Refining our XGBoost model for top-tier sales forecasting. Real-time Personalization: Custom offers and real-time product recommendations. Dynamic Pricing: Real-time pricing for increased revenue. Inventory Management: Predictive analytics to manage stock efficiently. Marketing: Fine-tuning marketing based on model insights. Customer Segmentation: Custom strategies for better conversions.Cross-selling: Promoting related products and upgrades.A/B Testing: Constant strategy optimization.Loyalty Programs: Building programs for lasting customer relations.Multi-Channel Integration: Merging online and offline shopping experiences.Supply Chain: Collaborative and prompt product supply.Sustainability: Eco-friendly sales practices. Driven by our XGBoost model, these strategies ensure precision, customization, and efficiency for Black Friday.

THANK YOU!