DMML ASSIGNMENT 1

Name: Malde Dharmi
Roll No: MDS202333

Task 1: Customer Churn

Preprocessing:

Before building the classifiers, the following preprocessing steps were applied to the dataset:

Scaling: Some variables like AnnualIncome, TotalSpend, AvgTransactionAmount were scaled using Standard scaling technique to ensure that all features are on a similar scale. This step is essential for algorithms sensitive to the scale of features, such as Adaboost.

One-Hot Encoding: Gender was encoded using one-hot encoding to convert it into a numerical format suitable for machine learning algorithms. This step helps in handling categorical variables.

Label Encoding: Some categorical variables with ordinal information like LastPurchaseDaysAgo, EmailOptIn, Churn were encoded using label encoding to convert them into numerical format while preserving the ordinal relationship between categories.

Classifier Models:

For the Adaboost classifier, the following hyperparameters were selected to prevent overfitting:

n_estimators: 5
base_estimator (Max Tree Depth): 1

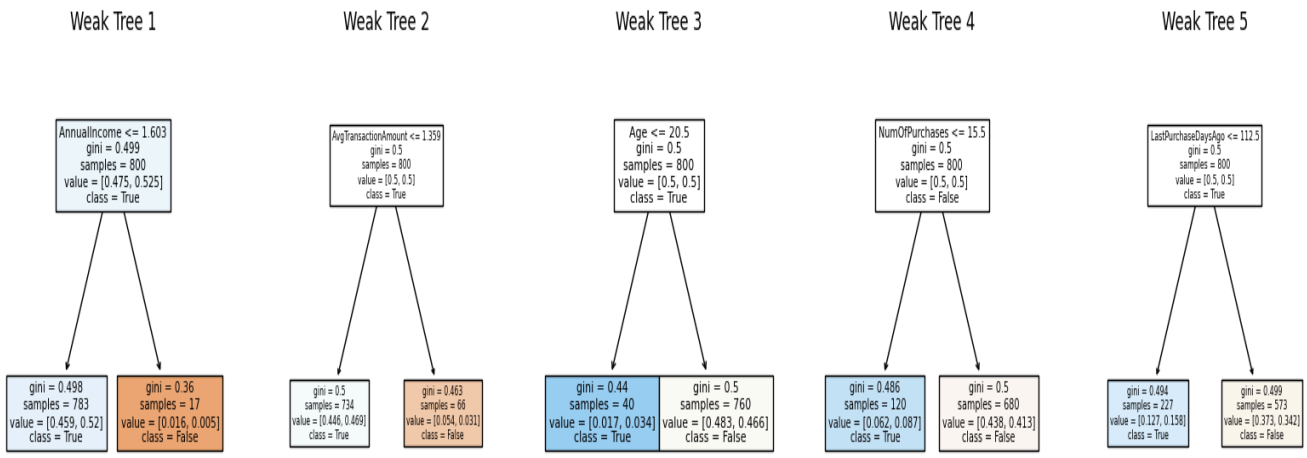
Results:

Adaboost Classifier Accuracy: 51%
Time Taken to train model**:** 0.0257 seconds

Adaboost Analysis:

To understand why the model performs poorly and gain insights into its behavior, the following steps were taken:

Visualizing Individual Trees: Every tree in the Adaboost ensemble was visualized to examine its structure and decision boundaries. These visualizations provide insights into the complexity and diversity of the individual trees.



Importance of Each Tree: The importance of each tree in the Adaboost ensemble was assessed to understand its contribution to the final prediction. Surprisingly, it was found that every tree was equally important, indicating a lack of diversity in the ensemble.

Feature Importance Analysis: An analysis of feature importance was conducted to identify which features were most influential in making predictions. We found that LastPurchaseDaysAgo, AnnualIncome, AvgTransactionAmount.

For the Random Forest algorithm, the following hyperparameters were selected:

n_estimators: 30
oob_score: True

Out-of-Bag (OOB) Error:

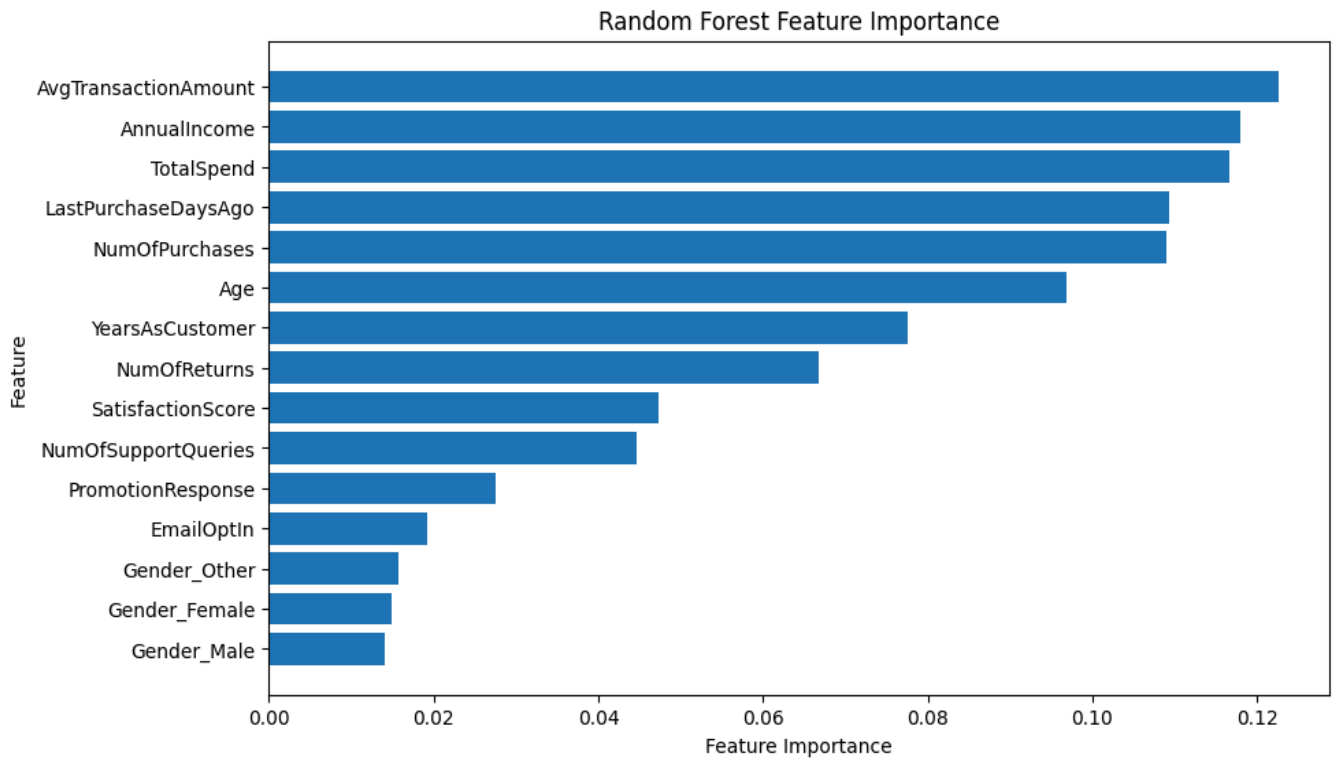
The out-of-bag (OOB) error was utilized to evaluate the performance of the Random Forest model during training. This approach provides an estimate of the model's performance on unseen data without the need for a separate validation set.

Results:

OOB_score: 0.5
Time Taken for Training: 1.3412 seconds

Random Forest Analysis:

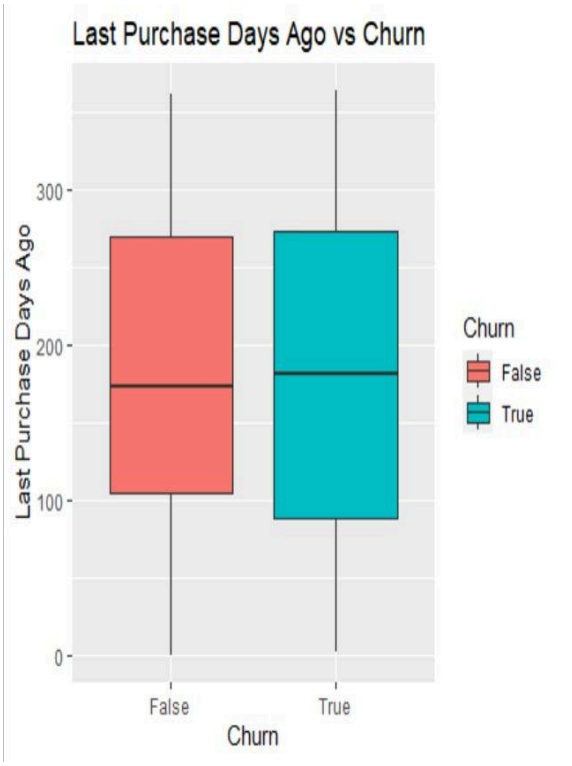
To gain insights into the importance of different features in the Random Forest model, a feature importance graph was plotted. This graph provides a visual representation of the relative importance of each feature in predicting the target variable (churn).

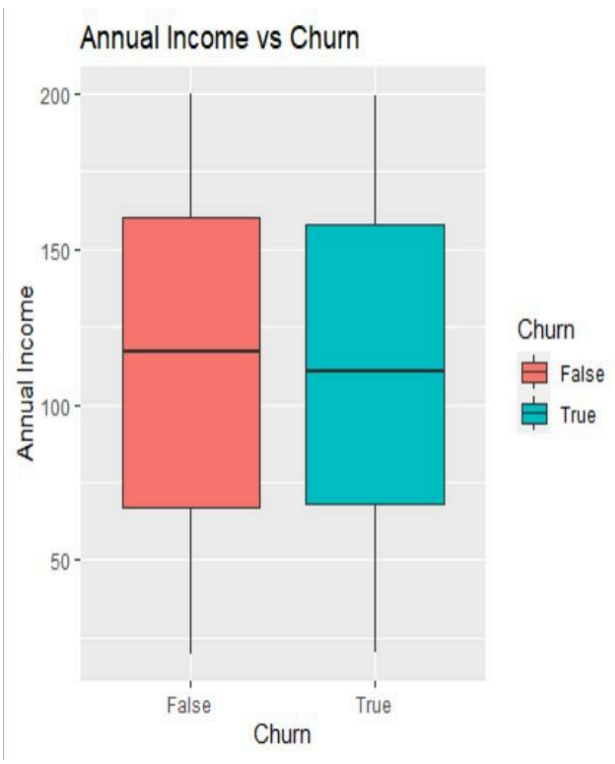


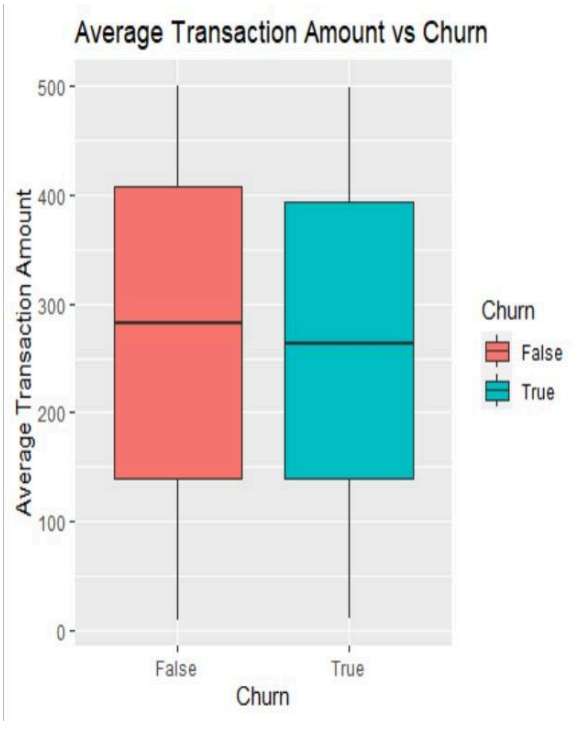
Evaluation Metric:

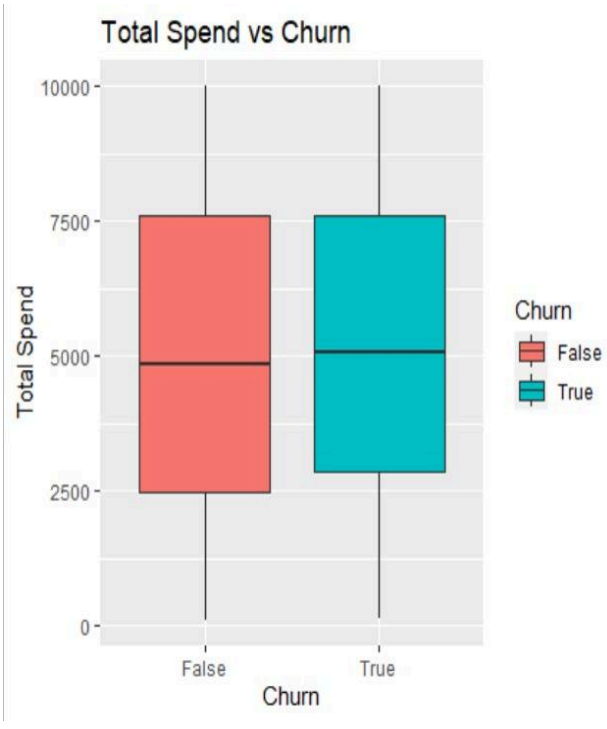
Accuracy score was used as the evaluation metric to assess the performance of the classifiers. If we are introducing some scheme for churned customers than recall might be a better evaluation matrix but since the data is well balanced and we don’t have proper application information we are using accuracy

Below are the box plot of important features by Random Forest and Adaboost









Despite the model identifying certain features as important, it was observed that these features did not significantly contribute prediction. This discrepancy raises

questions about the reliability of the model's interpretation of feature importance and its implications for decision-making.

Potential issues

Data Quality: Issues with data quality, such as missing values, outliers, or noise, may have influenced the model's interpretation of feature importance and its ability to accurately identify relevant features.

Feature Engineering: The features selected for modeling may not adequately capture the underlying patterns in the data or may be poorly engineered, leading to misinterpretation of feature importance by the model.

Comparison:

- Both models achieved similar accuracy scores, indicating comparable performance in predicting customer churn.
- Adaboost model trained significantly faster than Random Forest, with training time being approximately 50 times faster.
- Both models showed issues with feature importance interpretation, where certain features identified as important did not significantly contribute to prediction.
- While Adaboost utilized fewer estimators and had faster training time, Random Forest considered a larger number of estimators and incorporated out-of-bag (OOB) error during training.

Task 2: Supermarket Sales

Preprocessing:

Before building the classifiers, the following preprocessing steps were applied to the dataset:

Scaling: Some variables like UnitPrice, Tax, Total were scaled using Standard scaling technique to ensure that all features are on a similar scale. This step is essential for algorithms sensitive to the scale of features, such as Adaboost.

One-Hot Encoding: CustomerType, Branch, PaymentType, PaymentType were encoded using one-hot encoding to convert it into a numerical format suitable for machine learning algorithms. This step helps in handling categorical variables.

Gender Prediction:

For predicting Gender, I used a Decision Tree Classifier and a Random Forest Classifier.

For the decision tree, the following hyperparameters were selected to prevent overfitting:

Criterion: 'gini',
Min_impurity_decrease: 0.005

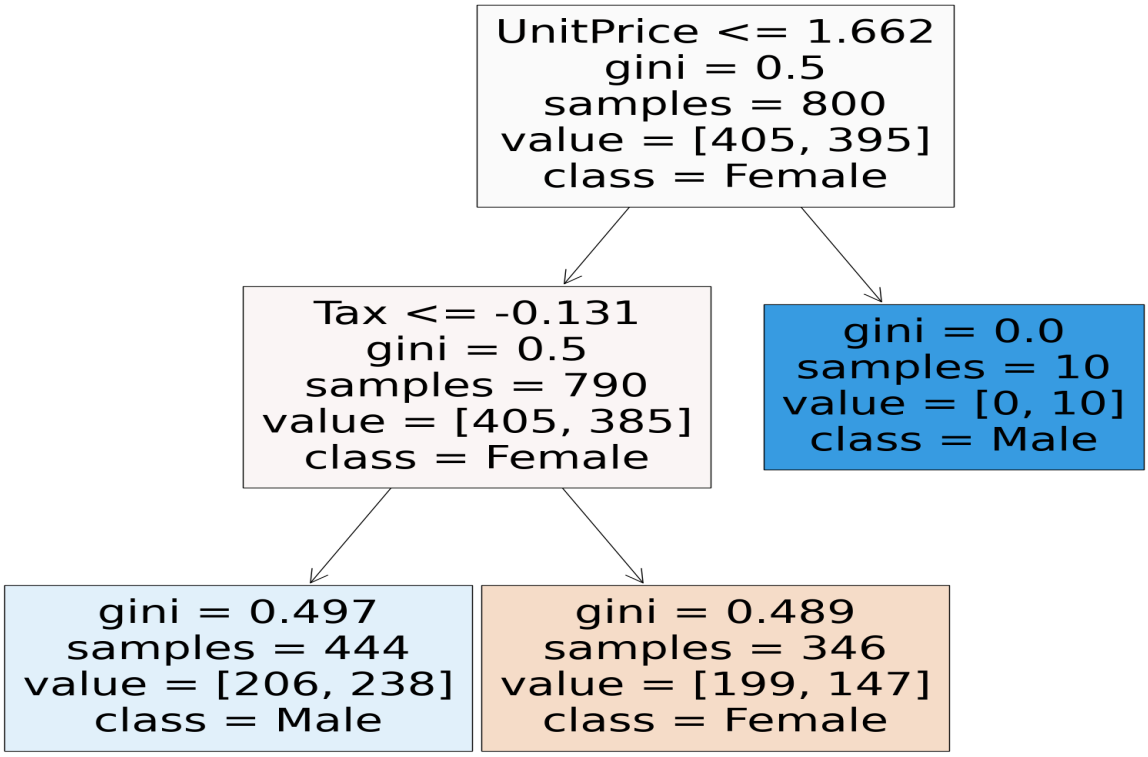
Results:

Decision Accuracy: 49.5%
Time Taken to train model**:** 0.012521 seconds

Decision Tree Analysis:

To understand why the model performs poorly and gain insights into its behavior, the following steps were taken:

Visualizing Trees: To delve deeper into the decision-making process of the Decision Tree Classifier for gender prediction, visualization of the decision tree was conducted. The decision tree graph portrays the hierarchical structure of decisions made by the model based on the input features.



For the Random Forest algorithm, the following hyperparameters were selected:

n_estimators: 20
oob_score: True

Out-of-Bag (OOB) Error:

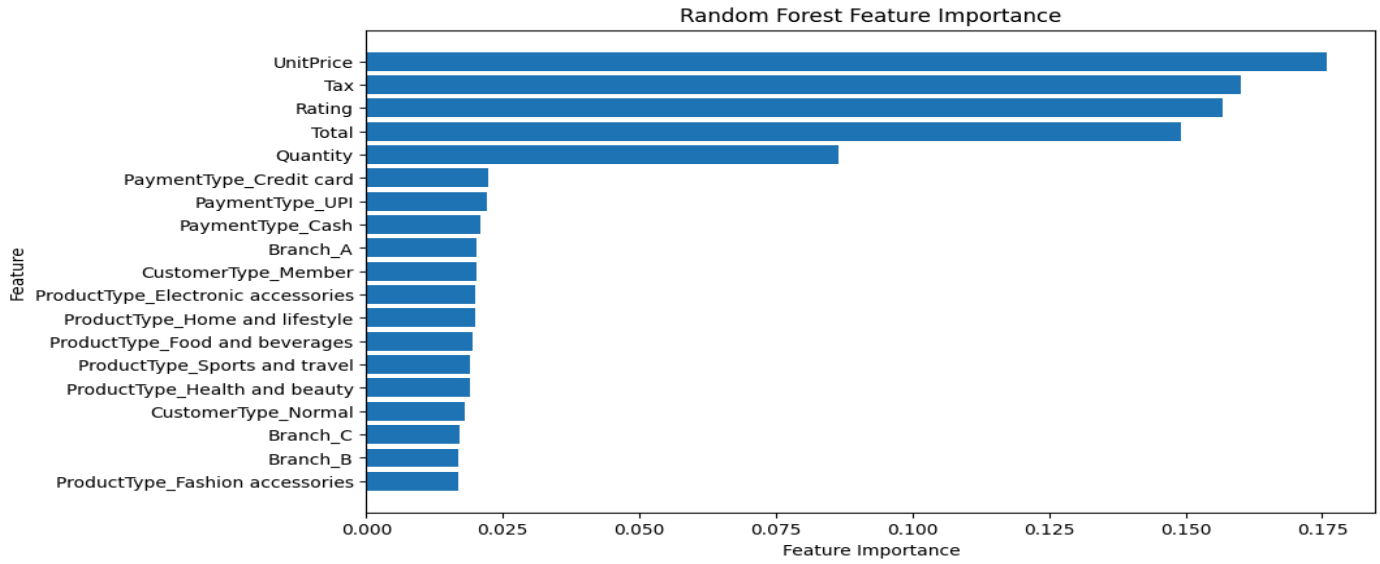
The out-of-bag (OOB) error was utilized to evaluate the performance of the Random Forest model during training. This approach provides an estimate of the model's performance on unseen data without the need for a separate validation set.

Results:

OOB_error: 0.484
Time Taken for Training: 0.12413 seconds

Random Forest Analysis:

To gain insights into the importance of different features in the Random Forest model, a feature importance graph was plotted. This graph provides a visual representation of the relative importance of each feature in predicting the target variable (Gender).



Evaluation Metric:

Accuracy score was used as the evaluation metric to assess the performance of the classifiers.

Conclusion:

Both models exhibited limitations in accurately predicting gender, with the Decision Tree Classifier performing suboptimally and the Random Forest Classifier showing a modest improvement. Despite visualizing the decision trees and analyzing feature importance, further investigation into the dataset's characteristics and potential feature engineering may be necessary to enhance model performance. Additionally, considering alternative evaluation metrics or experimenting with different algorithms could be beneficial in addressing the challenges encountered in gender prediction.

Rating Prediction:

For predicting Rating, I used a Decision Tree Regressor and a Linear Regression Model.

For the decision tree, the following hyperparameters were selected to prevent overfitting:

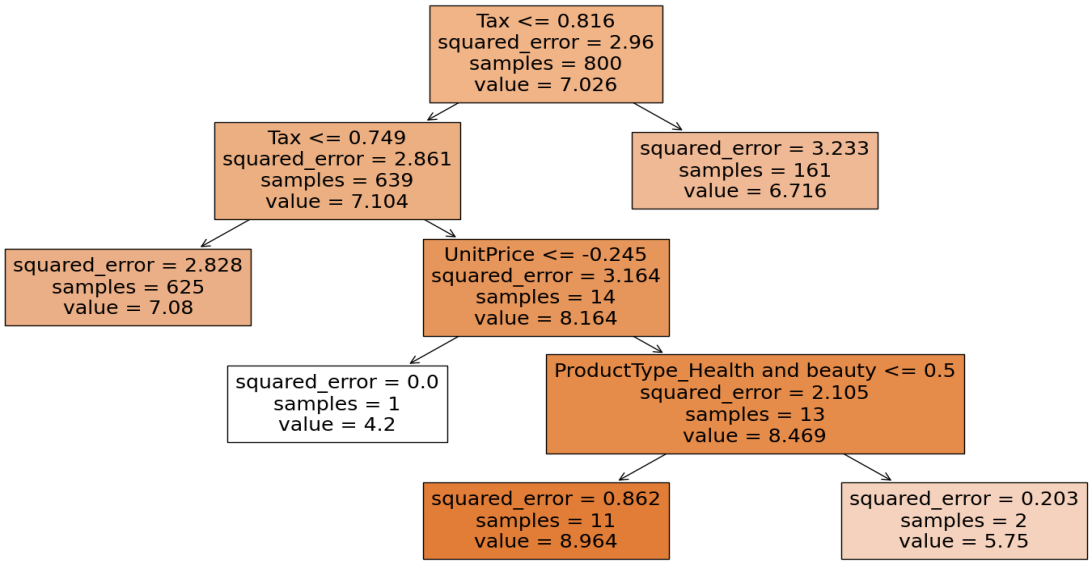
Min_impurity_decrease: 0.02

Results:
Decision Tree Regressor MAE: 1.5005725271597965
Time Taken to train model**:** 0.025320768356323242 seconds

Decision Tree Analysis:

To understand why the model performs poorly and gain insights into its behavior, the following steps were taken:

Visualizing Trees: To delve deeper into the decision-making process of the Decision Tree Regressor for Rating prediction, visualization of the decision tree was conducted. The decision tree graph portrays the hierarchical structure of decisions made by the model based on the input features.



For the Linear Regression model:

Results:
Linear Regression MAE: 1.483268299566657
Time Taken for Training: 0.04983091354370117 seconds

Linear Regression Analysis:

After training the Linear Regression model, we derived the following equation to represent the relationship between the features and the predicted rating:

Rating = (0.18 * UnitPrice) + (0.05 * Quantity) + (-0.16 * Tax) + (-0.16 * Total) + (-0.02 * CustomerType_Member) + (0.02 * CustomerType_Normal) + (-0.13 * ProductType_Electronic accessories) + (0.10 * ProductType_Fashion accessories) + (0.02 * ProductType_Food and beverages) + (0.06 * ProductType_Health and beauty) + (-0.09 * ProductType_Home and lifestyle) + (0.05 * ProductType_Sports and travel) + (0.04 * PaymentType_Cash) + (-0.05 * PaymentType_Credit card) + (0.01 * PaymentType_UPI) + (-0.02 * Branch_A) + (-0.08 * Branch_B) + (0.10 * Branch_C) + (-0.00 * Gender_Female) + (0.00 * Gender_Male) + 6.72

Evaluation Metric:

MAE is preferred as a performance metric for regression models due to its interpretability, robustness to outliers, and equal weighting of errors regardless of their magnitude.

Conclusion:

Both models displayed limitations in accurately predicting ratings, with the Linear Regression Model exhibiting slightly better performance compared to the Decision Tree Regressor. Further analysis and refinement may be necessary to enhance the predictive accuracy of both models. This could include exploring additional features, refining existing features, or experimenting with alternative algorithms. Additionally, considering the interpretability and robustness of the

Linear Regression Model, further investigation into its feature importance and coefficient significance could provide valuable insights for improving rating predictions.