Dataset

The dataset includes:

- **Customer Demographics:** Age, Gender, Marital Status, City Type, and Stay in Current City.
- **Product Details:** Product ID, Product Categories.
- **Purchase Amount:** Total purchase amount from the last month.

Key Columns:

- `Gender`: Customer's gender (encoded as 0 for Male and 1 for Female)
- `Age`: Age of the customer
- `Occupation`: Customer's occupation
- `Stay_In_Current_City_Years`: Number of years the customer has stayed in the current city
- `Marital_Status`: Marital status of the customer
- `Product_Category_1`: Primary product category
- `Product_Category_2`: Secondary product category
- `Product_Category_3`: Tertiary product category
- `Purchase`: Total purchase amount (target variable)

Model

Algorithm Used: Random Forest Regressor

Why Random Forest Regressor?

- **Handles Non-Linearity:** Capable of capturing complex relationships between features and target variables.
- **Robustness:** Reduces overfitting and provides reliable predictions with ensemble learning.
- **Feature Importance:** Helps identify the most influential features affecting purchase amounts.

. Why Random Forest Regressor for this project?

- **Handles Non-Linearity**: Random Forest can handle non-linear relationships between features, which is important because the relationship between customer demographics, product categories, and purchase amount may not be linear.
- **Feature Importance**: It provides insights into which features are most influential in predicting the purchase amount. This could be beneficial for understanding customer behavior. In this case it is Product_Category_1.
- **Less Prone to Overfitting**: Random Forest uses multiple decision trees and averages their predictions, which makes it more robust and less prone to overfitting compared to individual decision trees.
- **Handles Missing Data**: While RandomForest does not automatically impute missing values, it is still more tolerant of missing data than many other models.
- **Scalability**: Works well with large datasets and high-dimensional data, which fits a dataset with 783,667 rows and many features.
- **Default Robustness**: It usually performs well without needing extensive hyperparameter tuning right away.

Performance Metrics

Model Evaluation:

- **Mean Absolute Error (MAE):** $2,221 (23.98% of Average Purchase Amount)
- **Root Mean Squared Error (RMSE):** $3,051 (32.94% of Average Purchase Amount)
- **R-squared (R^2):** 0.629

Interpretation:

- **MAE (23.98%):** The model's average prediction error is about 24% of the average purchase amount, indicating a moderate level of accuracy.
- **RMSE (32.94%):** The error in squared units is about 33% of the average purchase amount, suggesting some significant deviations in predictions.
- **R-squared (0.629):** The model explains approximately 63% of the variance in purchase amounts, which is a moderate fit.

Acceptability:

- For marketing purposes, the current error rates may be acceptable, but improvements may be needed for financial forecasting or other high-precision applications.