## **Black Friday Purchase Prediction Walkthrough**

This is a walkthrough of a Python code that predicts purchase behavior in a Black Friday sales dataset. Here's a breakdown of the steps involved:

**1. Importing Libraries:**

* Necessary libraries for data manipulation, visualization, modeling, and evaluation are imported.

**2. Loading the Data:**

* The training data (train.csv) is loaded into a pandas DataFrame (df).
* We confirm that the data contains information about various features and a target variable named "Purchase" that needs to be predicted.
* We check for missing values (isnull().sum()) in each column.

**3. Visualization:**

* Exploratory data analysis is performed to understand the data distribution and relationships between features.
* Count plots show the distribution of categorical variables like gender.
* Bar plots visualize purchase trends across different occupations and how gender interacts with occupation for purchases.
* Boxplots identify potential outliers in purchase amounts and product categories.

**4. Pre-Processing:**

* Data cleaning and transformation are crucial for model performance.
* The "Product\_ID" column is preprocessed by removing leading characters and scaling numerical values.
* The column "Product\_Category\_3" with many missing values is dropped.
* The missing values in "Product\_Category\_2" are imputed with the mean value.
* Label encoding is applied to convert categorical columns ("Gender", "City\_Category", "Age") into numerical representations for modeling.
* Inconsistent values in "Stay\_In\_Current\_City\_Years" are replaced for consistency.
* Data type conversion is applied to ensure compatibility with modeling.

**5. Distribution Plot:**

* Distributions of various features are visualized to identify potential skewness (asymmetry).
* Purchase amounts exhibit skewness, so a log transformation is applied to normalize the distribution.

**6. One-Hot Encoding:**

* The get\_dummies() function creates dummy variables for each category within categorical features, allowing the model to understand these features better.

**7. Train-Test Split:**

* The data is split into training and testing sets. The training set is used to build the model, and the testing set is used to evaluate its performance on unseen data.
* Standardization is applied to the training and testing sets to ensure features are on a similar scale.

**8. Modeling:**

* Three different models are evaluated for purchase prediction:  
  + **Linear Regression:** This is a baseline model that performs poorly due to the non-linear relationships in the data (R-squared score is low, and RMSE is high).
  + **Decision Tree:** This model performs better than linear regression, achieving a higher R-squared score and lower RMSE.
  + **Random Forest:** This ensemble model consisting of multiple decision trees outperforms the previous models, achieving the best R-squared score and lowest RMSE.

**9. Evaluation:**

* The performance of each model is evaluated using R-squared score, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

**10. Prediction on New Data:**

* The Random Forest model, being the best performing model, is used to predict purchase behavior on a new dataset (test.csv).
* The new data undergoes similar pre-processing steps as the training data.
* Predictions are made on the new data using the trained Random Forest model.

**11. Saving Predictions:**

* Predicted purchase values are saved in a CSV file named "BlackFridayPredictions.csv" along with corresponding User IDs and Product IDs from the new data.