Summary:

I read all dataset and merge them based on the user ID. The initial dataset has 2,845,930 rows and 29 columns and it includes string, float, and integer data types.
The first step of any data scientist project is Exploratory Data Analysis where we use different libraries to visualize the features and show statistical relationship between them. This analysis gives us a good insight into the data.
First, I found the percentage of missing values of each column. Some of the features had missing values more than 30% of data samples. I couldn't drop those feature as they had important information so I dropped all rows with missing values.

In the next part, I found different categories of features and I used Seaborn Countplot to show different categories and the number of data samples in each category. Some of the features had more than 4 categories. I made a separate group for all categories for the case where their data sample counts are not in the top 3 categories. That made it easier to convert these categorical variables to numerical variables and the size of the final dataset did not grow dramatically. It should be noted that I used get_dummies to transform the categorical variables to numerical ones. In addition, I studied the relationship between features and the target (can be found in the notebook) and I dropped the features with non-valuable and repetitive information.

Then I used boxplot and histogram plots, and I showed that the features are highly skewed and they had outliers. There are many ways to make the data more normalized such as using Box-Cox and Yeo-Johnson methods. I used Box-Cox method to show its performance in reducing the skewness of some features. We know that the performance of certain methods such as linear regression and logistic regression are affected by skewness and outliers in data, but research shows that the performance of tree-based methods such as Random Forest are less impacted by the outliers. In this project, I used Random Forest classifier which is more robust for outliers and it can be used for high dimension and nonlinear datasets.

It should be noted that our dataset was highly imbalanced. So, I used SMOTE and class-weights methods to make the dataset more balanced and to improve the performance of the ML model.

I compared the performance of regular Random Forest classifier, Random Forest classifier with SMOTE and RandomUnderSampler method, and Random Forest classifier with class-weights. To compare these three classifiers, I considered F1-score since it is a better metric when there are imbalanced classes. From the classification report, it was concluded that SMOTE combined with RandomUnderSampler model had a better performance. I also used the predict_proba to find the probability of users that are likely to spend money after finishing the tutorial and I found the feature importance scores of the model. More details and the conclusion can be found in the notebook.