Summary of Approach

I started off with exploring the data to identify variables that could be useful as input for the model, and whether it needs to be cleaned. I examined each variable to see its level of impact on whether a user is a spender. From exploring the data, I found a few additional variables that could be useful for prediction, for example, the number of intervals a user gained experience in out of the 4 time intervals.

I decided to leave out the redeemer and scribe action variables as they represent users' total actions since install. The question we want to answer is to identify the spend behavior for new users, so knowing the total amount might not work. Instead if we knew the number of actions at the time they finish the tutorial or early in the user journey that would be useful.

The model I used includes adding features, transforming the hour of the day and day of week to capture the cyclical nature of the feature, and standardizing the features.

I trained the model separately for users who have and have not completed the tutorial, as the behavior could be quite different. From exploring the data, users that did not complete the tutorial also had some substantial amount of experience within the first 4 time intervals.

I split the data into a training (60%) and testing (40%) sets, where model fitting is done with the train data set and evaluation on the test.

I subsampled the training data for spenders (total spend > 0) to try to improve precision and recall.

As a baseline, I used random forest with a selected number of features and had no other modifications. It uses the default parameters.

I ran with another random forest but with improved features and added more weight to the spender label to try to improve precision and recall.

Next step would be to perform a grid search on the random forest parameters and feature selection.