In the challenge two datasets are given - train.csv and test.csv.

train.csv contains 70% of the overall sample (243,787 subscriptions to be exact) and and reveal whether or not the subscription was continued into the next month (the “ground truth”).

test.csv dataset contains the exact same information about the remaining segment of the overall sample (104,480 subscriptions to be exact), but does not disclose the “ground truth” for each subscription.

Task : Using the patterns you find in the train.csv data, predict whether the subscriptions in test.csv will be continued for another month, or not.

Submission Format : A dataframe (prediction_df with two columns and exactly 104,480 rows (plus a header row). The first column is CustomerID so that we know which prediction belongs to which observation. The second column is called predicted_probability and should be a numeric column representing the likellihood that the subscription will churn.

Grading : To determine final score, they compared the predicted_probability predictions to the source of truth labels for the observations in test.csv and calculate the ROC AUC.