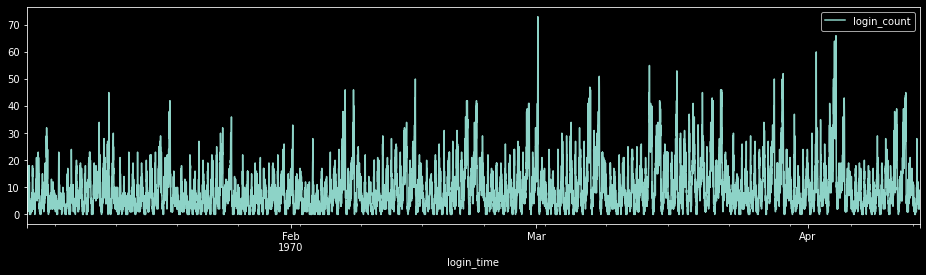
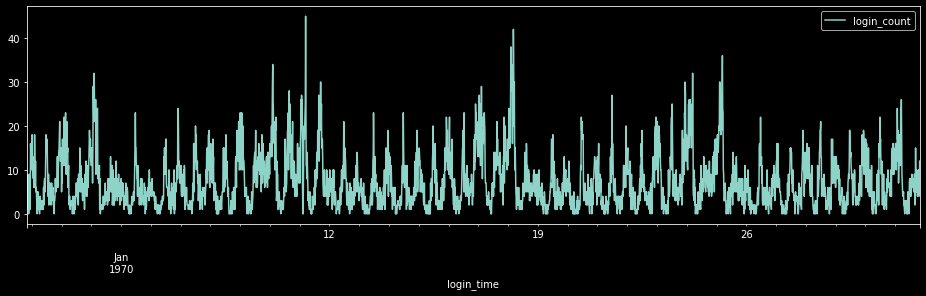
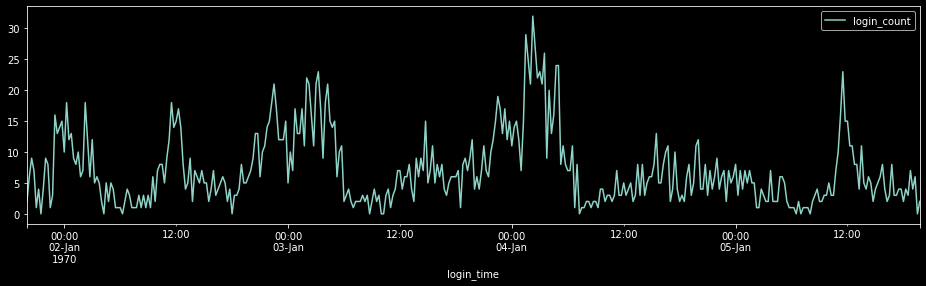
Ultimate Challenge Report

Part 1:

The time series data appears to be seasonal showing similar patterns every week, there seems to be more activity at the start/end of every week, and the hours of 12am and 12pm seem to be the most active hours.





Part 2:

1. The key measure of success for this experiment would be the difference in the number of times drive partners cross the toll bridge for business purposes after implementing a program to reimburse tolls. This experiment can be considered successful if there is a statistically significant positive difference between frequency of drive partners crossing the toll bridge before giving reimbursement for the toll and after.
2. For this experiment the frequency of on-the-clock driver partners crossing the toll bridge will be recorded for a year before implementing the toll reimbursement program and a year after as well as the date/time, home city, company, and any other relevant detail of the driver partner. The 1 year length of time is recommended to account for any seasonal variables. To determine whether the difference is significantly different, running a hypothesis test with the following hypotheses should suffice:

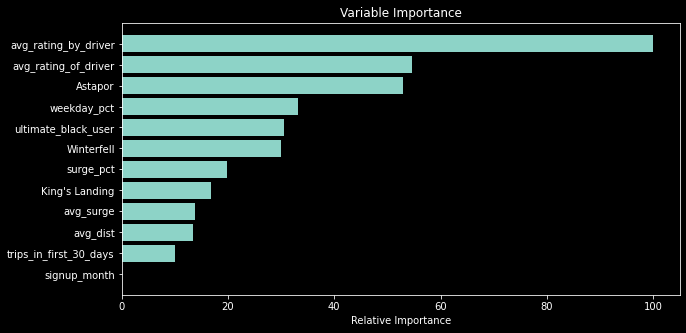
null hypothesis – *(# of times on-the-clock driver partners crossing the toll bridge* ***after*** *reimbursing tolls)/( # of times on-the-clock driver partners crossing the toll bridge both* ***before or after*** *reimbursing tolls)* = .5

alternative hypothesis - *(# of times on-the-clock driver partners crossing the toll bridge* ***after*** *reimbursing tolls)/( # of times on-the-clock driver partners crossing the toll bridge both* ***before or after*** *reimbursing tolls)* > .5

Part 3:

According to the instructions, ‘we consider a user retained if they were “active” (i.e. took a trip) in the preceding 30 days’. Without any further clarification, I assumed they meant 30 days preceding the latest ‘last trip date’ of the samples taken (which would also be the earliest date the data could have been collected). By this definition, before preparing the data for analysis, 548 of 50,000 users were considered to be retained. Because this dataset is severely imbalanced, any model trained will typically guess the majority case and not have any predictive validity. To balance this data, synthetic minority oversampling was used to create a balanced dataset. Additionally, missing values were removed. The resulting dataset contained 82,424 entries, half of which were ‘retained’ users (synthetically sampled).

A random forest algorithm was applied to the balanced dataset and its parameters were optimized with gridsearch. The resulting model attained a cross validated accuracy of .926 with both precision and recall of at least .9 for both retained and un-retained users. The following is a variable importance plot showing which variables had the most influence on the results:



This shows that the most important variable in retaining users by far is the average rating of the rider suggesting that driver opinions of their rider(s) should be investigated.