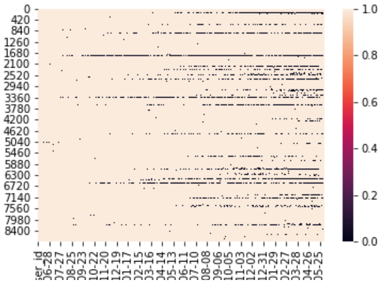
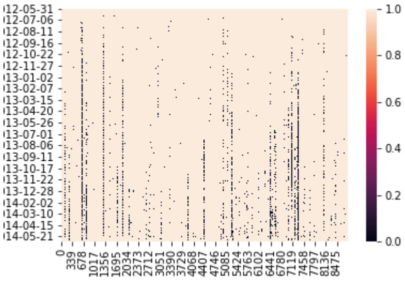
1 Target Creation

Prior to conducting any analysis, the type of problem needs to be defined. In this case, the problem is a binary classification. The goal of the project is to predict adopted users in the app, defined by users who have logged in at least three times over the course of a seven day window. In order to conduct the analysis the target column needs to be created. Properly defining the target column is the top priority as without it the analysis cannot be completed. This was the most difficult hurdle as using the weekday would not provide a rolling window. Considerable time was spent trying to achieve the proper calculation. Attempts to use a group by operation proved fruitless, but continuing with a pivot table yielded the results needed. By creating the pivot table and plotting a heatmap of the null values, the following plots were created. By using a transpose it was possible to get the date as the index and users as the columns. Once in this form, a for loop was written to calculate over a rolling 7 day window. The result found that 1601 of 8823 users adopted the app. This indicates that the problem is a binary classification with an imbalanced target class.





2 Feature Engineering

Dummy variables were created for the creation source column, resulting in 5 categories. Two features are created from the date columns. Account age was created by taking the latest date in the data set, 5-30-2014, and subtracting the the creation date, and account history is created by subtracting the creation date by the last session time. Null values in the last session creation time column are imputed with the most common date in the feature.

3 Results and Recommendations

Exploratory data analysis of the features did not provide any useful information in how they correlate to the target. We are presented with only categorical features and the count plots all show that in every case users were more likely to not adopt the app.

In the modeling stage, the random forest classifier was the chosen algorithm. Random forest was chosen due to the categorical nature of the features used, and performance of an ensemble model. The features used in the model were: opted into mailing list,enabled marketing drip, account history and account age, in addition to the five dummy columns created from the creation source feature. The resulting model performed poorly with an auc score of 48 percent, less than a random guess.

To improve the model, I would start with trying to balance the target class with a sampling technique. IN addition,looking at the feature importance of the model shows that account age and account history are the best indicators for predicting the outcome. 'Opted into mailing list','enabled marketing drip' provide little to no value. In order to increase user adoption, the recommendation would be to redesign and include new features to the app. With such low engagement it is apparent that the app is only useful to a handful of people. This is supported by the fact the opting in for mailing list and enabling marketing drip does not provide any value in the sense of getting users to adopt.