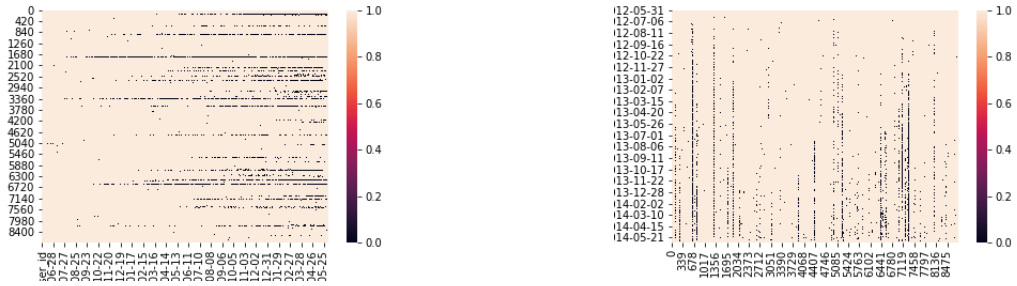
1 Target Creation

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

In order to create additional features, the email column was split to email type. There were 7 email types that accounted for more than 1 instance. The rest of the email types occurred only once and seemed to be generated by some random function. These random email types were collected into a 'rare' category. In addition, dummy variables were created for the creation source column, resulting in 5 categories. Lastly, an account age was created by taking the latest date in the data set, 5-30-2014, and subtracting the the creation date. In subsequent analysis, it was determined that the creation source dummy variables and the email types did not add any value to the model. These columns were then dropped. The last session creation time variable was also dropped due to there being the data quality issue of all the dates occurring in 1970.

3 Results and Recommendations

At 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 only features used in the model were: 'opted into mailing list','enabled marketing drip', and 'account age.' The resulting model performed poorly with an auc score of 50 percent, the same as a random guess. However, applying SMOTE techniques yeilded an auc score of 67 percent. An improvement, but still poor.

To improve the model, the first thing needed would be to gather the correct data for the last session creation time, the correct data would allow to engineer better features that contain a time component. Looking at the feature importance of the model shows that account age goes the furthest in 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.