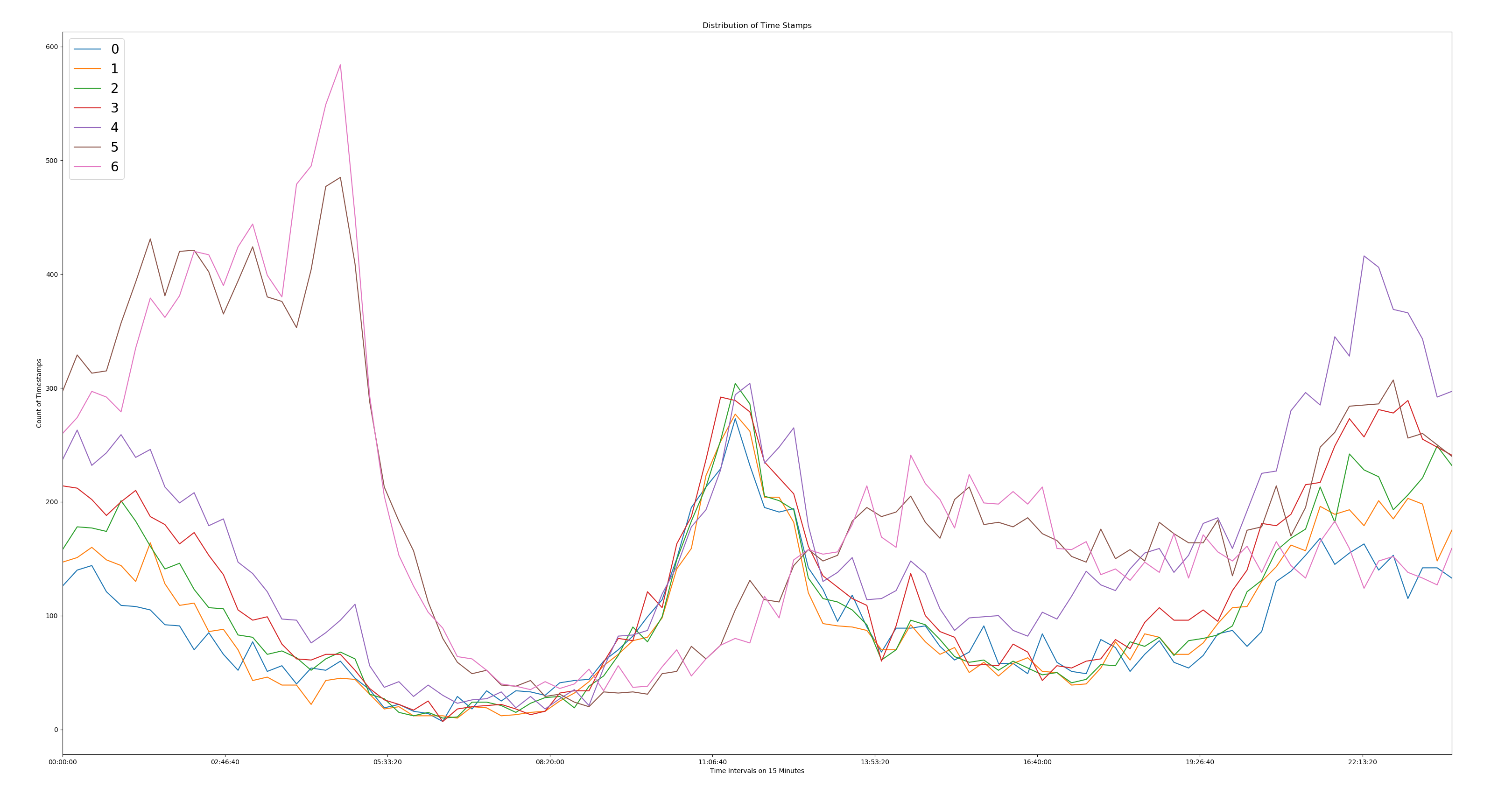
Part 1

1. On the weekends, demand lasted though out the early morning as in up to 5AM. In addition, lunch tends to become later on the weekends around 2PM, which makes sense as people who use this service came home late at night. On the weekdays, things were typical of the daily routine, such as the lunch surge around 11:30 and the going home surge around 9PM to 10PM. The surprising part was the surge at midnight during the weekdays.

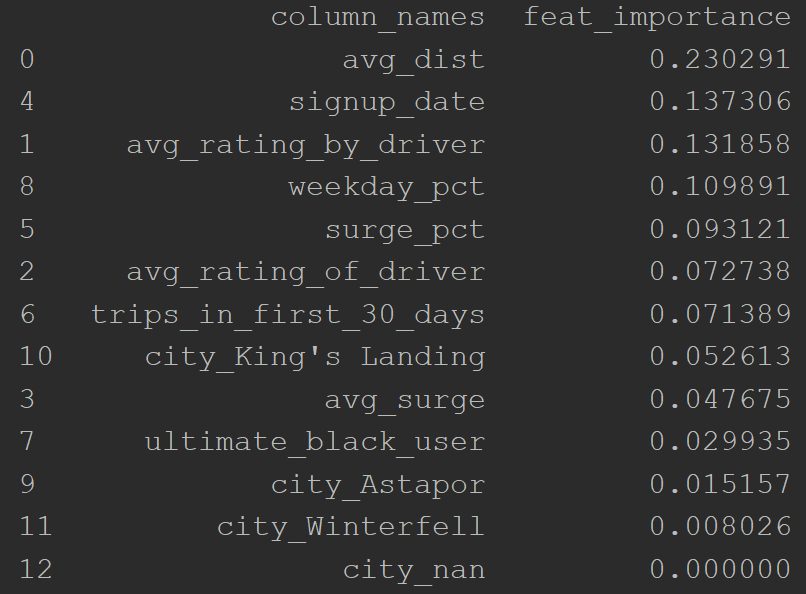


Part 2

1. There are 2 metrics that we should track to see if this is a successful program. The first metric will be the number of visitors that visit both cities or will be eligible for the reimbursement program. The second would be the tax revenues collected from businesses since the implementation of the program to see whether the visitors are helping businesses in both cities. I believe that those 2 metrics will determine if it is a successful program.
   1. We would implement the program by gathering data on visitors who paid the 2 way toll like time spent, number of visitors on the days of the week, time of when they passed the toll way, etc. In addition, I would like recommend starting a survey to understand the reasoning behind paying the 2 way toll and asking the propensity for them to take advantage of the program. Lastly, we would have a 90-day test period to implement the program, while also partnering with the finance team to measure the tax revenues generated during the time of this program.
   2. I would conduct a 2-sample z-test to determine if the change of visitors on days of the week is positively significant. And I would also conduct a 2-sample z-test to determine if the change of tax revenues is significant.
   3. If it is significant, I would recommend the continue support of the program. I would also gather the reasoning behind the visitors to support the businesses and show them how they can increase their business with the new visitors into their cities.

Part 3

1. The data set was fairly clean besides a few missing values. Due to the low amount of missing value, I exchanged the NaN values with 0 because they were different than the other values within the feature. I also needed to change the date from string to datetime in order to label the “Active Users”. After the cleaning and exploratory steps were completed, it turned out that 37.6% of users that joined in January 2014 were still active in July.
2. After cleaning the dataset, I converted all of the data besides the object data types into integer, and the one hot and encoded the object data type. After I used the Min-Max Scaler to normalize the data set to help with the wide differences in the data like the % of users to number of rides within the first 30 days. Lastly, I dropped the date values because they were not providing any real answers to the problem. This is because the client wanted to understand why were continued to use their service. Overall, the random forest generated a F1 Score of .63. While it is not the highest score, there are some validity to the data which will be explained below.
3. Below are the best features based on the model:



The major insights would be discussing the average distance. Users tend to use this service on longer routes than shorter. Another insight is the weekday\_pct, which seems like the users see a benefit to using it during the weekday. On the negative side, black users tend to not be an indicator of active users and the city does not have a large impact as well, so targeted marketing based on cities may not be a useful idea. The other surprising feature is that usage within the first 30 days does not have a large impact being an active user in the long term. Therefore, I would suggest doing further analysis on the longer trips and why are they more likely to be active users with longer trips. On the other hand, what is not encouraging users that used the service within the first 30 days to continue to use the service?