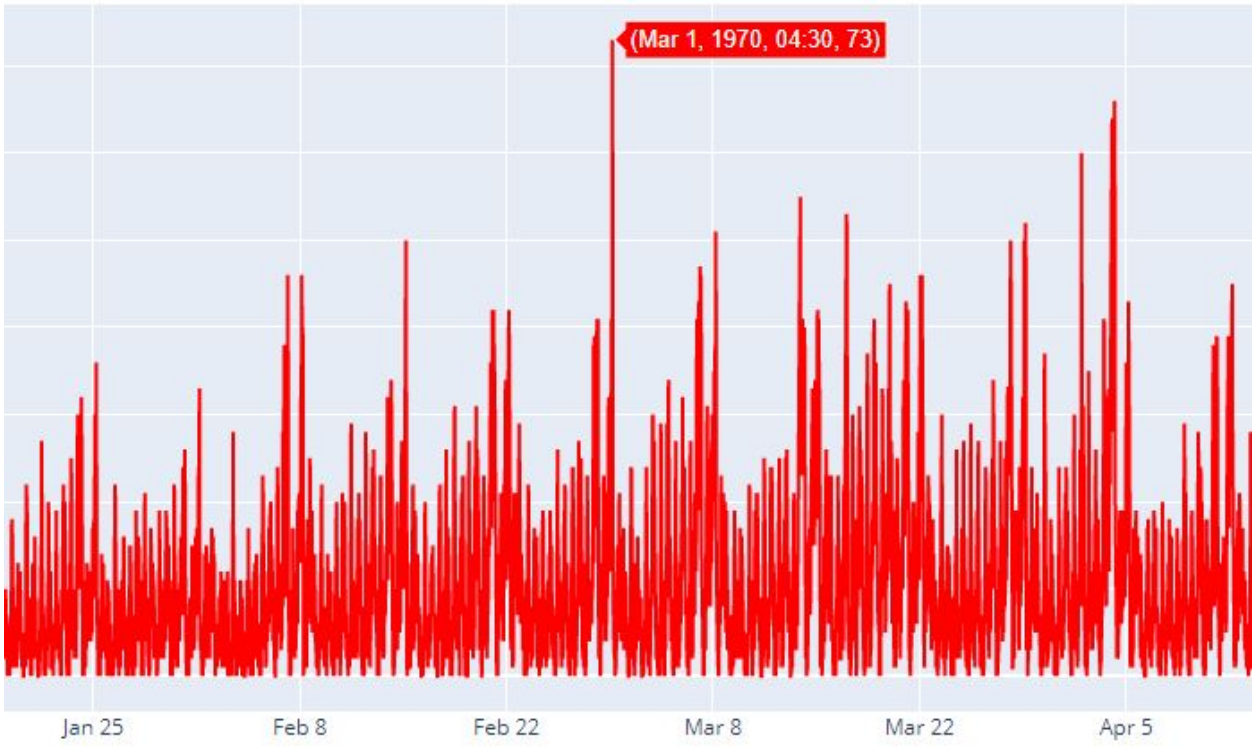
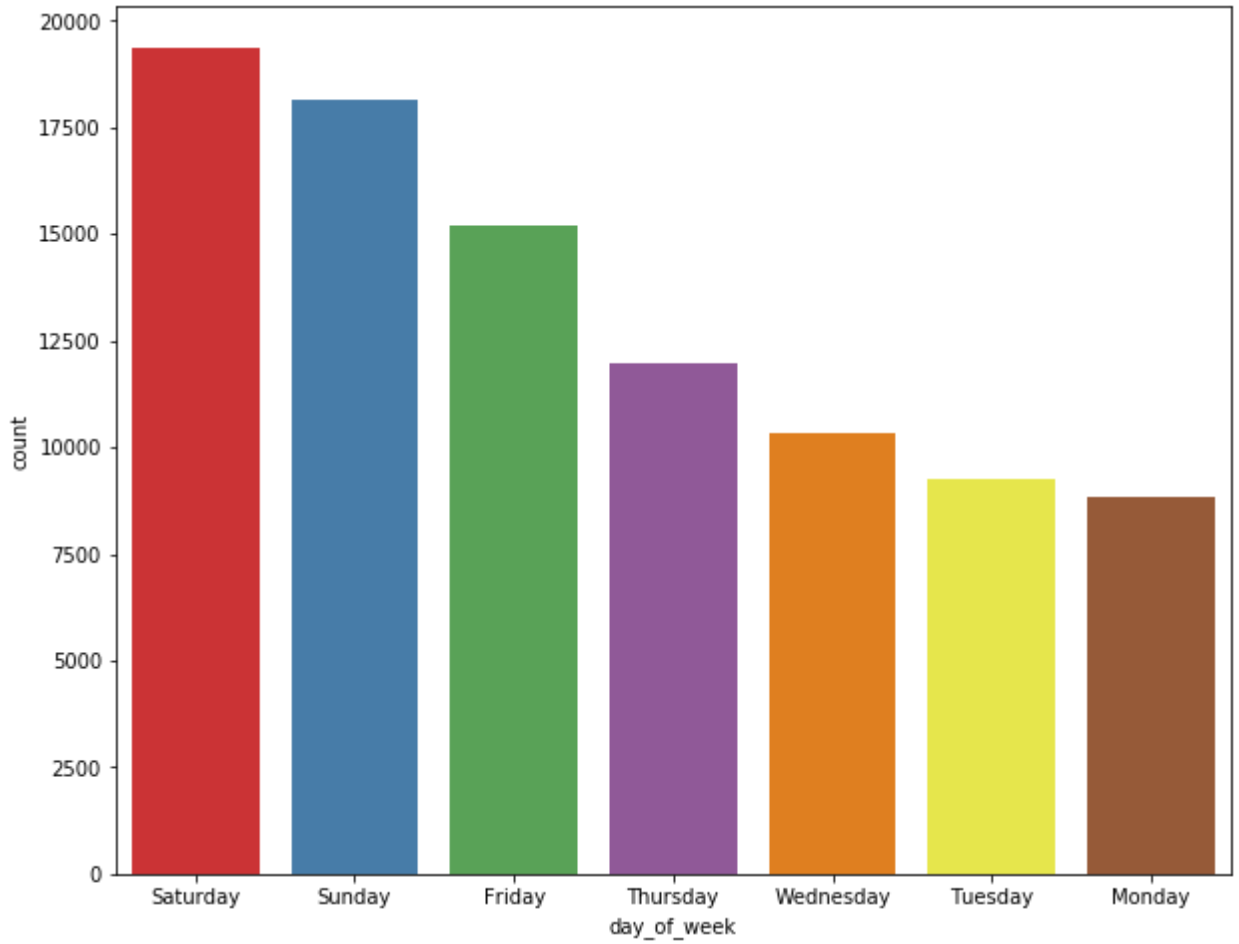
Part 1 - Exploratory data analysis

The attached logins.json file contains (simulated) timestamps of user logins in a particular geographic location. Aggregate these login counts based on 15 minute time intervals, and visualize and describe the resulting time series of login counts in ways that best characterize the underlying patterns of the demand. Please report/illustrate important features of the demand, such as daily cycles. If there are data quality issues, please report them.

My findings:



March 1st at 4:30 AM had the most login counts. The number of user logins seems to be increasing over time but it is suddenly dropping after April 4th. We should investigate the drop in login counts. The cycle of going up and down could be explained by the weekends vs weekdays.



People tend to login more going into the weekend starting from Thursday and then dropping back down at the start of the week on Monday. A value count shows Saturday and Sunday have the highest login counts. Biggest peak time on the weekend was at 4:30 AM while the biggest peak time on the weekday was at 11:30 AM and at 10:30 PM.

The data was in timestamp format and was converted using pd.to_datetime() for the analysis.

Part 2 - Experiment and metrics design

The neighboring cities of Gotham and Metropolis have complementary circadian rhythms: on weekdays, Ultimate Gotham is most active at night, and Ultimate Metropolis is most active during the day. On weekends, there is reasonable activity in both cities.

However, a toll bridge, with a two way toll, between the two cities causes driver partners to tend to be exclusive to each city. The Ultimate managers of city operations for the two cities have proposed an experiment to encourage driver partners to be available in both cities, by reimbursing all toll costs.

1. What would you choose as the key measure of success of this experiment in encouraging driver partners to serve both cities, and why would you choose this metric?
2. Describe a practical experiment you would design to compare the effectiveness of the proposed change in relation to the key measure of success. Please provide details on:
   a. how you will implement the experiment
   b. what statistical test(s) you will conduct to verify the significance of the observation
   c. how you would interpret the results and provide recommendations to the city operations team along with any caveats.

The neighboring cities of Gotham and Metropolis have complementary circadian rhythms:

- On weekdays, Ultimate Gotham is most active at night, and Ultimate Metropolis is most active during the day.
- On weekends, there is reasonable activity in both cities.

My Answer to Part 2:

- One metric that we would try to get and measure would be the increase profit a driver can make in a day during the weekdays by being able to serve in both cities during the peak hours. We would most likely need to measure it by profit by distance because a driver would still need to waste time crossing the bridge to be able to serve both cities during a day on the weekday. We would also want to see if there is an increase in client rate for driver partners. If all drivers can now serve both side clients, there might be a problem of supply being greater than the demand with drivers being in one city at certain times.

- To implement the experiment we would need to be able to have access to when the driving partner logged into the application that Ultimate provides. This way we would be able to track the distance that the driver has traveled throughout the day.
- Then of course we would want to compare the number of active vs inactive driving partners during both peak times in both cities.
- Also from Downtown Metropolis to Downtown Gotham is a 60 to 90 minute drive depending on traffic. https://www.quora.com/DC-Comics-How-far-is-Gotham-from-Metropolis

- We can do a t-test. For example a difference of means. Depending on whether the sample data that we collect is normal or not. If it is not normal then we would perform a non-parametric test to see if the distributions are different.
- But assuming they were normal and hoping this program could benefit both Ultimate and the driving partners, the null hypothesis would be that there is no difference in profit/distance for driving partners after the toll bridge program. While, the alternative hypothesis would be profit/distance is greater for driving partners after the program.
- Similarly for active vs inactive cars. H0 would be there is no mean difference for active and inactive cars after the toll bridge program. HA would be there is a difference.
- If there is an increase in profit/distance for the driving partners vs the mean of before the program then it would be a success for the driving partners.
- It is important to also look at vacant vs non vacant. Many drivers perhaps hearing of the program would all rush to one city location and this would cause chaos between drivers. It could also cause a shortage in the other City which isn't at peak hours. This perhaps could hurt revenue for us as there might be more vacant drivers who are wasting 90 minutes of potential client ride time from crossing the bridge.

Part 3 - Predictive modeling

Ultimate is interested in predicting rider retention. To help explore this question, we have provided a sample dataset of a cohort of users who signed up for an Ultimate account in January 2014. The data was pulled several months later; we consider a user retained if they were “active” (i.e. took a trip) in the preceding 30 days.

We would like you to use this data set to help understand what factors are the best predictors for retention, and offer suggestions to operationalize those insights to help Ultimate.

The data is in the attached file ultimate_data_challenge.json. See below for a detailed description of the dataset. Please include any code you wrote for the analysis and delete the dataset when you have finished with the challenge.

1. Perform any cleaning, exploratory analysis, and/or visualizations to use the provided data for this analysis (a few sentences/plots describing your approach will suffice). What fraction of the observed users were retained?
2. Build a predictive model to help Ultimate determine whether or not a user will be active in their 6th month on the system. Discuss why you chose your approach, what alternatives you considered, and any concerns you have. How valid is your model? Include any key indicators of model performance.
3. Briefly discuss how Ultimate might leverage the insights gained from the model to improve its long term rider retention (again, a few sentences will suffice).

A few things we can do here:

1. explore why avg_rating_of_driver has a lot of null values. One possibility is the lack of user motivation to leave a review.
2. phone missing values, could be because person used a pc
3. avg_rating_by_driver is also missing, driver might have been too lazy to leave review

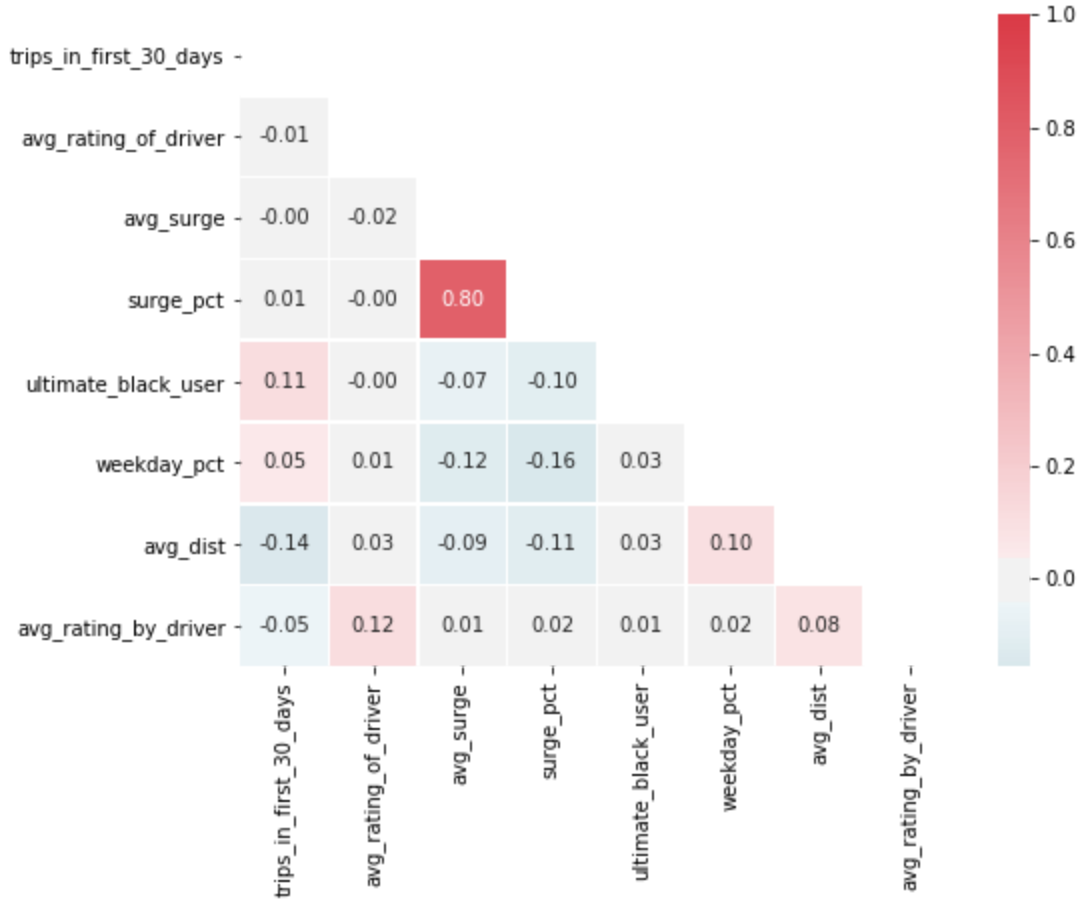
A few things to note after looking at .info():

1. city can be changed to categorical
2. signup_date should be changed to date time
3. last_trip_date should be changed to date time
4. phone should be categorical. iPhone, Android, ...
5. need to transform to numerical with onehotencoder or columntransformer
6. avg_dist - is it in miles or kilometers?

My Answer to Part 3.1:

1. Perform any cleaning, exploratory analysis, and/or visualizations to use the provided data for this analysis (a few sentences/plots describing your approach will suffice). What fraction of the observed users were retained?

For the cleaning part, we had to change the signup date and last trip to datetime format. Also, we created another column ‘signup_lasttrip_diff’ to see the difference between sign ups and last trip dates.
There are quite a few NaN values in review and phone columns. For the review columns, this could be because the user or the driver simply forgot to leave a review. The missing values for the phone column could be explained by the user having to use a computer or another device to make the reservation. We removed the review column rows that have more than 0 value. We also replaced the NaN value in the phone column with 0. Now, we can ask the company to investigate this or do a promotion to get users and drivers encouraged to stay on top of giving reviews/ratings.



There isn't a strong correlation with any of the columns. Besides average surge and percentage surge which are the same thing.

1. Doing a bar graph (using Bokeh) of device, city and ultimate users vs non.
    a. iPhone was the more popular device to make bookings on
    b. The most amount of bookings was made in Winterfell

    c. There were more non-ultimate-users so more people did not use Ultimate black in their first 30 days.

2. Looking at the scatter graphs we can see the following:
    a. Drivers who rated the user higher tended to travel more distance on average
    b. Drivers with 4-5 stars would get more mileage, however there are a few 1 star drivers with high average distance.

3. By doing a .describe() on the review columns we can confirm that most users who got worse ratings tend to travel much less (almost half) than users with 5 stars. This can be surveyed to see why these drivers gave the users such a low rating.

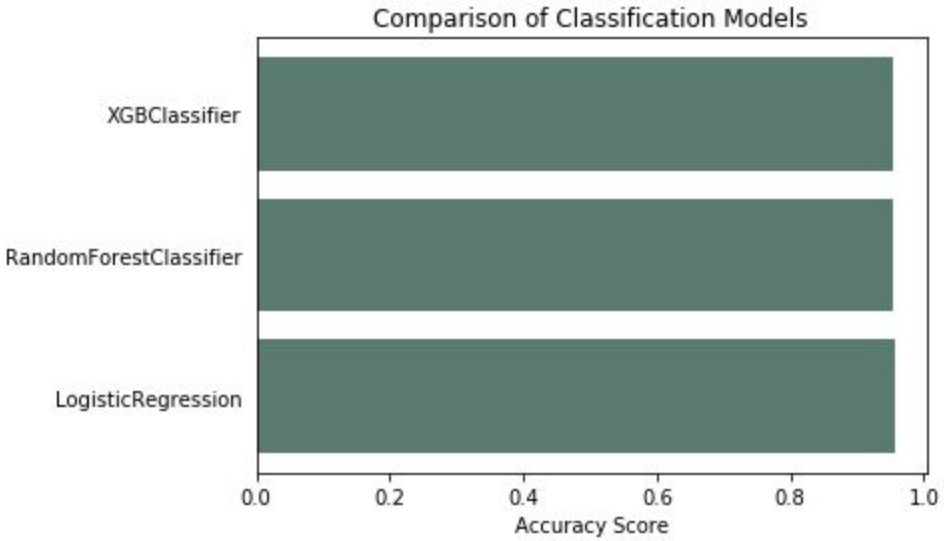
My Answer to Part 3.2:

Logistic Regression, Random Forest and Xgboost models were used to try to predict if users would still be active in their 6th month. We only have information of an active user using Ultimate dataset in the last 30 days. The problem with just using the last 30 days is a user could just be active in the last 30 days but before that the user was inactive the entire time. They should try to collect this information.

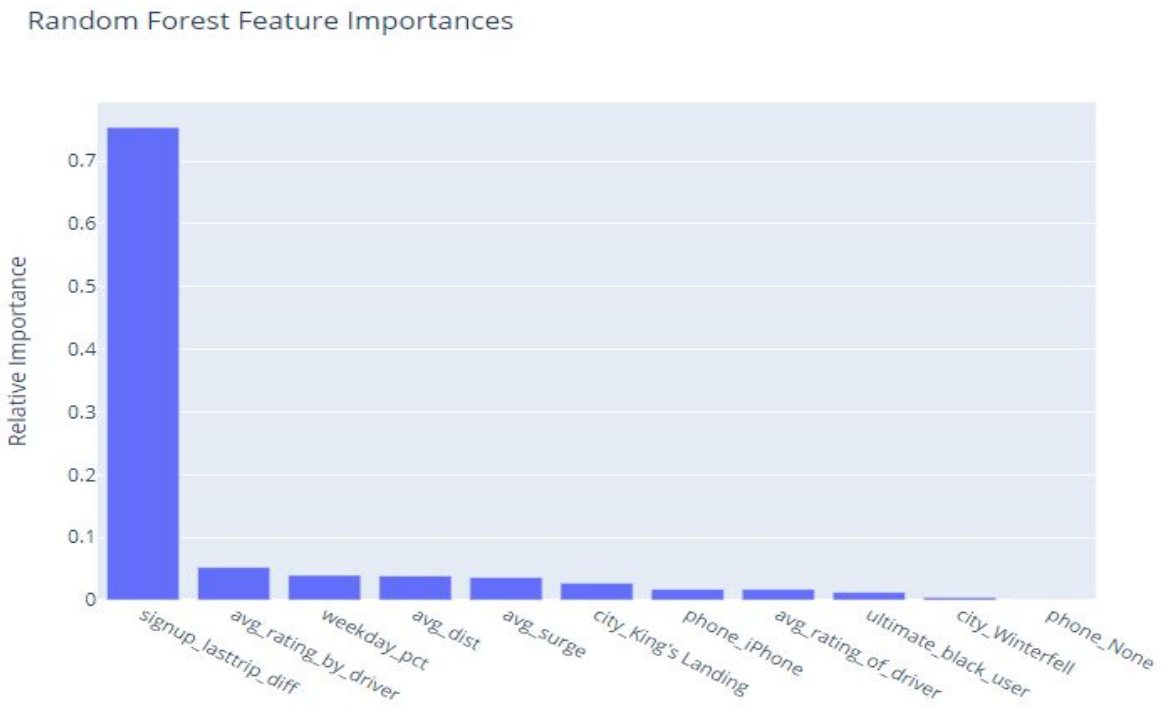
To build our predictive model we used the following columns:

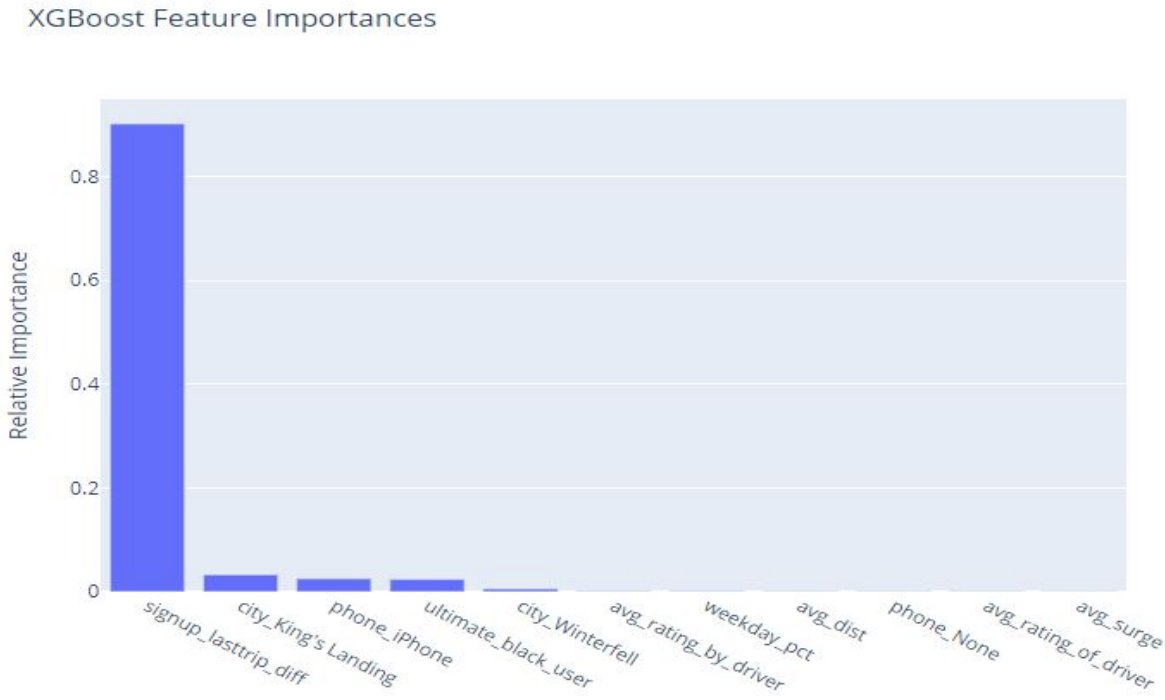
- 'city',
- 'avg_rating_of_driver',
- 'avg_surge',
- 'phone',
- 'ultimate_black_user',
- 'weekday_pct',
- 'avg_dist',
- 'avg_rating_by_driver',
- 'signup_lasttrip_diff'

```
LogisticRegression Model: 0.9556088782243551
RandomForestClassifier Model: 0.9545290941811637
XGBClassifier Model: 0.955128974205159
```



The Logistic Regression model which is much faster to fit, gave the best score of 0.9556 accuracy. However, XGBoost scored an accuracy of 0.9551. If we were to deal with more data later on instead of just 50,000 observations that we started off with, we would recommend using Logistic Regression as the accuracy is higher and the speed of the model would be definitely faster.





Both the XGBoost and the Random Forest models confirmed that ‘signup_lasttrip_diff’ was the most important feature.

My Answer to Part 3.3:

We saw that the feature that was most important was days between the sign up date and the last trip. However, we might need to redefine what we classify as active instead of just using the last 30 days activity. Ultimate needs to pull up user activities daily, weekly, monthly. Knowing a true active user is more important than a one time user in the last 30 days.

From feature_importances_ of the Random Forest and the XGBoost models, we would want to encourage drivers to be more active. Promotional events can be used to encourage both the drivers and the users. Promotions like discounted gift cards or free points give users if they charge their account with certain amounts. For example, a $100 dollar account charge would give the user $10 free.

We also might want to investigate why there were so many non Ultimate users and these users still used ultimate service. We should conduct an A/B test to see if we can improve the app that might make it more user friendly or inviting so these users would join the Ultimate users.

Another way to attract more users is to offer discounts for companies and their employees. This would encourage brand recognition of Ultimate company.