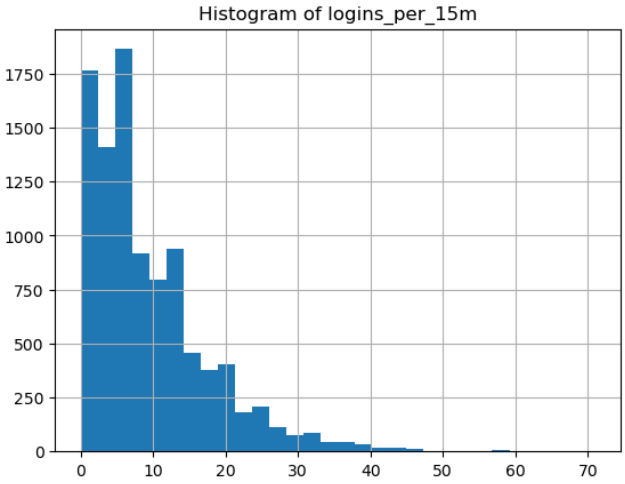
Data Analysis Interview Challenge

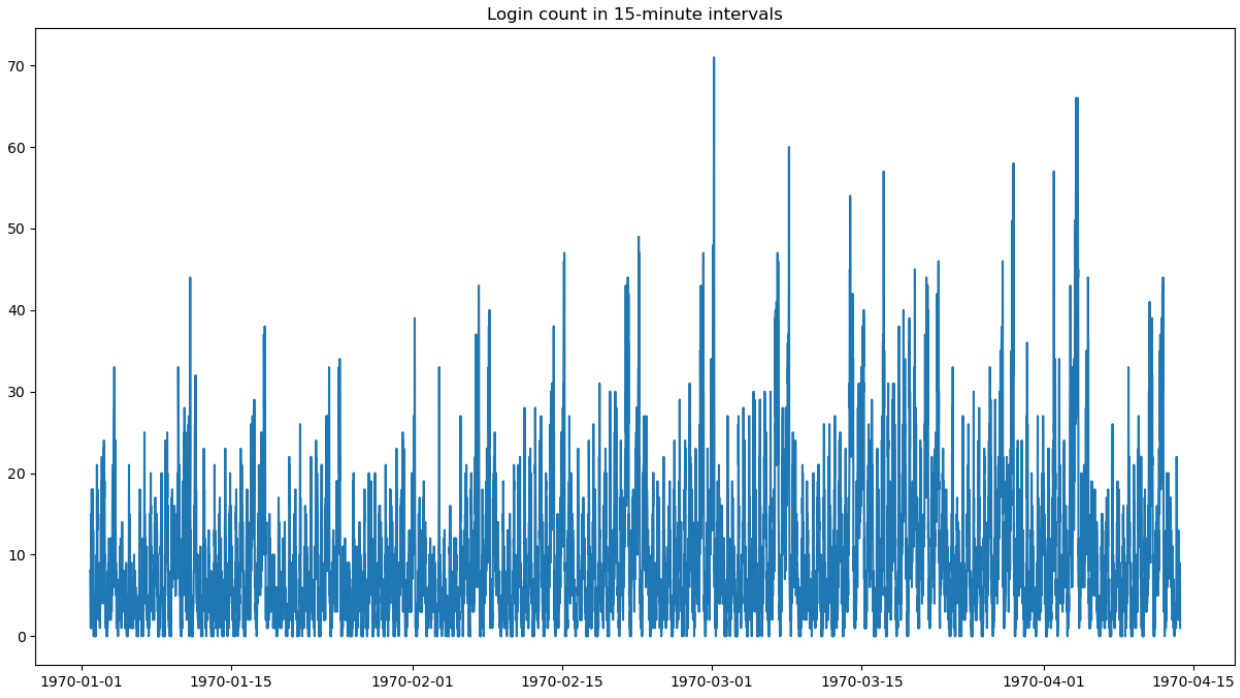
Rachel Chiang

Part 1 – Exploratory data analysis

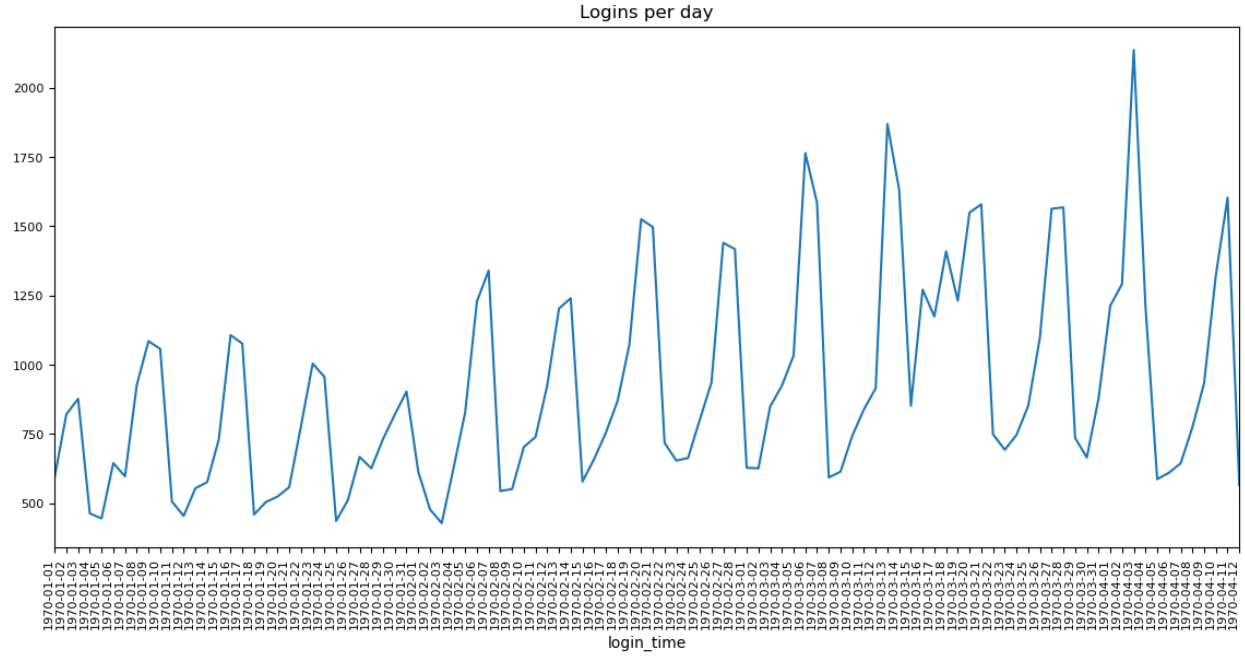
The attached logins.json file contains (simulated) timestamps of user logins in a particular geographic location. Aggregate these login counts based on 15minute time intervals, and visualize and describe the resulting time series of login counts in ways that best characterize the underlying patterns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

The logins were aggregated in 15-minute interval, resulting in a dataset with 9788 entries, where the mean was 9.5, the median was 7, and the range was 71.

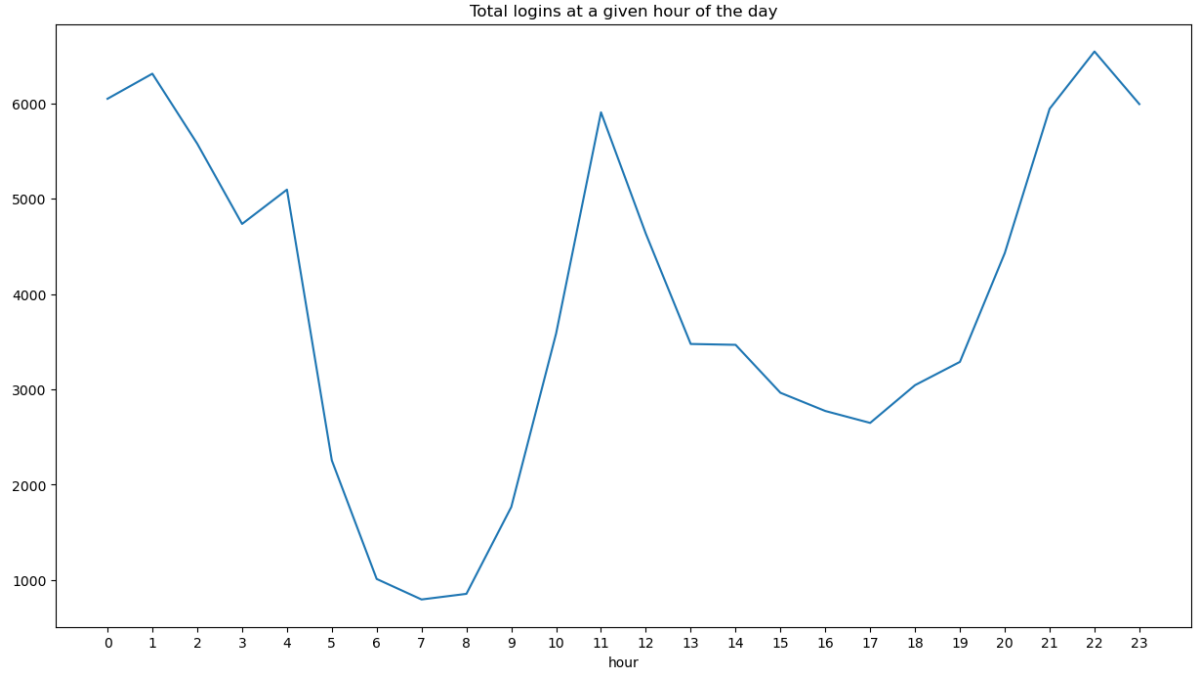




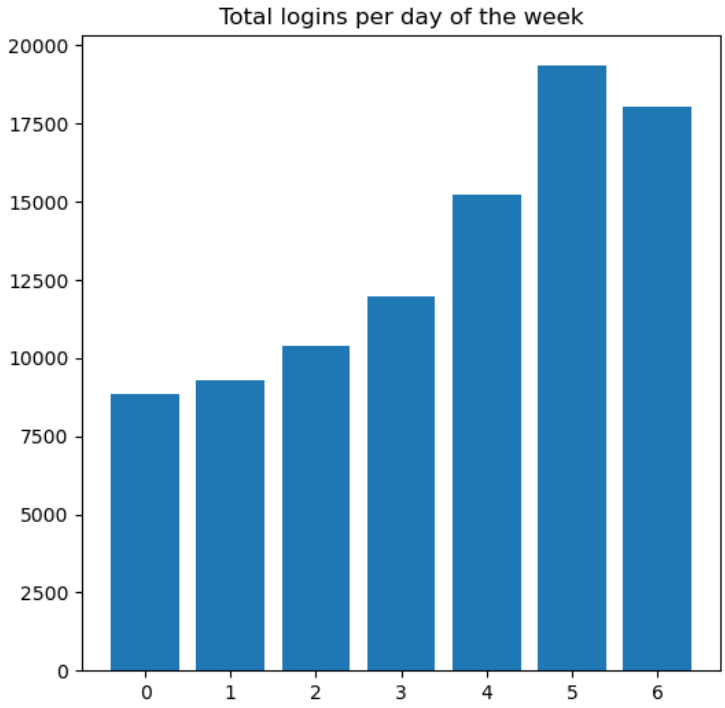
The month of March seemed to have a lot of activity, which can be confirmed by grouping the logins by month; March had almost 10000 more logins than February, the next highest month. When graphing by logins per day, it looks like there may be a trend and seasonality, but they do not look very strong, and it is hard to tell with certainty from the graph.



The logins are a bit perplexing because they don't match up with my assumption about when people might be active (or even awake). People seem to log in at late hours of the night and at midday, but not so much in the hours of 6-8 AM.



We can also see that Mondays and weekends had relatively fewer logins, whereas Saturday and then Sunday had the most logins.



To test for stationarity, both the Augmented Dickey Fuller (ADF) and Kwiatkowski-Phillip-Schmidt-Shin (KPSS) tests were used on the `logins_per_15m` series. The ADF found that the series was stationary or had no unit root, whereas the KPSS found that the series was non-stationary or had a unit root. To make it stationary, the difference was taken.

Github link: https://github.com/raechiang/Springboard/blob/main/Unit_27-2-2_Ultimate_Technologies_Take-Home_Challenge/1-exploratory-data-analysis.ipynb

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?

I would choose the average weekly revenue from the driver's trip minus the reimbursement of the toll costs as the key measure of success of this experiment in encouraging driver partners to serve

both cities. In other words, the success metric is the weekly average of $(total\_trip\_price - total\_toll\_cost)$ over the course of some time.

The prompt mentions only that it wishes to encourage driver partners to serve both cities, so the minimum definition for "success" would be whether the driver has served users in both, but this alone feels too simplistic to me because the driver could arbitrarily serve users at a bad time of day—for instance, if that city does not have a lot of activity at that time—or the driver serves only one user in one of the cities. I am assuming that the end goal is that the company wants this experiment to eventually or hopefully result in better profit. This means we would want our drivers to purposefully select which cities to serve depending on the time without bouncing back and forth too much for poor reasons, and so we could try to to maximize the $total\_trip\_price$ and minimize the $total\_toll\_cost$.

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

[A] The experiment would involve splitting the drivers into two groups: a control group which would continue to serve users as they have been (that is, exclusively serving one city) and a second group which would serve both cities with the user activity tendencies in mind (Gotham during the night and Metropolis during the day on weekdays, and either of the cities during the weekends).

[B] We could do a hypothesis test. Our null hypothesis, H_0, is that our exclusive drivers would generate the same amount of or more weekly revenue as our multi-city drivers; the difference in the weekly revenues of the one-city drivers and multi-city drivers is greater than or equal to 0 $(avg\_weekly\_rev\_one - avg\_weekly\_rev\_multi \geq 0)$. The alternative hypothesis, H_A, is that our multi-city drivers would generate more weekly revenue than our one-city drivers; the difference in weekly revenues of the one-city drivers and multi-city drivers is less than 0 $(avg\_weekly\_rev\_one - avg\_weekly\_rev\_multi < 0)$. Then we can use the p-value to test for significance. We should also test for assumptions, such as if our data is normally distributed, because it would help us determine what sort of test to proceed with.

[C] If the p-value is less than or equal to 0.05, we would reject the null hypothesis, so we would reject that the exclusive one-city drivers generate the same amount of or more weekly revenue as the multi-city drivers. We could recommend the city operations team to consider implementing the activity-based service strategy to all the drivers. If we cannot reject the null hypothesis, then the drivers may be able to continue as they have been.

[C] Certain further implementation details may have to be worked out. How would the exclusive, single-city drivers' cities be distributed? Should the drivers be randomly selected, or should we have an even distribution of Metropolis and Gotham drivers? As for the multi-city drivers, should the drivers be strictly assigned to a city depending on the time and day, or should drivers have more freedom and flexibility with time windows? (I suspect that there may be more convenient,

profitable, or cost-effective trips for drivers depending on where they and a potential user are, but I admit that I don't have much knowledge on this subject.)

[C] There may be a bit of an ethical dilemma, depending on how the drivers are paid. If a driver is paid depending on the number of users they service and we knowingly only assign some portion of our drivers to a strategy that could yield much better profits for them, is it right for the single-city drivers to be excluded from these gains?

Part 3 – Predictive modeling

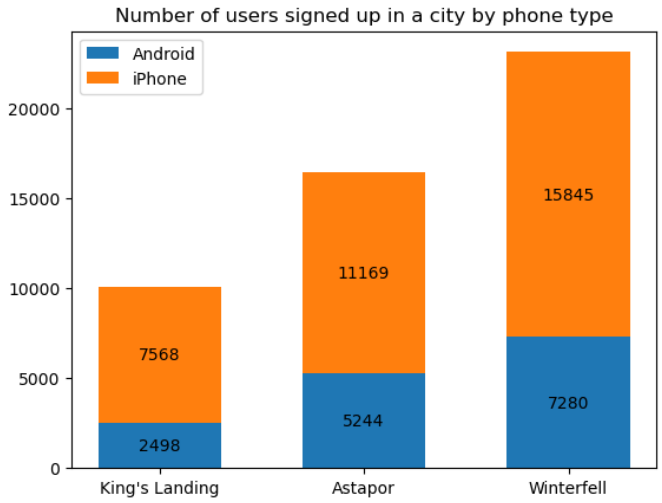
Ultimate is interested in predicting rider retention. To help explore this question, we have provided a sample dataset of a cohort of users who signed up for an Ultimate account in January 2014. The data was pulled several months later; we consider a user retained if they were "active" (i.e., took a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

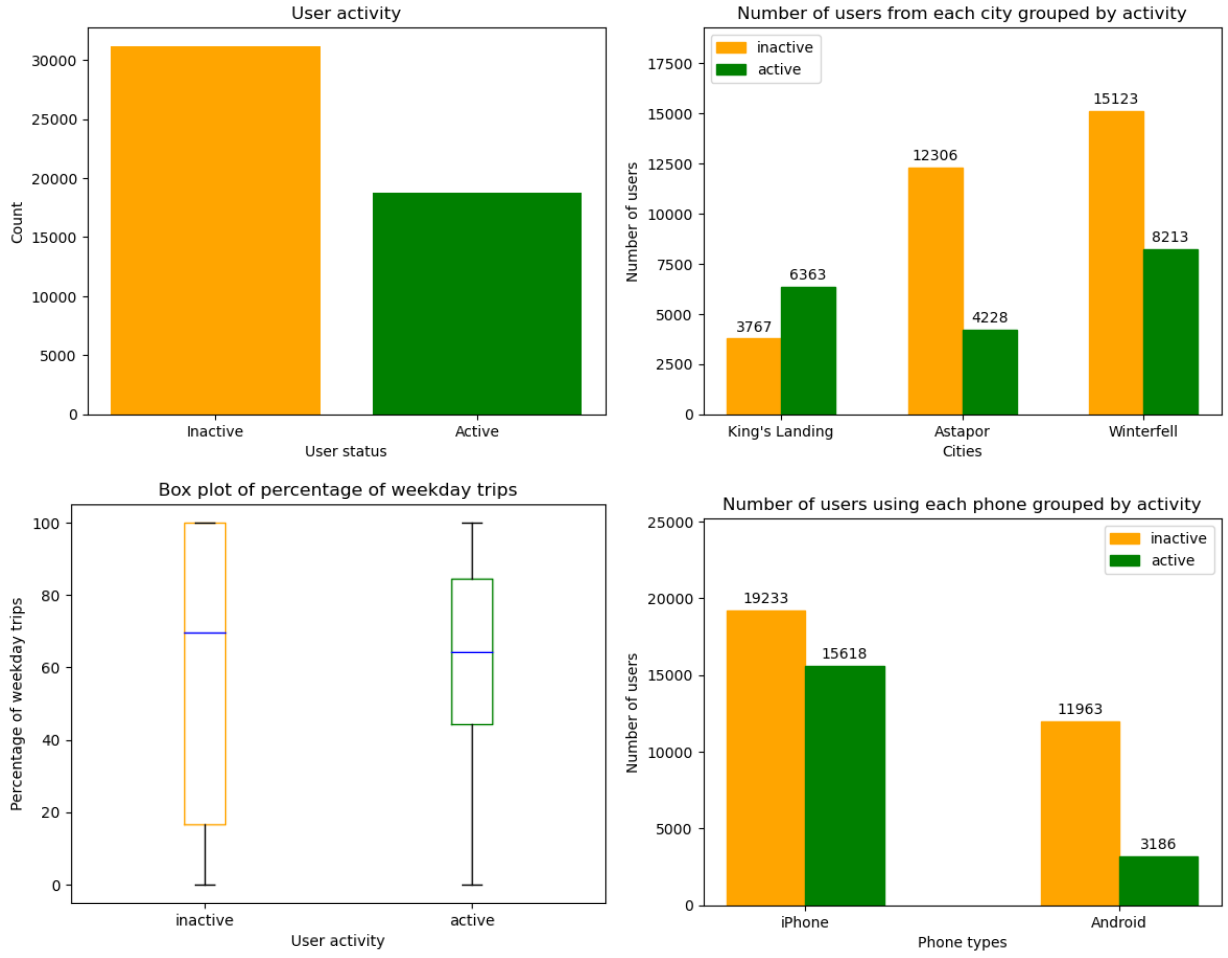
1. Perform any cleaning, exploratory analysis, and/or visualizations to use the provided data for this analysis (a few sentences/plots describing your approach will suffice). What fraction of the observed users were retained?

In the data wrangling step, there did not seem to be too many errors. Two columns (the `avg_rating_of_driver` and `avg_rating_by_driver`) with missing values were imputed with their respective means. Missing `phone` values were randomly assigned to retain the original distribution of phones. Many distributions were skewed, with the occasional outlier, but they did not seem meaningless, so they were left untouched. We have a lot of iPhone users, and most of our users signed up from Winterfell.



The response variable, `active`, was determined using the `last_trip_date` and was defined according to the project specification: a user is considered retained if they took a trip within the last 30 days. The fraction of users retained was 18804/50000 (37.6%), which means that our dataset is mildly imbalanced. (When I tested a few models with a down-sampled dataset to balance the variable, there was not much difference in performance, and it was even noticeably detrimental to a couple of the models, so I kept the original dataset intact for modeling.)

I explored each feature in relation to our response variable. My initial impression was that the `city`, `weekday_pct`, and `phone` would be indicative of user retention, but there seemed to be some hint of a pattern in at least eight of the features, which I discuss a bit more in the notebook.



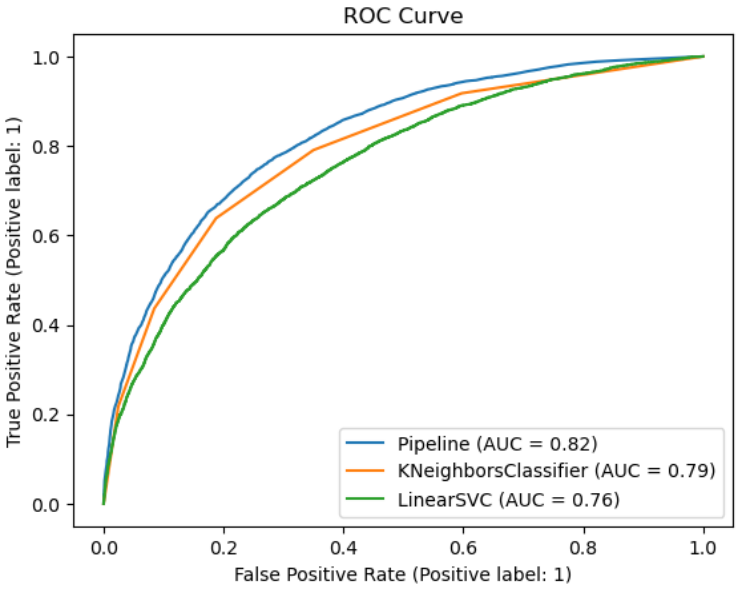
2. Build a predictive model to help Ultimate determine whether or not a user will be active in their 6th month on the system. Discuss why you chose your approach, what alternatives you considered, and any concerns you have. How valid is your model? Include any key indicators of model performance.

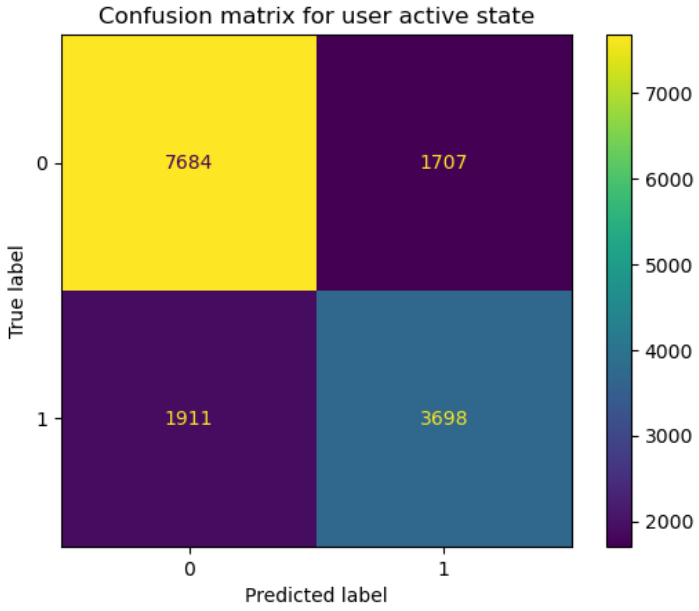
I tested a handful of models out of the box for an idea of a baseline and where to start: LogisticRegression, LinearSVC, KNeighborsClassifier, and RandomForestClassifier. More

experimentation can be done on them because they were quick defaults, but I chose the RandomForestClassifier to tune further in this notebook.

Name	Acc Score
RandomForestClassifier	0.7563
KNClassifier	0.7473
LinearSVC	0.7177
LogisticRegression	0.7161

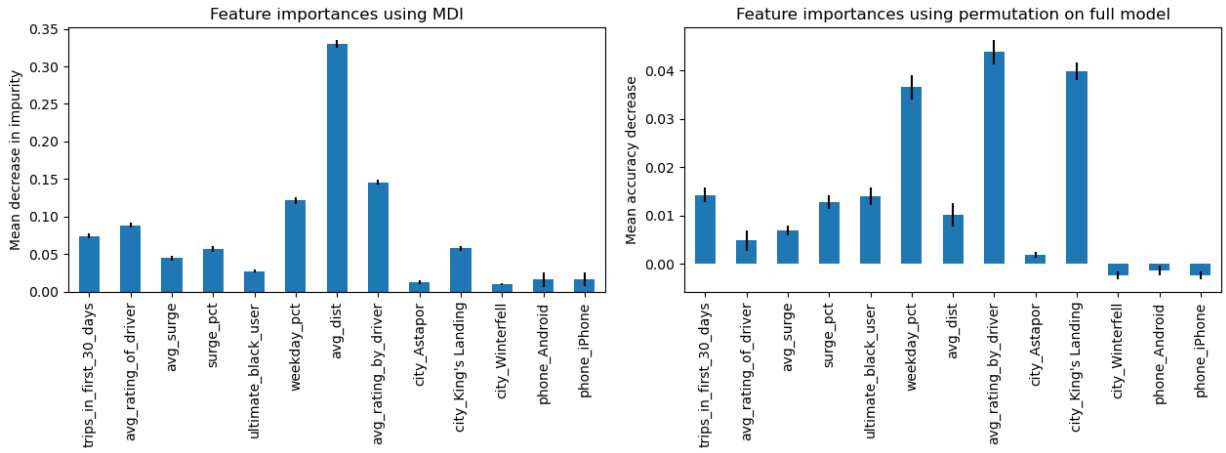
Tuning did not improve performance by very much, only about 0.0025, with a score of 0.7588, and the AUC came out to be 0.82, which is only 0.03 higher than that of the next best (out of the box) model. The RandomForestClassifier is a bit slow to train, but we don't have too many features, so it was not too bad. When looking at the confusion matrix, the class imbalance is clear.





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).

I also examined the feature importances. The random forest classifier's feature importances' top three features were the `avg_dist`, `avg_rating_by_driver`, and `weekday_pct`, whereas the top three using permutation importances were found to be the `avg_rating_by_driver`, `city_King's Landing`, and `weekday_pct`. When thinking about the EDA done earlier, these make some sense, because the distributions indeed looked different between the two activity groups.



For long-term rider retention, finding a way to encourage frequent, short travels may be a good thing. An example off the top of my head could be a reward system for an accumulated number of trips a user ordered. I think the `avg_dist`, `avg_rating_by_driver`, and `weekday_pct` may hint at the behaviors of retained riders. I speculate that the lower `avg_rating`_by_driver` for active users probably actually has to do with the frequency that these active users have ordered trips, which is not a feature that is explicitly present in the dataset, but I think would be very useful information. The shorter `avg_dist` and higher `weekday_pct` hint to me that these retained riders are probably going for high frequency, short distance rides, so they might be routine users, who use the app for mundane transportation, such as work or school. As for the `city`, I suspect "King's Landing" could either have a more affluent demographic that can afford to buy rides as opposed to using normal public transit, or the city could be a more metropolitan environment, where ridesharing would be more convenient than owning a vehicle, or perhaps even the overall age of the group is younger so they are more likely to order Ultimate rides. Whatever the reason may be, it may be worth it to research more into King's Landing or to try to increase the user base in that location, since King's Landing has the least number of users of the three cities (though again, this could relate to the location's characteristics that I am not aware of).

Github link: https://github.com/raechiang/Springboard/blob/main/Unit_27-2-2_Ultimate_Technologies_Take-Home_Challenge/3-predictive-modeling.ipynb