Ultimate Data Challenge

Part 1 - Exploratory data analysis

https://github.com/bjnugent/Springboard/blob/main/ultimate_technologies_challenge/notebooks/part1_EDA.ipynb

In analyzing the data, I noticed distinct patterns between weekdays and weekends.

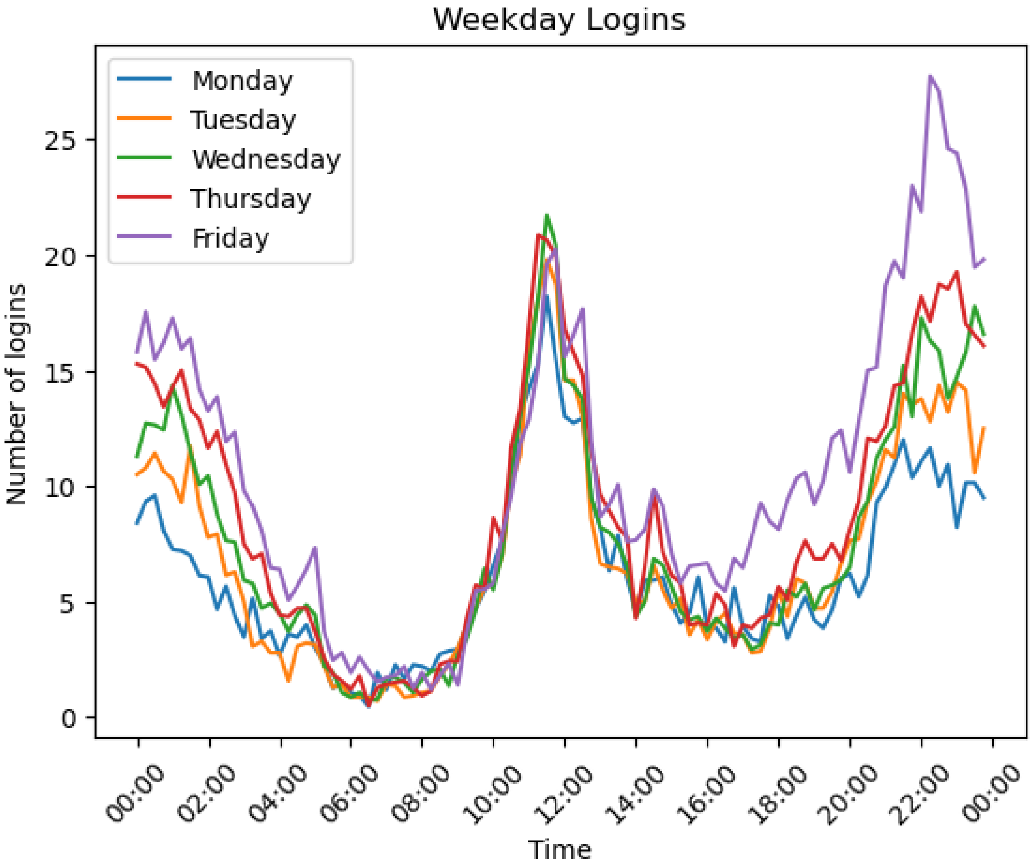

Figure 1. Average Weekday Logins for 24 hours

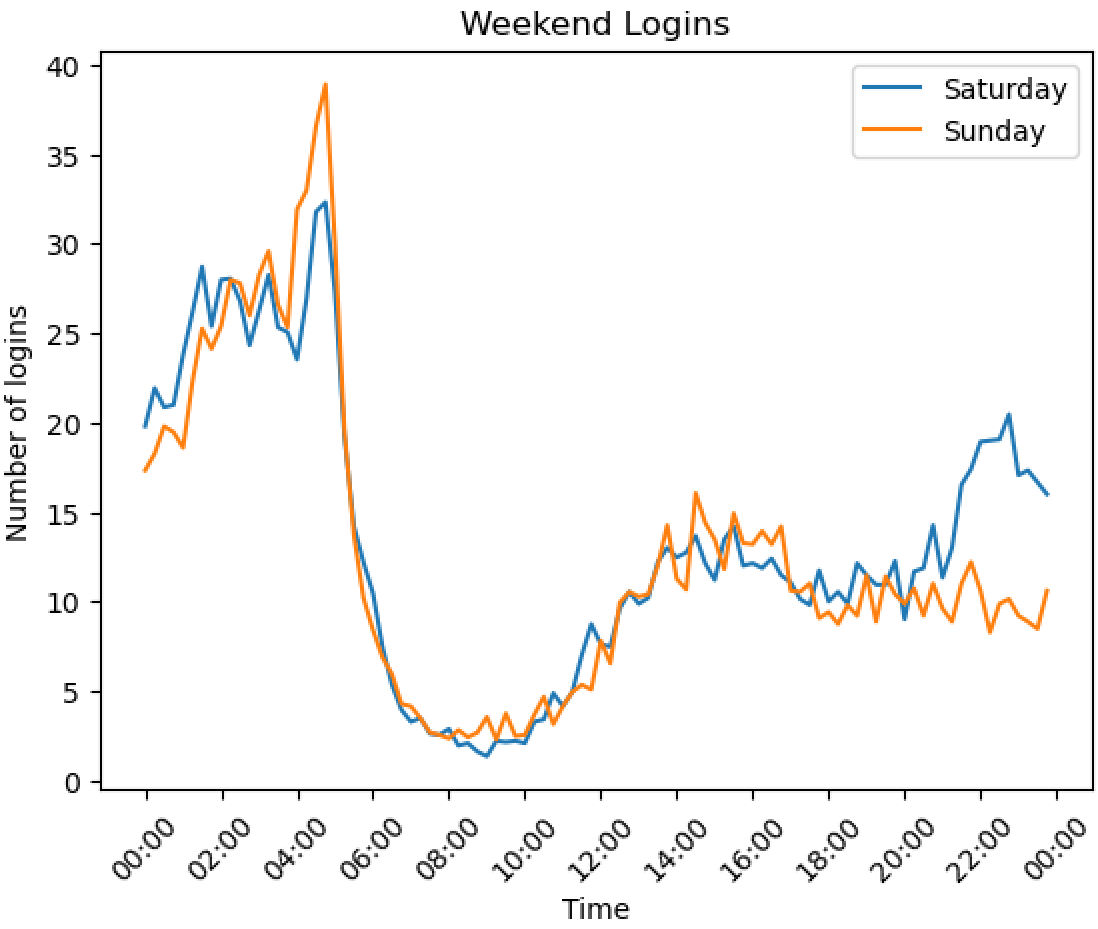

Figure 2. Average Weekend Logins for 24 hours

As shown in Figure 1, the average number of logins per 15 minutes show a daily recurring pattern including:

- Logins lull from 6:00 am to 9:00 am
- Logins spike sharply from 10:00 am to 1:00 pm
- Logins gradual rise from 6:00 pm to 10:00 pm
- Logins fall slowly from 10:00 pm to 6:00 am

One factor of data quality to investigate is the timezone recorded compared to the timezone of the users' login. At the current times, the lull at 6:00 am indicates that users are not logging in for a morning commute when this pattern may make more sense in the middle of the night when the vast majority of users are sleeping. Additionally, the time of highest demand is 5:00 am on Saturdays and Sundays, which makes more sense to align with people coming home late at night instead of early in the morning.

However, if the timezone is correct, weekdays give a clear pattern of peaks at lunchtime and at night at 10:00 pm. The lunchtime peaks are very consistent each day of the week, giving a constant predictable demand. The nightly peak on the other hand varies and increases each day with Monday being the lowest peak and Friday being the highest, showing an increasing demand as Friday approaches.

If the timezone has shifted, the peaks may instead align with commuting to and from work.

Saturday and Sunday show a distinct pattern from weekdays:

- The highest peak of logins occurs from 12:00 am to 5:00 am
- Logins lull from 6:00 am to 10:00 am
- Logins remain constant at 50 logins per hour between 2:00 pm and 9:00 pm

Overall, the weekend has higher demand than weekdays. By far the highest demand occurs at night and early morning on the weekends. This most likely reflects people going out and possibly needing a safe ride to get home. During the daytime, there is more of a constant demand and no spike around lunch, possibly due to users having a more flexible schedule on this day and not commuting.

Part 2 - Experiment and metrics design

The key measure of success in this experiment is whether the increased profit from potentially more rides outweighs the cost of reimbursing all tolls. While reimbursing all tolls may increase availability in the two cities, it may also cost the company more than that availability is making the company.

To explore this, I would assign a small percentage of drivers to the reimbursement group and reimburse their bridge tolls for a month. I would take data on the income of their rides each hour as well as the reimbursement amounts. I would compare that to a control group who are not being reimbursed. This data can also be separated to see patterns of weekday vs weekend as well as during the day and at night to investigate which specific times this reimbursement may be most impactful.

I would conduct a bootstrapped t-test as the sample size will be small to determine the percent confidence that the means of the reimbursement and control groups are different.

The results will yield which of the times of day and type of day would have a significant difference in profit, meaning that there is no overlap between the 95% confidence intervals between the two bootstrapped groups. I would use that to determine if the reimbursement group had increased profits compared to the control group for the following times:

- Weekdays during the day
- Weekdays at night
- Weekends during the day
- Weekends at night

Additionally, for the times of where this test was successful, compare the increased income to the cost of reimbursement. For the conditions where the increased income from reimbursements outweigh the cost of reimbursements, I recommend implementing a reimbursement policy for drivers.

Part 3 - Predictive modeling

https://github.com/bjnugent/Springboard/blob/main/ultimate_technologies_challenge/notebooks/part3_modeling.ipynb

In order to determine if users are active, I found the latest last trip date within the dataset (2014-07-01) and used that to determine that any user whose last trip date was after 2014-06-01 was active in the last month. Of the 50,000 users who signed up in January, 36.6% remained active after 5 months, with the largest number of last trips of inactive users being in the first month of use.

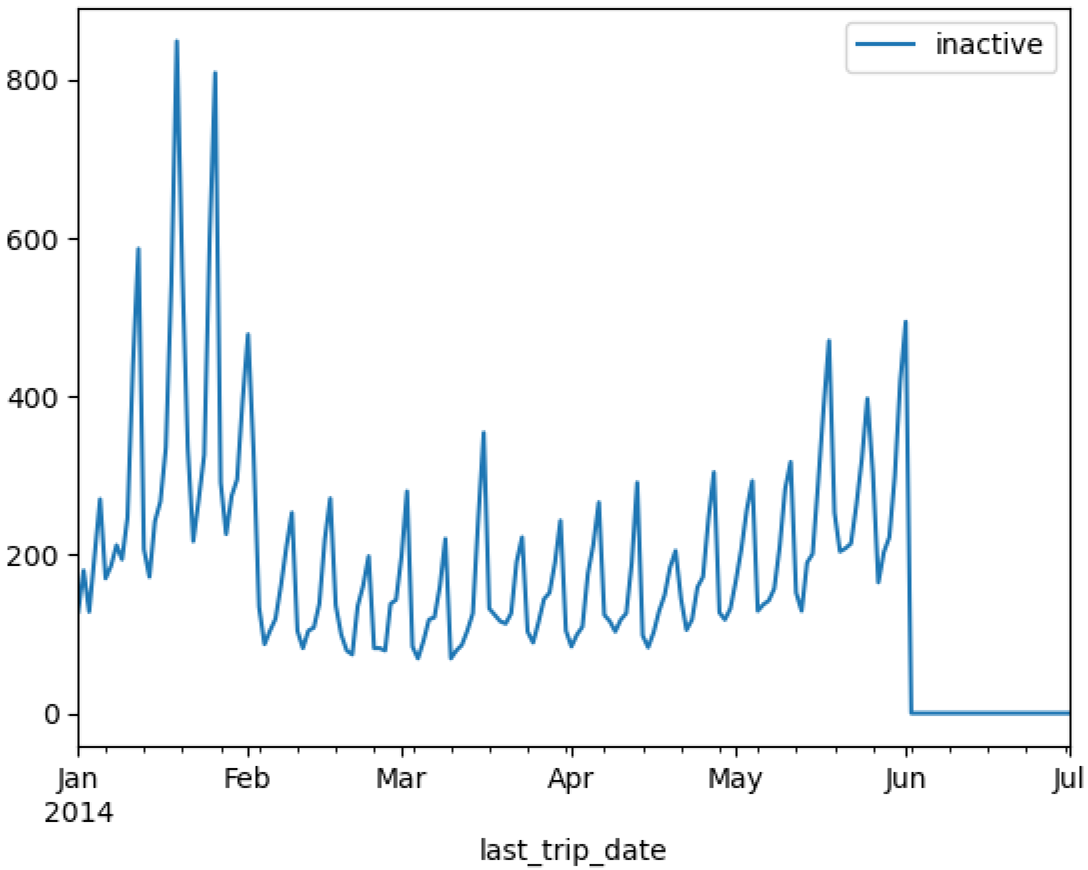

Figure 3. Last trip dates for inactive users

I searched for missing data and found the majority of missing data from the average rating of driver, most likely due to users not rating their drivers, which will be replaced with the median before modeling. I then plotted the categorical and continuous features individually and separated by user activity.

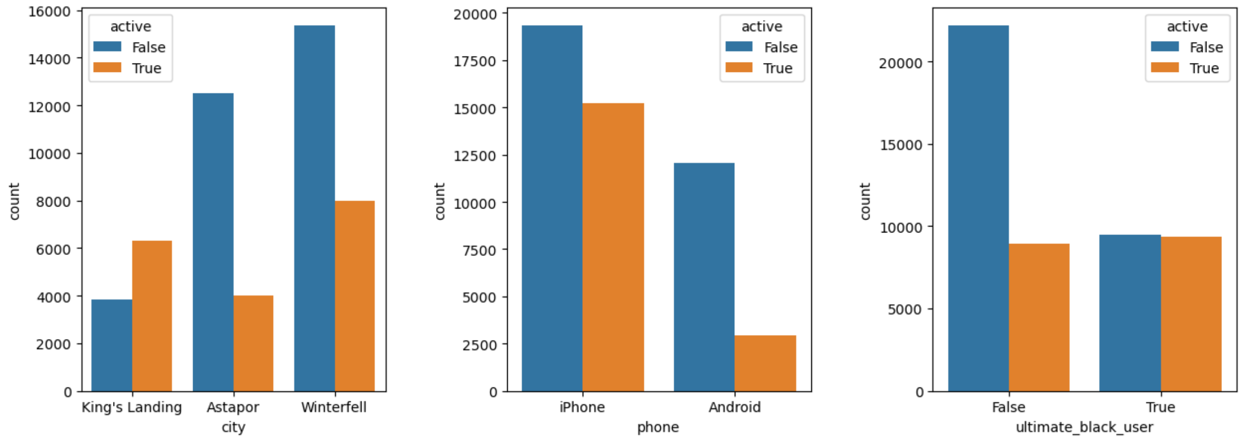

Figure 4. Categorical features by user activity

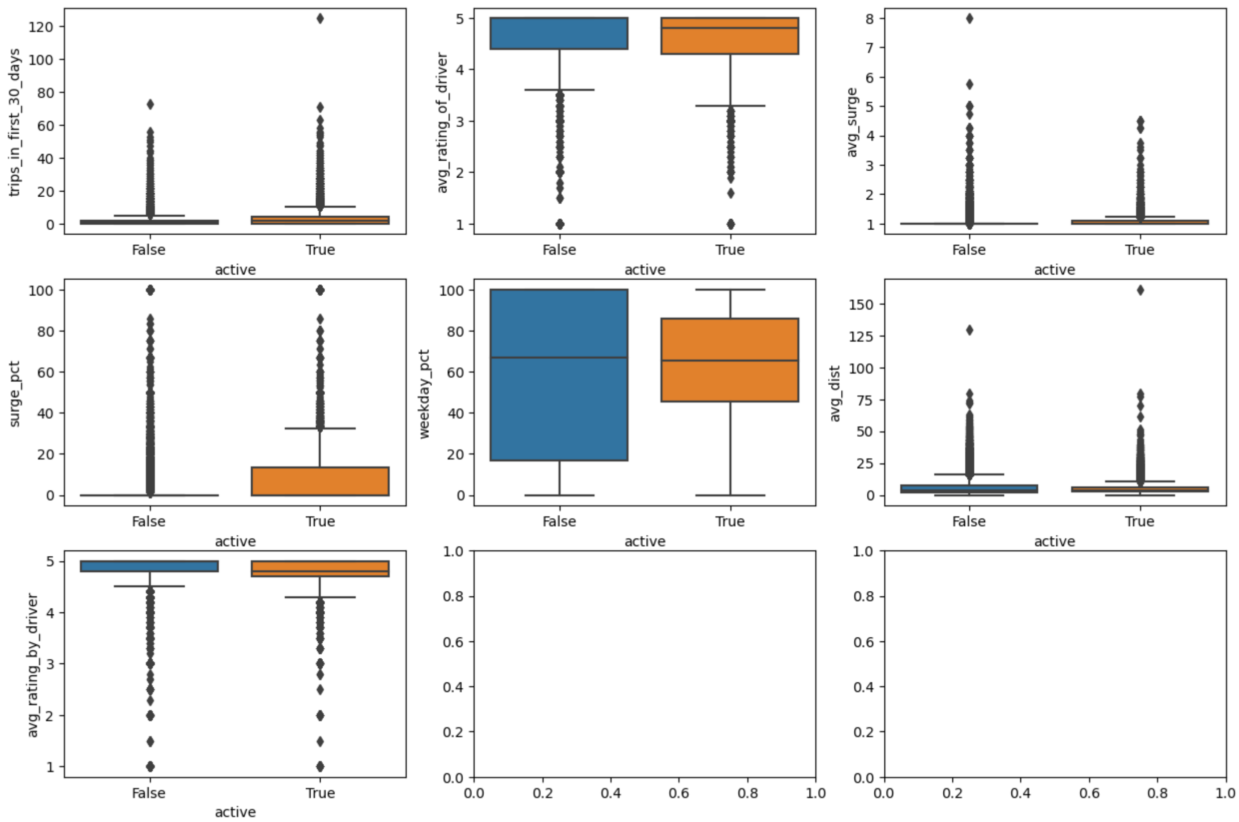

Figure 5. Continuous features by user activity

Users in King's Landing, users with iphones, and users who used Ultimate Black in the first 30 days were more likely to remain active. Active users also tended to have lower average ratings of drivers as well as lower ratings by drivers, most likely due to having taken more trips. Extreme overlap is seen within the continuous features and high variance.

I decided to build three models: a decision tree, a random forest, and a logistic regression classifier. Decision trees provide explainable results that can lead to company insights while an ensemble method like random forest tends to have higher accuracy due to reduced overfitting. Logistic regression also provides clearly understood weights to each feature. I also considered other ensemble methods including gradient boosting or ADA boosting, which may risk increased computation time. I prepared the data by one-hot removing the date columns, encoding the categorical features, converting the 'ultimate_black_user' feature to an int, and replacing missing values with the medians of the training sets. I conducted random grid searches for the three models optimizing for F1 score as both precision and recall are of high importance in this context. I evaluated the best models against the testing set, yielding the following results:

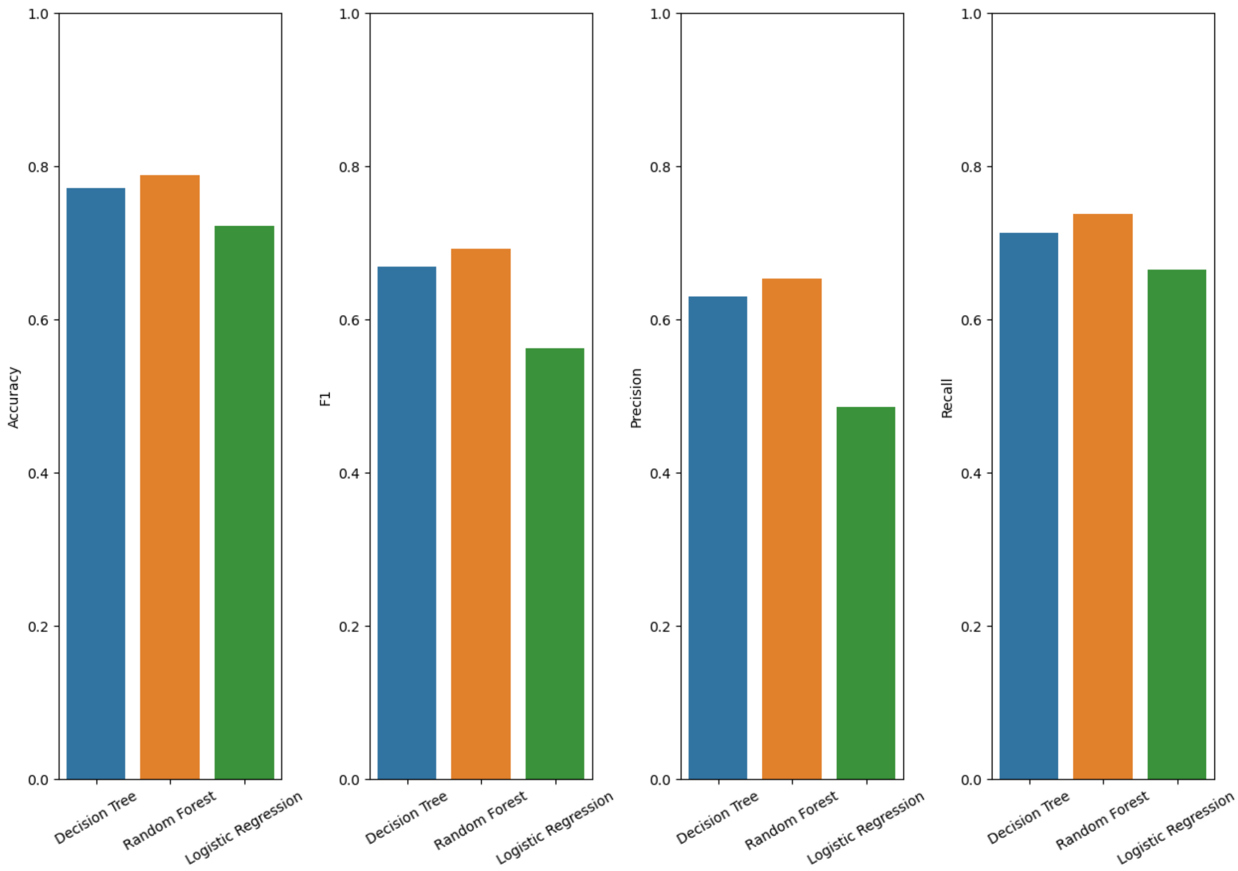

Figure 6. Model performances

The decision tree outperformed the logistic regressor for most metrics, but the random forest performed the best with the highest accuracy of 0.79, F1 of 0.69, precision of 0.65, and recall of 0.73.

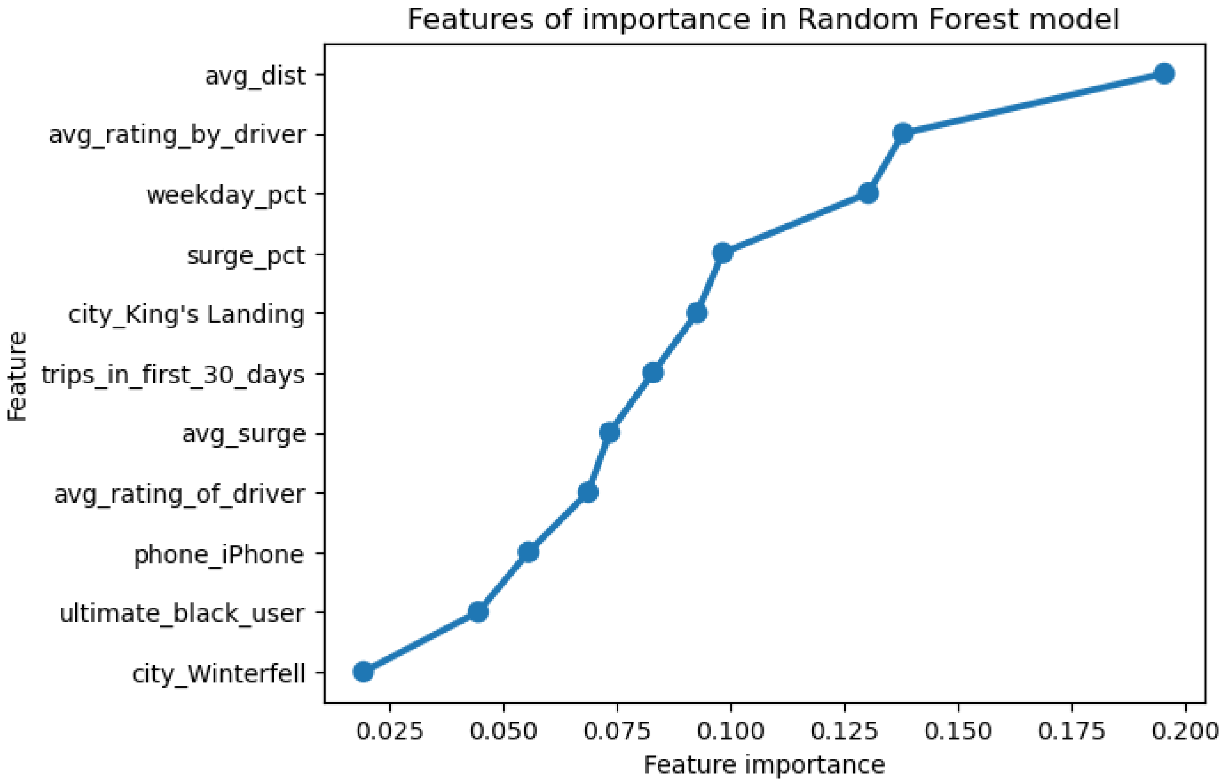

Figure 7. Feature importance for random forest model

The highest feature of importance is the average distance of trip made by users, with active users having slightly lower distances than inactive. I suggest investigating why users who travel longer distances are not retained and possibly encouraging shorter trips through discounts to users. The second most important feature was average rating by driver, which was a median of 4.8 for active drivers and 5.0 for inactive drivers. Because ratings by driver accumulate with more rides, I suspect this impact is due to small numbers of rides among inactive users. While the data included the number of trips in the first 30 days, potentially recording data for longer time intervals could reveal a greater pattern. Active users tended to have a lower percent of weekday rides, so determining if availability of drivers varies on weekdays may lead to insights of how this affects retention. With users in King’s Landing having lower retention, an investigation into the availability of drivers serving the city and competitor apps also serving the city can yield insights into increasing retention.