Jeremy J Jones Ultimate Challenge 5/2/2019

Part 1 – Exploratory data analysis

The challenge in this part is to explore user login timestamp data. Figure 1 shows the results of aggregating the total user logins per 15 minutes and hourly as a function of date and as a histogram while showing the daily user logins and weekly user logins as a function of date. There are a number of trends that can be seen here. First, the daily logins have a periodicity that becomes smoothed out when taking the weekly login count. This suggests that the total daily logins are a function of day of the week. Also, both in the daily and weekly login totals it is apparent that the total user logins are increasing over time (from January to April). This may be a result of more individuals going out as the seasons change from winter to summer.

Figure 2 shows the daily total logins for each day of the week and the hourly total login for each hour of the day. Notice, the weekend (Saturday and Sunday) have the highest user logins while the weekdays show a steady increase in user logins from Monday to Friday. The hourly logins peak around noon/early afternoon and late in the evening (midnight and early morning). This suggests that individuals go out for lunch and therefore login in the evening they want to go out on the town therefore the number of logins peaks at this time. Also, weekends and later in the week are better times to go out at night therefore the number of logins peaks during the weekend and steadily increases throughout the week. To check this theory the hourly total logins per hour of the day is plot for each day of the week, Figure 3. The it is noted that the evening travel totals increase during Friday and Saturday evenings into Saturday and Sunday morning. This confirms that a large amount of user logins is due to going out on the weekend nights. Also, the frequency of logins increases on Saturday and Sunday mid afternoon because people are off of work. Lastly, people login the least often between 6-10am in the morning.

Part 2 – Experiment and metrics design

1. The metric I would choose to look at for the measure of success would be the fraction of riders picked up from a particular city by a particular ride partner. The total number of riders or profit was considered as a metric too but this metric has an issue. After allowing the other driver partner to access the same city there would be sharing of the market and therefore providing easier access to the other city may not change the overall number of riders. Similarly, it would be difficult to use riders per time of day because the market is now shared and the total number of riders will be affected by this too. The one metric that would not be affected by now more often sharing the rider market in both cities is the proportion of riders picked up in a city. This would indicate whether drivers are more often traveling in between the two cities.
2. Practical Experiment
    a. One would need data from the same time period to avoid polluting the data with errors caused by temporal effects. Hence, one would need to do a type of A/B testing. One of the driver partners that was originally exclusive to one city would be asked to give half of their drivers a reimbursement for all tolls (group A) and the other half kept the same (group B). After implementing this change the rider information would be recorded for several weeks. To better select the length of the experiment more information is needed. After collecting rider information the fraction of riders from a particular city would be compared between the group A and B.

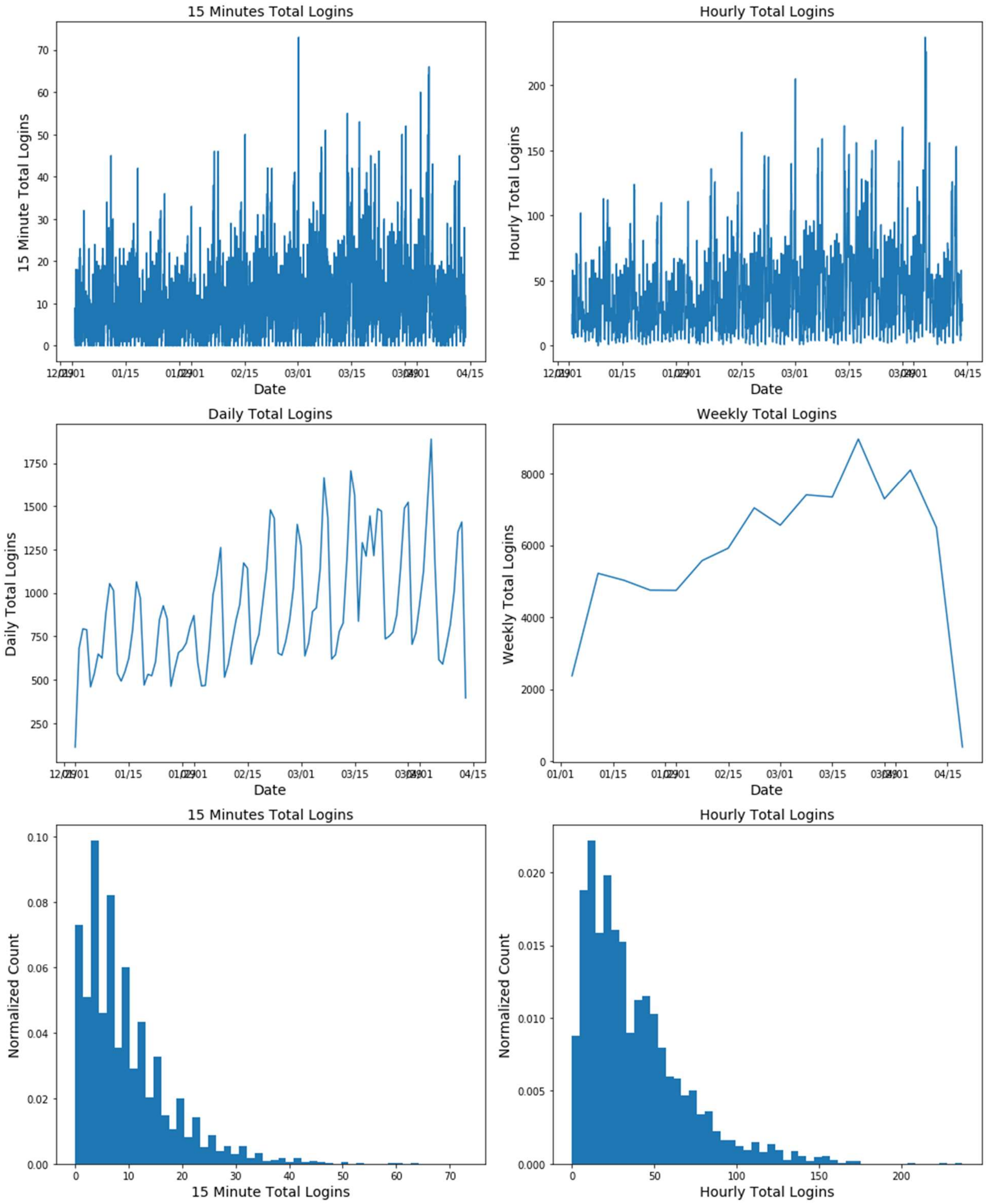


Figure 1 User login totals for 15 minute intervals (upper left and lower left), hourly user logins (upper right and lower right), daily (middle left) and weekly (middle right).

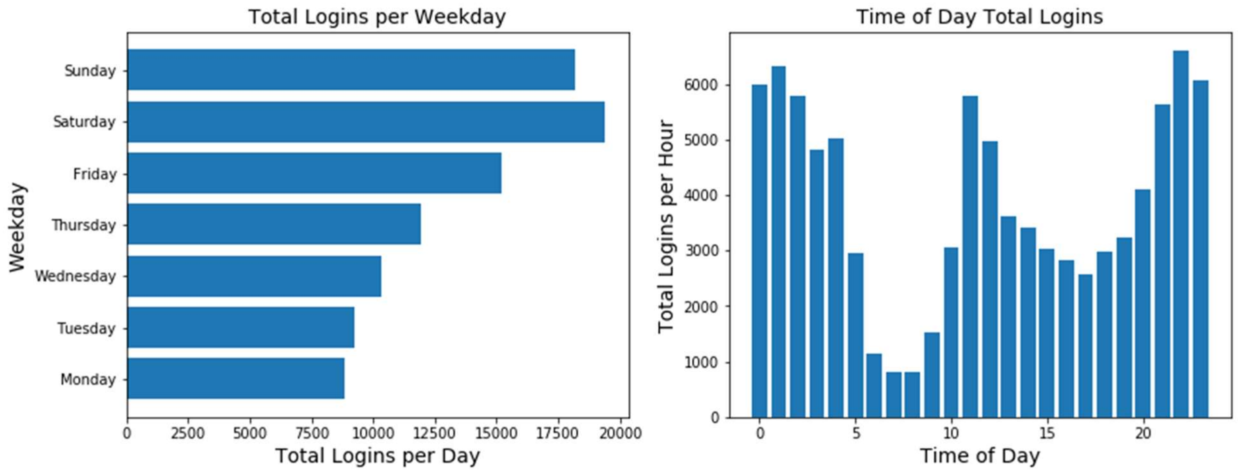


Figure 2 Total logins per day for different weekdays (left). Hourly total logins for each hour of the day (right).

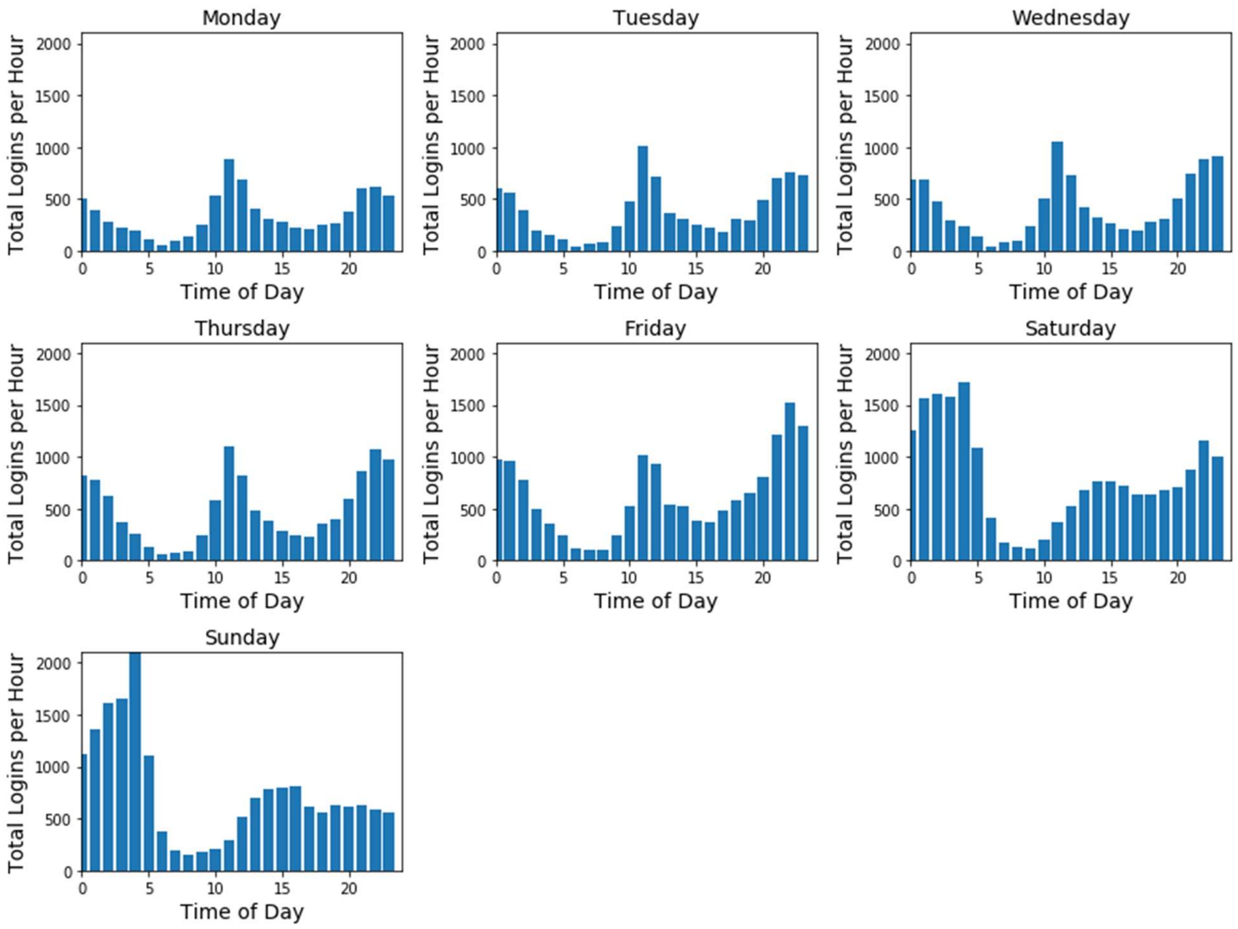


Figure 3 Hourly total logins per time of the day for different days of the week.

b. The proportion of riders pick up in a particular city would be compared and because we are using the entire populations of each group the statistical test used would be a z-test $(\hat{p}_A - \hat{p}_B) = z^*\sqrt{{\sigma_A}^2 + {\sigma_B}^2}$.

c. The hypothesis test would be: null hypothesis, $(\hat{p}_A - \hat{p}_B) = 0$ and alternative hypothesis $(\hat{p}_A - \hat{p}_B) \neq 0$. If there was a significant enough difference between the mean of the proportions such that the p-value was smaller than a cutoff point such as $\propto = 0.05$ then the null hypothesis would be rejected. If the key goal is to encourage drivers to serve both cities then rejected the null hypothesis would say there is a statistically significant argument that reimbursing tolls encourages drivers to serve both cities. Therefore, it is recommended to reimburse toll fares for drivers. If failed to reject the null hypothesis then there is no statistical evidence that reimbursing tolls encourages drivers to serve both cities. Hence, new ideas would have to be formulated to encourage drivers.

Part 3 – Predictive modeling

1. First, the first few rows of the data set were look at to understand the format of the data. Then the missing (null) values were found. There was a significant number of null values for the avg_rating_of_driver column. Therefore, the null values were treated in two ways. One, the rows with null values were completely eliminated. This method created a large difference in the fraction of retained users between the original dataset and the cleaned dataset. It was noted that null values in this the avg_rating_of_driver most likely indicated that the user did not rate any drivers. Hence, to retain the data the null values in this column were replaced with 0s. The remaining null values were eliminated. To check for outliers and see if they were realistic the describe() method was used on the data frame which gave statistics to choose which features may need to be looked at. A boxplot of different features is show in Figure 4. There seemed to be no outlier that seemed unreasonable or unrealistic therefore these points were not eliminated. Additional exploratory data analysis was done to help answer question 3 and therefore many of the plots are in that section. The retention rate for the datasets are:
```
Fraction of users retain, original dataset : 0.376
Fraction of users retain, clean dataset : 0.377
```
2. Before using a machine learning model, the features needed to be altered to be able to be used numerically. The city names, phone types and ultimate black user indicator were converted to numerical values. Because this is about the retention of riders at 6 months the problem is a classification issue. For a moderate amount of features a random forest model or a boosting algorithm would be a good choice. Here the RandomForestClassifier() and the XGBoostClassifier() are used to predict retention. Before predicting on the testing data, a grid search with 5-fold cross validation was done to optimize parameter values. The models give the following accuracies:
```
Random Forest Accuracy 0.787
XGBoost Accuracy 0.795
```

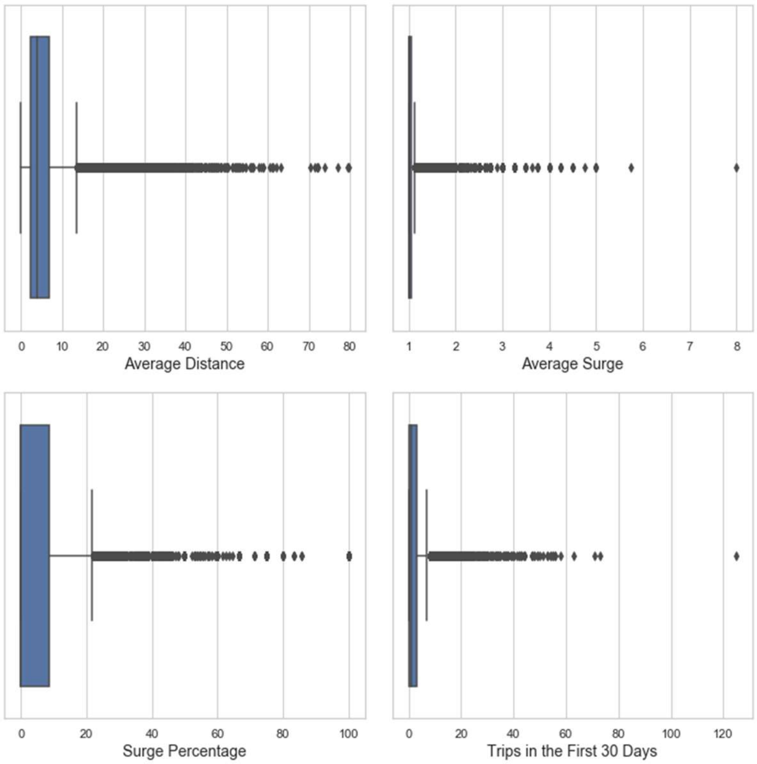


Figure 4 Boxplot for different features in dataset.

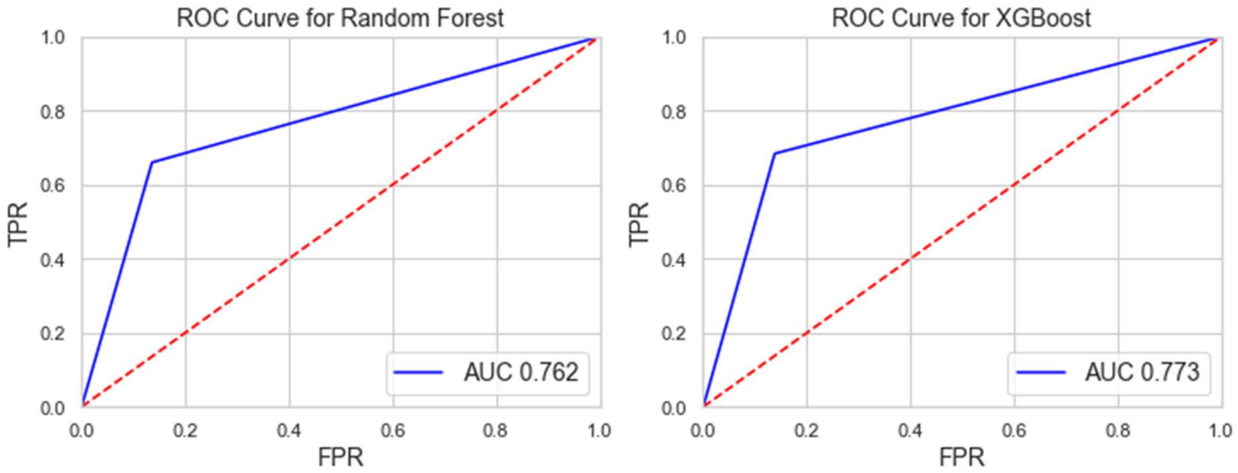
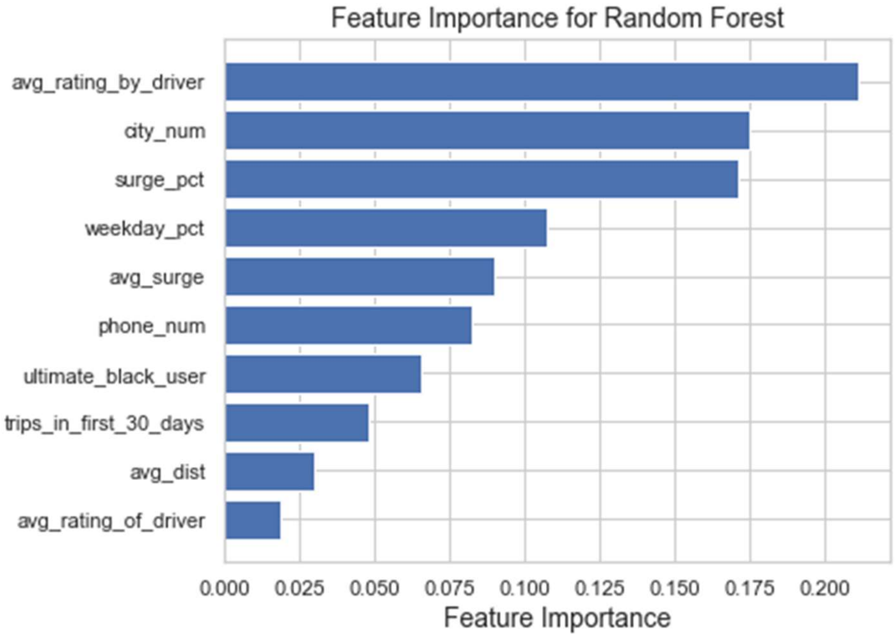


Figure 5 ROC curve for random forest and XGboost classifiers. In the bottom left hand corner is the AUC for each model.

Another way to interpret the model performance is by a ROC curve shown in Figure 5. A weight average ensemble method with these two models was attempted however it did not provide a better accuracy. A naïve Bayes approach, K-nearest neighbor and logistic regression approach was also considered. However, it was thought that these approached would not as effectively deal with the cross correlations between so many variables as well.

3. In Figure 6 the feature importance in each model is shown and in Figure 7 the retention rate for different features is shown. There are several features that are highly important in 6-month retention of riders. The average rating by driver, the city, the phone type, the ultimate black user and the surge percentage seem to play a part in the retention rate. The following recommendation can be made:
    i. Average rating by driver plays a major role. As seen in Figure 7 a rating below 4 shows a large drop off in retention. It is difficult to know what causes this therefore it is recommended that riders be surveyed to better understand if this is a factor that plays a role in retention or if it is an indicator about, they type of riders that will not be retained.
    ii. The phone type seemed to play a role. In particular iPhone users were retained more than Android users (Figure 7). Therefore, it is recommended that the Android app be revisited to see if there is a usability issue or explore if there are more competing options exclusive to Android.
    iii. Those who took an Ultimate Black in their first 30 days were more likely to be retained (Figure 7). Perhaps encourage riders to take an Ultimate Black early with advertisements or a promotional offer in the first 30 days.
    iv. Those riders who averaged a surge percentage around 10% were most likely to be retained. Hence, perhaps keeping the surge multiplier at 1 more often would encourage more riders.



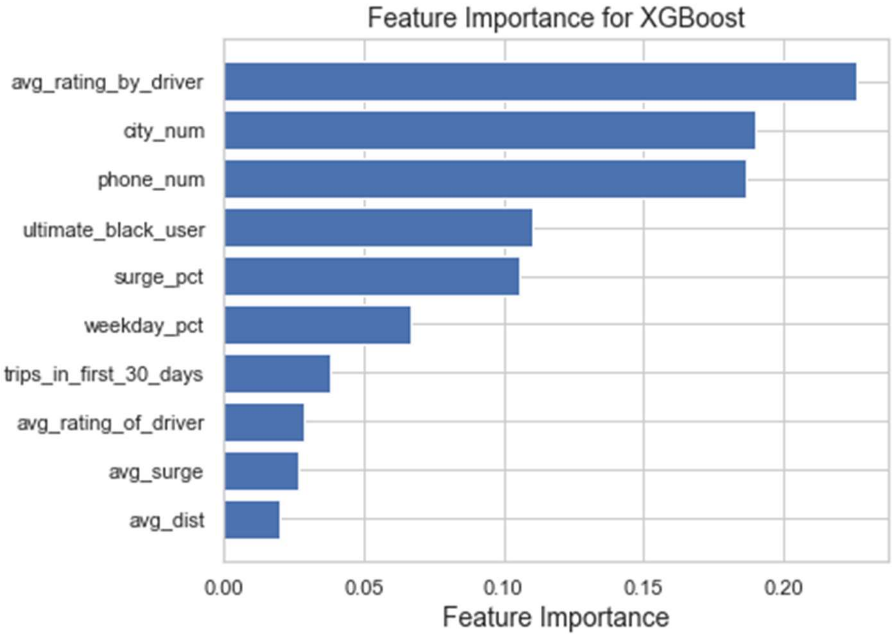


Figure 6 Feature importance in the random forest model (top) and the XGBoost model (bottom).

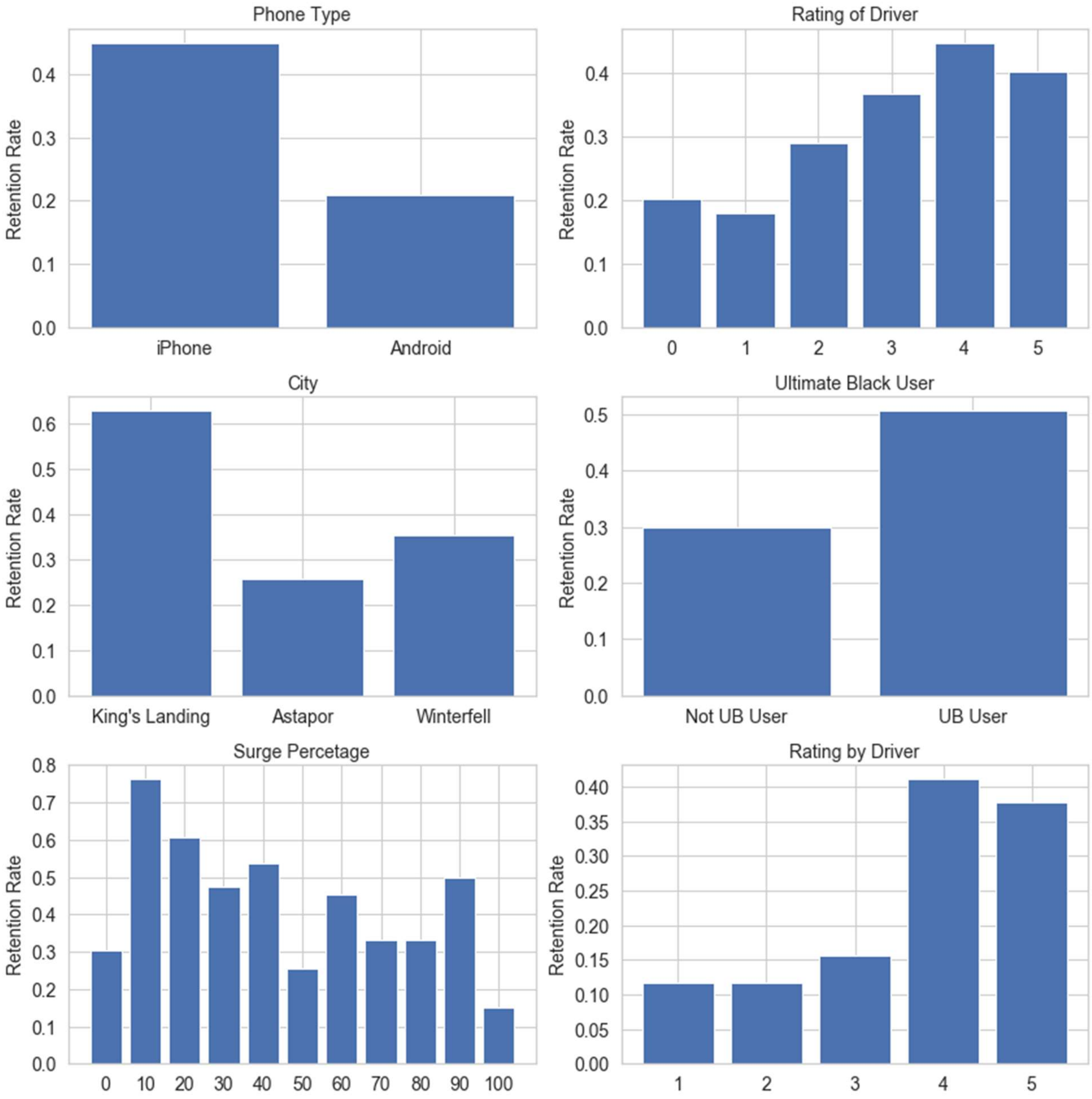


Figure 7 Retention rate for different features.