# Data Cleaning, EDA and Visualizations

We converted json data file into a pandas dataframe and sliced the dataframe by isnull condition to check for the missing data. 8,155 rows had no avg\_rating\_of\_driver\_rating. Also, there are 396 records with no phone classification into Android/iPhone. We dropped the missing values for predictive modelling.

The rest of the modifications included converting the dates to datetime format and adding the retained column to show 1 if a user was “active” (i.e. took a trip) in the preceding 30 days.

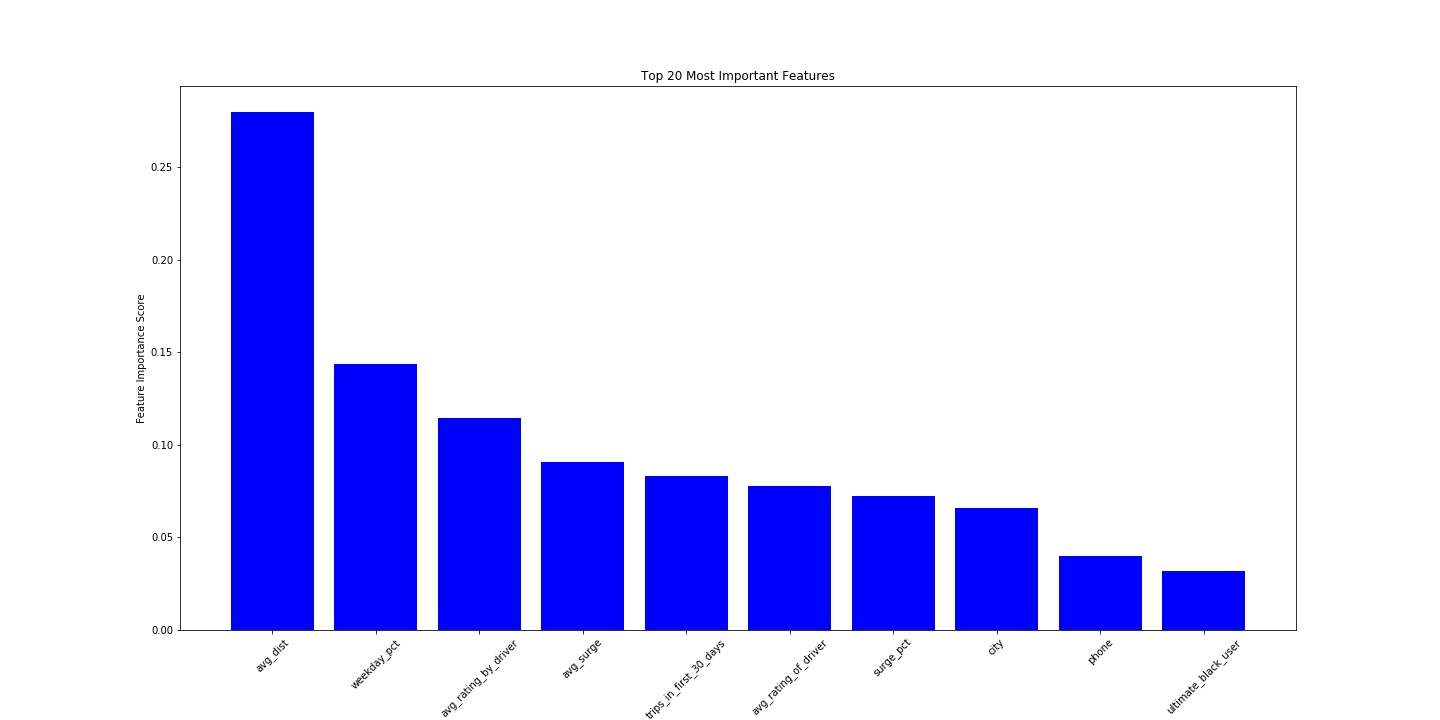
We grouped the dataframe in several ways to observe that the lowest % retained was in Astapor, where the users took on average 1.95 trips in the first 30 days, the lowest of all 3 cities, for the lowest average distance, and where the average surge percent was 10.14%, the highest out of 3 cities. In Astapor, like in all 3 cities, retention of Android phone users was problematic (only 12.3% Android users retained vs 31.8% for Iphone).

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **city** | **phone** | **Total users** | **Retained Users** | **% of total** | **avg\_rating\_of\_driver** | **avg\_surge** | **avg\_surge\_pct** | **number\_of\_ultimate\_black\_users** | **weekday\_pct** | **trips\_in\_first\_30\_days** | **Avg\_trip\_dist, miles** | **avg\_rating\_by\_driver** |
| Astapor | Android | 5244 | 645 | 12.3% | 4.65 | 1.08 | 9.68% | 33.0 | 60.07% | 1.73 | 5.73 | 4.82 |
| iPhone | 11169 | 3553 | 31.8% | 4.62 | 1.09 | 10.39% | 41.9 | 60.47% | 2.05 | 5.28 | 4.79 |
| King's Landing | Android | 2498 | 1092 | 43.7% | 4.74 | 1.07 | 9.38% | 36.5 | 63.99% | 1.94 | 6.12 | 4.85 |
| iPhone | 7568 | 5233 | 69.1% | 4.70 | 1.07 | 10.19% | 43.0 | 62.91% | 2.64 | 5.90 | 4.85 |
| Winterfell | Android | 7280 | 1409 | 19.4% | 4.57 | 1.07 | 7.43% | 30.0 | 59.41% | 2.08 | 6.11 | 4.73 |
| iPhone | 15845 | 6739 | 42.5% | 4.52 | 1.07 | 7.50% | 37.0 | 60.67% | 2.59 | 5.91 | 4.72 |

The fact that the average driver rating in Astapor is not the lowest in 3 cities makes us assume that the problem with retention is not due to the quality of drivers, but rather due to the surge multipliers.

# Predictive Modeling

We considered 2 models for predictions: Random Forest and Linear Regression. Random Forest model is known to be good in classifying complex dependencies and offering a reliable feature importance estimate. Linear Regression Model is more simplistic and less computationally intense. We chose Random Forest, because the ROC-AUC score of the fine-tuned random forest model was 85%, which is high in general, and it was higher than the same score for the Linear Regression Model (71%). We wanted to ensure that the model is correctly predicting the retained users, and we were happy with recall of 96%. The features that are the most predictive for user retention are the average distance in miles per trip taken in the first 30 days after the signup, the percent of trips taken on the weekend, the average rating by driver and the average surge multiplier for all of the user’s trips.



Ultimate will leverage the most predictive features by giving discounts for users to ride longer trips, ensuring that the surge multipliers are reasonable across all the regions, encouraging drivers to provide good rating for users, and encouraging riders to drive on the weekend.