Ultimate Take Home Challenge part 1

Weekdays all show similar patterns, a spike in logins between 11 am and noon around 300, and an increase between 9 pm and 11 pm which grows a little from Monday, around 150, to Wednesday, around 200. By Thursday that increase is about as high as the midday spike, at around 300, and by Friday that increase is greater than the midday spike, at around 400. Weekends show different trends, with Saturdays showing peak logins between 1 am and 5 am, around 400, a low peak between 1 pm and 4 pm, around 200, and a middle peak between 10 pm and 11 pm, just under 300. Sundays show the busiest peak of all, beginning at midnight, around 300, and growing until 5 am, around 600, with a small peak between 2 pm and 5 pm, around 200.

Ultimate Take Home Challenge part 2

1) If the only goal is to get drivers to work out of both cities, the obvious measurement would be number of tolls reimbursed. However, I would suggest this is a business decision, thus perhaps a better measurement would be increase in mean profit. It would be better to measure profit rather than revenue, because profit would take into account the money lost in reimbursed tolls, instead of just increased revenue. The underlying assumption is if drivers have more perceived freedom to pick up riders, the company's profit would increase.

2) Based on the suggested profit experiment, assuming drivers have a "home base" In one of the two cities, I would divide the drivers from both cities into two batches and offer toll reimbursements to one of the halves from each city. Thus, there are four batches of drivers, one from each city who don't get tolls reimbursed and one from each city who do get tolls reimbursed. Then, the experiment is run for however many days the company wants, 30, 60, 90, whichever, and the mean profit is measured each day for all drivers.

Once the experiment is concluded, a 2-tailed t-test can be run on the data to determine whether the mean profit is higher or lower with the tolls reimbursed than without. Specifically, run the t-test on the difference between mean profit with tolls reimbursed and without. The hypothesis test would be the difference between mean profits is 0, with the outcome of the test showing the evidence one way or the other.

Based on the p-value calculated, the company can decide if it is more profitable to continue reimbursing tolls or revert to the original operating procedure. Perhaps, the data can be further studied and tolls reimbursed only during specific times of day and/or in specific directions, since during the week, Metropolis is busier during the day and Gotham is busier at night.

Ultimate Take Home Challenge part 3

1) During cleaning, I found missing data in the driver rating, passenger rating and phone categories. I chose to impute the mean for the missing driver rating, because the majority of the data was similar, with a mean near 4.6 out of 5, and there was quite missing, roughly 8000 out of 50000 entries. The passenger rating and phone were missing very little, around 200 and around 400 entries, so I chose to simply remove the missing data from the data set. Here are the number on the retained customers:

```
Total number of users retained : 18671
Total percent of users retained : 0.37640109668575117

Total Ultimate Black users retained : 9410
Percent Ultimate Black users retained : 0.5055877928218354

Total users retained from Winterfell : 8148
Percent Winterfell users retained : 0.3523459459459459
Total users retained from Astapor : 4198
Percent Astapor users retained : 0.25577286297447144
Total users retained from King's Landing : 6325
Percent King's Landing users retained : 0.628352871051063

Total iPhone users retained : 15525
Percent iPhone users retained : 0.44893297090972184
Total Android users retained : 3146
Percent Android users retained : 0.2094261749434163
```

2) I chose to work with a random forest classifier, because I have had success with that model in the past. An alternative classifier would have been a logistic regression classifier, but I've had better success with the random forest. After using hyper-parameter tuning, the model provided the following classification report:

```
Classification Report :
             precision    recall  f1-score   support

          0       0.80      0.81      0.80      7733
          1       0.68      0.66      0.67      4668

avg / total       0.75      0.75      0.75     12401
```

As shown, the model performs relatively well predicting customers who will likely not be retained, and slightly less well predicting customers who wi ll be retained. The precision score is the accuracy, the 0 row represents customers who will not be retained and the 1 row represents customers who will be retained.

3) According to the summary information and the top graph, 75% of reta ined customers travel less than 6 miles per use, so targeting short trip c ustomers seems to be a good idea. The weekday usage is less clear, with th e mean use around 60%. 75% of retained customers show weekday use up to 85

%, and the graph show the largest group of customers use the service only on weekdays, with 2 smaller groups using only weekend service, and split e venly between both, so targeting weekday users may also be favorable. Incl uding the earlier visualizations, Ultimate's most retained customers appea r to be mainly weekday users who travel short distances, especially those who originated in King's Landing, use iPhones, and participate in the Ulti mate Black program.