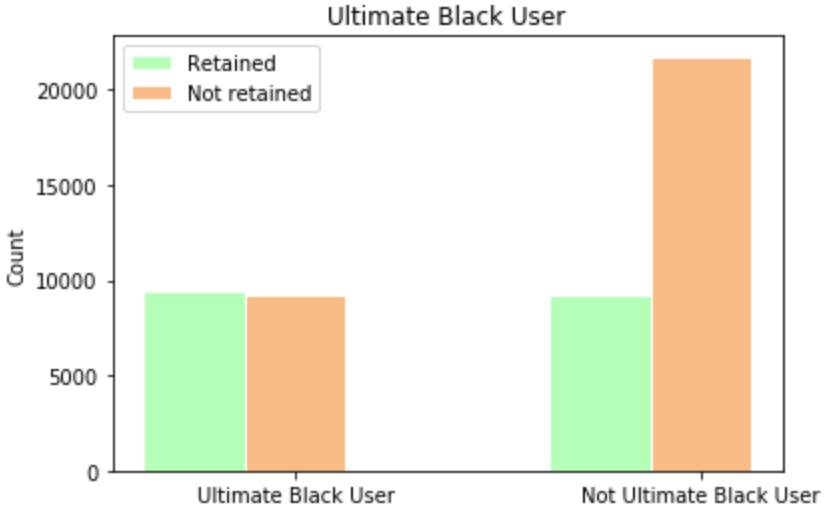
Melanie Malinas
Ultimate Challenge
Springboard

Part 3: Predictive Modeling

1. Cleaning, exploratory analysis, and visualizations

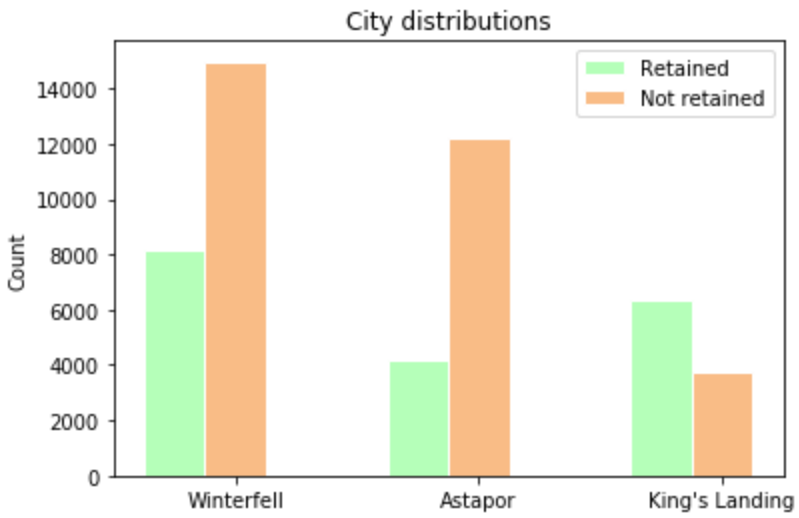
I loaded the data into a Pandas dataframe using the json.load() function. I then sorted the dataframe by last trip date and found out that the last trip was on July 1st, 2014. I then would consider a rider to be retained if they took a trip after June 1st. I created a datetime object for June 1st, 2014 and then created a column called “retained” that was True if the last_trip_date was after June 1st and False otherwise. I then counted up the number of Trues and Falses and found out that **37.6% of the riders were retained**.

I also made several visualizations and calculated some statistics to try and understand the factors that would cause someone to be retained or not retained. For example, I looked at how many people were Ultimate Black users and how many of those people were retained or not retained.



As you can see, the proportion of Ultimate Black users retained is much higher than the proportion of non-Ultimate Black users retained.

I also looked at the distributions for the different cities:



As you can see, Winterfell and Astapor had way more people not retained than retained, but King's Landing had more people retained than not retained, which is interesting.

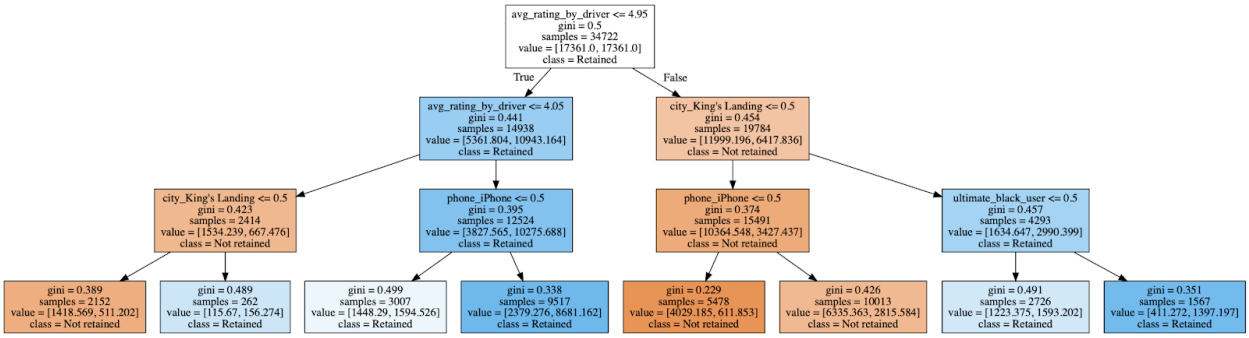
2. Predictive model

I built a predictive model to help Ultimate determine whether a user will be active in their 6th month. I used my "retained" column as the y, and the other columns as the X. I also dropped last_trip_date from the dataframe because we only want to know whether someone was retained or not.

I also converted the city and phone columns to numeric columns using pd.get_dummies() and converted the True/False columns to numeric values of 0 and 1.

I then tried two different methods for the model. I first tried logistic regression, because it is a very simple method for classification. Using logistic regression, I got an F1 score of **0.62**. I then tried random forests, because random forests are known to be a strong method for classification. With random forests I did better, getting an F1 score of **0.69**.

I also created a decision tree for this model in order to better understand the features in the model that were contributing most strongly.



It looks like the factors that most affect the model are the average rating by the driver, whether the city is King's Landing, whether the phone is an iPhone, and whether the person is an Ultimate Black user. For example, people with an average rating by their driver of greater than 4.95, who are from King's Landing, who are Ultimate Black users, are more likely to be retained (Gini gain is lower for this group).

3. How might Ultimate leverage the insights gained from this model

I think there are a couple of things that Ultimate can do with this information. For one, they need to find out what they're doing in King's Landing that is so successful. The other two cities have way worse retention rates, so Ultimate needs to more closely examine what's happening in King's Landing and try to replicate it.

The second thing that I think they need to look at is Ultimate Black users. Clearly they are doing well with users of Ultimate Black, so they should probably do a survey of their Ultimate Black users and see what it is that they like about Ultimate Black. This might help them to improve.