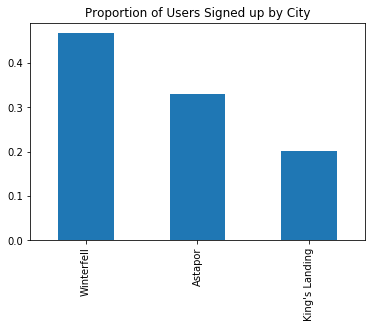
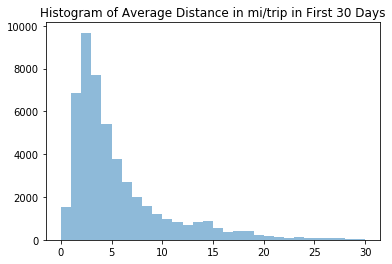
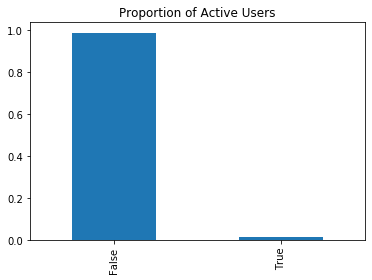
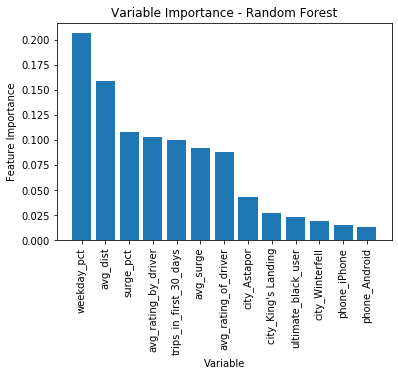
1. As far as data cleaning, there were a some missing values in 3 of the columns. The columns were average rating by driver, average rating of driver and phone type. For the two driver rating averages, I just used the average of the overall column to fill in missing values, as I felt like population averages were probably pretty accurate/appropriate and this was a better solution than throwing all these rows away. For phone type, there were more iPhone than Android user (~70/30 split), but I just front filled the data, because the population averages should fill in the phone types correctly, with some variance of course, but again I felt like this was better than throwing away data points. After cleaning the data, I just performed very basic exploratory data analysis to make sure that population frequencies and histograms looked fairly normal. For examples of this, I’ve included a bar graph of users by the city they signed up in and a histogram of the average distance per trip of a user in their first 30 days.   One very important observation was regarding the target variable (the prediction variable of interest). It was highly imbalanced as about 99% of riders were considered inactive after 6 months with Ultimate. This is displayed below: 
2. When I went to build my original models, I realized they quickly had issues predicting the minority class (those riders who were still active after 6 months). Whether out of the box or tuned, most models failed to predict even one active user, while even if a few users were predicted as active, the PR AUC scores were horrible. The best ROC AUC score was .77 and PR AUC score was just .03 and these were produced by the random forest model. Accuracy was 99% but that is irrelevant as that is just the sample mean of inactive users and all predictions were of inactive users. The final model recommended to Ultimate after my analysis would be my tuned AdaBoost model that accounted for class imbalance. I handled class imbalance using a method called Random Undersampling (RUS). The idea is to keep all the samples in the minority class and randomly draw samples from the majority class to have an equal population of both active and inactive users. Once I modified the dataset in the following way, the ROC AUC score improved to .80 and the PR AUC score improved to 0.77. Overall accuracy was 72%, but this model dealt with a population that had a 50/50 split of active and inactive users, so this was still a decent result. The model seems quite valid, but one caveat is that the total dataset was reduced to 1,096 riders, instead of the original 50,000 due to having to account for class imbalance.
3. Since the random forest model had decent performance (ROC AUC of 0.78, PR AUC of 0.73, and accuracy of 73%), it makes sense to extract feature importance from this model. A visualization of this is given below: 

As we can see from the feature importance visualization, ‘weekday\_pct’ is the most important feature in the model. This represents the percent of user’s trips occurring during a weekday. A little more digging yields that the averages trips during a weekday for an active user is 70%, while that drops to just 60% for inactive users. This might suggest that active users are more likely to use Ultimate for their commute to work. As such, maybe there would be a marketing push to promote advertising to commuters to offer things like monthly packages or discounts for taking a certain number of rides per month or week to try and incentivize commuters to make Ultimate their main means of transportation to and from work.