Part 3. Rider retention

By the end of the tracking period (July 1, 2014), 36% of the original tracked cohort were still active, per the definition of having taken a trip in the last 30 days.

I used a random forest classifier with the target variable as a simple yes/no on whether the user was still active at the end of the 6 month period. I tried both logistic regression and KNN models as well. While performance was similar on all three models, random forest was the most effective model and also ran the fastest. I also considered doing a regression model since I had the number of days active for each user. However, the stated output desired was simply whether or not a user was still active, so in the end I went with the classification options.

My model is able to predict with roughly 84% accuracy whether a person was NOT active, but it had a much harder time classifying the people who were still active. However, this is not necessarily a bad thing, as the people a company wants to target are the ones who are most likely to drop out of the program.

My biggest suggestion would be to reach out to people who have gone on at least two rides but have not rated the drivers. They are the most likely population to go inactive, so targeting them with incentives to stay would likely be the most productive use of any targeted advertising or promotions.