w207 Kaggle Project Progress Report

Keri Wheatley · Nicholas Chen · Robert Mulla · Zach Ingbretsen
March 30, 2017

The goal of the competition that our group selected is to predict the number of bike rentals in a bike share program. The data are provided by Capital Bikeshare; the bike share program in Washington D.C. Each observation in the data is one hour of the day. The target variable is the number of total rentals across the city. The features that are part of the dataset are temperature, date and time, season, 'feels like' temperature, a flag indicating whether the day is a holiday, a flag indicating whether the day is a working day, weather, humidity, and wind speed. The main target variable is the total number of rentals. The target variable is also split into two separate variables; casual rentals which are rentals made by individuals who are not registered in the program and registered rentals which are rentals made by individuals who have previously registered.

As the first step, our group conducted exploratory data analysis on each of the components of the dataset. We looked at distributions of each of the variables in the dataset and correlations between the features and the target variables. We noted that the distributions of casual rentals and registered rentals looked different across levels of some features in the data. For example, holidays seem to have a positive effect on casual rentals while they seem to have a negative effect on registered rentals. We also noted that the number of total rentals appears to be growing over time. As a result, we feel it will be important to include a program age variable in each of our predictors.

Next, we began fitting models to the data. We have tried a variety of models including linear regression, decision trees, and random forests, but have primarily focused our attention on random forest models. Initially, we thought that predicting casual rentals and registered rentals separately, then adding the results to get total rentals would be more accurate than predicting total rentals directly, but our initial results taking this approach were less accurate than our initial results predicting total registrations directly.

A large proportion of the observations in the dataset have a very low total number of rentals. In our error analysis of one of our initial models, we noticed that the random forest model that we've fit seems to do a much better job with low rental observations than high ones. As a result of discussions with Zach during office hours, we will try fitting a sequence of models as follows to alleviate this issue: first, we will have a classifier that categorizes each example as a 'low' example or a 'high' example and next we will try fitting separate models to predict 'low' rentals and 'high' rental cases. We also plan to look into other classifiers such as Poisson regression, Adaboost, and XGboost.