Team G

Marcos Antonio de Souza Barrozo Filho

My final approach was to use a regression tree, with a final Root Mean Squared Logarithmic Error (RMSLE) of 0.85. To give that number some perspective, the benchmark submission (where you just predict the mean demand) had a score of 1.58 and the winning submission scored 0.338.

It was a long process, one which involved trying multiple failed approaches. I began by trying linear and poisson regressions with LASSO shrinkage. The poisson approach is said to work well with count data, but in my case I ended up getting scores around 3.07; that is, worse than just predicting the mean! With LASSO linear regressions I had a bit more luck, with an initial score of 1.37 using mostly variables related to weather – I was having trouble manipulating the *datetime* strings. Then I tried some regression trees with the same variables, also getting scores around 1.30.

But I was not satisfied, after all, my scores were not that much better than the mean prediction. So I started searching through the discussion boards on Kaggle to see how people manipulated the time variables. I found one submission in particular which used a lot of time variables, and I replicated some of its approaches. It basically divided the day into 4 groups, and predicted (of course) low rental counts for really odd hours. By just adding that, my trees then scored 0.85 and the linear LASSO went down to roughly 1.10.

If I were to do this all over again, I would spend more time on data manipulation, getting the variables in a format which is suitable for prediction. Different approaches, like LASSO and the trees, have different effectiveness in different situations, but my predictions only got to a decent level once I started spending more time playing with the data themselves. I think there is still a lot I could do with the timestamp, but the time crunch amidst final exams limited the scope of my exploration. It was, however, a great learning experience!