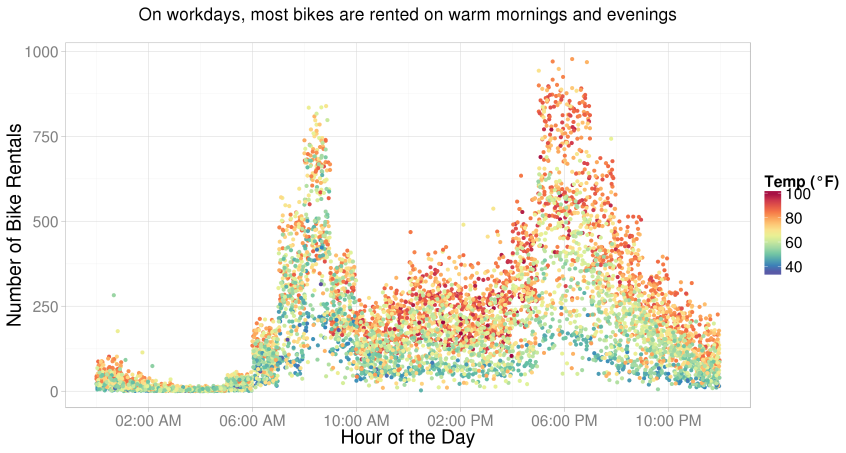
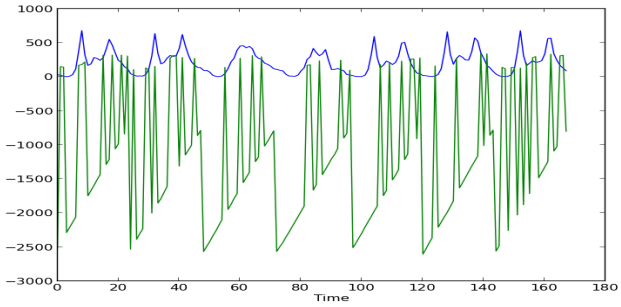
Benjamin Chen
a take-home interview exercise from DataRobot for Data Science intern
http://www.kaggle.com/c/bike-sharing-demand

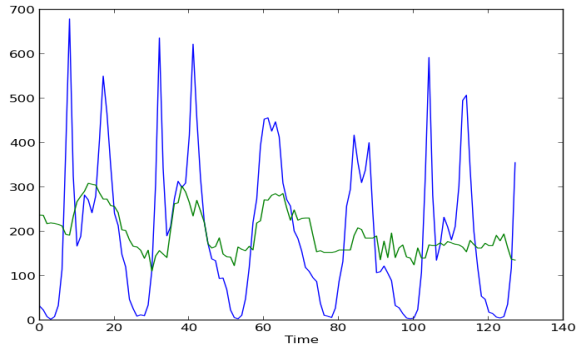


The data is composed of almost 2 years of bike rental information, of which the last week was reserved for the test set. This application of forecasting new bike rentals calls for this instead of sampling for the test set randomly. Below are sample training sets plotted against time. The blue is ground truth and green is the predicted point. 0 on the time axis is Thursday midnight.

Fitted Week 1 of Model using Date and Hour as separate variables



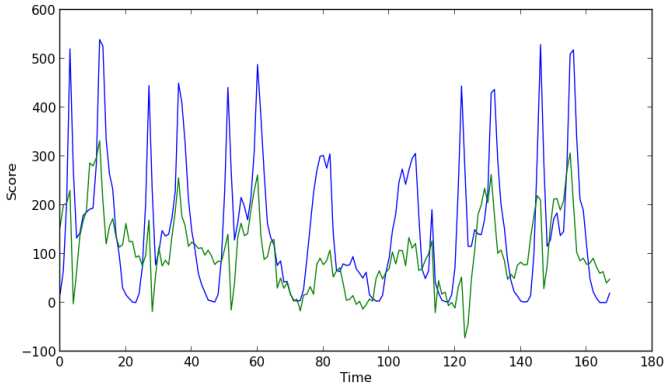
Fitted Week 1 of Model using a unified Time variable



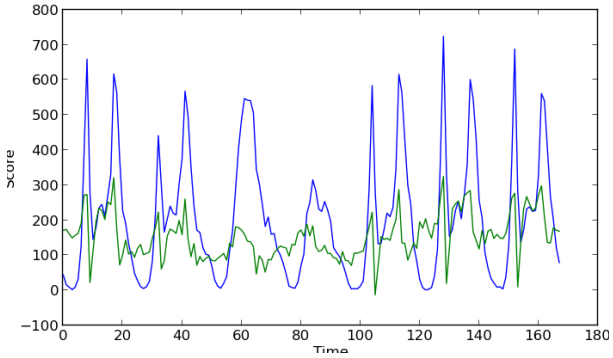
We can see that using date and hour as separate numeric variables, Google does not learn that those two variables are related. Clearly, with the zigzagging pattern, it is using a linear component to hour of the day. Using a unified time variable (ex: 1337.5 being noon 1337 days after 2011-01-01), we see a better model that seems to somewhat understand the cyclic component of the data on Thursday through Saturday, but captures neither the spikes around 9am and 5pm, nor the dips between days.

Taking a different direction, encoding time as the number of rentals from one hour prior and one week (168 hours) prior. We see that it is more representative of the data, but has similar problems when trying to generalize. The weekends in most predictions are particularly poor. I suspect this is because the regularization parameter was lower than desired. Further work would be to assess the hour of the week as a category and analyze the residuals with respect to the rest of the input.

Fitted Week 35 using a Model Including Last Hour and Last Week



Sample Predictions of Model including Last Hour and Last Week on Test Data



Overall, I am unsatisfied with the results but am unsure Google Prediction API is the right tool for this particular job. Considering its opaqueness and how much processing needs to occur beforehand, I question how much work it is actually saving, if any.