Spencer Guerrero

# DataRobot - Take Home Test

Do I think Google's prediction API, results are good?

A similar question:

Say won't it be fun to test electrical demand prediction models using Google's Prediction API? Answer: Not as much as it could be, because your data will break it and result in the API returning the same number no matter what input.

Some background:

After reading the documentation for the prediction API—there are two types of models the API can create: Classification (categorical) or Regression. Given the scope of this project (a couple of hours) I decided to test the prediction API's regression model on my electrical load data, I'm always interested in testing data using a new technique.

I decided to train a model on 2 years of hourly electrical load data and independent variables consisting of temperature, dew point, time and various dummy variables representing calendar effects. Using R I would test a multiple linear regression model against the Prediction API results.

Well for my first model the API kept returning 28464.879287. I thought after the first model, that maybe the dummy hour variable was to blame. So for the second model I took an hourly subset of the data. This model's output was 30609.222668. A couple of days later, I was wondering if the day of year variable was to blame, so I decided to remove it, and at the same time, go big or go home and introduced lagged variables. I used 48 lags each of electrical demand, cooling degree hours and heating degree hours. This set of data also resulted in the "only one number for output" problem but gave very good results in R for next hour forecasts.

At this point I was pretty confused, given multiple linear regression in R was giving acceptable output. So I decided to run a trivial model just to make sure it was the input the prediction API was struggling with. I ran a perfect linear model experiment and the prediction API returned acceptable results. The next step would be to test if the monthly dummy variable and the heating/cooling degree hours are causing problems. If I couldn't narrow it down to "the API struggles with dummy variables or variables with lots of 0's" then I would have to start removing variables one by one until I located the problem. (Assuming contacting support isn't an option—I'm not sure how fast support would get back to me).

In the end, "Do I think Google's prediction API results are good?" Maybe but I haven't been able to test it yet.

On a more serious note, not being able to examine the trained models in this case is a very serious problem. It could be a deal breaker for major projects but maybe they would be able to contact Google for support.