Competency challenge for Arundo Analytics

Johan Holm

September 7, 2017

The goal of the challenge is to predict the number of engineering support requests that a set of field sites request on any given day. I am tasked with building a model that can predict the request count in the test data set and answering a few questions about the data which can be found under the Discussion section.

1 Methods

1.1 Collection of data

Data used for this analysis are two datasets that comes from Arundo Analytics. Usually when dealing with data it is important to make sure that we know any eventual selection bias. I will assume there are no selection bias in the data, but I mention it so that it is known that I am aware of selection bias and to make sure I am aware of it so that the data can be weighted correctly to avoid skewing the results.

1.2 Exploratory analysis

Exploratory analysis was performed in order to identify the quality of data (such as identifying and excluding missing values) and to determine the proper terms for the regression model in order to model a relationship between request count and other variables. Exploratory analysis consisted of

1. looking at the raw data tables
2. transforming the raw data (e.g., identifying missing values and converting variables into numeric and factor formats, as needed)
3. studying data plots
4. determine likely approaches to modelling, which might best yield a predictive function

1.3 Statistical modelling

In order to identify the relationships between request count and other variables of interest, several regression models was performed. Because the task is to predict values it is most reasonable to use regression models.

The choice of the model was informed by comparative analysis of several different regression methods and based on adjusted RMSE (root mean square error) as a 'Goodness of fit' value.

1.4 reproducibility

All analyses have been reproduced in a Python file, which can be provided upon request.

2 Results

Dataset used in this analysis contains 153 observations and 8 variables:

- date: yyyy-mm-dd format
- calendar_code: 0 or 1 (a code describing certain calendar events)
- request_count: an integer (the number of support requests received on that date)
- site_count: an integer (the number of sites operating on that date)
- max_temp: a float (max temperature for that day in degrees Celsius)
- min_temp: a float (min temperature for that day in degrees Celsius)
- precipitation: a float (millimeters of precipitation on that date)
- events: a string (description of weather events on that date)

One column of dates and one categorical coulmn. The date column was excluded from the training (I assume I could have gotten day of the week from it, but that has to be delivered the coming week if I have time.)

I first tried with regular regression techniques, Linear regression, Support vector machine regressor, random forest regressor, decision tree regressor, but none of them seemed to work despite my efforts and tweaking with kernels.

I then tried Lasso, ElasticNet, Ridge Regression, but none of them seemed to have any results that looked promising. As I am unemployed and don't have access to a very powerful computer I want to avoid deep learning algorithms so I started on the Ensamble regressors and iterated through certain variations with a python script until I ended up with a composite estimator with Extra Trees Regressor and then Cross-validated Lasso, using the LARS algorithm.

Eventually I ended up with the best possible RMSE I was able to.

3 Conclusions

This analysis suggests that there is a statistically significant association between the request_count and the provided data. I unfortunately did not have time to answer the 3 questions