Exploring Forest Fire Burn Area with Linear and Nonlinear Models

Tobias Iven | Jacob Oliver | Aungadt Walia

June 11th, 2018

STAT 331-74

Overview of the Data

The Forest Fires Data Set, courtesy of UCI's Machine learning repository, consists of 517 observations,13 variables, and has no missing values. The input variables consist of spatial, meteorological, and periodical data. These inputs are used to predict the burned area of a forest with in the Montesinho Natural Park, a reserve in the northeast region of Portugal. The predictor variables include the X and Y spatial coordinates with in the park, the month and day of the observation. Four measures from the Fire Weather Index including the Fine Fuel Moisture Code (FFMC), the Duff Moisture Code (DMC), the Drought Code(DC), and the Initial Spread Index (ISI). Additional meteorological information such as the temperature (in degrees celsius), the relative humidity (as a percentage), wind speed (in kilometers per hour), and rain (in millimeters per square meter) were recorded on the day of the observation. The data was made available by two professors at the university of Minho in 2008.

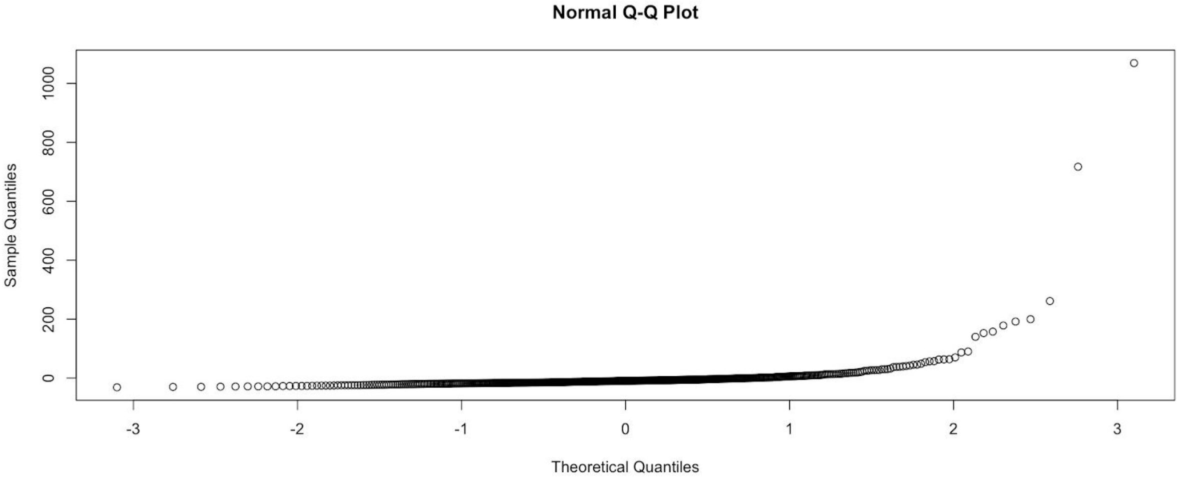
Goal Of Applet

We wanted to demonstrate how there are some regression tasks that linear regression are just not up to par for. Even with some transformations we can see that the path down the linear brick road is not going to be a fruitful one, as demonstrated by our extremely large p values and our small global F - statistics.

We wanted to show how using a non-linear approach like neural networks could produce a better representation of the data.

Statistics Review

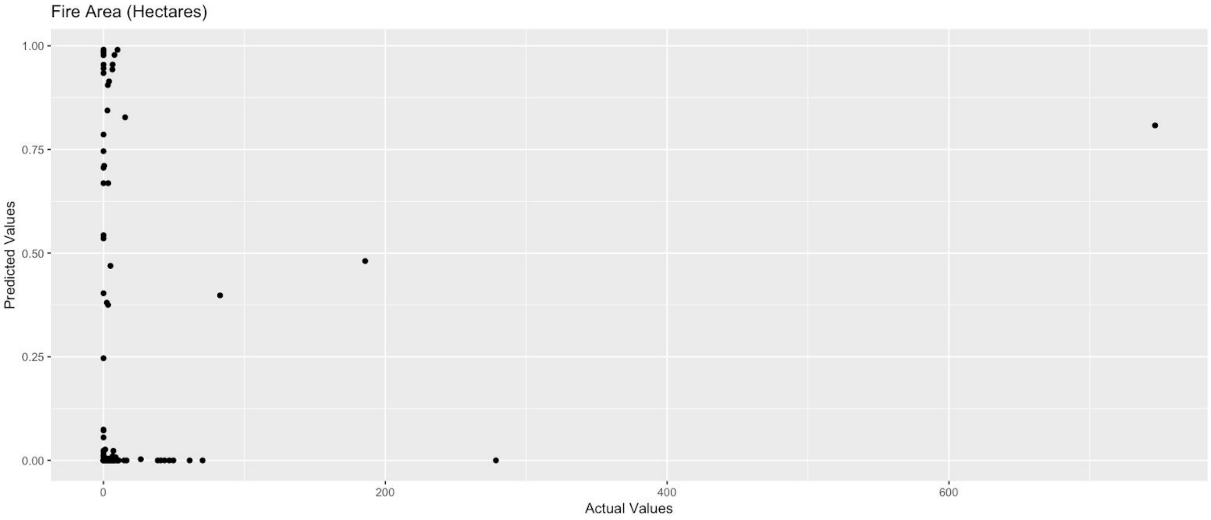
As we can see in the linear regression output, the approach of regression is fundamentally flawed. Even with several transformations on y it is very hard to get any sort of good model for the data. What's more, it seems that in every instance we still have assumption violations for regression, and therefore cannot even trust our already weak results. (As we can see in the normal quantile plot the residuals from the multiple regression are horrendously non-normal).



The progression to neural networks is natural as a starting point for non-linear models. The visualization of the neural networks are a tool to see the progression of each stage of the neural net in the matrix multiplication and moving through the net. The black arrows and their labels represent the weights, and the circle represent the nodes along with activation function. The sole output node is our response variable area.

As noted by comparing the MSE of the neural net and the regression (for the same training observations as the neural net) we do considerably better with the neural net. This must be qualified however with the fact that the neural net is still not cross validated, and is merely tested and trained. Further work should be done to cross validate the neural network to get a better result.

As you can see in the observed versus predicted plot for the validation data set, the neural network still does a fantastically bad job of predicting the data. We believe that even stronger non-linear approaches with even less assumptions like a support vector machine would be more effective at predicting the area of the forest fires.



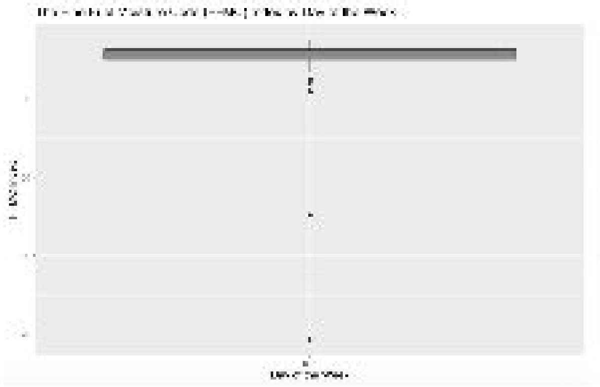
Using the Application

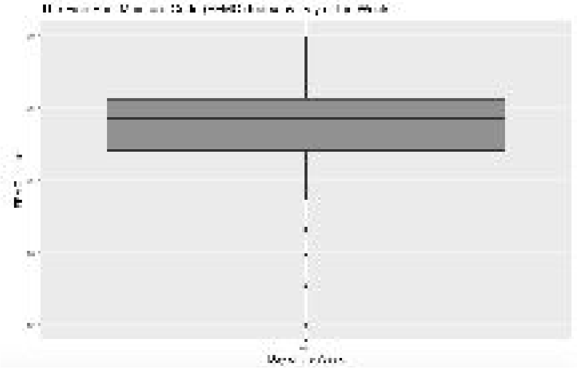
The application has three interactive graphs, a heat map of the park by month, a scatterplot of the ISI, relative humidity and temperature, and finally a box plot that shows the distribution of the FFMC index by the day of the week.

The *Month of the year* slider in the control panel will show corresponds to the first graph heat map of Montesino Park. The graph illustrates the intensity and spatial distribution of temperature in the park. The X and Y axises of the graph respectively represent the X and Y spatial coordinates on the park map. The color indicates temperature, the pale yellow represents low temperatures. The closer a given area is to color red, the higher the temperature is. The slider allows the user to change the month that they would like to observe in the heat map. The numeric values on the slider represent the months in chronological order. For example setting the slider to seven will display the intensity and spatial distribution of temperature in the park during the month of July.

The second graph is a scatter plot of the relative humidity, by the Initial Spread Index, and colored and grouped by temperature. The temperature checkboxes in the control panel allow the user to change which groups they would like to plot. The there are seven groups ranging from 0-5 to >30 degrees celsius. Once the user has plotted the groups they would like to observe, they can look to the legend on the right to comprehend the temperature of a given point. The closer an observations temperature is to 0 degrees the the darker it’s plotted point will be, the higher the temperature the lighter the color. The user can select any number of groups to observe between zero and seven.

The last graph illustrates the distribution of the FFMC index for the day of the week that the user selects. To observe a specific day the user selects that day from the radio buttons in the control panel. With this tool the user can see if the day of the week effects the FFMC index. Below is the distribution of FFMC scores on Sundays as well the distribution of FFMC scores on Wednesdays.





Conclusions Regarding the Data

We struggle to conclude much about the data, except that this regression task needs very ad hoc nonlinear approaches to see an accurate result. We believe that cross validation instead of test/train splitting, a support vector machine, and model tuning would yield a much better result than we have seen. The most interesting exploratory aspect of the data to us was to see the heat map of fire by area over months, it was intriguing to see which parts of the park are most at risk during certain parts of the year. It warrants further investigation to overlay the heat map with physical phenomena to see if areas with more forest fires have more growth and less moisture, or otherwise what sort of physical aspects give rise to large fires.

This data-set has long plagued machine learning experts and we are not exempt from its vexing non linearity and overwhelming 0 response rate. We will leave it to the professionals to better tune neural networks to this task [1, 2, 3] .

References

1. Y. Safi, A. Bouroumi and A. Bouroumi, "A neural network approach for predicting forest fires," *2011 International Conference on Multimedia Computing and Systems*, Ouarzazate, 2011, pp. 1-5.
2. Onur Satir, Suha Berberoglu & Cenk Donmez (2016) Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem, Geomatics, Natural Hazards and Risk, 7:5, 1645-1658
3.Emre Aslan, Yunus & Korpeoglu, Ibrahim & Ulusoy, Özgür. (2012). A framework for use of wireless sensor networks in forest fire detection and monitoring. Computers, Environment and Urban Systems. 36. 614–625. 10.1016/j.compenvurbsys.2012.03.002.