Forest Fires: An Analysis of the Initial Spread Index

December 12, 2017

Abstract

1 Introduction

- overview
- why are fires important: fires are bad, but if we can stop them early, then they are less bad
- inference vs prediction (we are doing inference)
- why is inference better for this problem

  This paper is organized as follows. In Section 2 we give a summary of our reference paper and restate our goal in this context. In Section 3 we outline our analysis, including our initial exploration of the data, all feature engineering and variable selection, and our final modeling decisions. In Section 4, we apply our final model to our holdout set in order to make unbiased inferences. We conclude our project in Section 5 with a discussion of our project. Finally, relevant code may be found in the attached Appendix.

2 Background

- summary of reference paper
- how data were collected
- goal
- restate prediction problem

3 Modeling/Analysis

3.1 Exploratory Data Analysis

- initial ideas: what should matter, types of covariates (spatial, temporal, index)
- first plots
- identify problems (colinearity, skewness)

3.2 Feature Design

- transformations
- creations
- One interesting feature class in this problem is geo-spatial. The raw data contains X and Y coordinates corresponding to a grid that has been overlaid on the map of Monteshino Natural Park. Each coordinate ranges from 1 to 9, therefore there are 81 total boxes in the grid. Of course, the first attempts at capturing any geo-spatial signal involved looking at the raw X and Y coordinates, as well as their interaction. Unfortunately, many of these boxes were sparse and the 81 degrees of freedom necessary for the raw grid were detrimental to the modeling process. Instead, we designed several features based on these values. This resulted in three candidate features. First, we created new coordinates $X2$ and $Y2$ that were created from the following algorithm

    1. sum the first row, last row, first column, last column
    2. combine the row or column with the lowest sum with its neighboring row or column
    3. repeat steps 1 and 2 until every box contains at least 1% of the data

  The algorithm was written under the belief that your neighbors are most similar to you and therefore should be the first candidates when grouping spaces together. Applying this algorithm, we reduced the number of boxes from 81 to 12. We note that $Y2$ only has two levels, which is unsurprising since Monteshino Natural Park is wider than it is tall. The other two engineered features integrated outside information. For these, we found a topographical map of Monteshino Natural Park on Google Maps and overlaid the original $X - Y$ grid. The result is Picture ###. From this, we created *forest_ind*, which is a binary variable that takes the value 1 when the box is mostly covered in trees and 0 otherwise, and *grid_group*, which identifies five major mountain ranges and groups the boxes that cluster around these mountain ranges. Both of these variables are intuitively appealing, since obviously a forest fire needs trees and generally travelers hike and camp on a single mountain range.

3.3 Variable Selection

- FFMC
- LASSO
- tradeoff between interpretability for inference vs prediction
- creation of competing models - with and without FFMC
- The three engineered geo-spatial features from Section 3.2 were all considered as covariates. Unfortunately, none were selected during variable selection, as LASSO zeroed their coefficients are they were not significant under further scrutiny. However, we firmly believed that there should be signal from these geo-spatial features. Our next attempt was to fit a Mixed Model with *grid_group* or $X2 : Y2$ as the random intercepts. Again, this did not provide an improvement. Similarly, we tried weighting by the number of fires in *grid_group* and $X2 : Y2$ to no avail. We concluded that without further refinement or data collection, all of the geo-spatial signal was being picked up by the class of climate features, and therefore X, Y, and the derivative variables were discarded for our final model.

3.4 Modeling

- types of model we considered: weighted, mixed intercept, good ol lm
- someone please bootstrap something

- best 2 models
- model comparison
- our choice and why

4 Prediction

- apply best model to test set
- report diagnostics
- report p-values
- MSE or some loss function
- final inferences

5 Discussion

6 Appendix