Forest Fires

Byungkwon Han, Taylor Han, Samruddhi Hande, Nathan Orenstein

Background

The dataset examined is titled “Forest Fires,” and can be found in the UCI Machine Learning Repository.* The data consists of 517 samples (i.e, 517 individual forest fires), each measured by thirteen attributes:

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. Month - month of the year: 'jan' to 'dec'
4. Day - day of the week: 'mon' to 'sun'
5. FFMC - an indicator of the relative ease of ignition and flammability of fine fuel.
6. DMC - an indicator of fuel consumption in moderate duff layers and medium-size woody material.
7. DC - a useful indicator of seasonal drought effects on forest fuels and the amount of smoldering in deep duff layers and large logs
8. ISI - combination of the effects of wind and the FFMC on rate of spread without the influence of variable quantities of fuel
9. Temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. Wind - wind speed in km/h: 0.40 to 9.40
12. Rain - outside rain in mm/m2 : 0.0 to 6.4
13. Area - the burned area of the forest (in ha): 0.00 to 1090.84

The label for each fire is the burned area left behind, measured in Hectares (equivalent to 10,000 square meters or 107,639 square feet). Hence, the burned area is the output variable.

Research Question: Can the burned area resulting from each forest fire be predicted from each attribute?

*http://archive.ics.uci.edu/ml/datasets/Forest+Fires

Methodology

Linear regression was used to answer the research question as it enabled discovering if there are, if any, relationships between the variables and the output variable, i.e., to see if the amount of area burned from a forest fire could be predicted via a multitude of variables.

Before we start using the linear regression, we had to make sure the data is clear and solid to use.

1. Convert the categorical inputs to numerical inputs
2. Checked the heatmap for the correlation
   a. None of the variables have a clear correlation with the burned area
   b. Some variables have a correlation to one and another
3. Dropped all the rows with Area = 0
4. Assumed it is better to not scale the data for our Linear Regression
5. Dropped two rows with extreme outliers of Area Column
6. Checked if we can predict between variables that had a correlation
   a. They could be predicted with using the different order model
   b. High order models seemed not useful for our dataset
   c. Not directly related to answering the research question
   d. Just making sure our knowledge of Linear regression is correct

To apply the knowledge in our multivariate linear regression:

1. Used K-Fold Cross Validation to make sure every data points are tested
   a. Also, randomized these cross validation for better representation of regression
2. Used w =np.linalg.lstsq(A,y)[0] to find weights
3. Find the average of SSE_{train}, SSE_{test}, and Weights for each models
   a. Since we did 5 fold cross validation, we had to add each factors and divide by 5
4. Compare the SSE with M0, which is the null hypothesis model
   a. To check if models have lower SSE than M0
5. Visualize the result we have found

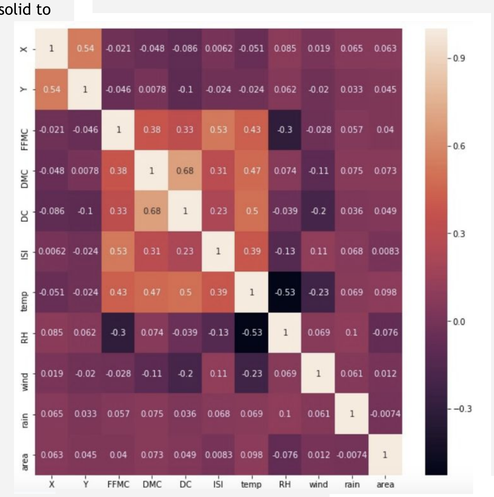


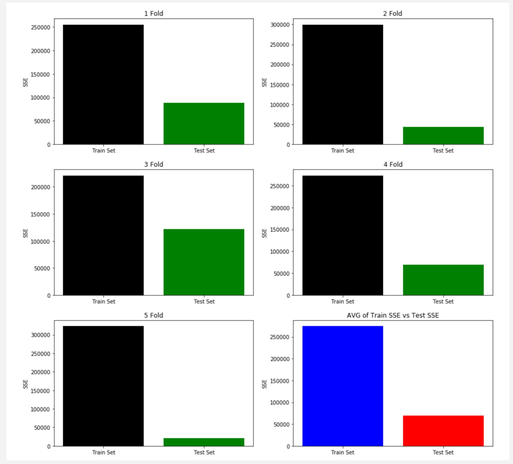
Fig 1. This heat map allows checking for correlation.

Models used:

- Model M0: Area = 17.93
- Model M1: Area = -36.92 + 0.62 * FFMC + 0.06 * DMC + -0.01 * DC + -1.5 *ISI + 0.54 * temp + -0.04 * RH + 0.67 * wind +-4.78 * rain
- Model M2: Area = 15.19 + 0.06 * DMC + -0.01 * DC
- Model M3: Area = 14.24 + 0.30 * RH + -0.05 * temp
- Model M4: Area = -1.28197212e-02 * FFMC*ISI + 7.09065252e-03 * FFMC*temp + -1.55554378e-05 * DMC * DC+ 3.54693718e-03 * DMC*temp + -7.25556199e-04 * DC*temp + 1.73331552e-03 * temp*RH

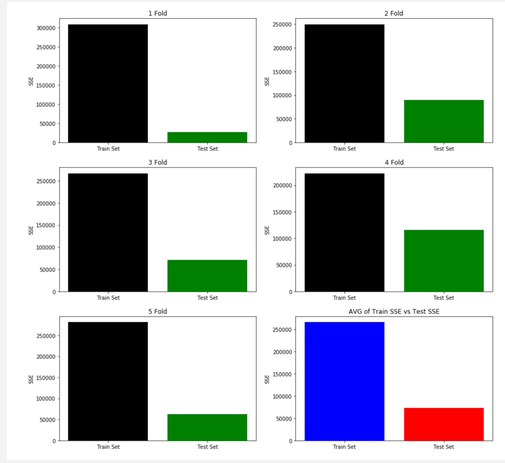
Results

Model 0: Null Hypothesis



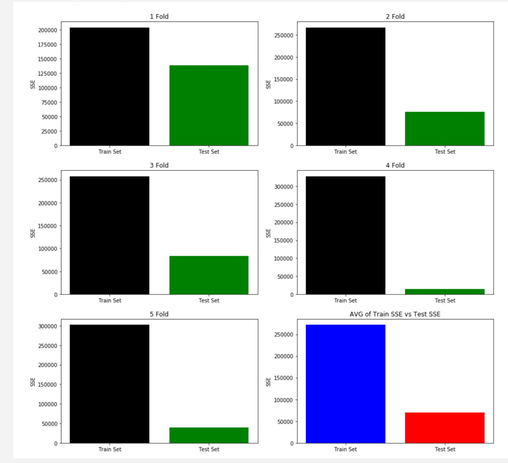
AVG Train SSE: 274366.157632 AVG Test SSE: 69044.7159938

Model 1: Using all variables to predict the burned area



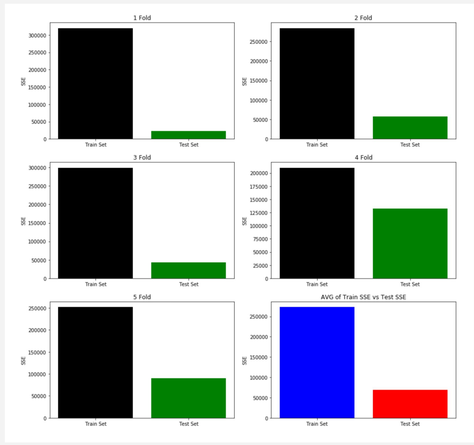
AVG Train SSE: 265724.336723 AVG Test SSE: 73258.4453942

Model 2: Using DMC and DC (highest positive correlation variables) to predict the burned area



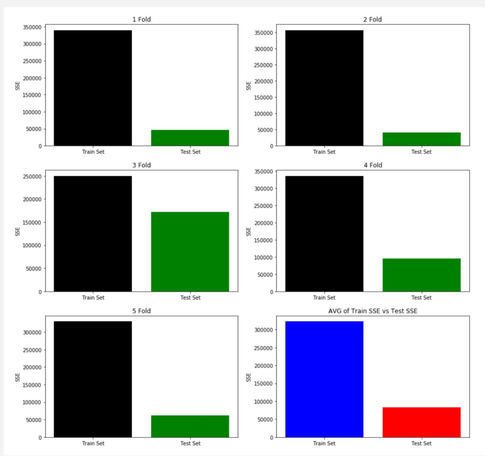
AVG Train SSE: 271690.167714 AVG Test SSE: 70530.5579636

Model 3: RH and temperature (highest negative correlation) to predict burned area.



AVG Train SSE: 272983.458753 AVG Test SSE: 69270.8423442

Model 4: Multiplying correlated variables to predict burned area.



AVG Train SSE: 322655.163145 AVG Test SSE: 83549.0283639

Interpretation of Results

Based on the pictured models, we are not able to show that the proposed models can predict the resulting burned area of a forest fire. This is because we found no such model that shows there is a smaller SSE_{test} than the null hypothesis SSE_{test}.

M4 was our experimental model with multiplication of correlated variables. As shown in the result, it had higher SSE_{test} than the null hypothesis SSE_{test}. Therefore, the model did not work at all.

Linear regression worked the best to answer the research question pursued using this dataset. It allowed for the examination of correlation between the output variable and the other variables (i.e., area burned vs each attribute). This worked better than K-means clustering would have, since that would have only shown differences amongst clusters of similar data points, instead of answering our research question.

Conclusion

The variables used in this supervised machine learning study seem to correlate based on their definitions. However, using multivariate linear regression, the hypothesis that burned area could be predicted from the various attributes was not able to be supported. The models above demonstrate a similar SSE to the null hypothesis SSE, which shows that the models are not predictive.

A linear regression model using a logarithmic function would have potentially worked, which the curator of the dataset suggests, but we were unable to properly implement such a model. If we were able to implement this model, we could have potentially rejected our null hypothesis, but this was not the case.