Predicting Fire Behavior Using Weather & Meteorological Data

Edwin Ramirez
University of the Pacific
School of Engineering and Computer Science
San Francisco, CA, 94103
e_ramirez23@u.pacific.edu

Darshil Desai
University of the Pacific
School of Engineering and Computer Science
San Francisco, CA, 94103
d_desai1@u.pacific.edu

***Abstract*—The abstract goes here. On multiple lines eventually.**

INTRODUCTION

Our regressional model will focus on utilizing previous forest fire and weather data sourced from the UCI Machine Learning Repository to predict the Initial Spread Index (ISI) of fires in Portugal and any countries using the Canadian Forest Wildfire Index System (CFWIS). We found this project to be of relevance considering the recent devastating droughts and fires in Northern Europe throughout Greece, Portugal, and Spain within the last year. Additionally, the relevance of the recent California fires also give support to the necessity of predictive analytics in this area. Hence, by utilizing a regressional model, we can predict the early behavior of a fire. Portugal utilizies the Canadian Forest Fire Weather Index System in order to track fuel moisture and wind speed to determine the intensity of a fire. Thus, by predicting the initial spread index of potential fires, we'd be able to predict the danger of a fire based on how quickly it would spread. The ISI scale begins at 0, where a 10 indicates a high rate of spread after ignition, and 16 or higher indicates an extreme rapid rate of spread.

The research question can be hypothesized as follows:

H_0 : None of the predictor variables in the dataset are useful in making predictions about the Initial Spread Index.

H_a: At least one of predictor variables in the dataset is useful in making predictions about the Initial Spread Index.

DATASET

View Dataset

The UCI dataset comprises of the following 13 variables:

- **X** : coordinate within the Montesinho park. Ranges from 1 to 9
- **Y** : coordinate within the Montesinho park. Ranges from 1 to 9
- **month**: month when the fire frst occured
- **day**: day of the given month when the fire occured
- **FFMC (Fine Fuel Moisture Code)** : a numeric rating of the moisture content of litter and other cured fine fuels. This code is an indicator of the relative ease of ignition and the flammability of fine fuel.
- **DMC (Duff Moisture Code)**: A numeric rating of the average moisture content of loosely compacted organic layers of moderate depth. This code gives an indication of fuel consumption in moderate duff layers and medium-size woody material.
- **Area**: area of forest burned
- **temperature**: temperature in Celsius
- **RH**: relative humidity
- **wind**: wind speed
- **rain**: cm of rain
- **DC (Drought Code)**: A numeric rating of the average moisture content of deep, compact organic layers. This code is a useful indicator of seasonal drought effects on forest fuels and the amount of smoldering in deep duff layers and large logs.
- **ISI (Initial Spread Index)**: expected rate of fire spread.This variable will be our response variable and we will try and establish a linear relationship between the myriad of weather and meterological factors of the forest experiencing fires and the future expected area burn. [^2]

METHODOLOGY

In order to draw a relationship between the various weather, meteorological variables in the dataset and the response variable (Initial Spread Index), it is vital to transform and preprocess our data to succesfully represent this data in a linear relationship.

Our methods to do the same involve the following:

1. Deleting redundant data that will fail to enhance the model's predictive power
2. Representing categorical data (months) as binary numerical data.
3. Transforming our response variable
4. Examine multicolinearity

Redundant Data

Several variables in the given dataset do not assist in creating the linear relationship between with the response variable (Initial Spread Index). Briefly:

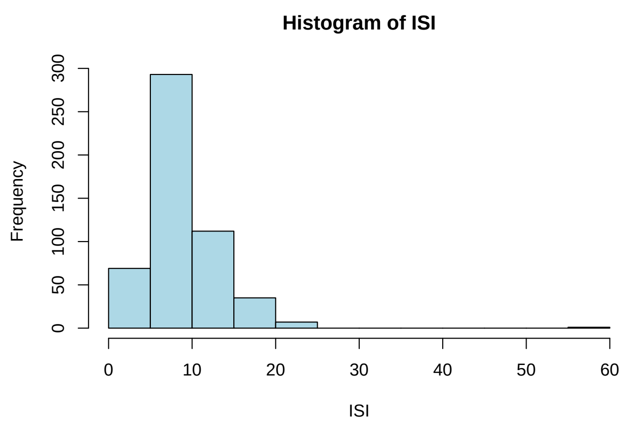
- **X,Y**: Coordinate data provided by the data can be considered redundant due to the coordintes being confined to the specific area of Montesinho Park in Portugal. This data can be removed because the goal of our model is to predict fires within countires that utilize the Canadian Fire Weather Index System.

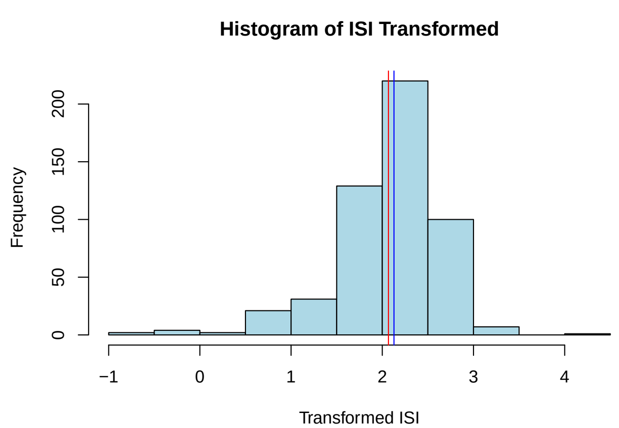
- **Days**: In isolation, the numerical value of a day does not provide any significant relevance to a datapoint

Categorical Data

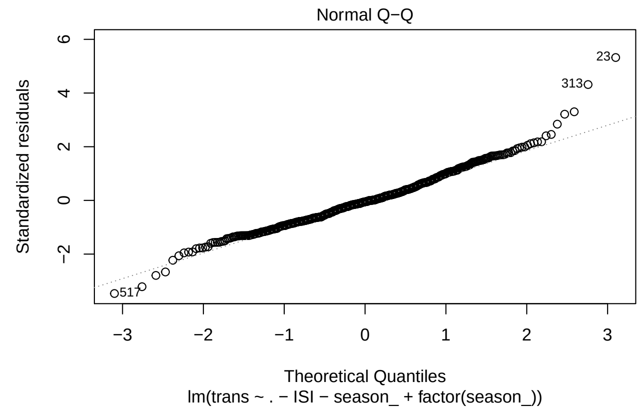
The variable of **months** can be grouped to create a categorical variable that signifies **seasons**. Thus, each season will be represented as nominal variables in our regressional model.

Normality the Response Variable





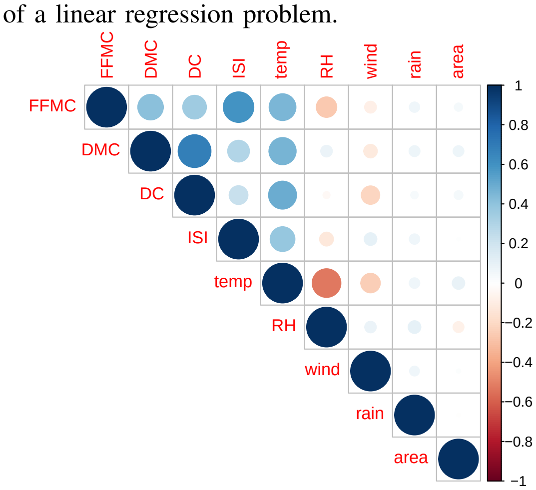
The response variable of ISI was originally right-skewed. To prevent a normality violation in the regressional model, the data is transformed by taking the natural log of all ISI data. A QQ-normal plot is shown below to illustrate that the residual data is normally distributed after being transformed.

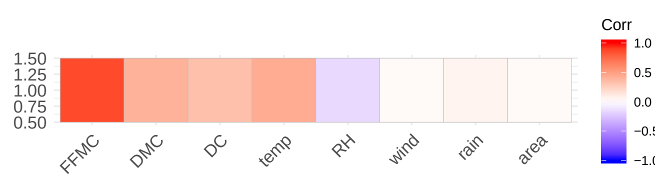


Examining Multicolinearity

Multicolinearity occurs when predictor variables are linearly correlated with the other. This implies that a change in any one of the predictor variables would entail a change in another highly correlated predictor variable.

Before we proceed to fit our model it is important to perform two vital checks. Firstly, there needs to be a reasonable correlation between the predictor variables and the response variable. An absence of any linear pattern does not warrant the use of a linear regression problem.





The visual analysis above shows us that there are several predictor variables that possess a significant correlation strength with the response variable (transformed ISI). Furthermore we also see that there exists no significant multicolinearity between the predictor variables. In general there always lies the possibility that variables within the same domain will correlate with one another to an extent.

However to further confirm our claim, we will employ the use of the Variance Inflation Factor analysis. This analysis allows us to support / reject our claim using numerical proof. It is important to note that the lower the VIF (lowest being 1), the less multicolinearity exists in our dataset. A VIF of 5 represents industry standard acceptance rate for multicolinearity as a small value indicates that the standard deviation of the respective variable parameter will remain relatively stable when other predictor variables are added into the regression equation.

```r
#Calculate the Variance Inflation Factor for each of the predictor variables
model <- lm(trans~ fire$FFMC + fire$DMC + fire$DC + fire$temp + fire$RH + fire$wind + fire$rai
#summary(model)
vif(model)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## fire$FFMC                1.459696  1        1.208179
## fire$DMC                 3.101415  1        1.761083
## fire$DC                  9.060482  1        3.010063
## fire$temp                3.730699  1        1.931502
## fire$RH                  2.198147  1        1.482615
## fire$wind                1.122068  1        1.059277
## fire$rain                1.058029  1        1.028609
## fire$area                1.023608  1        1.011735
## factor(fire$season_) 14.581963  3        1.563037
```

Based on the data above, it further supports that there exists no significant multicolinearity in the dataset. Most of the VIF values are under 2.5 far below the industry threshold of 5.

RESULTS

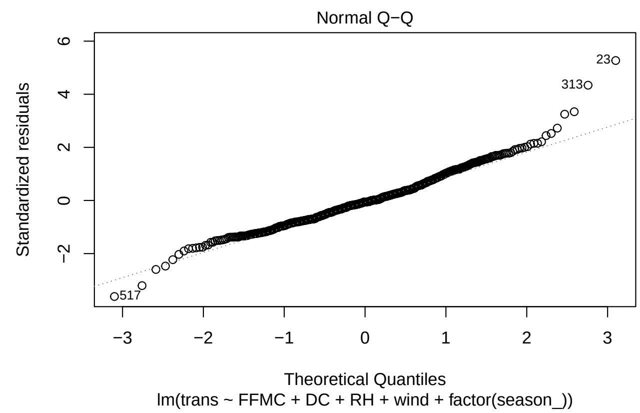
After analyzing the correlation of the predictor variables in the dataset, the next steps comprise the following: 1. Feature Selection

2. Model fitting using Training Data
3. Analysis

Feature Selection

Using stepwise regression, we iterate through different possibilities of a linear model and choose one with features yileding the best (lowest) error metric.

The QQ-normal plot below illustrates the distribution of our data after utilizing stepwise regression for feature selection.



Model Fitting

A portion of the data is separated to be utilized as a training set, while the remaining portion will be utilized as a test set.This will allow the accuracy of the model to be measured. Next, the training data is fitted to the model with the selected features. After removing features not selected from stepwise regression, the model includes `factor(season_)`,`FFMC`, `temp`, `RH`, and `wind`

Analysis

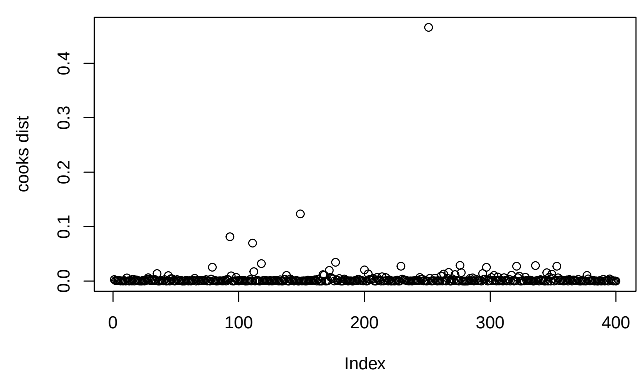
The following conclusions are the results of our model:

- The overall P-value of the model is <2.2e-16. Being far below the indsutry standard of 5% significance level, it can be concluded that at least one of the features is useful in predicting the response variable (ISI). Therefore the null hypothesis H_0 is rejected.
- The model yields an r-squared value of 0.76, indicating that it would likely be moderately useful in predicting the initial rate of spread of a fire.

Cook's Distance

After fitting the model, cook's distance is utilized to analyze if any influential data points affect the regression line. If any exist, this decreases the model's ability to generalize.

Based on the plot below, it can be confirmed that the training dataset one outlier point (>1). This outlier point is also influential in affecting the regression line of the model



Testing the Model

The model can be tested, and resulting mean squared error support our conclusion that the model is moderately useful.

Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank… Dr. Zarei

References