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*—This paper focuses on the predicting forest fire behavior using a variety of weather and meteorological data. The study incudes data from 517 forest fires in the Montesinho Natural Park in Northern Portugal from Janurary 2000 to December 2003. We look to draw a linear relationship between predictor variables (weather and meteorlogical data) and the Initial Spread Index (ISI) as our response variable. The Inital Spread Index, ranging from 0 to 56, refers to the rate of spread of a fire. Serving as an efficient measure of potential fire danger, a high ISI also indicates higher difficulty of fire control in a given region. In this paper we will test for multicolinearily between the predictor variables, examine outliers, transform data, and stepwise regression to select the most optimal features for a linear regression model.**

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). Previous studies have attempted to utilize the same dataset in order to create a regression model that could predict the potential area that would be burned in a fire, but these models proved to be unsucessful. Ultimately, 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 (1cm-4cm deep). 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 (5cm-10cm). 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 (10cm-20cm deep). 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:

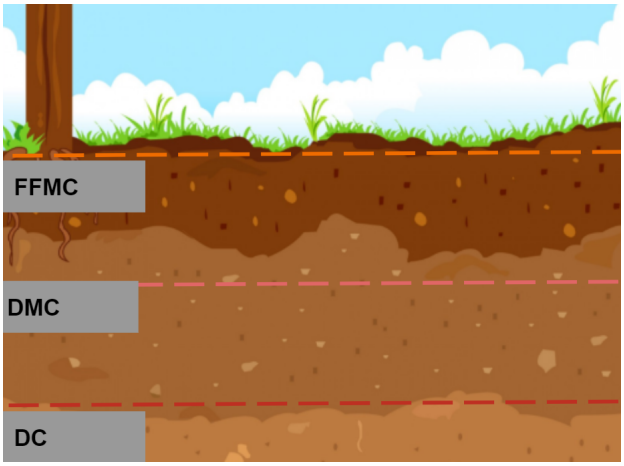


Fig. 1. Moisture Variables

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

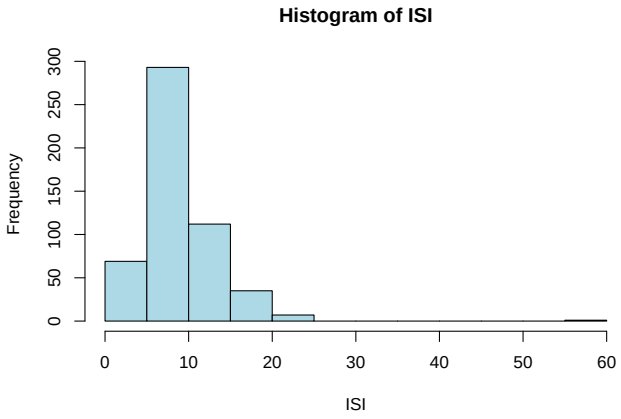
Several variables in the given dataset do not assist in creating the linear relationship between with the response variable (Initial Spread Index). Briefly:

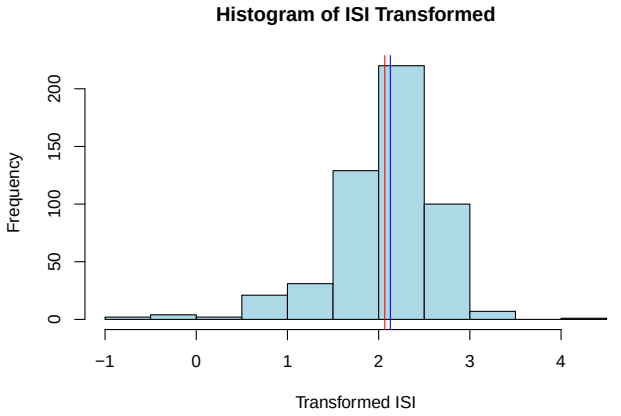
- **X**,**Y**: Coordinate data provided by the data can be considered redundant due to the coordintes being confined to the specific area of Montesinho Park in Portugal. This data can be removed because the goal of our model is to predict fires within countires that utilize the Canadian Fire Weather Index System.
- **Days**: In isolation, the numerical value of a day does not provide any significant relevance to a datapoint

Categorical Data

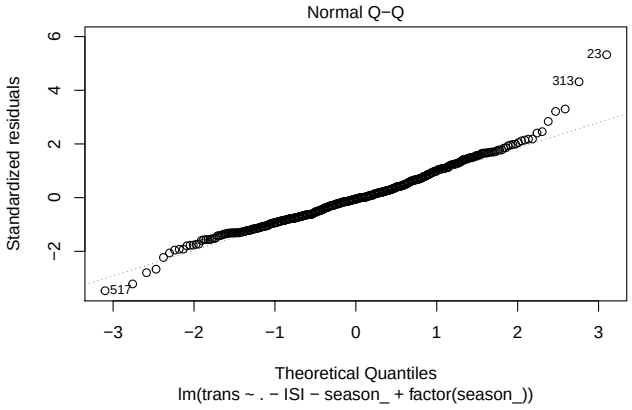
The variable of **months** can be grouped to create a categorical variable that signifies **seasons**. Thus, each season will be represented as nominal variables in our regressional model.

Normality the Response Variable





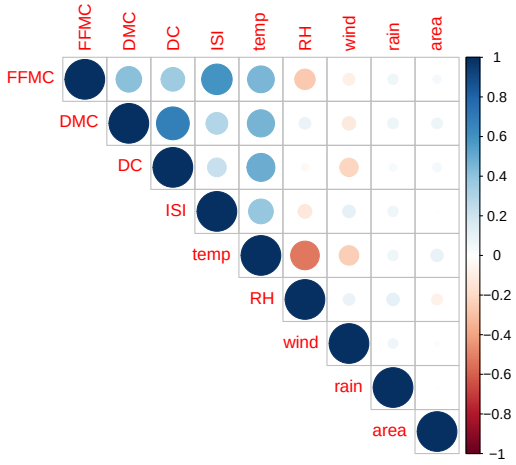
The response variable of ISI was originally right-skewed. To prevent a normality violation in the regressional model, the data is transformed by taking the natural log of all ISI data. A QQ-normal plot is shown in the top-right to illustrate that the residual data is normally distributed after being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

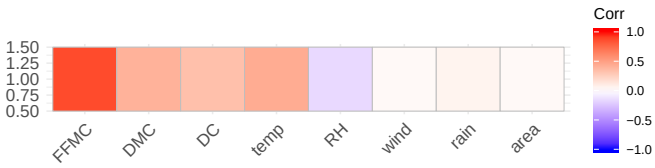


Examining Multicolinearity

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.

	GVIF	Df	GVIF^(1/(2*Df))
FFMC	1.459696	1	1.208179
DMC	3.101415	1	1.761083
DC	9.060482	1	3.010063
temp	3.730698	1	1.931502
RH	2.198147	1	1.482615
wind	1.122068	1	1.059277
rain	1.058029	1	1.028605
area	1.023608	1	1.011735
factor(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

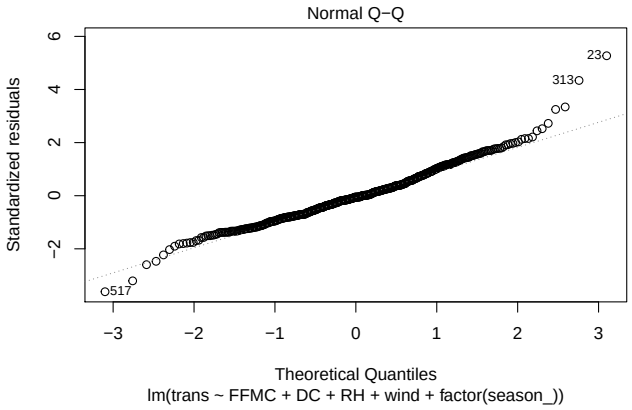
After analyzing the correlation of the predictor variables in the dataset, the next steps comprise the following:

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`. Therefore, our regressional model can be written as the following:

$$ISI = \beta_0 + \beta_1 season + \beta_2 FFMC + \beta_3 temp + \beta_4 RH + \beta_5 wind$$

Analysis

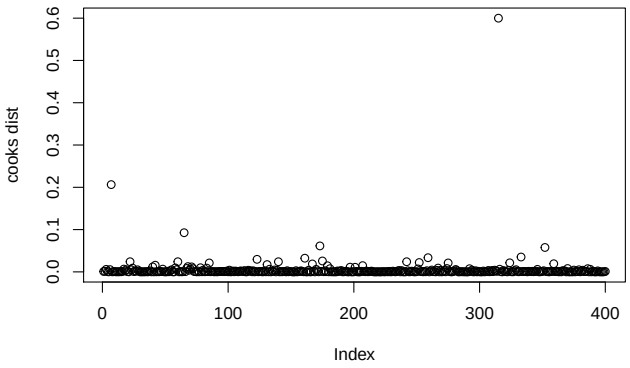
The following conclusions are the results of our model:

- 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 has no influential outlier points (>1), thus not influencing the regression line of the model.



Testing the Model

By testing the model, the resulting mean squared error supports our conclusion that the model is moderately useful. Because the mean squared error is 0.09 we can conclude that on average the actual Initial Spread Index, given the predictor variables in the test data, deviates from the true regression line by approximately the rate of 2.46 , in ISI units.

CONCLUSION

In conclusion our multiple linear regression model is successfully able to predict the Initial Spread Index given the various weather and meteorological predictor variables in the data. Therefore, we can reject the null hypothesis H_0 that none of the predictor variables would prove to be singificant. It is vital to point out that several other studies utilizing the same dataset focused their attention on predicting the area of the forest burned. Our study however deems the Initial Spread Index more suitable as it combines the effects of wind and the fine fuel moisture code. These two vital variables are present in any given forest environment, therefore strengthening the model and confirming its internal validity. All of our selected predictor variables occur naturally and completely independently from each other. This allows us to successfully predict the initial spread index of a fire to determine the rate of spread, and ultimately predict the behavior of fires.

If this model was to be used only regionally, (i.e. for forests in the general European ecosystems from where the data was collected) we would have also included predictor variables such as the forest coordinate data in order to take into consideration common forest areas that have previously caught on fire at a faster rate than the others. However, doing so would weaken the external validity of the model and not generalize well to forest ecosystems around the world that utilize the Canadian Fire Weather Index System. Hence aiming for this model to apply well in other settings, we have only considered weather and meteorological variables that are mutually present in all forests and grasslands.

Another opportunity to enhance our model's predictive power lies in expanding the time period the data was limited to. The dataset expanded to around three years time frame (January 2000 - December 2003). However, we believe the model would generalize better if we increased the size of the training data to account for previous years as well. Furthermore the analysis could then also extend to focusing on varying rates of Initial Spread Index over the years and examine its causes.

ACKNOWLEDGMENT

The authors would like to thank... Dr. Zarei

References

("Europe Heatwave: Spain and Portugal Struggle in 40C+ Temperatures" 2018) Cortez and Morais (2007) Han Shutting (1986)

Cortez, Paulo, and Anibal Morais. 2007. "Forest Fire Data Set." *UCI Machine Learning Repository*. https://archive.ics.uci.edu/ml/datasets/forest+fires.

"Europe Heatwave: Spain and Portugal Struggle in 40C+ Temperatures." 2018. *BBC News*. https://www.bbc.com/news/world-europe-45070498.

Han Shutting, Jin Jizhong, Han Yibin. 1986. "The Method for Calculating Forest Fire Behavior Index." *Fire Safety Journal - IAFSS* 19: 1–6. https://www.iafss.org/publications/aofst/1/77/view/aofst_1-77.pdf.