Predicting Algerian Forest Fires: Regression Analysis and Insights

Jessica Rumsey
Department of Statistics
Texas A&M University
College Station, Texas
jessica_rumsey@tamu.edu

I. Introduction

The Algerian Forest Fires dataset consists of 244 rows with half of the rows dedicated to the Bejaia region of Algeria and the remaining dedicated to the Sidi Bel-abbes regions of Algeria. Please note, for the following project, a subset of the Algerian Forest Fires dataset was used that contains only the 122 rows of information related to the Bejaia region. Tasks associated with this dataset are classification and regression.

The dataset measures daily weather conditions from June 2012 to September 2012, namely, temperature, relative humidity, wind speed, rain, fine fuel MC, duff MC, drought code, initial spread index, buildup index, and fire weather index. The variable type in the first three listed columns are integers while all others are continuous. An additional column contains the target classes (i.e. if a forest fire did or did not occur that day) and is the only column with a categorical variable. The web address to locate the Algerian Forest Fires dataset is listed here: https://archive.ics.uci.edu/dataset/547/algerian+forest+fires+dataset.

The objectives of this project are to learn of the identified factors, which are most related to each other and subsequently, are the most deterministic of a fire starting in the Bejaia region of Algeria. Given the nature of this dataset and its variables, multiple linear and logistic regression are performed to achieve these goals. Additional techniques are applied such as ridge and lasso regularization to best improve the models.

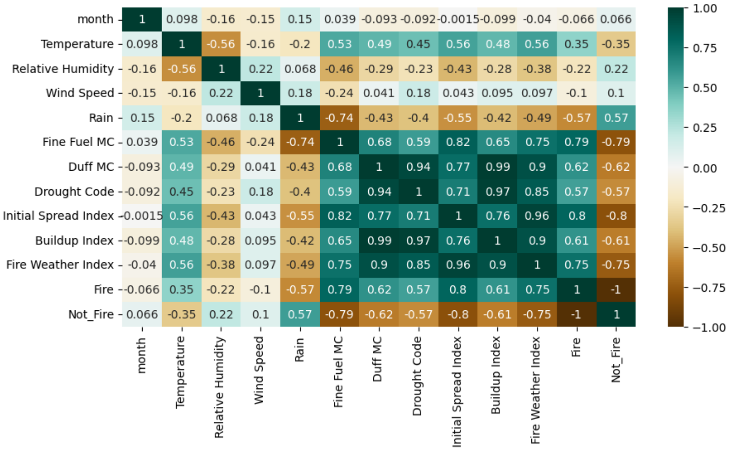
II. Methodology

A. Data Pre-processing

The subset of the Algerian Forest Fires dataset titled “Bejaia Region Dataset” is loaded into a Pandas DataFrame, and the columns are renamed to improve readability. The target classes column contains 59 instances of a fire and 63 instances of no fire occurring. One hot encoding is used to handle this single, nominal categorical variable and create two columns with binary entries for whether a fire occurred or did not occur. One hot encoding is selected to prevent falsely creating any relationships between categories.

B. Exploratory Data Analysis

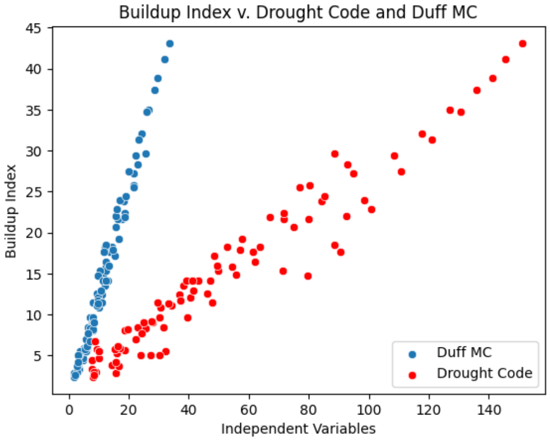
The dataset has no missing values or duplicate rows. Box plots are used to visualize outliers for each feature, and interquartile range calculations parse the data to exclude noticeable outlying values from these visualizations. The numeric columns, which include the two new columns from encoding, are used to formulate the correlation matrix shown below.



III. Experimental Results And Discussion

A. Hypothesis One

Motivated by the values in the correlation matrix, the first hypothesis addresses the impact of drought code (DC) and duff MC (DMC) on buildup index (BUI) where DC and DMC are the predictor variables, and BUI is the dependent variable. By definition, the BUI is a numeric rating of the total amount of fuel available to a fire and is a combination of DC and DMC. This definition explains the strong correlation of these features displayed in the matrix above. A scatterplot of these variables confirms a linear relationship between them.



The formal hypothesis is as follows.

Null Hypothesis (H0): Drought code and duff MC do not influence buildup index.

Alternative Hypothesis (H1): At least one of the variables (drought code or duff MC) significantly influences buildup index.

Multiple linear regression is an extension of simple linear regression to accommodate more than one predictor variable and a single dependent variable. Multiple linear regression relies on the data upholding assumptions of linearity, independence, homoscedasticity, normality, and no multicollinearity. Violation of any of these assumptions may hinder its effectiveness. The goal of multiple linear regression is to model the relationships between the independent and dependent variables and is done so with the equation below.

$$y = b + w_1 x_1 + w_2 x_2$$

The model is evaluated using mean squared error (MSE) and R-squared metrics. MSE is used to infer how much error the model makes in predictions while R-squared explains how much variance in the dependent variable can be explained by the independent variables. Multiple linear regression is selected to evaluate this hypothesis.

Relevant data for this model is split into training and testing sets with 30% of the data dedicated to testing. The model is fit on training data and evaluated on test data. MSE and the R-squared score (coefficient of determination) are calculated and return values of ~0.4699 and ~0.9950 respectively. Interpreting the R-squared score, approximately 99% of variance in BUI can be explained by DMC and DC. The weights applied to each predictor variable are 0.826 and 0.103 for DMC and DC. The corresponding equation is

$$y = -0.227 + 0.826x_1 + 0.103x_2$$

Therefore, the null hypothesis is rejected because the weight applied to DMC is close to 1 and hence, has a significant influence on the BUI measurement.

To confirm the accuracy of this model, assumptions of linearity, independence, homoscedasticity, normality, and multicollinearity are checked. The Variance Inflation Factor (VIF) is used to detect multicollinearity with the general guidelines that a VIF value greater than five implies high correlation. The data returns a VIF value for DMC and DC of 22.63. Thus, there is multicollinearity present in the data that violates the assumptions of the linear regression model.

To accommodate the assumption violation, ridge regression is applied to the data which passes a penalty term with L2 regularization to the loss function, effectively shrinking less important features. Ridge regression is originally selected over lasso regularization because, since lasso performs feature selection, it may drive the coefficient of either DMC or DC to zero. However, since BUI is a combination of both predictor variables, both should be maintained. The Scikit-Learn function, RidgeCV, performs ridge regression with built-in cross-validation of the alpha parameter. Interestingly, the best alpha value found through cross validation equals 0.0001. This means the ridge regression is almost identical to the multiple linear regression already performed. This is further confirmed by the same outputs for error metrics and coefficients. Therefore, ridge regression does not enhance the model, so lasso regularization is tested next.

Cross-validation is also used to find the best alpha parameter for lasso regularization. Lasso is applied and after calculating error metrics, proves to perform slightly worse than both multiple linear and ridge regression. The MSE increases to 0.555 and the R-squared score decreases in the thousandth decimal place.

Therefore, multiple linear regression with no regularization techniques adequately evaluates this hypothesis, and from the computed coefficients for each predictor variable, the null hypothesis is rejected.

B. *Hypothesis Two*

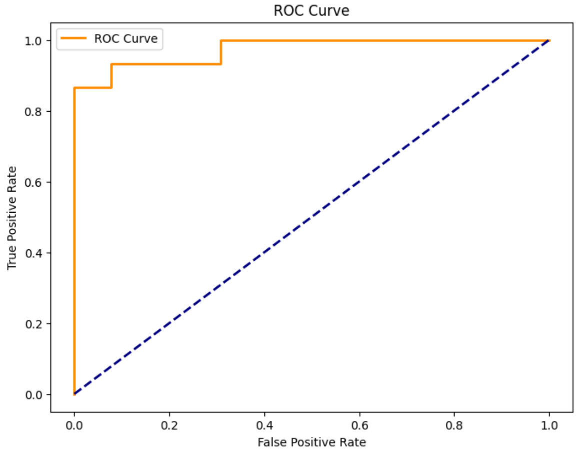
Understanding that some features are combinations of others such as DMC, DC, and BUI, we can infer not every feature is an important predictor in the occurrence of a forest fire. The next hypothesis addresses this and is stated below.

Null Hypothesis (H0): Every weather condition captured in the Bejaia Region Dataset does not affect the likelihood of a forest fire.

Alternative Hypothesis (H1): Every weather condition captured in the Bejaia Region Dataset significantly affects the likelihood of a forest fire.

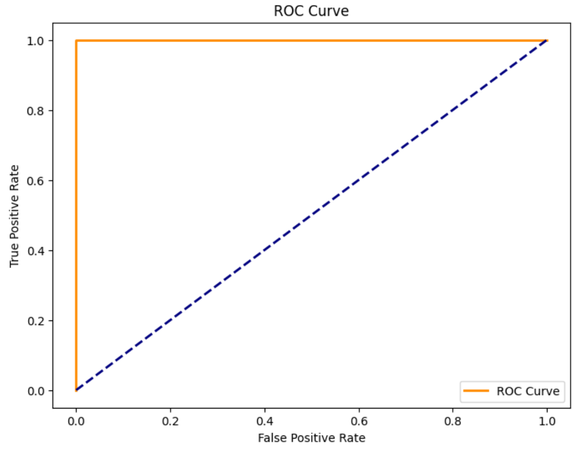
Logistic regression is used to evaluate this hypothesis and handle the binary classification of whether a forest fire did or did not occur. Logistic regression is used when the desired dependent variable is categorical. It utilizes the sigmoid function for application in classification methods to bound outputs between 0 and 1. Logistic regression is considered supervised learning. The goal is for the predicted class labels to be as close as possible to the true labels passed into the model. Logistic regression finds the optimal weights, or coefficients, that should be applied to each feature to most accurately perform classification.

The relevant data is once again split into testing and training sets with 30% of the data dedicated to testing. Then, StandardScaler() is applied to X in the training set to ensure all features are scaled the same. The logistic regression model is fit with the scaled training set, and predictions are made on the test set. The classification metrics are calculated for the model and return accuracy of 0.893, precision of 0.875, recall of 0.933, and F1-score of 0.903. The ROC curve is plotted and shown below. The dotted blue line represents a random classifier, and the orange line (the ROC curve) represents the model being evaluated. If the ROC curve is close to the top-left corner (0,1), the model can decently distinguish between positive and negative classes.



The area under the curve (AUC) is 0.947 where 1 represents a perfect model. However, this logistic regression is performed with every measured factor of the dataset included, and the hypothesis aims to discover which are the most important predictors. Therefore, lasso regularization is used to perform automatic feature selection.

The features selected after performing lasso regularization are printed and listed here: temperature, relative humidity, wind speed, fine fuel MC, duff MC, drought code, fire weather index. Notice, DMC and DC are included but BUI is not. Performing logistic regression again with only the features selected from lasso regularization, the error metrics claim this is a perfect model with accuracy, precision, recall, F1-score, and AUC of 1. The new ROC curve after regularization is below.



The coefficients output by the model for each feature are -0.424, 0.221, -0.125, 2.219, -0.303, 0.525, and 1.599 listed in the same order as their counterparts above. The coefficients applied to temperature, wind speed, and DMC are negative, so these features are inversely related to the event of a fire occurring. Additionally, fine fuel MC and fire weather index have the most weight applied to them.

The null hypothesis is accepted because lasso regularization proves not every weather condition is a significant indicator of whether a forest fire occurs, and a perfect model can be achieved with appropriately selected features.

To incorporate the results of the first hypothesis, logistic regression is performed a third time to compare how the model performs when BUI replaces DMC and DC as an independent variable. With this change in features from the optimal solution given by lasso, the model outputs an AUC of 0.971. In general, using both DMC and DC as opposed to only using BUI is a better metric for classification of whether a fire occurred.

Conclusion

The objectives of this project are met using multiple linear regression, logistic regression, and both L1 and L2 regularization techniques. Buildup index has a strong linear correlation with its predictor variables duff MC and drought code. Additionally, the binary classification of whether a fire occurs or not is dependent on a select few factors as opposed to every feature presented in the dataset. The most important factors for this classification are temperature, relative humidity, wind speed, fine fuel MC, duff MC, drought code, and fire weather index. The Bejaia Region dataset, taken as a subset of the Algerian Forest Fires dataset, accurately performs regression and classification tasks and provides valuable insights into predictions of these natural disasters. This study is limited to one region of Algeria, so cannot be accurately generalized to predict forest fires in other parts of the world. This may be reexamined in the future by compiling data from geographically diverse areas and seeing if the established relationships hold.

References

[1] Yoonsung Jung, “Lecture 28-Regression02-Multiple Linear.”
[2] Yoonsung Jung, “Lecture 30-Regression04-Lasso.”
[3] Yoonsung Jung, “Lecture 31-Regression05-Ridge.”
[4] Yoonsung Jung, “Lecture 33-Regression07-Logistic Regression.”