**Predicting Forest Fires with Data Science**

**Forest Fires Data Set**

(https://archive.ics.uci.edu/ml/datasets/Forest+Fires)

**Multiple Linear Regression using R and Python**

Tess Anderson, [tanderson4@bellarmine.edu](mailto:tanderson4@bellarmine.edu)

Jalin Roberts, [jroberts12@bellarmine.edu](mailto:student@bellarmine.edu)

**ABSTRACT**

Up to 150 word summary of your project.

1. **INTRODUCTION**

In this project, we use a dataset from the UCI Machine Learning Repository which contains information about forest fires. With this dataset, we perform analysis on the data using Exploratory Data Analysis to explore the variables of the dataset and their interactions with one another. Then, we will fit a machine learning Multiple Regression model in Python and analyze the fit of the data to the model for more insight into relationships between variables. The purpose of this analysis is to understand how other measurable factors are related to the size of the burning area during forest fires. The Multiple Regression Model that we used is imported from SciKit-Learn in Python.

1. **BACKGROUND**
   1. *Data Set Description*

The dataset is titled “Forest Fires Data Set” and it comes from the UCI Machine Learning Repository. More directly, their source is Paulo Cortez and Anibal Morais – both from the Department of Information Systems at the University of Minho in Portugal. The dataset has 12 variables containing meteorological information and 1 variable ‘area’ which is the burned area of the forest and the independent variable in this set. The meteorological information was collected to predict the size of the burned area so that researchers can understand why forest fires occur, as well as how they behave and spread.

* 1. *Machine Learning Model*

Multiple Linear Regression is a Machine Learning Model that provides large amounts of mathematical and statistical information about the relationships between multiple dependent variables and one independent variable. The idea behind Multiple Linear Regression is that each variable can be given a coefficient so that when the values are multiplied with the coefficients, the output of the equation will match the output of the real-world situation with those same conditions for variables. For categorical variables that can’t be simply multiplied by a coefficient, we transform them into binary variables having the value either 1 or 0, a 1 indicates that variable is true or present and a 0 indicates that it is false or not present. Then, they can be multiplied in the same way as numerical variables.

The coefficients are calculated by the Model during the *training* stage. During the *testing* stage, we use part of the original dataset that we didn’t use to train the model and we compare the actual results from the testing set with the predicted results given by the trained model.

1. **EXPLORATORY ANALYSIS**

Rather than providing tabular statistics and plots for each variable, provide only statistics and plots that seem unusual. For example, if one or two variables have significant missing values or the distribution of the variable is skewed or looks unusual note that. **Provide the unusual statistics or plots in this section. Provide any appropriate plots (e.g. correlation matrix, heatmaps, bar charts, etc.) that you deem necessary.**

This dataset contains 517 entries with 13 columns with various data types. There are no missing values.

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| X (x-axis spatial coordinate) | Int64 |
| Y (y-axis spatial coordinate) | Int64 |
| Month (jan to dec) | Object |
| Day (mon to sun) | Object |
| FFMC (Fine Fuel Moisture Code) | Float64 |
| DMC (Duff Moisture Code) | Float64 |
| DC (Drought Code) | Float64 |
| ISI (Initial Spread Index) | Float64 |
| Temp (in Celcius) | Float64 |
| RH (Relative Humidity) | Int64 |
| Wind | Float64 |
| Rain | Float64 |
| Area | Float64 |

1. **METHODS**
   1. *Data Preparation*

The data was already normalized when it was collected and published by Cortez and Morais. We didn’t decide to drop any variables because they are all relevant to the Independent variable – they are weather and fire index values that are important in predicting the area of the burned forest. We checked the dataset for null values in Python and there are none. There was little work necessary to prepare the data.

* 1. *Experimental Design*

We ran the model several times with different parameters to find the best fit for the data with the least error. We tried different ratios of splitting the data into training, testing, and validation sets. The table below shows all the different settings we tried for the parameters. The third setting provided the best results with the least errors – this makes sense as it uses the largest proportion of the data for training.

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | All thirteen (13) raw features with 80/10/10 split for train, validate, and test |
| 2 | All thirteen (13) raw features with 70/15/15 split for train, validate, and test |
| 3 | All thirteen (13) raw features with 90/5/5 split for train, validate, and test |

* 1. *Tools Used*

The following tools were used for this analysis: Python running in Anaconda environment with Pandas, NumPy, Matplotlib, Seaborn, and SciKit-Learn. R-Studio utilizing caTools and tidyverse.

1. **RESULTS**
   1. *Mean square Error and R-Square calculation*

The MSE for this dataset (using our model in R) is approximately 4238.69, and the R-Squared is 0.04914 (Adjusted R-Squared is –0.009336). The R-Squared value is not very indicitive of a good model. The MSE for our dataset (using our model in Python) is approximately 1226.85, and our R-Squared is –0.22, which is not indicative of a good model.

* 1. *Discussion of Results*

We believe that our best model was our model used in R-Studio, with a splitratio of 0.8, this gives the best MSE and R-Squared values.

* 1. *Problems Encountered*

No project goes perfectly smooth. Discuss any problems you had with obtaining the data, preparing the data, implementing the model, or evaluating the model. **It would be highly unusual to indicate that you had not problems.**

Obtaining and preparing the data was relatively easy, with almost no problems occurring along the process.

* 1. *Limitations of Implementation*

The limitations of the R model we used is it doesn’t tell us if our model is good or bad, also it does allow us to interpret or understand if our data and/or predictions are biased.

* 1. *Improvements/Future Work*

We would like to have more variety in our dataset, with more values to implement, change around as necessary. Most of our values are contingent on area as it is what is mainly affected. This dataset worked well for our predictions, however we would certainly use more variables (independent and dependant) in future projects.

1. **CONCLUSION**

Ultimately, we wanted to better understand how forest fires can be affected by other factors such as temperature, wind, rain, area, etc. The overall purpose of Multiple Linear Regression is to estimate the relationship between seeveral independent variables, and one dependent variable. What we are looking for is the association between our independent variable, and the outcome that holds our other variables constant. Some variables can have confounding results, so having Multiple Linear Regression allows us to take a more detailed look at the data, as apposed to simply finding the correlation.

**REFERENCES**

List any websites, books, articles, etc. that you found useful while you worked on this project. It is not necessary to cite the references in the paper unless you specifically mention it in the text.