IST 718: Big Data Analytics



Forest Fires Prediction System

By

GROUP 15

Aman Awana

Anjali Chintam

Zhiwei Wang

Under

Prof. Williamson

## Introduction

Forest fire is a crucial environmental issue because it can cause incalculable damage to the environment and economy while also risking human lives. For effective fire control and resource preparation, it is necessary to predict fire occurrences in advance and estimate the possible losses caused by fires. It is not possible to completely prevent forest fires from happening, but fast detection can definitely help to minimize damage. Traditional forest-fires detection methods usually involve satellite-based, infrared/smoke scanners and local sensors, which require high equipment installation and maintenance costs. In the study, we aim to use the machine learning approach to predict the forest fires occurrence and damaged area with meteorological data (temperature, rain, wind, etc.), which is easy to obtain.

**1.1 Data Overview**

The study examines machine learning approach to forest fires prediction using a specific dataset, which is acquired from the UCI Machine Learning Repository.[[1]](#footnote-0) The dataset contains forest fire occurrence, forest fire weather index (FWI) components in Montesinho Natural Park, which is located in the northeast region of Portugal. Weather observations were collected by Braganca Polytechnic Institute and integrated to the forest fire dataset. The park was divided into 81 distinct locations by placing a 9×9 grid onto the map of the park. There are a total of 517 records, which are collected from 2000 to 2007.

There are 13 different variables involved in our data set, varying between weather factors, geographic location, and the severity of the fire. The location of the fire was recorded on a grid map of Montesinho Park. X is the x-axis spatial coordinate within the park map, and Y is the same, but with the y-axis special coordinate. The month and day that the fire occurred is recorded as well. The Fine Fuel Moisture Code (FFMC) denotes the moisture content surface litter and influences ignition and fire spread. The Duff Moisture Code (DMC) and Drought Code (DC) represent the moisture content of shallow and deep organic layers, which also affect fire intensity. The Initial Spread Index (ISI) is a score that correlates with fire velocity spread. Certain weather factors were also recorded, including: temperature in Celsius degrees, relative humidity in percent, wind speed in kilometers per hour, and outside rain in millimeters to meters squared 30 minutes before the ignition of a wildfire. Lastly, the area of the burned section of the forest is recorded in hectares. The bulk of the data involving weather characteristics is pulled from the meteorological station database. The variables can also be summarized in the following table.

|  |  |
| --- | --- |
| Variable Name | Meaning |
| X | x-axis spatial coordinate within the Montesinho park map: 1 to 9 |
| Y | y-axis spatial coordinate within the Montesinho park map: 2 to 9 |
| month | Month of the year, a categorical/string variable from ‘jan’ to ‘dec’ |
| day | Day of the week, a categorical string variable from 'mon' to 'sun' |
| FFMC | Fine Fuel Moisture Code (numerical) |
| DMC | Duff Moisture Code (numerical) |
| DC | Drought Code (numerical) |
| ISI | Initial Spread Index (numerical) |
| temp | temperature in Celsius degree (numerical) |
| RH | relative humidity in % (numerical) |
| wind | wind speed in km/h (numerical) |
| rain | rain volume in mm (numerical) |
| area (output variable) | the burned area of the forest in hectare (numerical) |

It is worth noting that there is no missing value in the dataset.

**1.2 Prediction Task**

The study aims to find the best machine learning model to predict the intensity (burned area) of forest fires, what parameters should be, and what variables should be included in the model.

**1.3 Inference Task**

The study aims to perform two inference tasks. First, we explore which variables have the most significant impact on the variance of burned area in forest fires. Second, we explore which variables contribute to a higher risk or a lower risk of forest fires, or put in another way, which variables positively or negatively correlate to the burned area.

**1.4 Packages**

In addition to the pyspark, we also perform basic data visualization with numpy, pandas, seaborn, and matplotlib packages.

## Exploratory Data Analysis

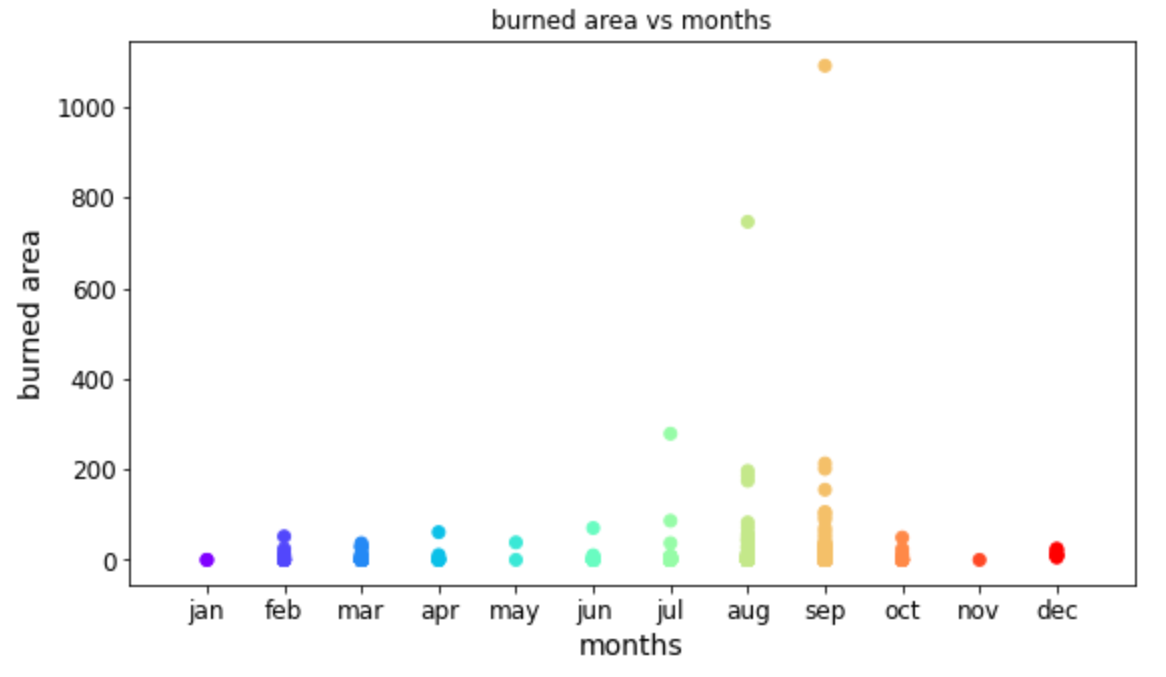
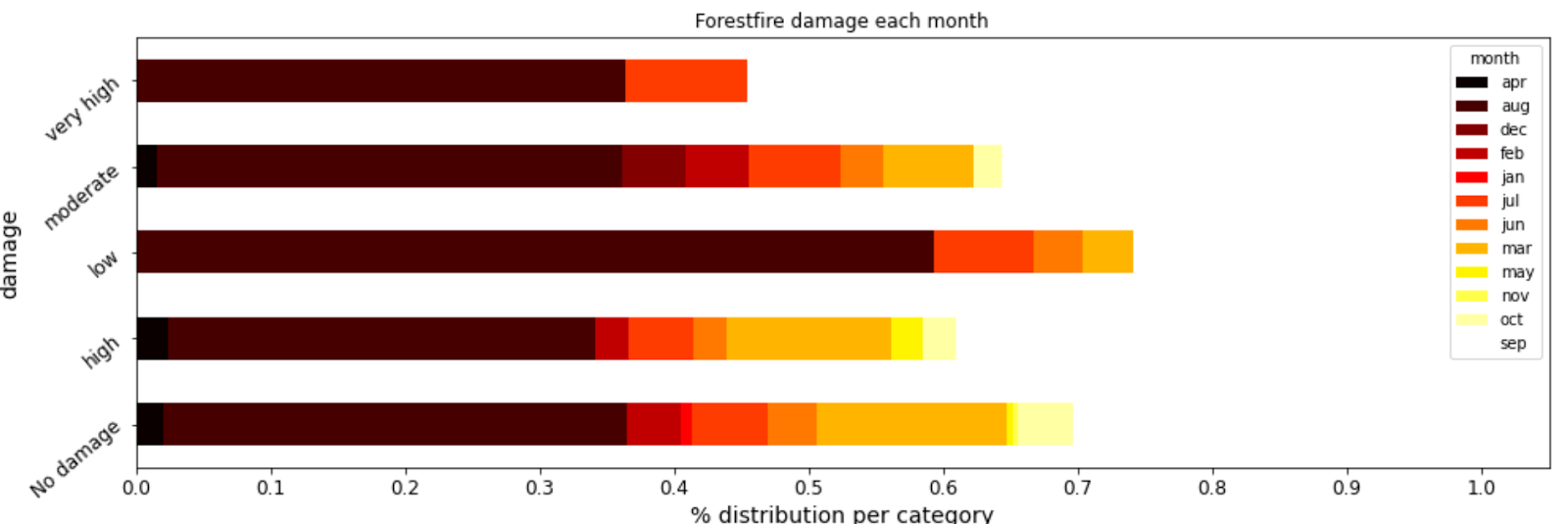
Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypotheses and to check assumptions with the help of summary statistics and graphical representations.

To use linear regression for modelling, it's necessary to remove correlated variables to improve your model.



Dark shades represent positive correlation while lighter shades represent negative correlation.

The correlation between ISI and FFMC is the highest and temperature being the second highest. Wind and rain are slightly correlated while relative humidity interestingly is negatively correlated. Even though RH is negatively correlated, we will still use it in our linear model.



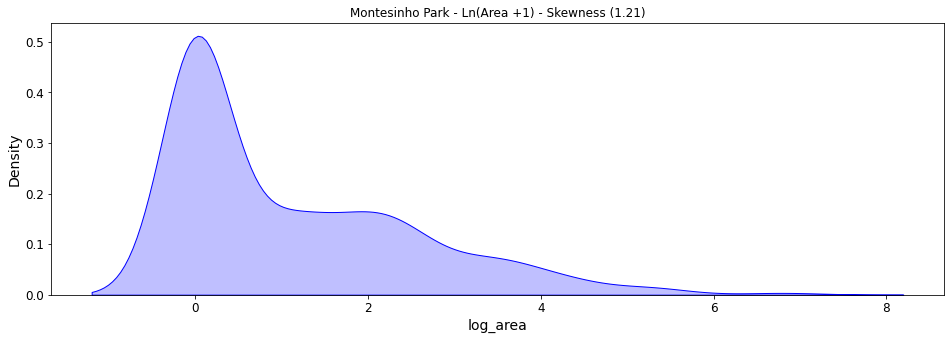
It is interesting to see that an abnormally high number of the forest fires occur in the month of August and September. And from the above plot of the month, we can understand a few things, most of the fires in August were low (< 1 hectare). The very high damages(>100 hectares) happened in only 3 months - August, July and September which is a typical summer in Portugal.

## Data Transformation

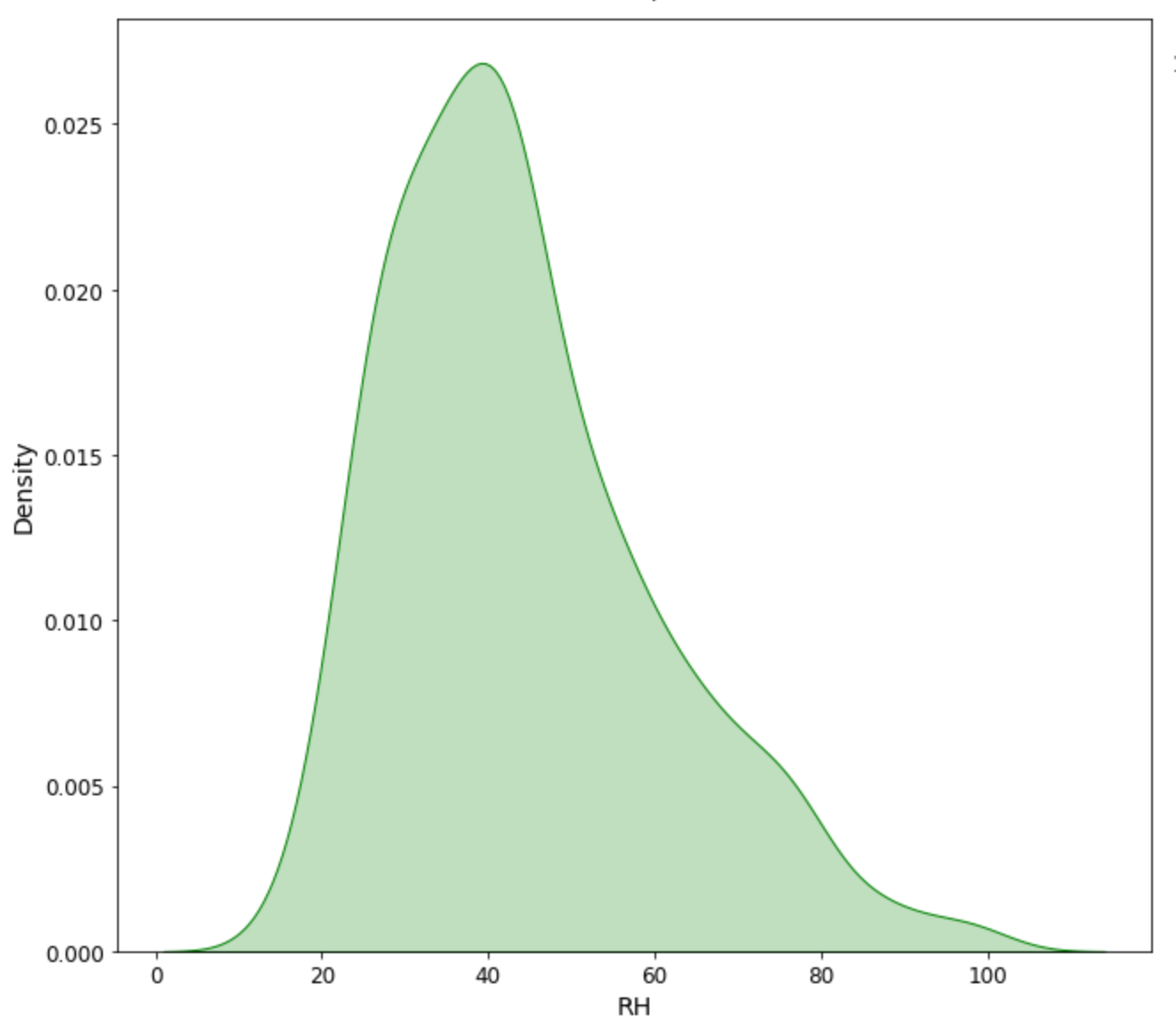
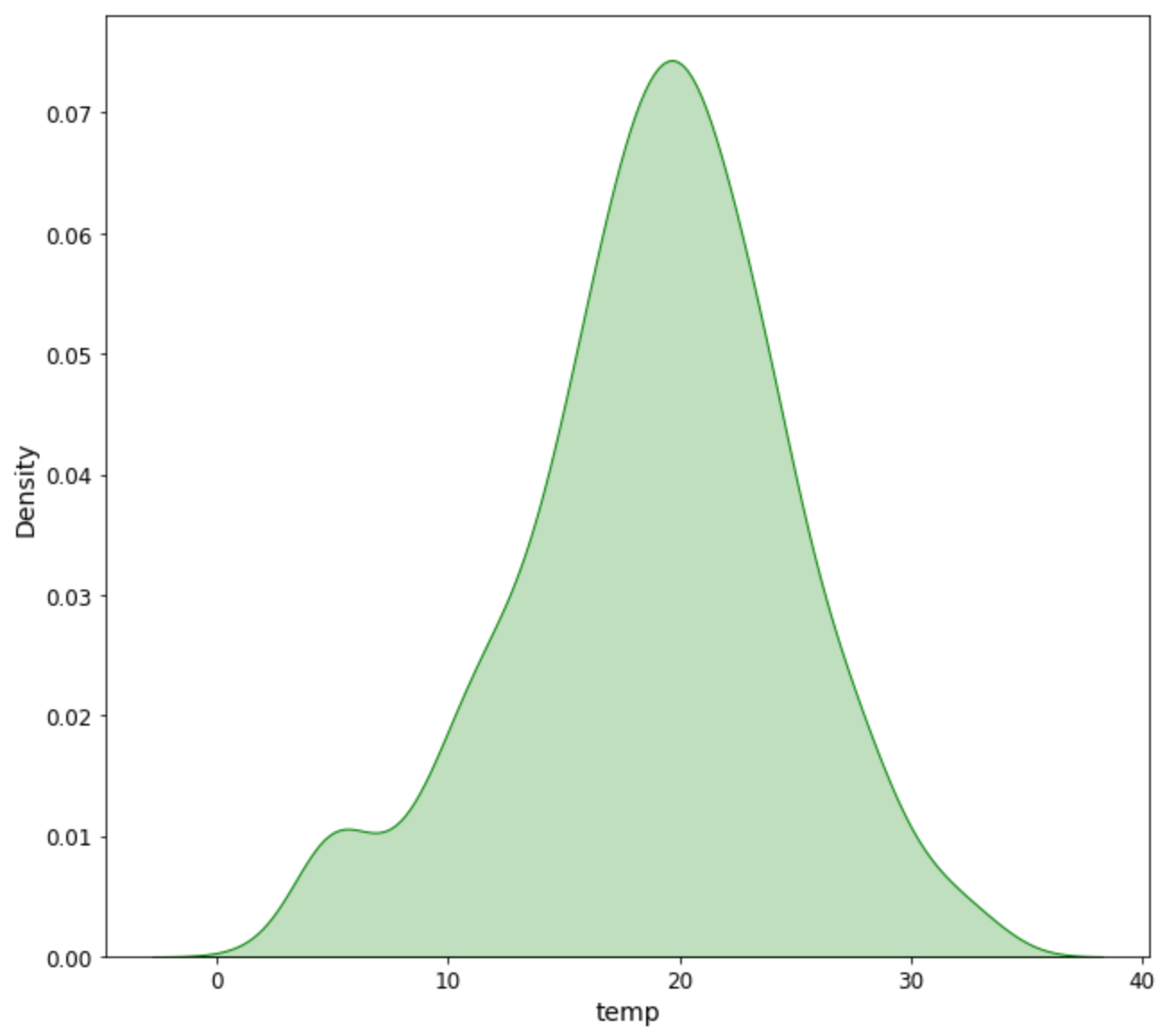
While the dataset contains no missing value (n.a.), it is still necessary to examine and transform the data to prepare for the model.

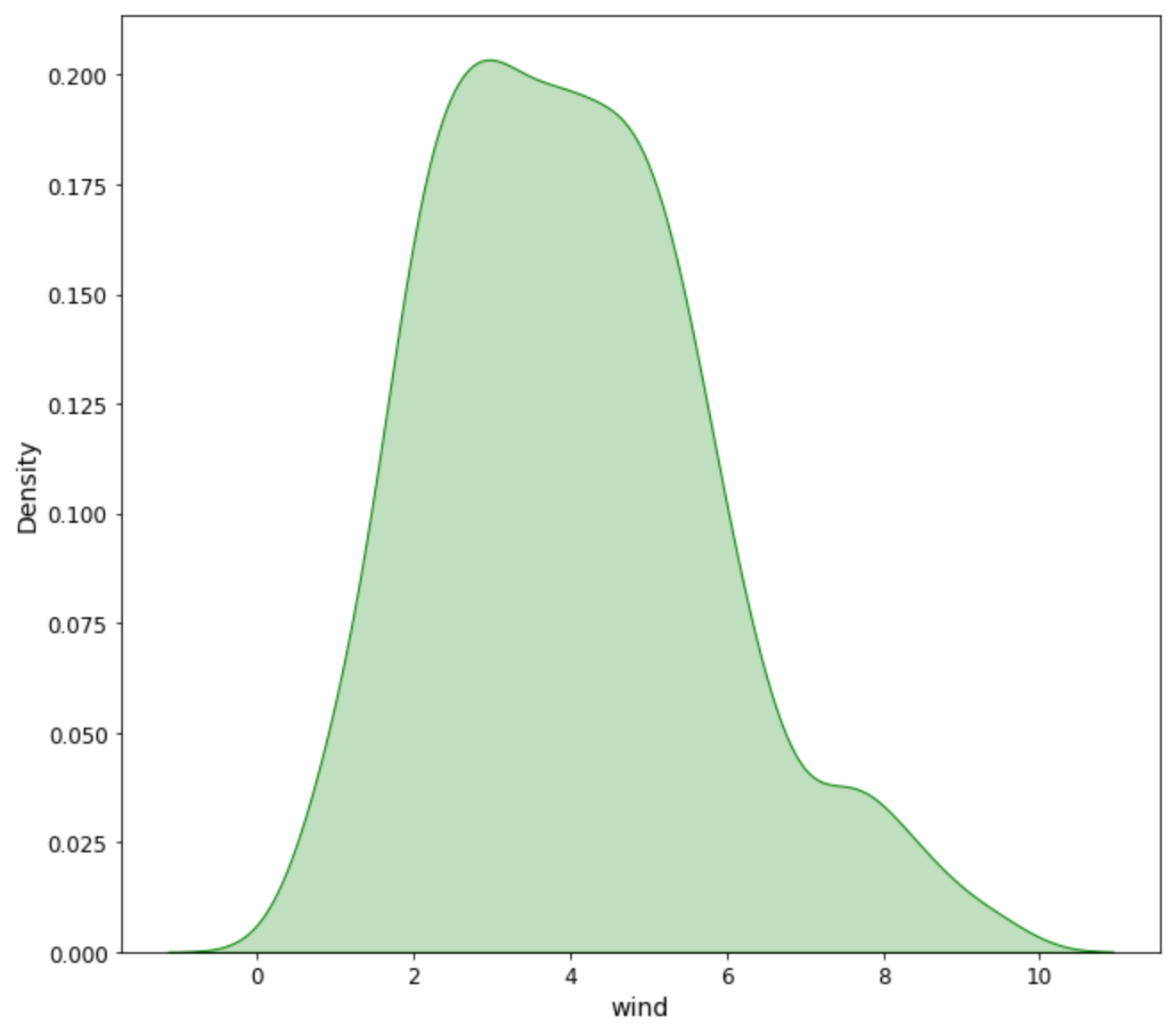
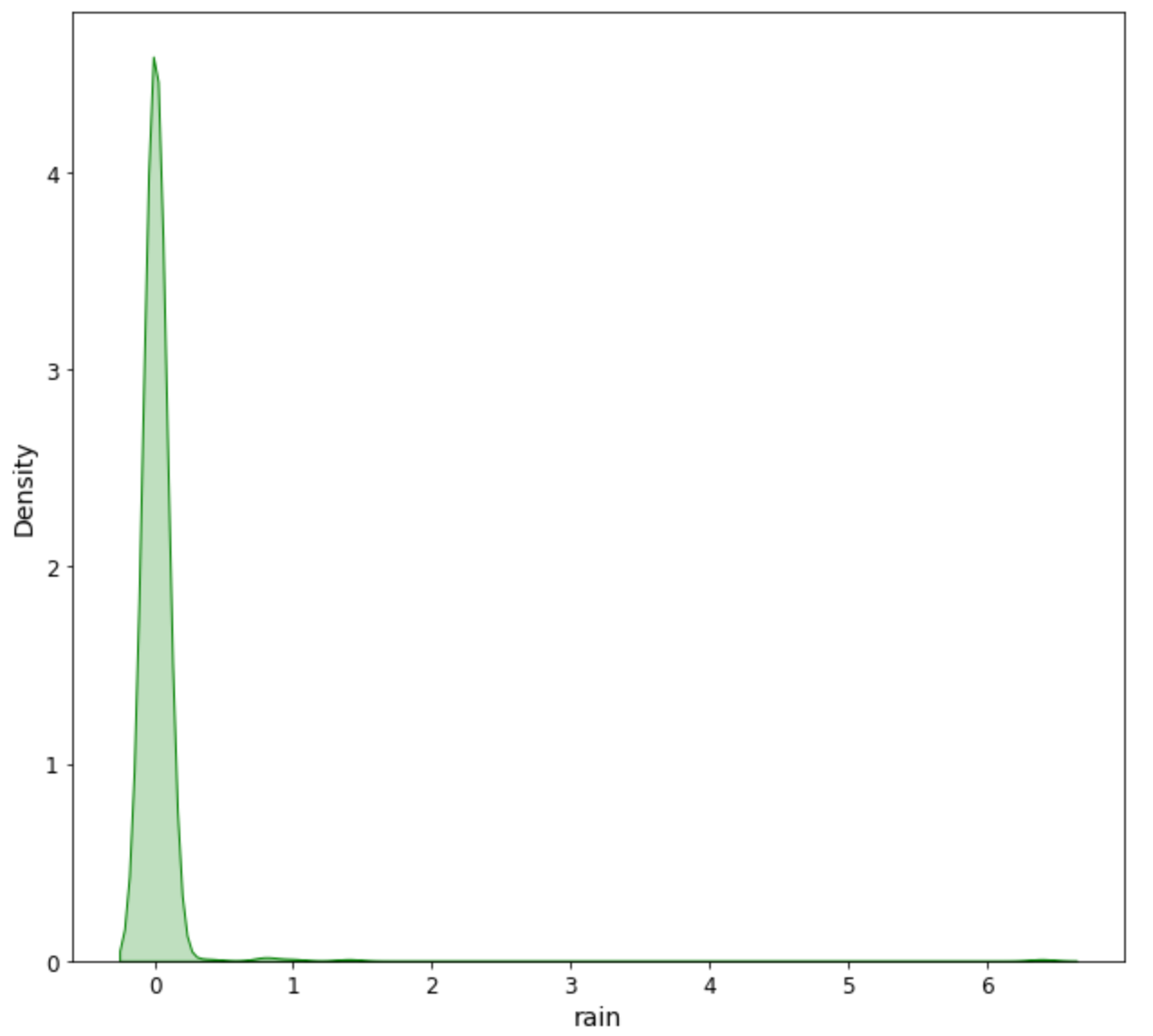
First, we need to solve the problem of skewness. As we can see from the distribution graph, the output variable ‘burned area’ is highly skewed towards 0. The reason for skewness is straightforward: there is no or very small-scale fire in most days, but once a forest fire starts, the burned area can grow into a very large scale. The skewness of data, especially the skewness at this scale, can have a significant impact on the prediction model. The data points that are outside of the normal range may make a disproportionate effect on the parameter estimates. In the study, we solve the skewness problem by performing log transformation of the variable ‘burned area’. We build a new variable ‘log\_area’ as the new target variable, which is the natural logarithm of (‘burned area’ +1). By performing the transformation, the skewness of the data sharply decreases from 12.81 to 1.21.





Second, we face the problem of outliers. As we can see from the graphs, three of the four predictors (‘temp’, ‘RH’, ‘wind’, and ‘rain’) have roughly bell-shaped distribution. The only exception is the ‘rain’ variable at the bottom-right. When we take a closer look at the ‘rain’ variable, we find that the skewness is a result of only one data point. There is only one observation that has a value of 6.4, while the rest of the data are below 1.5. Since there is only one outlier that we need to worry about, we decide to filter the outlier manually instead of performing any transformation. By removing the outlier, the skewness of ‘rain’ decreases from 19.76 to 11.48.

## Regression Models

The output variable in the study is the “burned area”, which is a numerical variable. Therefore, we need to build multiple regression models to make predictions. In the study, we select three models and compare their performance: linear regression, decision tree regression, and Gradient-boosted Trees (GBTs) regression.

We decide to keep only four predictors (‘rain’, ‘RH’, ‘temp’, ‘wind’) in our model for three reasons. First, we don’t include the geographical and temporal variables because we only aim to explore the relationship between forest fire intensity and weather. Second, we don’t include the weather index variables because we want to avoid the problem of collinearity. Variables like ‘FFMC’ and ‘DMC’ are actually secondary variables, which are the combination of basic weather variables like rain and temperature. Third, the model with only basic weather variables has a better interpretability.

**3.1 Linear Regression**

We start with the multi-linear regression model. We evaluate the performance through cross validation and optimize the hyperparameters with the grid search method. To increase the interpretability of the coefficients of the linear regression model, we also standardize the predictors using the standard scaler.

Best Hyperparameter:

|  |  |
| --- | --- |
| Hyperparameters | Optimized Value |
| regParam | 2.0 |
| elasticNetParam | 0.0 |

Model Performance:

|  |  |
| --- | --- |
| Evaluation Metrics | Value |
| Mean Squared Error (MSE) | 1.87 |
| Root Mean Squared Error (RMSE) | 1.37 |
| Mean Absolute Error (MAE) | 1.14 |

**3.2 Decision Tree Regression**

Decision tree regression is another machine learning algorithm for predicting numerical variables. We construct a decision tree with the method of recursive partitioning. The decision tree regression method is included in the pyspark machine learning package. Similarly, we evaluate the performance through cross validation and optimize the hyperparameters with the grid search method.

Best Hyperparameter:

|  |  |
| --- | --- |
| Hyperparameters | Optimized Value |
| maxDepth | 2 |
| maxBins | 26 |
| minInfoGains | 0.05 |

Model Performance:

|  |  |
| --- | --- |
| Evaluation Metrics | Value |
| Mean Squared Error (MSE) | 1.87 |
| Root Mean Squared Error (RMSE) | 1.37 |
| Mean Absolute Error (MAE) | 1.13 |

**3.3 Gradient-boosted Trees (GBTs) Regression**

Gradient Boosting Decision Tree is another regression algorithm that is included in the pyspark machine learning package. It is constructed as an ensemble of decision trees, which improve the models by combining weak learners.

Best Hyperparameter:

|  |  |
| --- | --- |
| Hyperparameters | Optimized Value |
| maxDepth | 2 |
| maxBins | 26 |
| minInfoGains | 0.05 |

Model Performance:

|  |  |
| --- | --- |
| Evaluation Metrics | Value |
| Mean Squared Error (MSE) | 1.87 |
| Root Mean Squared Error (RMSE) | 1.37 |
| Mean Absolute Error (MAE) | 1.13 |

1. Results

**Model Comparison**

When all three models are tuned in the best parameters, the performance is basically the same. The decision tree regression is slightly better, with the lowest mean absolute error. Nevertheless, in order to perform inference study, we select the linear regression model for the study.

|  |  |  |  |
| --- | --- | --- | --- |
|  | MSE | RMSE | MAE |
| Linear Regression | 1.87 | 1.37 | 1.14 |
| Decision Tree Regression | 1.87 | 1.37 | 1.13 |
| Gradient-boosted trees (GBTs) | 1.98 | 1.41 | 1.15 |

**Inference**

|  |  |
| --- | --- |
| Coefficients Value | Variable |
| 0.036 | temp (Temperature) |
| -0.054 | rh (Relative Humidity) |
| 0.020 | wind (Wind Speed) |
| -0.024 | rain |

Based on the coefficient value of the four predictors, we can draw some preliminary conclusions. First, the temperature and wind speed contribute positively to the size of burned area during a forest fire, while relative humidity and rain volume contribute negatively to the burned area. Second, among four predictors, relative humidity has the highest absolute value of the coefficient, which may suggest that relative humidity contributes the most to the variance of burned area, since all four predictors have been standardized in the model.

Nevertheless, it is also worth noting that the r squared values of all three models are a little bit disappointing because they are both slightly larger than 0 and smaller than 0.2. It means that the model built of weather variables as predictors can only explain a very small proportion of the variance of the occurrence and size of forest fires. In another word, even if we have the weather data, forest fire is still a comparatively random event to us. Therefore, we cannot replace the current detection method of forest fires with the machine learning method.

1. Future Work

First, we can transform the project into a classification problem. Currently, we build regression models to predict the size of forest fires. Since the ‘burned area’ variable has the value of 0 in around half of the records, we can also construct a categorical variable “fire damage” with two classes (“no damage” (area =0), “damage” (area >0)) or three classes ( “no damage” (area =0), “small damage (area >0 and area <=0.5”, and “large damage (area >0.5)”. We can try to solve this problem by using multiple classification algorithms.

Second, we can also perform clustering analysis of forest fires. In order to minimize the cost and time required for the firefighters to respond to fires, the forest department needs to identify where firefighting assets should be staged so that they are as close as possible. If we have enough data, we can partition the locations of past burns into clusters whose centroids can be used to optimally place heavy fire fighting equipment as near as possible to where fires are likely to occur. K-means clustering is an option for this problem.

Three, on a theoretical level, we can also dig deeper into the issue of evaluation metrics of the regression model. Normally, we use R squared to evaluate the predictability of a model. Nevertheless, in some cases (like in our project), the R squared value may not be very ideal (barely above zero in our project). Therefore, we need to consider what other metrics that we can use to evaluate the performance of a model.

# 

1. <https://archive.ics.uci.edu/ml/datasets/forest+fires> [↑](#footnote-ref-0)