Vishwakarma Government Engineering College

Data Science Mini Project

Project Name: **Forest fire prediction using meteorological data**

Branch: **Information Technology**

Semester: **5**

Team Members:

1. Vishvaa Chhatrara 180170116005
2. Priya Naika 180170116022
3. Urja Patel 170170116040

Forest Fire prediction using Meteorological Data

Objective:

Forest fires help in the natural cycle of woods' growth and replenishment. They Clear dead trees, leaves, and competing vegetation from the forest floor, so new plants can grow. Remove weak or disease-ridden trees, leaving more space and nutrients for stronger trees.

But when fires burn too hot and uncontrollable or when they're in the "wildland-urban interface" (the places where woodlands and homes or other developed areas meet), they can be damaging and life threatening.

In this kernel, our aim is to predict the burned area (area) of forest fires, in the northeast region of Portugal. Based on the the spatial, temporal, and weather variables where the fire is spotted.

This prediction can be used for calculating the forces sent to the incident and deciding the urgency of the situation.

Link of Dataset used: https://archive.ics.uci.edu/ml/datasets/forest+fires

Variable exists in dataset:

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: "Jan" to "Dec"
4. day - day of the week: "Mon" to "Sun"
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20 - Fine Fuel Moisture Code
6. DMC - DMC index from the FWI system: 1.1 to 291.3 - Duff Moisture Code
7. DC - DC index from the FWI system: 7.9 to 860.6 - Drought Code
8. ISI - ISI index from the FWI system: 0.0 to 56.10 - Initial Spread Index
9. temp - temperature in Celsius degrees: 2.2 to 33.30 - Temperature
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in hectare -- 1 ha = 10000 m2): 0.00 to 1090.84 (this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

Variables used from Dataset: All of the above mentioned.

List of independent variables: total = 29

1. *Already existing in dataset*: subtotal = 12
    a. X
    b. Y
    c. month
    d. day
    e. FFMC
    f. DMC
    g. DC
    h. ISI
    i. temp
    j. RH
    k. wind
    l. rain
2. *Added during execution*: subtotal = 17
    a. month_jan
    b. month_feb
    c. month_mar
    d. .....
    e. month_dec
    f. day_sun
    g. day_mon
    h. .....
    i. day_sat

List of dependent variables: total = 1

` 1. area

Model used:

1. Statistical Model for linear regression
2. Machine Learning Model for Linear Regression

Accuracy of the models:

1. Statistical Model

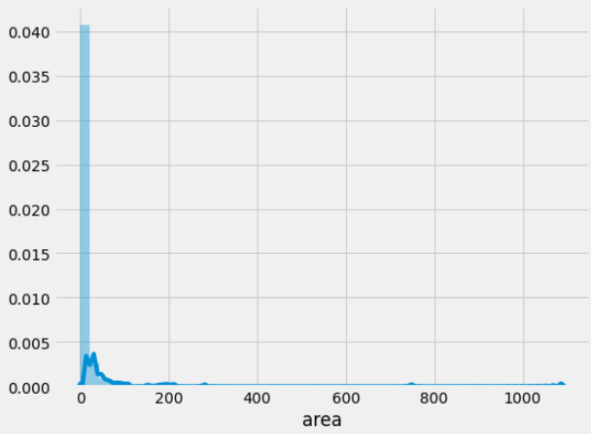
    R^2 score: 0.437

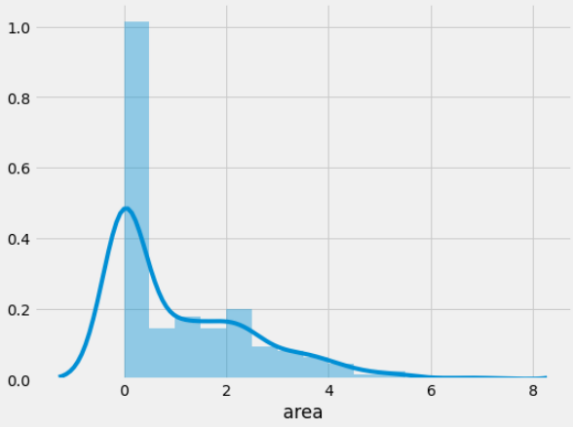
2. Machine Learning Model

    R^2 score: 0.05472

Steps implemented with little description:

1. Import basic dependencies that are useful in visualizing the dataset.
2. Then we import the dataset. The target variable in our dataset is 'area' which is distributed in following graph. To get more sense, we have taken a log and 2nd graph is for that.

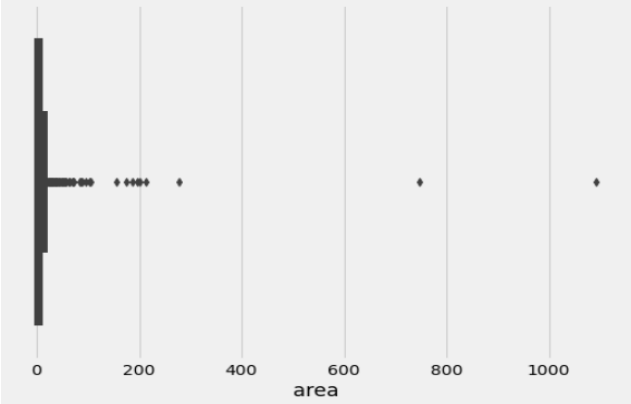




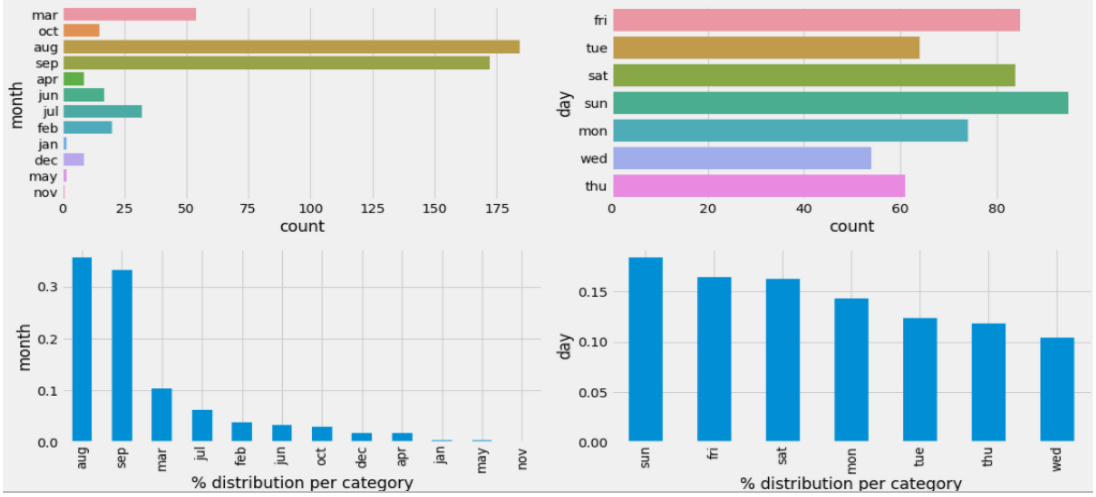
3. We looked at the correlation between attributes and passed our dataset through missing value treatment.
4. In this step, we have done 3 type of exploratory data analysis.
    a. Univariate
    b. Bivariate
    c. Multivariate

    Some of the result of data analysis are as follows:

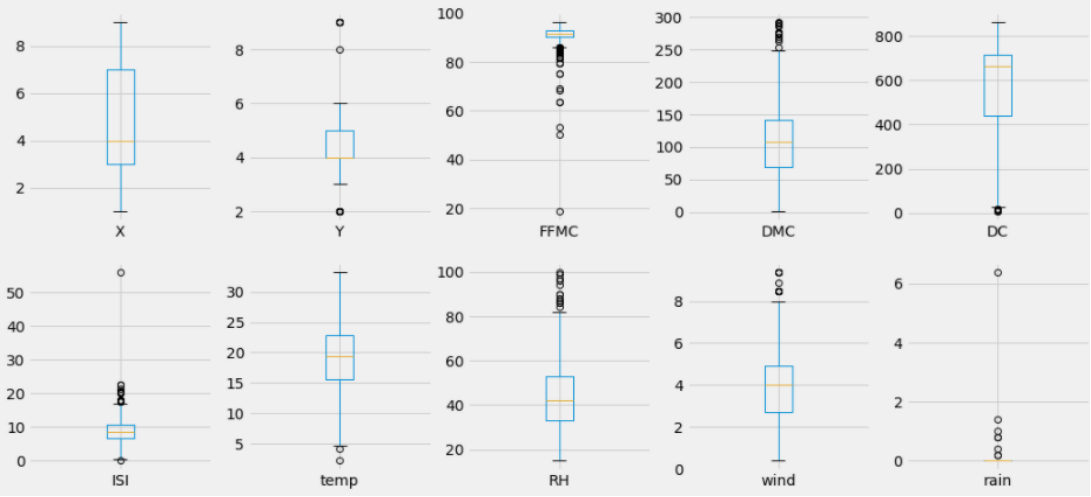
    a. Area: Univariate Analysis



b. Categorical Columns: Univariate Analysis



c. Numeric columns: Univariate Analysis



Some observations from above graphs are:

- Data of area is skewed with the high kurtosis of 194 and skewness of +12.84
- Most of the damaged area in fire is under 50 hectares land.
- There are 4 outliers in our area column but it will be decided whether to drop them or not.
- It is interesting to see that abnormally high number of the forest fires occur in the month of August and September.
- In the case of day, the days Friday to Monday have higher proportion of cases. (However, no strong indicators)
- Outliers, Skewness and kurtosis (high positive or negative) was observed in the following columns:
    - FFMC
    - ISI
    - rain
- 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.

5. This is the data modelling preparation step. Here we will add convert categorical columns into numeric using dummy variables and will add some columns corresponding to it.

   Also, we will perform some z-score and log transformation on the columns having very high skewness and kurtosis.

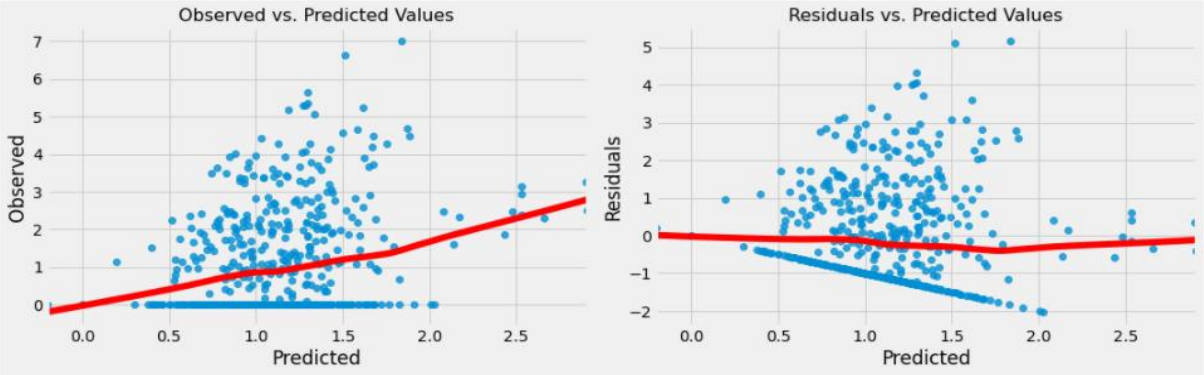
6. Statistical modelling

   We need to check 5 assumptions of linear regression, while going through statistical approach.

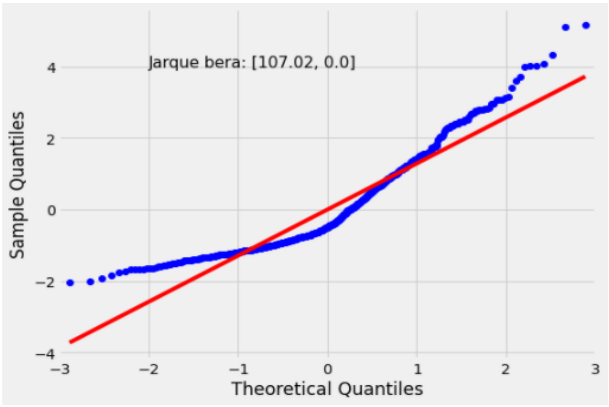
   a. Linearity of the model
   b. Normality of the model
   c. Homoscedasticity
   d. Autocorrelation
   e. Multicollinearity

   Our model is created using Ordinary Linear regression model of statsmodel.api. After performing various tests like. We get following results:

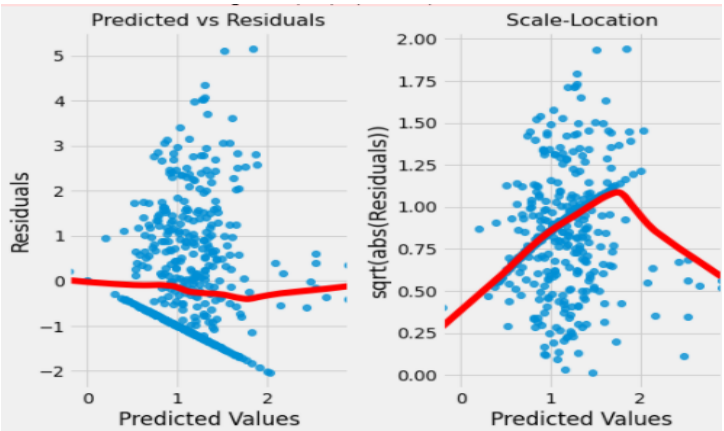
   - Rainbow Test: (fstat = 1.2832659, pvalue = 0.02704889)



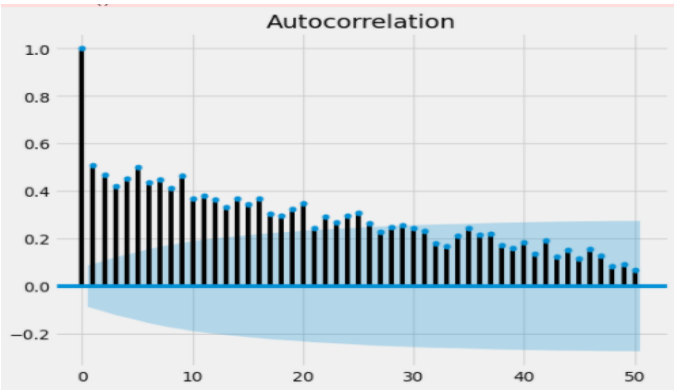
   - Jarque-Bera: (107.018, pvalue= 0.0)



- Goldfeld-Quandant test: (fvalue = 0.900533, pvalue = 0.7860123, result = 'increasing')



- Durbin-Watson Test: (0.979)



Observation form statistical learning is that our model does not preserve Linearity and Normality. Homoscedasticity, Autocorrelation and Multicollinearity are present in the model.

7. Machine learning modelling

   Here, we are using LinearRegression model of sklearn library. After fitting our x and y dataset -> training features and training result into the model, we get R^2 score as 0.076986153382917

8. Now we need to Improve our Stat and ML model. To improve stat model, we will check the variance inflation factor of all the columns and will discard a column having high vif one by one and check accuracy continuously. At one point we can see the clear improvement in R^2 and other factors. Our older model was giving R^2 score of 0.077 which got improved later to 0.437.

9. To improve our ML model, we go through various Feature Selection techniques. The resultant R^2 score of these are:
   a. Recursive Feature Elimination: 0.05542824436801641
   b. Forward Selection: 0.05472198851070975
   c. Backword Selection: 0.01821264603161099
   d. Ridge Regularization: 0.06055519668815157

10. From all of the above steps and output, we come to following conclusion:
    a. In our model we are using around 10 to 15 features. And though we can predict the area of burn, the accuracy of model is not sufficient.
    b. We can continue to optimize our model and discard or add (vegetation & firefighting intervention) some features.
    c. Nevertheless, our model is still useful in resource management.
    d. This offline learning, technique was applied after data was collected.
    e. Indeed, in the future we intend to test the proposed approach by using an on-line learning environment
    f. feedback from the firefighting managers.
    g. Another interesting possibility would be the use of weather forecasts, in order to build proactive responses.
    h. Finally, since large fires are rare events, outlier detection techniques.