Predicting Forest Fires

William Hahn

23 April 2025

Forest Fire Background Info

Forest fires have long been a major concern, especially in recent years in the drier regions of the United States. The impacts of these fires can be quite devastating because they can be very difficult to predict. Therefore, in this project, we will be creating 4 different types of prediction models (a tool that makes predictions from data): linear model, random forest model, decision tree model, and support vector machine model. We will compare the accuracy of each model to see which model is the best to use when predicting forest fires. We will also be using meteorological data from the data set below to create and test those models. The goal of this project is to create an accurate prediction model that can also be improved upon by others who are interested in predicting forest fires.

Table 1: First 6 Rows of Forest Fire Dataset

X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0
7	4	oct	tue	90.6	35.4	669.1	6.7	18.0	33	0.9	0.0	0
7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0.0	0
8	6	mar	fri	91.7	33.3	77.5	9.0	8.3	97	4.0	0.2	0
8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0
8	6	aug	sun	92.3	85.3	488.0	14.7	22.2	29	5.4	0.0	0

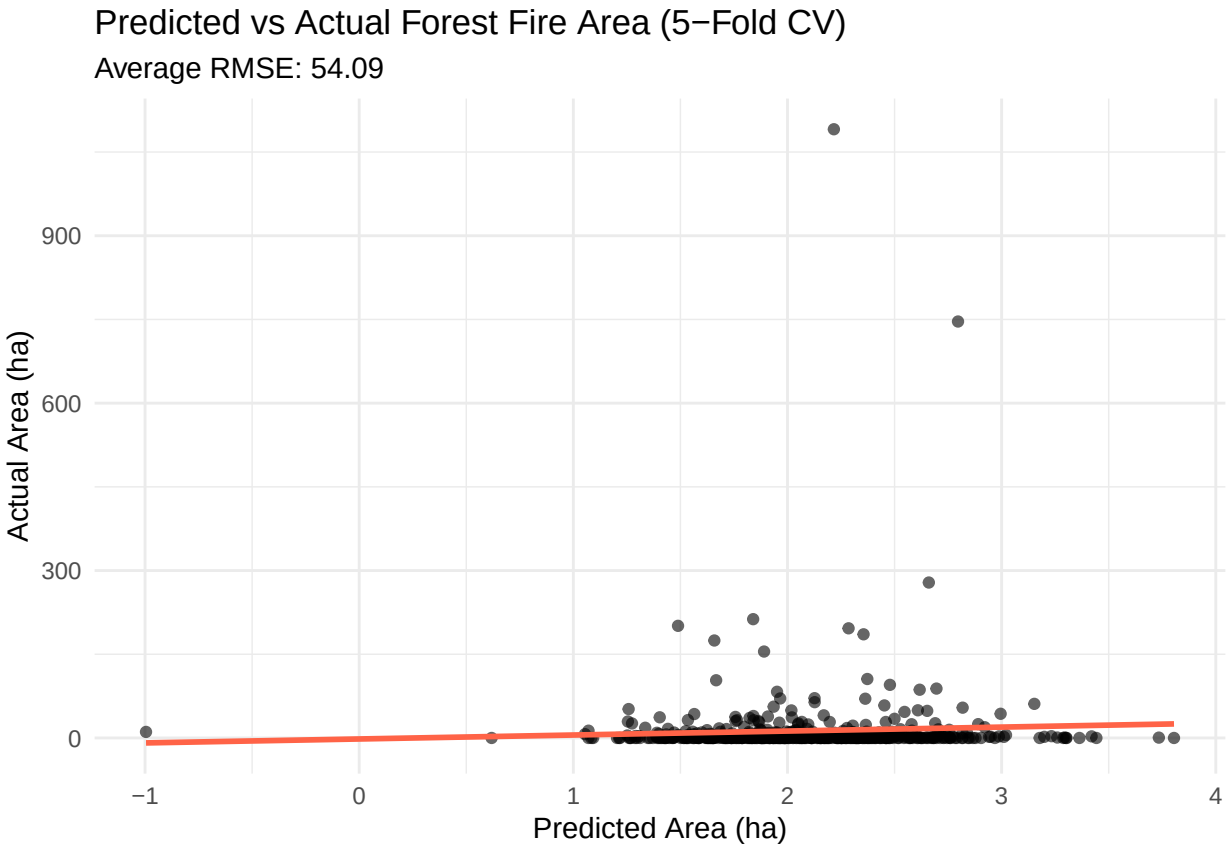
Explaining the Data Set/Methods Used

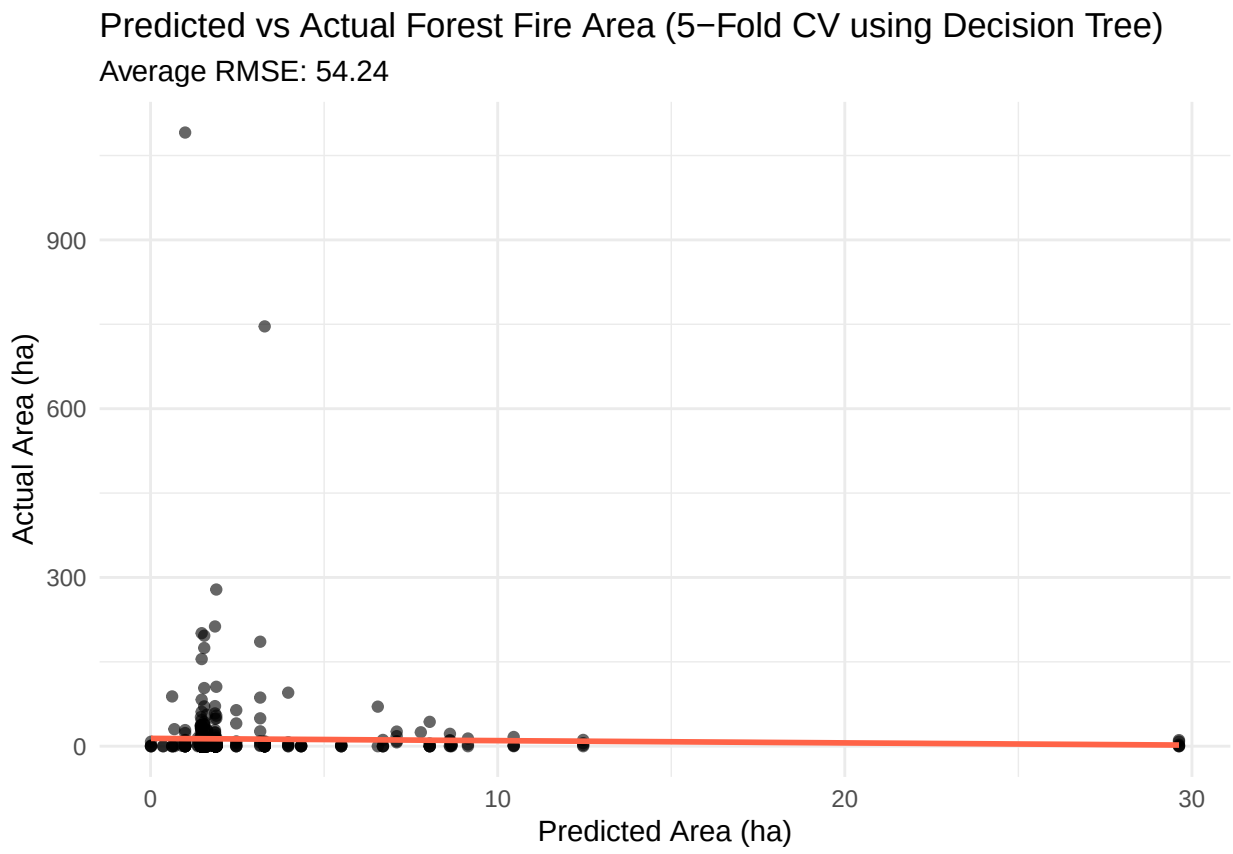
The meteorological data above was collected from Montesinho Natural Park, Portugal (2000–2003). The data set above was publicly posted on Kaggle by Paolo Cortez and Anbal Morais. The indepednent variables that we will be focusing on for our model building are: temp (measured in Celsius), RH (Relative humidity measured as a percentage), wind (measured in km/h), and rain (measured in mm/mm^2). The dependent variable that I will be using is area (measured in hectares). The only preprocessing step involved was applying a logarthimic transfomration on the area variable due to the skewness of the data. I will be using 4 types of models: linear regression, random forest regression, decision tree regression, and support vector machines. The advantages of using a linear regression model are: it's easy to understand, it's fast to train, and it works well the relationship is linear. The disadvantages of using linear regression are: it can't capture complex patterns, it doesn't work well for outliers, and it always assumes a linear relationship. The advantages of using decision tree regression are: it's easy to visualize, it captures nonlinear patterns, and it handles both numerical and categorical data. The disadvantages of using decision tree regression are: it can over fit the data, it can be unstable, and it is usually not as accurate as other models. The advantages of using random forest regression are: it is very accurate, it handles non-linear data, and it handles outliers very well. The disadvantages of using random forest regression are: it's harder to interpret and it is slow to train and predict. The advantages of using support vector machines are: it is captures high-dimensional data really well, can capture very complex relationships, and it works even when the data isn't completely separated.

The disadvantages of using support vector machines are: it's very slow on large data sets and it is very hard to explain. I will also be splitting the data into training and testing sets to determine the accuracy of each model. I will also be incorporating k-fold cross validation to further test the accuracy of each model. I will be looking at the mean squared error, root mean square error, and r squared value to determine the accuracy of each model. In addition, I will also be creating visualizations of each model. Additionally, I will not be using any third party libraries. If you are interested in looking further into the data, then please incorporate the rest of the columns for further testing. You can also incorporate different modeling techniques as well if you choose to do so.

Conclusion/Demonstration

The two visualizations below still need many improvements. The random forest model and the support vector machines model are also still in progress. The intended conclusion is to determine which model performs best in terms of accuracy. The visualizations below show that linear regression is not a good model to use due to the numerous outliers in the data. The visualizations also show that decision tree modeling is somewhat accurate.





References

https://www.kaggle.com/datasets/elikplim/forest-fires-data-set