Han Lee
CSC 2730

Predicting Area of Forest Fires

For the final project, I decided to analyze data collected from forest fires in a northeast region of Portugal. This data includes many environmental factors collected at the instance of each fire. Attributes include:

- 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: 'jan' to 'dec'
- day - day of the week: 'mon' to 'sun'
- FFMC - Fine Fuel Moisture Code index from the FWI (Fire Weather Index)system: 18.7 to 96.20
- DMC - Duff Moisture Code index from the FWI : 1.1 to 291.3
- DC - Drought Code index from the FWI system: 7.9 to 860.6
- ISI - Initial Spread Index from the FWI system: 0.0 to 56.10
- temp - temperature in Celsius degrees: 2.2 to 33.30
- RH - relative humidity in %: 15.0 to 100
- wind - wind speed in km/h: 0.40 to 9.40
- rain - outside rain in mm/m2 : 0.0 to 6.4
- area - the burned area of the forest (in ha): 0.00 to 1090.84

Here a screenshot of a sample of the data:

X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
8	6	aug	fri	90.1	108	529.8	12.5	21.2	51	8.9	0	0.61
1	2	jul	sat	90	51.3	296.3	8.7	16.6	53	5.4	0	0.71
2	5	aug	wed	95.5	99.9	513.3	13.2	23.8	32	5.4	0	0.77
6	5	aug	thu	95.2	131.7	578.8	10.4	27.4	22	4	0	0.9
5	4	mar	mon	90.1	39.7	86.6	6.2	13.2	40	5.4	0	0.95
8	3	sep	tue	84.4	73.4	671.9	3.2	24.2	28	3.6	0	0.96
2	2	aug	tue	94.8	108.3	647.1	17	17.4	43	6.7	0	1.07

I deemed the attributes X, Y, and day as irrelevant to my specific analysis of this problem, so I removed them from the dataset. Also, to make everything numerical, I replaced the months from objects to integers 1-12. I also removed two outliers where the area of forest burned was significantly higher than the others.

I used LinearRegression from sklearn as everything was numerical, and I assumed factors would be linear and have constant variance.

I measured performance of each attribute's relationship to area in LinearRegression using the mean squared error (MSE). I thought it would give me the best and worst correlation, but it just proved that LinearRegression was not an accurate model to use for this dataset.

Attribute	**MSE**
Month	743.1210019996515

FFMC	746.0029682678808
DMC	743.2483863710979
DC	745.1416672743895
ISI	745.9481507830296
Temp	743.0834580592666
RH	745.4452306358706
Wind	746.594431858591
rain	746.5601390614528

I tested the accuracy of LinearRegression compared to 2 other models, and the other models showed significantly better accuracy, showing that LinearRegression was not a good model for this dataset. To improve this analysis, I should've used other models. To measure accuracy, I calculated their explained variance scores and MSEs. RandomForestRegressor showed significantly better variance score and MSE, and ExtraTreesRegressor shows even better score and MSE. The names of models also coincidentally have the word 'forest' and 'trees' in them.

Model	**Explained Variance Score**	**MSE**
Linear Regression	0.020378	731.390087
RandomForestRegressor	0.833701	124.442509
ExtraTreesRegressor	0.984253	11.756586