CLASSIFYING FOREST FIRES

Mason Tilghman

TABLE OF CONTENTS

Why forest fires?

Forest fires are a major environmental issue worldwide. The damage they cause is both economical and ecological - and the costs are trending upward. These increased costs are dramatic. The United States Forest Service now spends around $2 billion a year in attempts to manage forest fires - a 600% increase from 2 decades ago.

Our Dataset

- Our dataset contains information on 517 forest fires in Northern Portugal.

- There are 12 columns, containing a mix of location data, meteorological data, and Fire Weather Indexes.

- We have created our target column ‘burn area’, and classified fires as either small, medium, or large.

DATA EXPLORATION

NULL VALUES

The data had already been prepared for us, so there were no null values.

DISTRIBUTION

The distribution of most variables is non normal, particularly with the target value ‘area’. We log transformed our data to help alleviate issues with this.

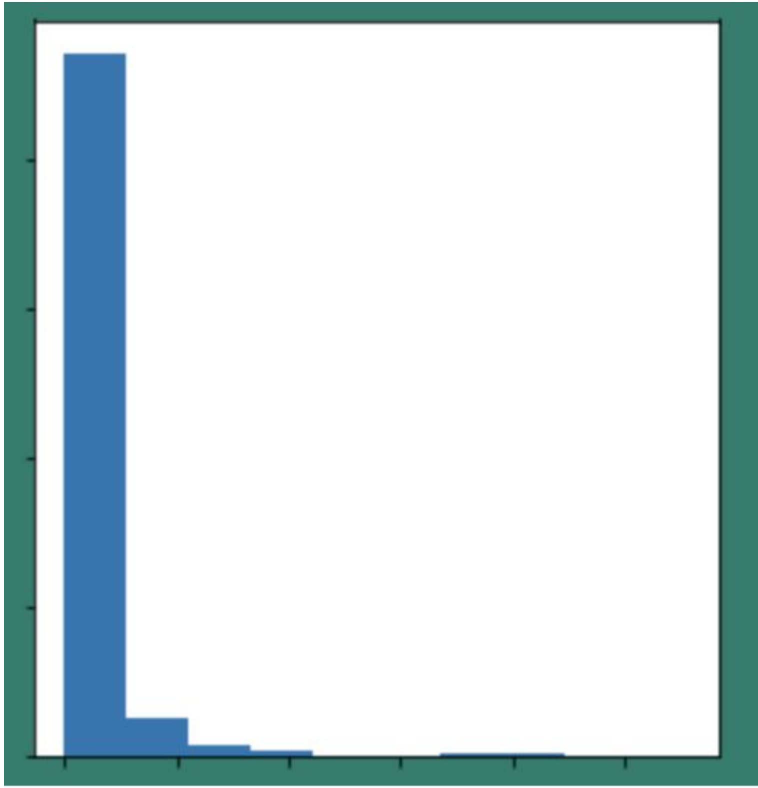
TARGET VARIABLE

We have created a column ‘burn_area’ and classified fires as either small, medium, or large.

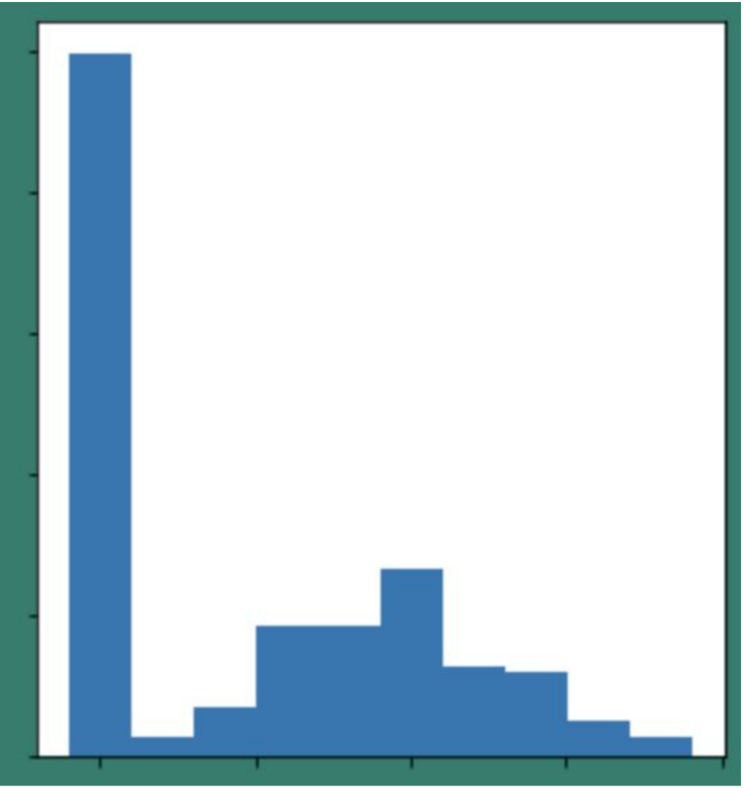
CLASS IMBALANCE

There is a class imbalance within our target column that we need to consider. In total there were 274 small fires, 92 medium, and 149 large.

area



area (log-transformed)



MODELING

K-NEAREST NEIGHBORS

The first model we implemented was a KNN classifier. This was chosen for its speed and to get a baseline accuracy.

RANDOM FOREST

Second, we implemented a random forest model using the best features of the KNN model (humidity, temp, wind, rain).

XGBoost CLASSIFIER

Lastly, to push the limits of our accuracy we implemented boosted decision trees. We started to see diminishing returns here.

KNN

	KNN		
	predicted		
actual	large	medium	small
large	36	1	112
medium	11	4	77
small	24	2	248

	precision	recall	f1-score	support
large	0.51	0.24	0.33	149
medium	0.57	0.04	0.08	92
small	0.57	0.91	0.70	274
accuracy			0.56	515
macro avg	0.55	0.40	0.37	515
weighted avg	0.55	0.56	0.48	515

- Our KNN model helped us round down our feature set to only the meteorological inputs
- Relatively much faster than other models
- Cross validated accuracy was 49.50%

Random Forest

	Random Forest		
	predicted		
actual	large	medium	small
large	12	0	137
medium	1	2	89
small	1	0	273

	precision	recall	f1-score	support
large	0.86	0.08	0.15	149
medium	1.00	0.02	0.04	92
small	0.55	1.00	0.71	274
accuracy			0.56	515
macro avg	0.80	0.37	0.30	515
weighted avg	0.72	0.56	0.43	515

- Overall accuracy improved slightly
- Correctly classified all but one small fire
- Cross validation accuracy was 53.60%

XGBoost

	XGBoost		
	predicted		
actual	large	medium	small
large	12	0	137
medium	1	3	88
small	1	0	273

	precision	recall	f1-score	support
large	0.86	0.08	0.15	149
medium	1.00	0.03	0.06	92
small	0.55	1.00	0.71	274
accuracy			0.56	515
macro avg	0.80	0.37	0.31	515
weighted avg	0.72	0.56	0.43	515

- Near identical to Random Forest model.
- Computationally expensive
- Cross validation accuracy was 51.44%

RESULTS

OVERALL

To be blunt, these models do a poor job predicting the size of fires.

SMALL FIRES

The final model did nearly a perfect job when predicting small fires. This is the most ‘useful’ function of our model.

IMPROVEMENTS

More data on how each fire was managed (response time, detection time, etc) could help further classify our fires. Using a random subset to fix the class imbalance could also improve results.

QUESTIONS?

Do you have anything you’d like to ask me?