CAPSTONE PROJECT PROPOSAL

Mariolys Rivas.
January 10, 2021.

Proposal

Predicting wildfire causes from wildfire size, location, and date.

Domain Background

Wildfires are a major issue in the United States and Canada, and are particularly damaging to humans and nature in the West Coast, where I reside. Once a regular and sometimes beneficial natural phenomenon, wildfires have become more destructive and disruptive in the past decade.

Climate change continues to be the most important environmental issue of our generation. The global average temperature of Earth has continued to grow rapidly over the past few years, accelerating the onset of drier and hotter conditions that help make natural disasters such as wildfires and flooding more frequent and catastrophic than ever before.

The reason I considered this project is because I am passionate about the environment and the planet we live in. It is my duty to use my skills to research and hopefully help mitigate some of the direst consequences of our own human activity.

Problem Statement

The problem that I am trying to solve was inspired by Kaggle's 1.88 Million US Wildfires dataset. Under the "Inspiration" header, there is a question:

"*Given the size, location and date, can you predict the cause of a fire wildfire?*"

Solving this problem may be useful in wildfire prevention efforts, and to educate local residents on the subject. If for example in a particular location wildfires are caused mostly by human activities, then campaigns can focus on that in the future.

Dataset and Inputs

I will use Kaggle's 1.88 Million US Wildfires dataset, which contains 24 years worth of US wildfires data (1.88 million data points). The independent variables I will be using for prediction are the following:

DISCOVERY_DATE = Date on which the fire was discovered or confirmed to exist.
DISCOVERY_DOY = Day of year on which the fire was discovered or confirmed to exist.
DISCOVERY_TIME = Time of day that the fire was discovered or confirmed to exist.
FIRE_SIZE = Estimate of acres within the final perimeter of the fire.

The target variable is the following:
STATCAUSECODE = Code for the (statistical) cause of the fire.

The column name below will be used for understanding the target variable codes.
STATCAUSEDESCR = Description of the (statistical) cause of the fire.

Solution Statement

As a first part of my solution I will start preprocessing the data and familiarizing with it, looking at the meaning of the different cause codes and analyzing the most common causes. I will also drop any unnecessary columns, and keep the ones I am interested in. I will explore how imbalanced the data is. Once the preprocessing is done I will go ahead and compare three different machine learning models, namely decision trees, SVM (Support Vector Machine) and MLP (multi-layer perceptron). I will compare them and choose the best one, according to the metric mentioned later.

The process above will be done by training on a predefined training set, and comparing results on a validation set. The final result will be reported on a preselected test set.

Bechmark Model

I will define a benchmark model that always picks the cause that is more frequent in the training data.

Evaluation Metrics

I plan to use the F_1 score and accuracy as metrics. I like to use the F_1 score as it provides a good balance between precision and recall. Specially if the cause codes are imbalanced. I will look at the confusion matrix as well.

Recall that the F_1 score is defined as follows:

$$F_1 = 2\frac{Pr \times Re}{Pr + Re}$$

Where Pr represents precision and Re represents recall. Notice that this measure provides a single value for each output label. These values can be averaged over all labels in order to compare two different models.

Project Design

The project design is as follows:

- Divide the data into three different sets: training (80%), validation (10%) and testing (10%).
- Data prepossessing, familiarization and cleaning, completing nan values (if any), obtaining histograms that reflect the balance/imbalance of the target variable, and segmenting the data by location. Preprocessing strategies are then noted for reuse in the test set.
- Model training under three different methods: decision trees, SVM and MLP, since I will be dealing with a multi class classification problem. I will compare these models according to evaluation metrics above.
- Choosing the best of the model and fine tuning it for refinement. If time permits, this may include fine tuning under different geographic segments. This model is then tested on the final test set, to avoid data leakage.