Can we predict fire with weather data?

Harsh (hp444), Yi Yao (yy899), Murali (mt788)

December 8, 2019

Abstract

Contents

1 Introduction **1**

2 EDA **2**

2.1 Covariates 2

2.2 Basic Statistics 2

3 Model Building and Accuracy Tuning **4**

3.1 Preliminary Analysis 5

3.1.1 Primitive Model Fitting 5

3.1.2 Under-sampling 5

3.2 Accurarcy Tuning 5

3.2.1 Boosting Models Fitted with Balanced Sub-samples 5

3.2.2 Clusters of Many-hot Encoding 5

3.2.3 DBScan Clustering as Feature 5

3.2.4 Stacked Ensemble 5

3.2.5 Adding XGBoost to the Ensemble Model 5

4 Conclusion **5**

1 Introduction

The weather has literally been the hot topic of discussion in recent times. One aspect of weather data is that, when it is combined with other data sets, it can be used to predict varied different aspects related to the society. For example, one particular aspect could be how does traffic conditions change based on the weather in a particular city or if there is a correlation between crime rates and weather conditions. In this project, we are trying to address a more important issue of predicting wildfires in different cities based on weather conditions. Since the future weather data is readily available these days, a good model would help predict these wildfires in advance and take necessary precautions to completely avoid it. This would help avoid human and wildlife loss. It would also help contain the impact of these wildfires on the environment which increases the content of poisonous gases in the environment, making the nearby uninhabitable.

Wildfires are growing in frequency and intensity by the day. The 2019 wildfire season of California has more than 6800 fires recorded by the US Forest Service and approximately 250000 acres of burnt land. These wildfires could be the result of foul human activities but there has been a lot of speculation on the high correlation of these fires with the weather. Climate change is not only making the fire season longer but on average much more intense. The idea is to use the model to predict these fires in advance, thereby reducing such incidents.

2 EDA

2.1 Covariates

The features that we have in our Weather dataset are:

Table 1: Covariates

Covariate	Type	Description
Temperature		
Pressure		
Humidity		
Weather Description		
Need to check		

The features in the fire dataset is:

1. erer
2. erere
3. erer4555

2.2 Basic Statistics

The size of the dataset in the Fires table is: $number \times number$ and the size in the weather table is: $1M \times 11$. The following figures help illustrate what the dataset looks like

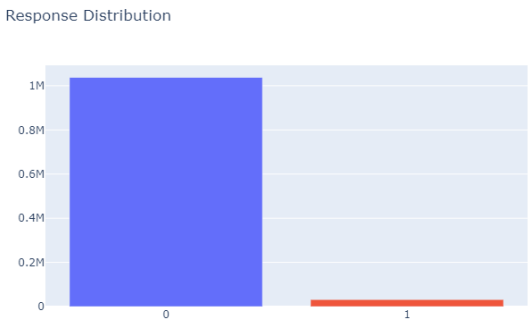

Figure 1: Response Distribution

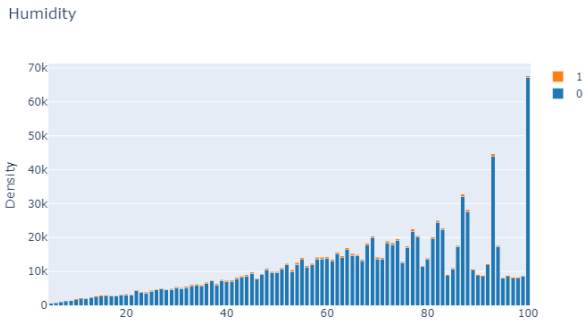

Figure 2: Histogram of Humidity

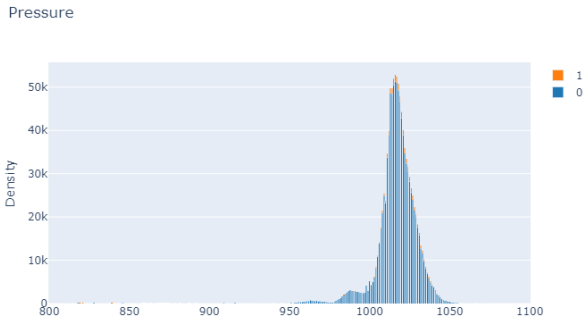

Figure 3: Histogram of Air Pressure

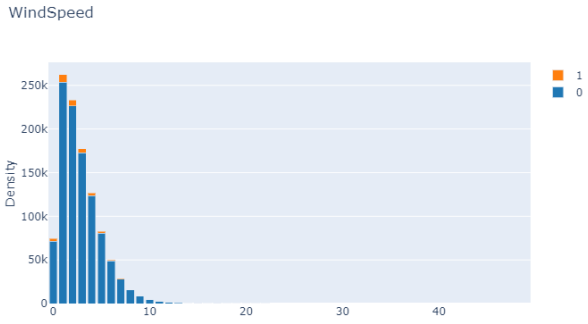

Figure 4: Histogram of Wind Speed

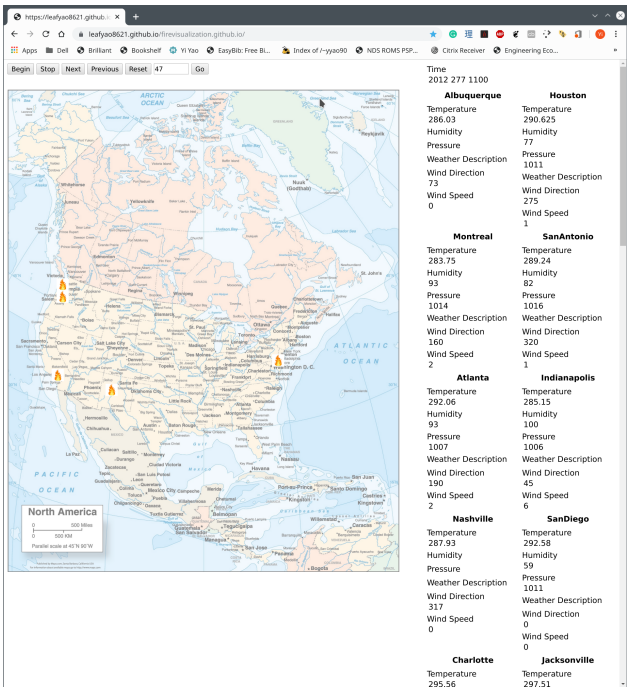

Figure 5: Screenshot from Visualization Tool

3 Model Building and Accuracy Tuning

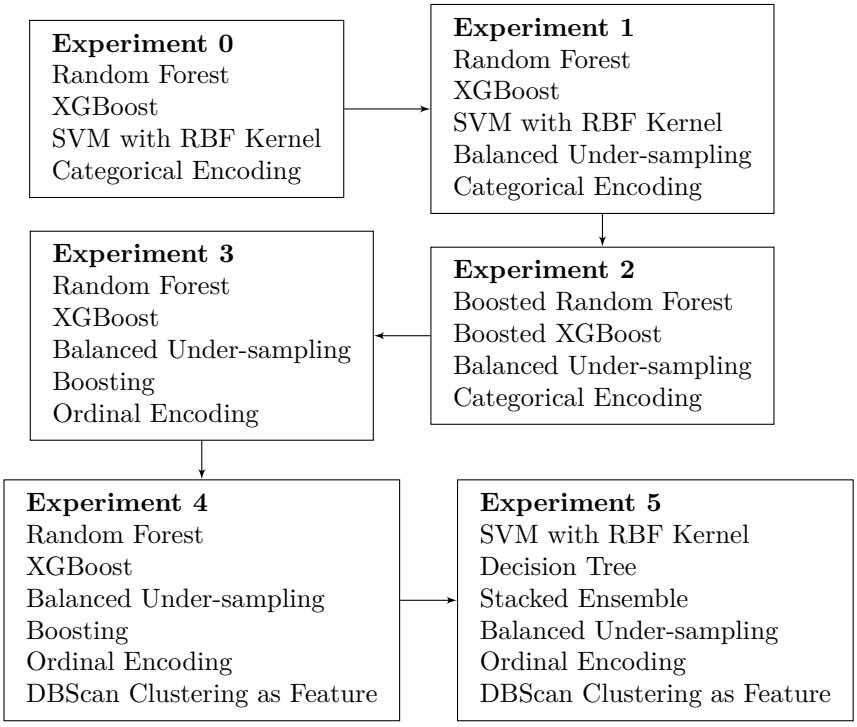

Figure 6: Flowchart of Experiments

3.1 Preliminary Analysis

3.1.1 Primitive Model Fitting

First, we fit XGBoost, Random Forest and SVM by using the entire dataset and found the accuracy to be quite low. The low accuracy could be because of the model overfitting. There are, afterall, 1M negative samples and roughly 25k positive samples.

3.1.2 Under-sampling

Recognizing the unbalanced nature of our dataset, we created a balanced training data set by combining all the positive samples with a random subset of all the negative samples. Using this balanced training data set, we fit XGBoost, Random Forest and SVM with RBF kernel. We observed a significant improvement in test accuracy in the XGBoost and Random Forest models; however, we have not observed any significant improvement in the SVM models.

3.2 Accurarcy Tuning

3.2.1 Boosting Models Fitted with Balanced Sub-samples

Having the possibility of underfitting due to under-sampling in mind, we have experimented with boosting to increase accuracy. By keeping all the positive samples and randomly resampling negative samples equal to the number of positive samples, we have kept the balance of the training data set while making better use of the data set we have available.

3.2.2 Clusters of Many-hot Encoding

Having observed some order in the covariate "Weather Description", we have experimented with many-hot encoding for ordinal variables. Since an overall could not be established by the team, we have turned to many-hot encoding on 3 clusters of levels and one-hot encoding on the remaining levels. As for models, we continued to use the boosted algorithms in the previous experiment.

3.2.3 DBScan Clustering as Feature

We have also experimented adding one-hot encoded output of an unsupervised clustering algorithm, DBScan as covariates to the dataset to be fitted with supervised models.

3.2.4 Stacked Ensemble

3.2.5 Adding XGBoost to the Ensemble Model

4 Conclusion