INGV - Eruption Prediction

Jonas Williams-Gilchrist -- CS 5665 -- ID A16

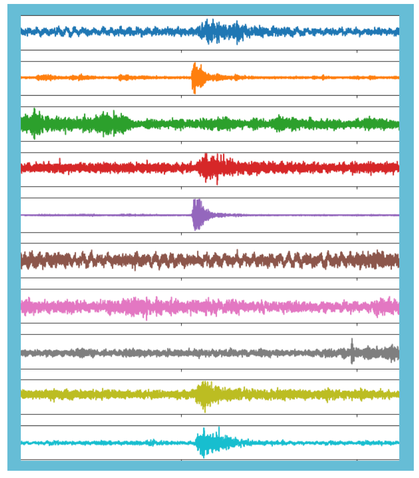
Problem Statement

- Volcanic eruptions are extraordinarily difficult to predict and there is a lot of variability in the accuracy of predictions.
- Similar to other weather-based phenomena, prediction accuracy is easier when data is closer in time to the actual event.
- Sometimes the early detection of volcanic activity and eruptions can be the difference between thousands of saved lives.

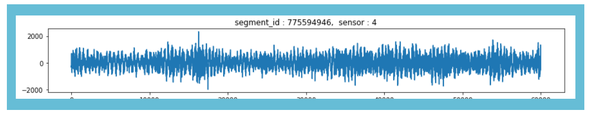
Combining these major parts of the problem, we need to develop some sort of technique to accurately predict volcanic eruptions from our data.

Data/Task

- Develop a data driven, machine learning based model for predicting volcanic eruptions.
- The data is collected from ten different sensors at varying proximity to the volcano.

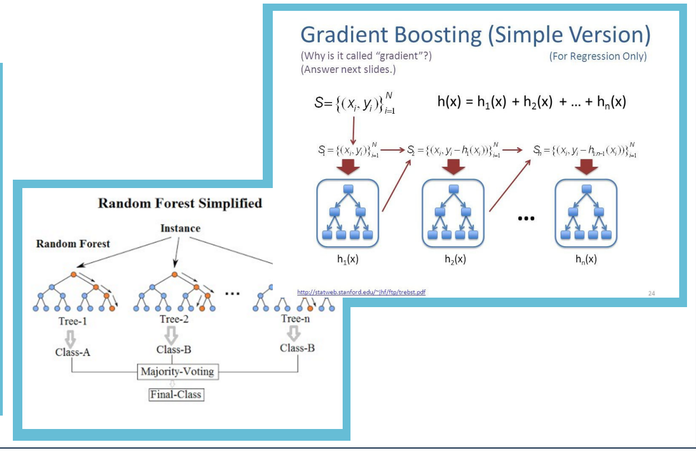


To the right is a visualization of the raw sensor data given to us. Each of these different colored wave-like distributions is a separate sensor, plotted over time where a larger amplitude represents more severe seismic activity. At the bottom right is a close up of a rather indeterministic sensor.



Appraoch

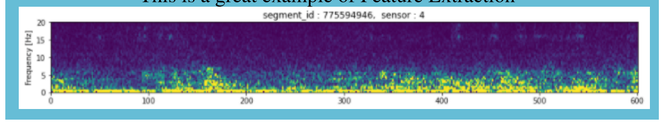
- I will be using a Gradient Boosting Machine, or GBM for short to process the multitude of sensor data.
- This is an ensemble learning model.
- The general idea of this model is similar to that of a random forest. A combination of a bunch of weak classifiers is much better than just one of that weak classifier.



Results and Analysis

- During this stage, hyperparameters were fine-tuned with lots of trial and error, the result is a model that can predict the time until a volcanic eruption.
- It is important to note that the GBM is a regression model, as such its accuracy is determined by Mean Squared Error, or MSE. This heavily penalizes results that are very far away from what's correct.

Below is the Fast Fourier transform of data from the specific sensor I showed previously. We can see here that while the new data is far from perfect, it is far more discriminative then it was previously.
This is a great example of Feature Extraction



Conclusions

To the left is my placement in the Kaggle competition for this project. At the time of submission, my model places 99th out of 358 total submission, the top 27.7%

- Generally, this project was a large success.
- I greatly enjoyed learning about a new type of regression algorithm, the GBM.
- I like how this relates to the other ensemble learning methods I've used in the past, primarily random forests.
- I think I was limited by my computing resources as well as my knowledge about this type of regressor, a lot of my time was spent figuring out how the models work with the data rather than optimizing the algorithms used.

	segment_id	time_to_eruption
0	1000213997	2.648216e+07
1	100023368	3.956723e+07
2	1000488999	2.061216e+07
3	1001028887	2.468681e+07
4	1001857862	1.515855e+07
...	...	...

An example of the first few elements in my submission file. This shows how my model uses the sensor's identification (labelled segment_id), connects it to the processed data, and eventually to a single value (time_to_eruption).

Github: https://github.com/Jandawg/CS-5665-INGV