INGV - Volcanic Eruption Prediction*

Maria Franchezka L. Sino

Machine Learning Engineer Nanodegree

Capstone Proposal

Domain Background

This Kaggle competition was launched by Italy's *Istituto Nazionale di Geofisica e Vulcanologia* (INGV) which focuses on geophysics and volcanology. The institute monitors seismic and volcanic activity all over the country.

Problem Statement

Scientists can predict the "time to eruption" based on the volcano and seismic activity but these patterns are hard to interpret. Their current way can predict for minutes in advance but not for long-term predictions. The competition aims to have a better model for long-term predictions of time to eruption.

Datasets and Inputs

The data provided by the competition organizers consist of ~30GB of data collected from sensors around the volcano. Data is split into training and test folders where each file contains 10 minutes of logs from 10 different sensors around the volcano with 100Hz sampling rate. These have been normalized within each segment. The file train.csv contains ID code of each segment, which corresponds to the file names from the train folder, and the "time to eruption". The file sample_submission.csv is similar to train.csv but it contains the IDs from the test folder and the "time to eruption" are all set to 0 since this will be where the predictions will be entered.

Solution Statement

This project would consist of data exploration and then pre-processing to ensure that the data would be fit for the model we would be using for training and predicting. As mentioned previously, our goal is to fit a model that would predict the volcano's "time to eruption" based on the sensor data logs from the different sensors on the volcano.

Benchmark Model

The specific model they're currently using was not mentioned on the competition page but it is mentioned that short-term prediction is possible, and they believe that seismic waves can provide more information on long-term prediction. A deep learning approach on the data might help in discovering this.

*current Kaggle competition

Evaluation Metrics

The submissions for the competition would be evaluated based on the Mean Absolute Error of the predictions using the test data.

Project Design

The project will be broken down but is not limited to the following tasks:

A. Exploratory data analysis*
    a. Exploring data metrics
    b. Data visualizations of the features
    c. Checking for missing/invalid values
    d. Looking at data statistics
B. Data preprocessing*
    a. Data aggregation
    b. Data pre-processing
    c. Feature engineering
C. Model Training and Prediction*
    a. Uploading the preprocessed data on AWS S3
    b. Splitting the train.csv to train and test sets
    c. Training a model (output: volcano's time to eruption)
    d. Predicting using the test set from train.csv
    e. Predicting on test.csv and uploading submission on Kaggle
    f. Further iterations on the model re-training as possible
D. Report writing

*will be done on my local machine
**will be done on AWS SageMaker