Capstone 2 – Project Proposal

Problem Statement:

Los Alamos National Laboratory and Purdue University are interested in scaling experimental data to forecast earthquake location, size and timing from seismic signals, helping to save lives and reduce incurred costs of infrastructure damage.

Context:

Los Alamos National Laboratory (LANL), an institution dedicated to developing technologies addressing threats to national security, has simulated many earthquakes in a laboratory setting collecting data on the events. The experiments are performed by applying a shear force to a sample equivalent of earth and rock, which contains a fault line. Acoustic signals from the experiment are collected in a single sequence during which several simulated earthquakes are observed. The simulated earthquakes can be analyzed in the train.csv data as the "time_to_failure" approaches zero.

The object of collecting this data is to reduce the impact of devastating earthquakes. Our goal as a data scientist is to develop a model to predict the time lapse before an earthquake hits. An important aspect of this project will be the ability of the results to scale and generalize to real world data. .

Success Criteria:

Develop a model to predict the time before failure of an earthquake to within 60 seconds in the laboratory setting. Predictions within 60 seconds of the laboratory earthquake are shown to scale to within an hour of a real-world earthquake. Given a time approximation within an hour and the location of the earthquake, aid services can prepare for possible injuries and prevent loss of life by 5x. Additionally, the advanced and precise notice allows preparations to be made, which can save costs in the range of $200,000-$5,000,000 depending on where the earthquake strikes.

Scope + Risks:

The model should focus on the time lapse between the initial signal and resulting seismic activity to answer when the earthquake is expected to strike. The experimental data may not properly scale to the physics of the real-world application.

Decision Maker/Other Stakeholders:

Lead Data Scientist at LANL

Data Scientist – Matt Miller

Constraints:

- Questions of location and size of the earthquake will remain.

- Impact reduction in some locations may not be possible with the data collected from this experiment.

- Limited experimental data is available; although the acoustic sensor has collected at a high rate, there are only sixteen observations of actual simulated earthquakes. This may become an issue as the geologists attempt to scale and generalize the results.

Data Sources:

Downloadable data from a past Kaggle competition. The data will be downloaded in the form of a train.csv set (size=9.56GB) and many test.csv subsets (size=~320kb). The data includes experimental data collected by LANL featuring the acoustic data, representing the seismic signal [int16], and time to failure data, which represents the time in seconds before the next laboratory earthquake [float64].