Machine Learning Engineer Nanodegree Capstone Proposal - LANL Earthquake Prediction

Praveer Nidamaluri

April 17th, 2019

Domain Background

The proposed capstone project is in the field of seismology, the study of earthquakes. Earthquakes are vibrations in the surface of the Earth that are triggered by releases of energy primarily at geological faultlines. Earthquakes can vary in intensity and frequency, but the strongest incidents can cause extreme destruction and loss-of-life. For example, the 2011 March 11th Tohoku earthquake in Japan resulted in $360 billion and almost 20,000 fatalities [1]. The December 26th, 2004 the Sumatra-Andaman earthquake and resulting tsunamis resulted in $15 billion and caused 227,898 fatalities [2].

Considering the disastrous nature of seismic events, any form of earthquake prediction would significantly reduce the human and economic impact. Seismologists are now able to estimate the frequency and likely magnitude of earthquakes in a particular region. Unfortunately, despite decades of attempts, there is still no reliable method of predicting exactly when an earthquake will occur.

However, recent research on laboratory earthquake models has suggested that analyzing the vibrations around faultlines using Machine Learning (ML) algorithms may successfully predict the remaining time until the next significant energy release [3]. The same ML methods were subsequently applied to real seismic vibration data from the Cascadia fault in North America [4]. Using the ML algorithm, researchers were able to extract previously unidentified features in the data that could reliably predict the overall surface motion.

To further advance the promising application of ML methods for earthquake prediction, the Los Alamos National Laboratory (LANL) is now sponsoring a Kaggle competition [5], which is the subject for this project proposal.

As a structural and seismic engineer by training, I design power-plant structures that can resist expected earthquakes. This method helps reduce the economic impact of earthquakes to a certain extent. However, it is highly dependent on the accuracy of our earthquake forecasts. The Earthquake Prediction Kaggle competition is a great chance to apply my newly learnt ML skills to gain a better understanding of earthquakes and hopefully improve future predictions.

Problem Statement

The LANL Earthquake Prediction Kaggle competition [5] provides a new set of laboratory experimental data that more closely models real earthquakes than past research from [3]. The task is posited as a supervised learning, regression analysis. The input to the regression problem is a seismic vibration time-series (V(t)) representing 0.0375s of vibrations. The output to the regression problem is the time in seconds until the next earthquake, (T_e). The training data provided by the competition contains approximately 88s of raw data that has to be converted to datapoints of (V(t),T_e) pairs. A separate set of test data, containing only the input time series, is also provided and is used to compare model performance against other competition entrants. The performance of the regression model is evaluated by the mean absolute error of the predicted time until the next earthquake (T_e’) to the actual time (T_e); error for each datapoint = |T_e’ - T_e|.

Datasets and Inputs

The dataset is available from the Kaggle competition webpage [5]. The training data for the competition consists of a 9GB csv file with two fields: ‘acoustic_data’ and ‘time_to_failure’. This is data recorded from a lab model earthquake experiment. The csv file contains approximately 88s worth of a raw vibration data time-series, with a time-to-earthquake time-series. The raw data will have to be converted into input/target pairs for the expected regression model: (V(t), T_e).

A converted input datapoint, V(t), is a 0.0375s long time-series of vibration data recorded at 4MHz; i.e. 150,000 values. T_e is a value that represents the elapsed time between the 0.0375s vibration recording and the actual model earthquake. An example of a data point is presented in Figure 1 below.

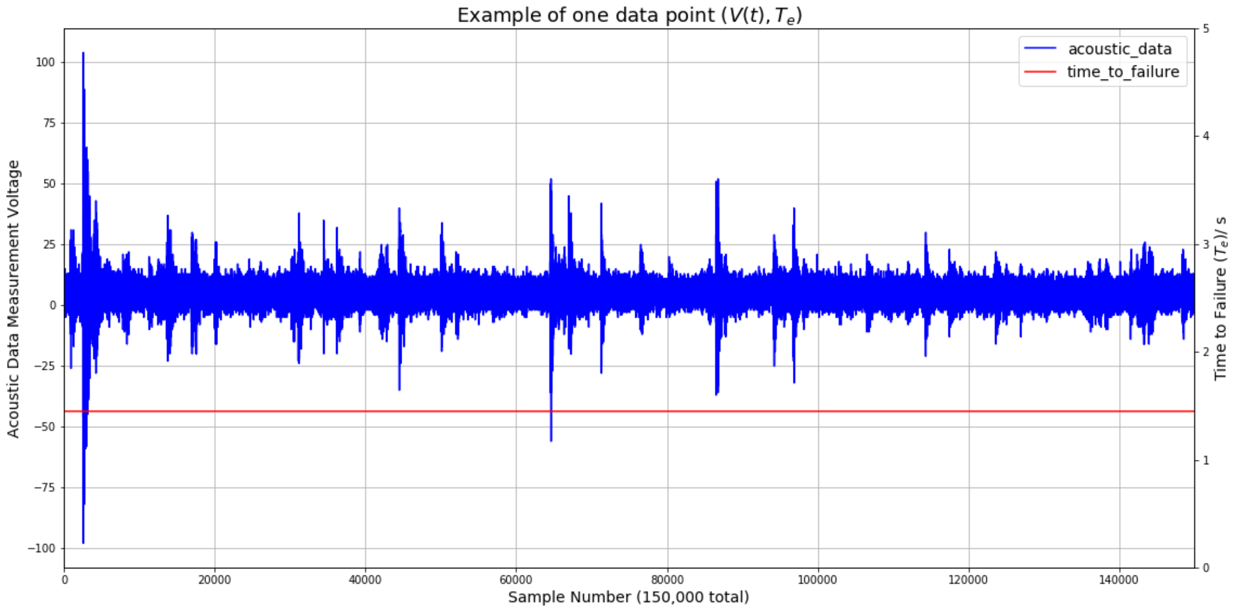


Figure 1 - Example of one data point (V(t), T_e)

The 0.0375s time-series within each datapoint are actually recorded from a single continuous laboratory experiment. It is therefore possible to plot a series of datapoints together (Figure 2) as continuous time-series of laboratory earthquake (labquake) acoustic data, and the time until the next labquake. The spike in the time_to_failure data in Figure 2 shows the occurrence of a labquake.

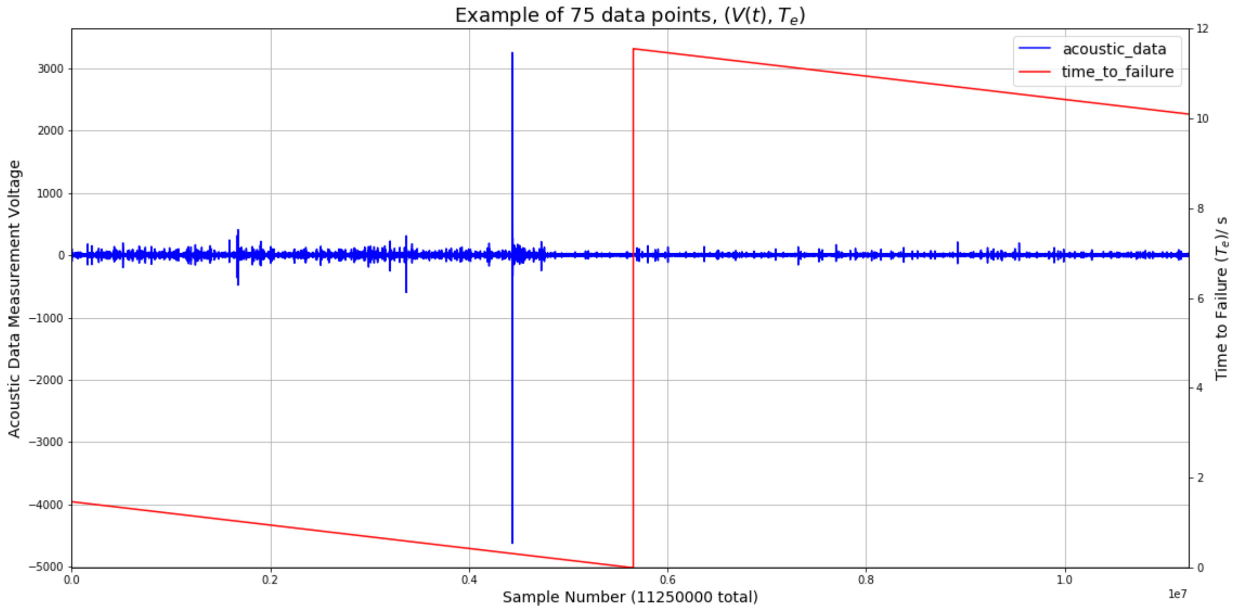
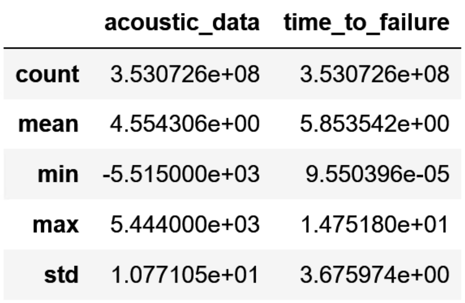


Figure 2 – Example of 75 datapoints, (V(t), T_e)

Key global statistics of the data are presented in Table 1.



	acoustic_data	time_to_failure
count	3.530726e+08	3.530726e+08
mean	4.554306e+00	5.853542e+00
min	-5.515000e+03	9.550396e-05
max	5.444000e+03	1.475180e+01
std	1.077105e+01	3.675974e+00

Table 1 - Global statistics of Kaggle competition training dataset

Solution Statement

The solution to the competition problem is the optimal regression model (f*) that can map 0.0375s labquake data segments (V(t)) to single predicted time-until-earthquake values (T_e):

$$f^*(\boldsymbol{v}) \approx T_e$$

Where **v** represents a vector of the 'n' most suitable features, that will be determined as part of the solution:

$$\boldsymbol{v} = (v_1, \dots, v_n)$$

The features will themselves be functions of the 150,000 acoustic data points within each 0.0375s labquake data segment. For example, a feature may be the mean, standard deviation, or an FFT coefficient of the input time-series segment:

$$v_i = v_i\big(V(t), \dots, V(t + \Delta t)\big)\ \forall i \in [1, n]; \Delta t = 0.0375s$$

The optimal regression model from the class of applicable models (F) will be determined based on an error function (C(f)) equal to the mean absolute error, as applied to the N datapoints within a cross-validation or competition test set:

$$C(f) = \frac{1}{N} \sum_{i=1}^{N} |\, f(\boldsymbol{v}) - T_{e_i} \,|$$

$$C(f^*) \leq C(f)\ \forall\, f \in F$$

In reality the optimal function will be chosen using model complexity graphs and learning curves to prevent over or under-fitting.

Producing the optimal model will ensure a good ranking in the competition, which is the primary aim of the project. In addition, the regression model should also ideally provide information about feature importances. This way, the final model will help ascertain which features from the seismic signals are most predictive of the labquake progression. The information can then inform future research with actual geological fault data and advance the state of the art.

Benchmark Model

A benchmark model can be made by following the machine learning algorithm suggested in previous research [3] from the sponsors of the Kaggle competition. The prior research describes a Random forest regression model trained on simplistic statistic features such as the mean and standard deviation of the input time-series. The output of the model is the estimated time until failure. This model represents the current state of the art. It can be recreated for the current competition dataset, and used as a benchmark with which to compare future models. The comparison will be quantified with the mean absolute error calculated on cross-validation or competition test sets, as detailed in the mathematical formulation above.

Evaluation Metrics

The competition evaluation metric is defined as the mean absolute error between the predicted and actual time-to-earthquake values [5]. Since this is a reasonable, quantified metric that will also be used to compare competition entrants, it will also be used to compare the benchmark, trial, and final solution models. The error is defined mathematically in the solution section.

Project Design

The theoretical workflow to reach a solution to the problem is summarized below:

1) Data exploration

   The input data will first be explored with simple statistics or visual representations. This is important for two reasons. First, exploring the data will help provide a better intuition and perhaps identify clear patterns that can then be expected in the results of more comprehensive models. Second, data exploration may help inform subsequent feature engineering, which may be a critical component of a successful model. Each input time-series datapoint contains 150,000 samples, which is a lot of data to process. Reducing the dimensions of the input will be essential.

2) Data preprocessing

   The input data is actually a 9GB csv file with a single long time series (approximately 88s) containing ‘acoustic_data’ and ‘time_to_failure’ values (Figure 2). This will be preprocessed into a series of $(V(t),T_e)$ input/target pairs, where: $V(t)$ is a 0.0375s segment of ‘acoustic_data’, and T_e is a single ‘time_to_failure’ value for that 0.0375s segment (eg, Figure 1).

   The input/target data pairs will have to be prepared such that the spike regions in the ‘time_to_failure’ values (Figure 2) are avoided. These segments represent the actual labquake.

   The simplest way to preprocess the raw data into input/target data pairs would be to split into as many 0.0375s segments as possible. This gives approximately 88s/0.0375s = 2,346 data points for training and cross-validation. This method ensures there is no overlap in the 0.0375s ‘acoustic_data’ segments across the input/target data points. However, an option to explore is if there can be overlaps between the input/target pairs. For example, if one data point is ($V(0s<t<0.0375s)$, $T_e(0.0375s)$), can another data point be ($V(0.02s < t < 0.0575s)$, $T_e(0.0575s)$)? Although this would generate more data points for training, it is also possible that the regression model would overfit to the data. This method would have to be explored.

3) Feature engineering

   Due to the large number of points within each input segment (150,000 points in each 0.0375s segment), some form of feature engineering to reduce dimensions is essential to reduce computational expense. Furthermore, feature engineering is a way to isolate the critical information from the data. That way, less training data is required to actually feasibly train a model. Finally, one of the aims of the research and competition is to help identify the artifacts in the acoustic data that are most relevant to predicting the labquake. This can then be used to advance actual earthquake prediction methods.

The first step in the solution workflow is to generate the benchmark model, which follows the methods in [3]. This method involves taking the mean, variance, maximum, and minimum of the V(t) acoustic data segments as features.

When more comprehensive models are then made, other features that can be explored are: more complicated statistical properties such as kurtosis and skew; fourier transform components; PCA components, etc.

The solution workflow will require multiple iterations of feature exploration, model training, and model comparisons.

4) Model generation

After generating a vector of features to represent each 0.0375s input acoustic data time series, the next stage is to choose a regression model, apply a suitable cross-validation strategy (K-fold cross-validation/Shuffle Split), and tune hyperparameters.

The benchmark model will be implemented first [3]; a Random Forest Regressor will be trained and tuned on the training data.

After implementing the benchmark model, the following models will be explored as they are known to be successful for time-series based, or general regression problems:

- Support Vector Regressor
- XGboost
- Long Short Term Memory (LSTM) models

The models will be tuned by evaluating model complexity graphs to prevent under/overfitting, and using grid search methods to optimize hyperparameters.

The model generation step will likely be iterated over multiple times to explore different feature sets, and to improve ranking in the Kaggle competition.

5) Submission and Presentation

The best performing model will be recreated in a Kaggle kernel and submitted to the competition. The path to reaching the final set of features and model will also be presented in the Capstone Project documentation.

References

[1] "Earthquake, Tsunami, Meltdown – The Triple Disaster's Impact on Japan, Impact on the World", Ferris, Solis, March 11, 2013, https://tinyurl.com/m2mwbbb/.

[2] *"Indian Ocean Tsunami – Economic Aspects". indianoceantsunami.web.unc.edu.*

[3] "Machine Learning Predicts Laboratory Earthquakes", Rouet-Leduc, Hulbert et al, 30 August 2017, https://doi.org/10.1002/2017GL074677.

[4] “Continuous chatter of Cascadia subduction zone revealed by machine learning”, Rouet-Leduc, Hulbert et al, 17 December 2018, https://doi.org/10.1038/s41561-018-0274-6.

[5] https://www.kaggle.com/c/LANL-Earthquake-Prediction#evaluation

[6] https://www.kaggle.com/inversion/basic-feature-benchmark