Earthquake Prediction

Mohit
Abhishek Rai Sharma
Vijay Karigowdara
Sai Santosh Kumar Ganti

Outline

1. Introduction
2. Dataset Description
3. Goal
4. Data Analysis
5. Feature Generation
6. Building Models
7. Results

Introduction

- Earthquake prediction has been thought of a nearly impossible task.
- It’s a step to save countless lives and caused damages.
- This is an active competition on Kaggle hosted by LOS Alamos National Laboratory.
- End Date for the competition is 3rd June 2019
- Earthquake forecasting focus on three key points: when, where, and magnitude.
- Using existing dataset of seismic activity to predict when the earthquake will take place.
- Two plates are put under pressure, resulting in sheer stress.

Dataset Description

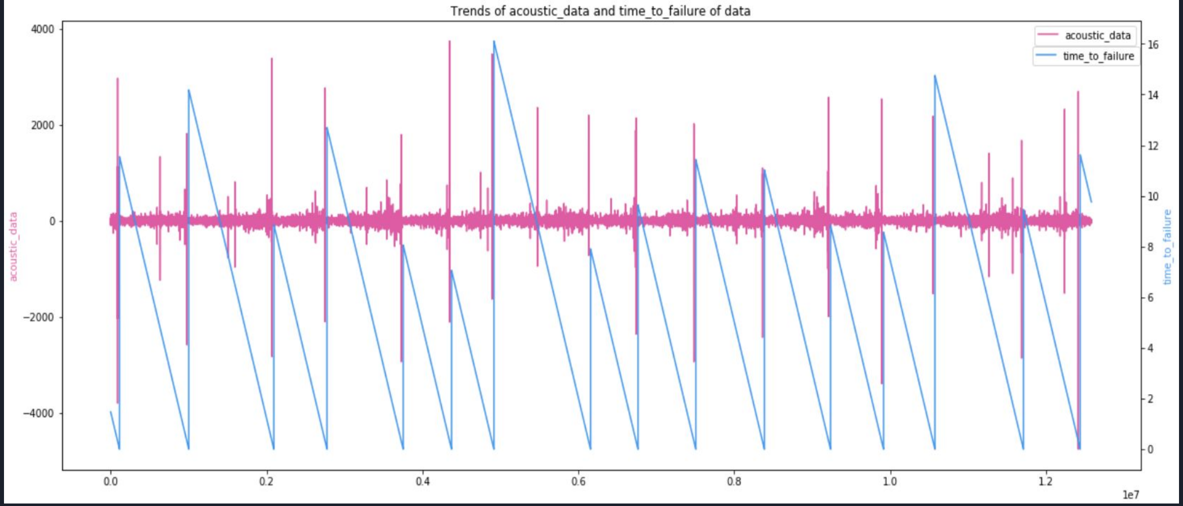
- Data is generated from a well-known experimental setup used to study earthquake physics.
- Train data - A single, continuous training segment of experimental data.
  - acoustic_data: The seismic signal[int16] .
  - time_to_failure: The time (in seconds) until the next laboratory earthquake [float64].
- Test data - A folder containing many small segments of test data.
  - seg_id: The test segment ids for which predictions should be made (one prediction per segment).
- The training data is a single, continuous segment of experimental data. The test data consists of a folder containing many small segments. The data within each test file is continuous, but the test files do not represent a continuous segment of the experiment; thus, the predictions cannot be assumed to follow the same regular pattern seen in the training file.

Goal

- The goal of this project is to use seismic signals to predict the timing of laboratory earthquakes.

Data Analysis

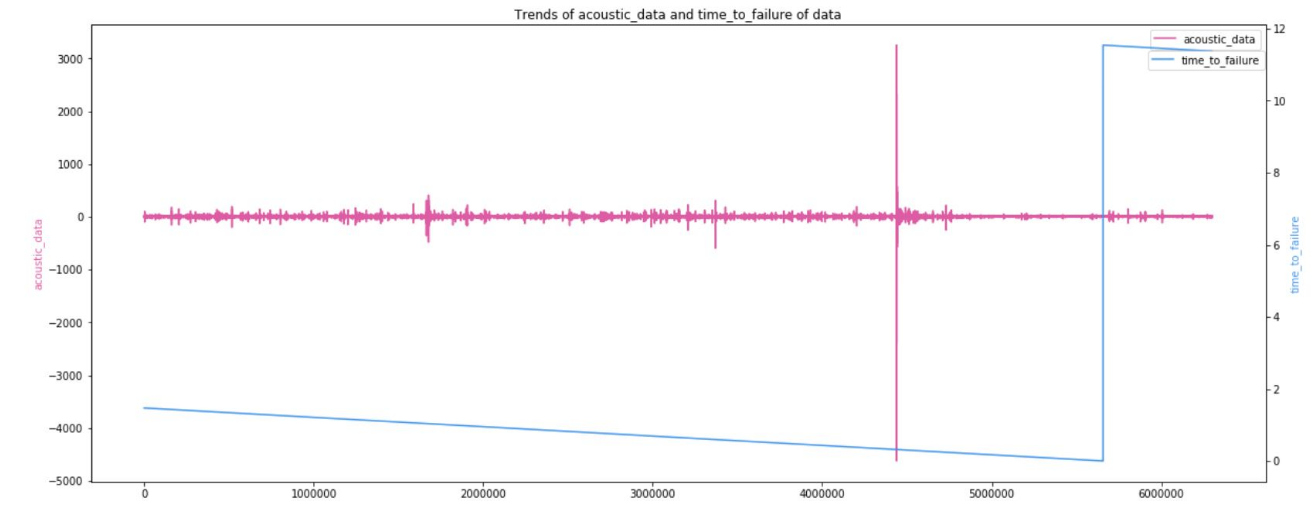
- We have 629 million rows and we plotted a sample of it.
- The acoustic data shows huge fluctuations just before the failure and the nature of the data is cyclical.
- Visually failures can be predicted as cases when huge fluctuations in signal are followed by small signal values.



Data Analysis Cont.

Here it seems that at first the signal has huge fluctuations for a short time, then it lowers and after some time the earthquake occurs. Which makes it difficult to distinguish target values properly

We will divide the data into blocks of 150000, so that we can get to know the patterns in the blocks.



Feature Generation

Since the given data only has two features, we have worked on feature generation from the given data and added them to the dataset using pandas.

- Usual aggregations: mean, std, min, max and kurtosis (heavy tailed or light tailed)
- Average difference between the consecutive values in absolute and percent values
- Absolute min and max values
- Aforementioned aggregations for first and last 10000 and 50000 values
- Max value to min value and their differences also count of values bigger that 500 (arbitrary threshold)
- Quantile features, Eg: 1, 05, 95, 99
- Rolling features : Window size- 10, 100, 1000
- Processing the given acoustic data signal to extract meaningful information.
  - Fourier Transform, Hilbert transform

Building Models

1. Light GBM Regressor:
    a. Light GBM Regressor is a gradient boosting framework that uses tree based learning algorithm. It is designed to be distributed and efficient with the following advantages:
        - Faster training speed and higher efficiency.
        - Lower memory usage.
        - Better accuracy.
2. Linear Regression:
    a. Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).
3. Extreme Gradient Boost:
    a. XGBoost (Extreme Gradient Boosting) belongs to a family of boosting algorithms and uses the gradient boosting (GBM) framework at its core. It is an optimized distributed gradient boosting library.

Building Models (Future work)

1. Support Vector Regression:
    a. The method of Support Vector Classification can be extended to solve regression problems. This method is called Support Vector Regression. The model produced by support vector classification (as described above) depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin.
2. CAT Boost Regressor:
    a. CatBoost is a recent open-source machine learning algorithm from Yandex.

Results

- We generated a dataset with predicted “time_to_failures” when an input signal is given.
- This submission data set is evaluated against kaggle test dataset.
- We get a mean absolute error between the predicted time remaining before the next lab earthquake and the actual remaining time.
- The current mean absolute error for our ensemble model of LGBM and XGBoost is 1.530

Thank You