A Predictive Model of Earthquakes Through Supervised Learning

Kristian Newell

December 6th, 2021

BrainStation Data Science

Introduction

This report has been prepared to provide a technical summary of this capstone project that's been built over the course of seven weeks. The overall objective of this capstone project is to investigate if earthquakes can be accurately predicted by supervised learning models. With supervised learning, the goal is to create a model that can account for the majority of the variance in a dataset and predict unseen data accurately.

Business Use Case

The business use case for predictive earthquake model is to act as an early warning system to areas that are seriously impacted by devastating earthquakes. The US Geological Survey reports that on average 20,000 people die from earthquakes annually. Although in 2010 this statistic jumped to over 200,000 deaths. Earthquakes additionally make up 5 of the 10 most deadly recorded natural disasters since 500 AD. Although human society has developed substantially since this time-period, we are no closer to removing the risk of earthquakes. If anything, as technology advances and more super structures such as large dams, nuclear powerplants, and megacities are constructed and integrated into all facets of life, the effects of the destructive power of earthquakes become even more apparent.

Earthquake prediction is becoming more necessary than ever to prepare for the disastrous outcomes of large-scale infrastructural failures. If someone were able to accurately predict the time and location of a serious earthquake before its occurrence it could allow people to evacuate the area or take shelter somewhere safer than, for instance, a large apartment building. Additionally, information on the estimated time and location of a serious earthquake could allow regional governments to allocate resources in reserve to ease the burden caused by the temporary or permanent destruction or disruption of local infrastructure.

Background

The information in my dataset consists of every earthquake of a reported magnitude over 5.5 since 1960. The dataset was provided by the US Geological Survey (USGS) in 2016 as part of a Kaggle competition. The USGS was created by the US Congress on March 3rd, 1879 in order to explore the geology and mineral makeup of the United States, but over its existence has become one of the leading agencies in the natural science world. This organization is one of the most well-known and credible sources of geological and seismic data currently in the world. As such, we can be confident that the information they have provided is accurate and credible.

Earthquake prediction is not a new topic of discussion in the natural sciences, as its implementation could result in tremendous prevention of loss of life and property damages. In fact, historically, the first attempt at recording earthquake magnitude was theorized in China in 132 AD. The first recorded attempts at modern earthquake prediction were in the late 1900s and consisted of trend analysis. For example in 1985, it was observed that the San Andreas fault underneath the Californian coast breaks with a moderate size earthquake (with magnitude 6) roughly every 20 years. This trend was then used successfully to predict that there would be another magnitude 6 earthquake roughly 20 years since the last. Even with modern technology there is no commonly accepted accurate earthquake predictor in use currently. The USGS has

a statement posted to their frequently asked questions page detailing that neither they nor any scientific community has ever successfully predicted a serious earthquake, nor do they expect to be able to in the near future.

The Data

The dataset that was used for this project consists of the date, time, latitude, longitude, magnitude, and depth of every recorded earthquake compiled by the USGS since 1960. The data was originally formatted as a csv file posted to Kaggle as part of a competition in 2016.

Upon downloading the dataset, it had 23,412 rows of information corresponding to 23,412 earthquakes across the world. There were originally 21 columns per row, however several of the less useful columns had as little as 300 observations.

Data Cleaning, EDA, and Feature Engineering

Before performing any feature engineering or modeling, the dataset was loaded, cleaned, and processed to its final form. The first step that was taken was to reduce the features list to those that I thought appropriate to the task of prediction. The USGS has stated that any predictive model of earthquakes is mandated to have the following: 1. Date, 2. Latitude/Longitude, and 3. Magnitude. Because of this I included these features as well as depth, type, and time, which were several of the few features that did not have a large amount of nulls. There were several features such as Status, Source, Location Source, and Magnitude Source that contained information about the stations that recorded each of the rows of information that did not pertain to earthquake magnitude, which were subsequently dropped. I then combined my Date and Time features to a single DateAndTime feature that I then converted to OrdinalTime in case any of my models were unable to read the data type datetime. In doing so I discovered 3 rows that contained letters in the date and time features. I considered imputing the date and time for these observations but since it was only 3 out of ~20,000 (0.01% of the data) I felt confident that my models would not be impacted by removing these rows.

After data cleaning, I inspected the distributions of my features and the relationships between my features to discover several noteworthy points. The first, being that 180 of my observations were not earthquakes, but other significant seismic events, such as nuclear explosions. I removed these rows and subsequently the Type feature, since after filtering to only earthquakes this feature was redundant. I additionally discovered no strong correlations between any features in my dataset, which is a good sign for modeling that each feature is independent. I did, however, discovered that my data would result in difficulty modeling due to the way a computer interprets longitudinal coordinates. For example, latitudes of 180W and 180E are equal on a globe, but a computer treats these two measures as the opposite ends of the longitude spectrum.

I began feature engineering by addressing the issue with regressing over longitude. I fixed this issue by isolating to a specific region of the world where latitude wrapping would not be an issue. In deciding the location to choose, I looked for a location with two main attributes. I determined that my model would have the most impact in an area that suffered from frequent, large earthquakes as well as an area that was under-prepared to manage the stress from these

earthquakes. Ultimately, I decided on South America for several reasons: the Pacific coast of South America falls on the “Ring of Fire” (the most active seismic zone on earth), the largest recorded magnitude for an earthquake ever was in South America, South America has significantly older and poorer infrastructure than in the US, and the higher population density of the area means that infrastructural failures result in greater scarcities. After this decision I filtered my dataset from its original 20,000 rows of information to only 88 representing the largest (magnitude 7 or greater) earthquakes to hit South America since 1960.

Modeling

In my modeling process I began by attempting to create a time series capable of predicting the pattern of time between serious earthquakes. I separated my data into two data frames indexed by time, one including all observations before 2010 and one including all observations after 2010, so that I could build a forecasting model. However, due to time constraints, I unfortunately was unable to create an accurate time series as my information was non-stationary.

I then began working on a model that, given the date and latitude/longitude of a predicted earthquake, could predict the depth and magnitude. For this modeling process I ran a basic linear regression to give myself a baseline dumb predictor to compare my future models to. This baseline model had an R squared value of around 16%, meaning it accounted for about 16% of the variance in the data. I then trained, optimized, and tested four regression models: a Random Forest Regressor, a Gradient Boosting Regressor, a LightGBM Regressor, and an XGBoost Regressor.

Findings

After testing each of my models I discovered that the Gradient Boosting Regressor was the most accurate at predicting my test set and accounted for 38.2% of the variance in the data, roughly double the basic linear regression. I then tested the residual plot of the predicted values and found that the model predictions were homoscedastic which is an excellent sign for the accuracy of the model in predicting on test data. Unfortunately, 38.2% is not a high enough R squared value for this model to go into production currently, and many more steps would need to be taken before this model would be able to predict serious seismic events.

Next Steps

Going forward from here, there are many steps I would take, time permitting, to improve this project. Firstly, I would reformat the problem that I am attempting to address; as the USGS states, the prediction of a single large earthquake down to a date and magnitude is not possible with current technology and knowledge of seismic activity. Instead, I would format my model to be able to predict the likelihood of a serious earthquake within a certain timeframe and geographical location. Additionally, I would ideally be able to integrate seismic sensors to my model to be able to record foreshocks and aftershocks to the most serious earthquakes in order to determine if there is any pattern that reveals insights on earthquake frequency. And finally, I was only able to create a model for a specific geographical region, I would like the opportunity to expand the scope of this project to all parts of the globe.

Sources:

https://www.usgs.gov/faqs/can-you-predict-earthquakes?qt-news_science_products=0#qt-news_science_products

https://en.wikipedia.org/wiki/Earthquake_prediction

https://www.usgs.gov/faqs/why-do-earthquakes-other-countries-seem-cause-more-damage-and-casualties-earthquakes-us?qt-news_science_products=0#qt-news_science_products