Identifying Key Predictive Indicators of Seismic Events

Ed Hollingsworth
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA, USA
eahollin@vt.edu

Abstract

This project explores various aspects of seismic event datasets created by the US Geological Survey (USGS) and other entities, with the goal of better understanding the potential threats posed by these phenomena. Significantly, we were able to create non-parametric machine learning models for predicting the magnitude of unseen seismic events based on the attributes of historical records. Work was also done around predicting the cause of an earthquake based upon the location, depth, and magnitude of historical events.

1 Motivation

Natural hazards are an ever-present threat to our planet and our collective civilization. The more we can learn about such hazards, the more capable we are of protecting ourselves and our planet from their catastrophic impacts. Although earthquakes typically fall into this category, there are, in fact, significant seismic events that are human-induced. In addition to fracking, "building construction, carbon capture and storage, nuclear explosions, geothermal operations and research experiments that test fault stress"[6] can also cause earthquakes. This project explores the following relevant aspects of seismic phenomena:

1. *Categorization*: Can the cause of seismic phenomena be reliably determined using machine learning methods that evaluate characteristics of historical earthquakes based on location, depth, magnitude and other factors? Can seismic events be classified as human-induced or natural based on such factors?
2. *Prediction*: Can the magnitude of earthquakes be predicted using machine learning methods, given the location and depth of such events? What are the key indicators that can provide predictive insight into when and where earthquakes occur, and how powerful they will be?
3. *Correlation*: Can human-induced seismic activity be reliably correlated to seismic events tracked by sensors used by the US Geological Survey (USGS) and other agencies? Does the presence of human-induced seismic activity increase the likelihood of "natural" earthquakes in that location?

2 Dataset analysis

Three datasets were used for this project. Datasets *1* and *2* are largely duplicative of one another, but as Dataset *1* contains more data over a longer time period than Dataset *2*, it is used as the primary focus of analysis. Dataset *3* is the HiQuake dataset, which was used in conjunction with Dataset *1* for further analysis around *Correlation* of seismic events.

Preprint. Under review.

2.1 Dataset 1

The first dataset[1] contains details on global earthquakes between 1930 and 2018 and was retrieved from Kaggle.com at the following link: `https://www.kaggle.com/gustavobmgm/earthquakes-for-ml-prediction1`.

Although the data was originally sourced from USGS' ANSS Comprehensive Earthquake Catalog (ComCat)[1], it appears to have been customized by the Kaggle user who posted it. Dataset *1* contains 797,046 rows and 22 columns.

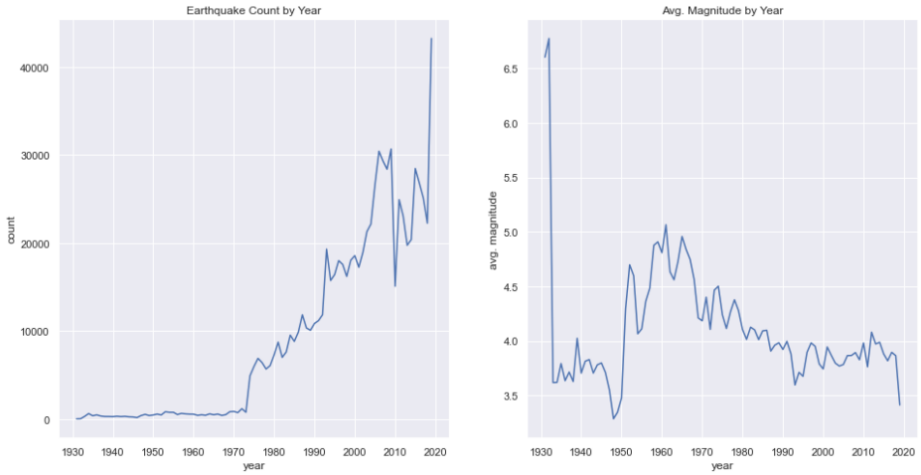

Figure 1: Event count by year, average magnitude by year.

Figure 1 plots the event count by year from Dataset *1*, as well as the average magnitude by year. Although these plots appear indicative of a trend in earthquake occurrence, as well as a shift in severity, I believe that they can mostly be explained by the increase in sensors and sensitivity, and just the accumulation of more seismic events than was possible in the earliest years of recording seismic activity.

2.2 Dataset 2

Dataset *2*[2] was published by the official USGS Kaggle account at `https://www.kaggle.com/usgs/earthquake-database`, and contains 23,412 rows and 21 columns.

Dataset *2* was used to explore the *Categorization* aspect of seismic phenomena by creating parametric machine learning models that categorize seismic events by their cause. Notably, 175 of the events in Dataset *2* were actually caused by nuclear explosions.

2.3 Dataset 3

The final dataset used for this project is the Human-Induced Earthquakes (HiQuake) dataset[7], which can be found at: `http://inducedearthquakes.org/`

Dataset *3* contains 1,235 rows and 36 columns. Table 1 provides the count of human-induced seismic events by Country for the top 10 counts.

3 Analysis and results

3.1 Geographic plots ("geoplots")

One of my earliest achievements with the datasets described in the previous section was the creation of geographic plots of the data. Figure 2 shows a plot of the nearly 800,000 seismic events in Dataset

[1] `https://earthquake.usgs.gov/data/comcat/index.php`

Table 1: Human-induced seismic events by country (top 10).

Country	Count
USA	540
China	156
Canada	104
Germany	37
Russia	27
Brazil	27
Australia	26
Japan	26
Netherlands	23
India	22

1, with transparent overlay of a Tectonic Plates reference map obtained from the National Geographic Society website.

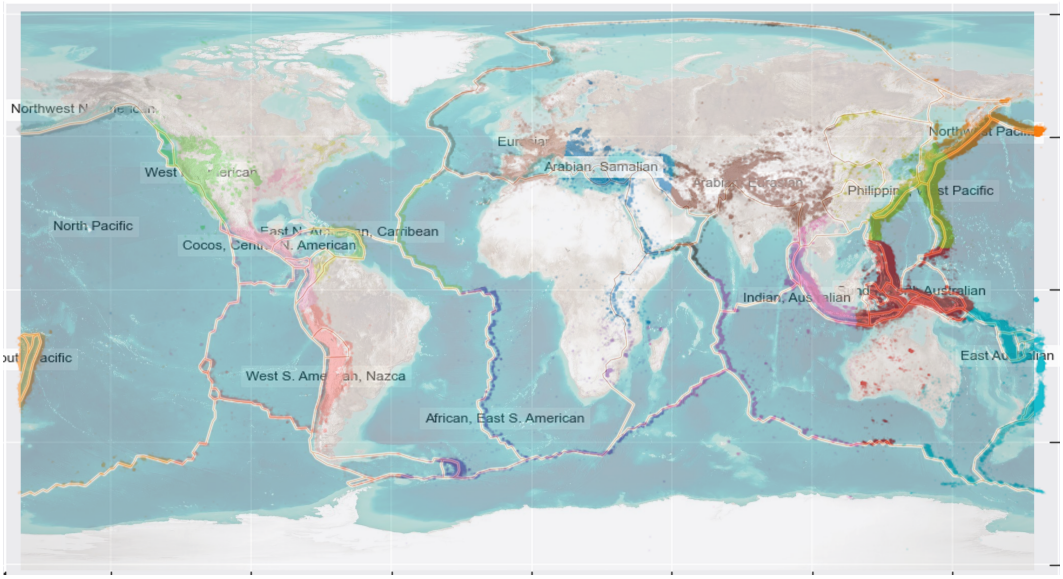

Figure 2: Geographic plot of dataset clusters, with tectonic plate overlay.

As can be seen, the seismic events correlate with striking accuracy to the tectonic plate boundaries. Obviously, this was an expected outcome, as there is a known association between plate boundaries and seismic activity, but it was still interesting to "prove" this correlation with nothing but raw data. I also produced a geoplot of the HiQuake dataset (Dataset *3*).

Both geoplots require overlay on a Equirectangular projection of the globe in order to line up the Latitude/Longitude coordinates properly. Additionally, I had to reverse the longitude coordinates to accommodate the east-to-west orientation of the values.

3.2 Clustering analysis

The color-coding and labels in Figure 2 reflect a K-means clustering that I ran against Dataset *1*, using a parameter of 16 clusters. I also ran the K-means clustering algorithm with a parameter of 8 clusters, but felt the 16 cluster version provided better results.

3.3 Categorization

In order to explore categorization of seismic events, specifically the prediction of cause (Earthquake, Nuclear Explosion, other Explosion) based upon historical attribute data found in Dataset *2*, I created

Table 2: Prediction Model Results

Model Attributes	MSE	R^2
1 Depth Only	0.60	0.28
2 Depth and Location (Lat/Long)	0.28	0.67
3 All Continuous Attributes	0.18	0.78

both a Decision Tree model, as well as a Naive Bayes model, both of which consistently demonstrated a 99% accuracy rate, using 5-fold cross-validation. The high degree of accuracy actually made me suspicious that I did something wrong, or at least fell victim to overfitting. Figure 3 presents a visual representation of the generated Decision Tree.

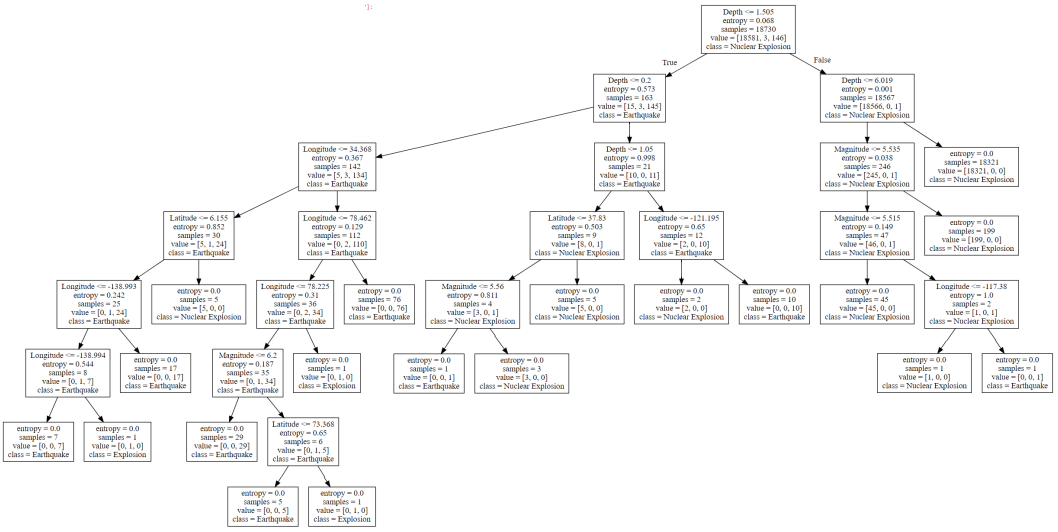

Figure 3: Decision tree to predict cause of seismic events.

3.4 Prediction

Using Dataset *1*, I developed a series of k-NN regression models to explore the predictive power of various sets of historical earthquake attributes, when used to predict the severity of unseen events. I attempted to use a linear regression model, as well, but the k-NN model consistently provided better results, so I placed the majority of my focus there. For the k-NN models, I used a StandardScaler to scale the attributes for uniformity.

Initially, I attempted to use the Latitude/Longitude coordinates and Depth of the historical events to create the k-NN model, and achieved underwhelming results. I then tested using only Depth to see if use of location was actually providing value, and proved conclusively that it indeed was. In an effort to achieve a higher accuracy rate, I then tried including all relevant, continuous attributes from the historical USGS record to produce a final model, which ended up producing the best results of all, ultimately arriving at a Mean Squared Error (MSE) of only 0.18. Table 2 lists the results for the three scenarios described above, and Figure 4 illustrate my findings for the final scenario.

In the referenced figure, the actual vs. predicted magnitude of the first forty test records are shown, side by side. The gap between the two is highlighted in either red (if the gap is greater than 0.5) or green (if the gap is less than or equal to 0.5). I used this cutoff to distinguish between what I considered a "bad" vs. "good" prediction. Unfortunately, since the first forty records represent an

arbitrary subset of the data, Figure 4 appears to show worse results than that of Latitude/Longitude coordinates and Depth, while overall for the entire test dataset, that is not the case.

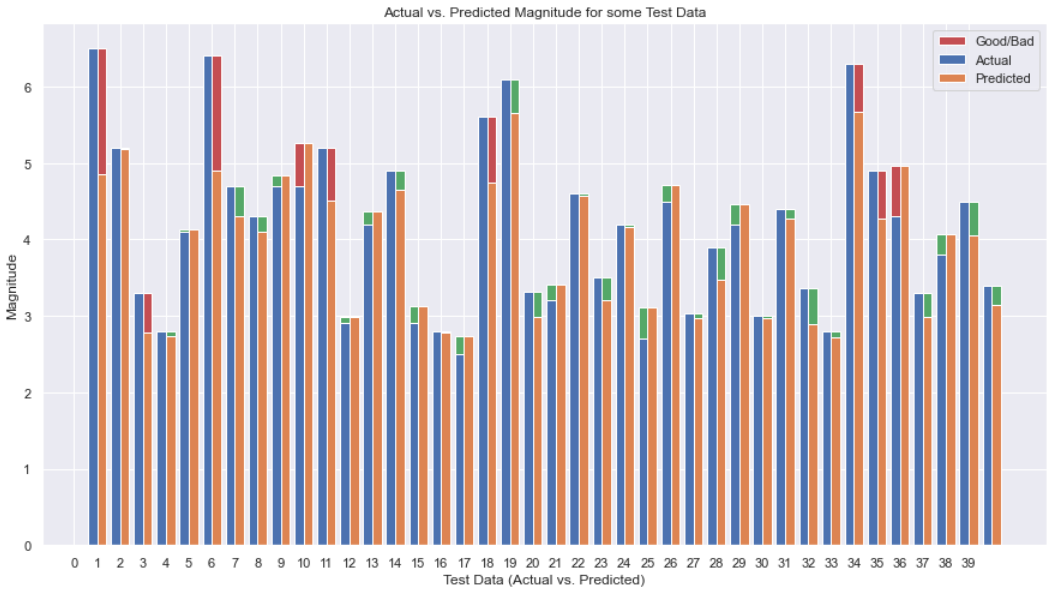

Figure 4: Actual vs. predicted severity, using all continuous attributes.

4 Conclusion

The most significant outcome of this project was the work done around prediction of earthquake magnitude, and what factors have the most predictive value. It was very clear through the experiments performed that depth alone is not a powerful enough determinant; at minimum, we also need to know where the event is physically occurring. This would lead one to the natural conclusion that the properties of the Earth's surface at various locations are such that two events occurring at the same depth may be more or less severe depending on where they occur.

Additionally, we found that the other attributes of the historical earthquake record also reinforce the severity prediction. When including all reasonable attributes from the USGS record, we were able to achieve an accuracy rate of .78 against our test data.

We also conclusively determined that a K-nn regression model produces far more accurate results than a Linear Regression model, when attempting to predict earthquake severity based on historical data.

5 Further research

I had also planned to explore the correlation between the events recorded in the HiQuake dataset and the events in Dataset *1*, with the goal of finding other events in Dataset *1* that are "nearby" the Human-induced events. While I was able to make some modest headway during the course of this project, ultimately, I did not have time to complete a thorough analysis of this area. This is one potential direction for additional research, though I'm sure there are many more, as well.

Seismic activity continues to be an area ripe for machine learning applications. The more we know about seismic phenomena, both natural and human-induced, the more prepared we will be to manage the risks they pose.

References

[1] Martins, Gustavo. (2021) "Earthquakes for ML prediction" dataset. Retrieved from kaggle.com, https://www.kaggle.com/gustavobmgm/earthquakes-for-ml-prediction,

[2] US Geological Survey. (2017) "Significant Earthquakes, 1965-2016" dataset. Retrieved from kaggle.com, https://www.kaggle.com/usgs/earthquake-database.

[3] OpenHazards.com. (2009-2022) "What is the significance of the depth of an earthquake?" Retrieved from OpenHazards.com, https://www.openhazards.com/faq/earthquakes-faults-plate-tectonics-earth-structure/what-significance-depth-earthquake.

[4] Martins, Gustavo. (2021) "Predicting Earthquakes using Machine Learning." Retrieved from Medium.com, https://medium.com/marionete/predicting-earthquakes-using-machine-learning-21689435dc52.

[5] US Geological Survey. (unknown) "Does fracking cause earthquakes?" Retrieved from usgs.gov, https://www.usgs.gov/faqs/does-fracking-cause-earthquakes.

[6] Shivni, Rashmi. (2017) "Human activity can trigger earthquakes, but how many? This number might surprise you." Retrieved from pbs.org, https://www.pbs.org/newshour/science/human-activity-can-trigger-earthquakes-many-number-might-surprise.

[7] The Human-Induced Earthquake Database. (2018) The Human-Induced Earthquake Database (HiQuake). Retrieved from inducedearthquakes.org, http://inducedearthquakes.org/.