Predictive Analysis of Earthquake Patterns in the USA

Hariharan Nadanasabapathi
UB ID: 50625272
hnadanas@buffalo.edu

Karthik Manjunath
UB ID:50625412
manjuna5@buffalo.edu

Naveen Manikandan
UB ID: 50625386
manikan2@buffalo.edu

I. PROBLEM STATEMENT

Earthquakes are unexpected natural disasters with the potential to cause severe damage to human life, infrastructure, and economic well-being of society. The sudden and violent nature of seismic movement of the earth's crust makes it very unpredictable, people get caught by its devastating impact. Analyzing previous trends in earthquakes and figuring out whether there are any future trends in seismic movement is crucial to reduce threats and prepare well. Our project's objective is to examine historical earthquake data to predict future seismic activity in the United States. Using the data from previous years, we aim to find trends, establish levels of risk, and guide decision-making in earthquake-prone regions.

A. Significance

This dataset is helpful as it provides a complete history of earthquakes in the US over years, enabling us to analyze seismic trends and patterns. By analyzing and predicting the future events we can:

1. **Understand long-term trends** of earthquakes
2. **Enhance predictive models** by studying data, and implementing statistical modelling techniques.
3. **Regional risk assessment** by identifying zones with high risk over years and enhance safety measures.
4. **Provide contribution** to the geology, seismology, and data science departments by working on advanced models to extract meaningful insights from large datasets.

B. Potential Impact

The major potential impacts of analyzing earthquake patterns over the years in the US includes:

1. **Improving** earthquake forecasting in order to increase accuracy and reliability.
2. Provide **early warnings** so that people can take cover and necessary actions to protect their homes.
3. Help government and rescue teams to **act quickly and effectively** on an issue.
4. Support global efforts to prepare and protect communities for **safer future**.

This not only contributes to earthquake pattern prediction but also sets the foundation for applying similar techniques to any other natural disaster.

II. DATA SOURCES

A. Source

The dataset used for this project is titled “**Earthquake**”. This dataset has been inspired from Kaggle but is sourced from USGS (U.S Geological Survey) website.

Link for the dataset:
https://www.kaggle.com/datasets/farazrahman/earthquake

USGS website: https://earthquake.usgs.gov/earthquakes/search/

The dataset we have chosen consists of merged data from the UGCS website covering earthquake data from January 2000 to February 2025, with magnitudes of 2.5 and above.

B. Description

The dataset comprises around 73,300 earthquake data. The dataset contains the following features, that are: **Time, Latitude & Longitude, Depth, Magnitude (mag), Magnitude Type (magType), NST -** the number of seismic stations that detected the event. **Gap -** largest gap in station coverage, affecting location accuracy. **Dmin -** minimum distance from the event to the nearest seismic station. **RMS -** root mean square error in the seismic data. **Net -** seismic network responsible for recording the event. **ID -** a unique id to an event. **Updated –** latest time the details were modified. **Place, Type -** the classification of the event. **Location Source, Magnitude Source (magSource) -** the organization that provided the magnitude estimate. **Horizontal Error, Depth Error, Magnitude Error, MagNst -** number of stations used in the magnitude calculation. **Status -** indicates whether the event’s data is preliminary or reviewed.

This dataset is a valuable resource for performing earthquake prediction analysis. It enables us to develop models using machine learning that are capable of trend analysis, risk-prone region identification, predictive insight generation, and so on.

III. MODEL EVALUATION

A. *Random Forest (RF) Regressor*:

We have chosen RF to predict continuous outcomes such as earthquake magnitude or depth. Random Forest is an ensemble of decision trees which trains random subsets of data and features, it then averages the results to improve accuracy and robustness. RF can capture non-linear relationships and handles tabular data well, making it suitable for the seismic dataset.

It can also provide an easy way to understand feature importance, helping us interpret which factors most influence predictions. We tuned hyperparameters such as the number of trees (estimators) and tree depth using grid search and cross-validation to prevent overfitting while maximizing predictive power.

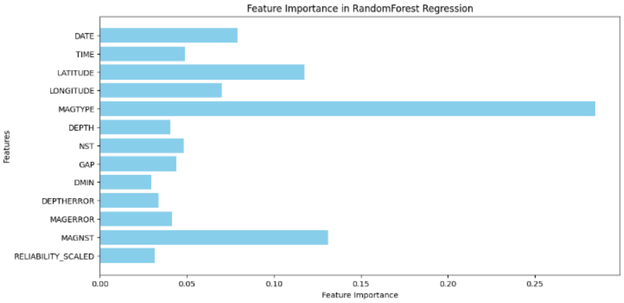


Figure 1: Feature Importance Bar Graph for RF

B.Extreme Gradient Boosting (XGBoost):

XGBoost is a powerful gradient boosting tree algorithm known for its efficiency and high predictive performance. XG has been the algorithm of choice in many data science competitions, often producing state-of-the-art results. We applied XGBoost for regression tasks (predicting Depth) given its ability to handle complex feature interactions and its built-in regularization to control overfitting. Key parameters (learning rate, max tree depth, number of estimators, L1/L2 regularization terms) were tuned using cross-validation. The justification for XGBoost lies in its superior accuracy on structured data and its scalability to large datasets, which in our case allows learning subtle patterns from tens of thousands of earthquake records. We understand that coordinates are important features for this model.

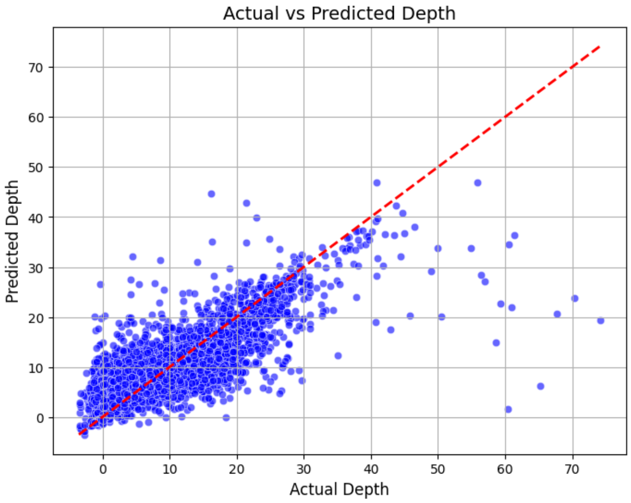


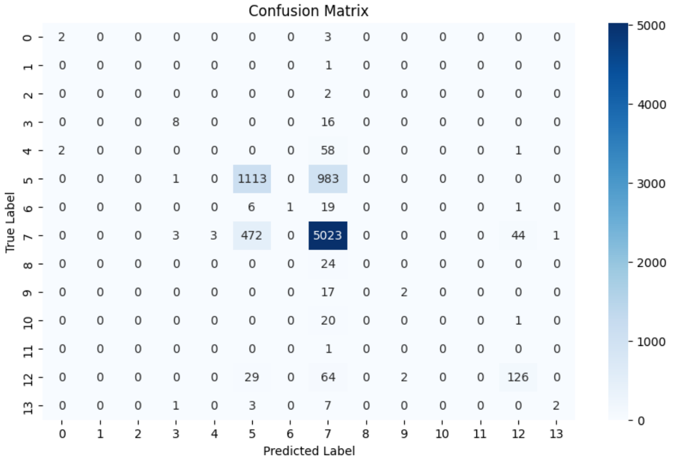
Figure 2: Feature Importance Bar Graph for XGboost

C. LightGBM:

LightGBM is yet another gradient boosting framework, which is optimized for speed and memory efficiency. We included LightGBM to compare against XGBoost and RF, as it can greatly speed up training (up to 20× faster) while achieving similar accuracy to traditional gradient boosting. LightGBM uses techniques like histogram-based splitting and exclusive feature bundling, which are advantageous for our dataset with many continuous features (like latitude, longitude, depth). We tuned similar hyperparameters (trees, learning rate, leaves) and observed training times much shorter than XGBoost, which is beneficial for iterative experimentation.

D. Logistic Regression:

For classifier tasks, we employed logistic regression as a baseline model. Logistic regression is a simple yet effective classifier that models probability of a binary outcome based on input features. In our context, we defined a classification problem such as distinguishing major earthquakes (e.g., magnitude ≥ 5.0) from less significant ones. Despite earthquakes being a continuous phenomenon, binarizing by magnitude threshold is useful for predicting the occurrence of damaging quakes. Logistic regression has been one of the most widely used models for binary classification problems since the 1970s. We included it to provide an interpretable reference point: its linear decision boundary and coefficients offer insights into how features like depth or location increase or decrease the odds of a major quakes. We optimized it with L2 regularization (ridge) to avoid overfitting and calibrated its probability threshold to handle class imbalance (since large quakes are relatively rare).



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	2	0	0	0	0	0	0	3	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
2	0	0	0	0	0	0	0	2	0	0	0	0	0	0
3	0	0	0	8	0	0	0	16	0	0	0	0	0	0
4	2	0	0	0	0	0	0	58	0	0	0	0	1	0
5	0	0	0	1	0	1113	0	983	0	0	0	0	0	0
6	0	0	0	0	0	6	1	19	0	0	0	0	1	0
7	0	0	0	3	3	472	0	5023	0	0	0	0	44	1
8	0	0	0	0	0	0	0	24	0	0	0	0	0	0
9	0	0	0	0	0	0	0	17	0	2	0	0	0	0
10	0	0	0	0	0	0	0	20	0	0	0	0	1	0
11	0	0	0	0	0	0	0	1	0	0	0	0	0	0
12	0	0	0	0	0	29	0	64	0	2	0	0	126	0
13	0	0	0	1	0	3	0	7	0	0	0	0	0	2

Figure 3: Confusion Matrix

E. Multi-Layer Perceptron (MLP) Regressor:

Exploring non-linear patterns beyond tree-based methods, we build a neural network regressor using a multi-layer perceptron. The MLP is a feed-forward artificial neural network that learns through backpropagation. Our MLP model consisted of one hidden layer (with 32 neurons) using ReLU activation, though we experimented with varying layers/neurons.

The MLP can in principle fit complex functions and interactions among features. We trained it to predict earthquake magnitude (as a regression) given features like location, depth, etc. The network was trained with mean squared error loss (appropriate for regression) and optimized with the Adam algorithm.

We standardized inputs as required, since MLPs are sensitive to feature scaling. The MLP model learns a non-linear mapping from inputs to output by adjusting weights via gradient descent. We expected the MLP to capture any intricate relationships that simpler models might miss, though at the cost of requiring more data and tuning (e.g., setting learning rate, number of epochs, and preventing overfitting with techniques like early stopping or regularization).

Effectiveness review: MLP neural network had comparable error (RMSE ~4.43, R² ~0.52), indicating it did not greatly outperform the simpler models; it may have been hampered by the relatively limited feature set and data quantity for training a complex model. As a baseline, a trivial model predicting the mean magnitude for all events yields an RMSE of ~0.45, so all our learned models improved upon this baseline to some degree.

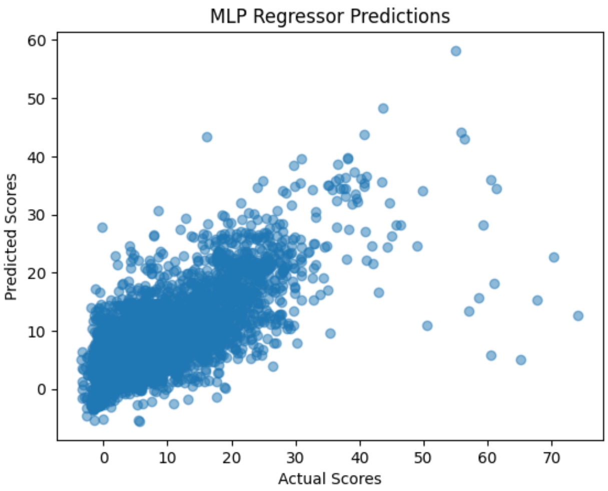


Figure 4: Predicted-Actual Depth Scatterplot

F.K-Means Clustering:

In addition to predictive modeling, we applied K-means clustering to the dataset for unsupervised pattern discovery. K-means aims to partition data points into K clusters such that each point belongs to the cluster with the nearest mean
.

We used features like geographic coordinates (latitude/longitude) and perhaps magnitude to cluster earthquakes into groups that might correspond to distinct seismic zones.

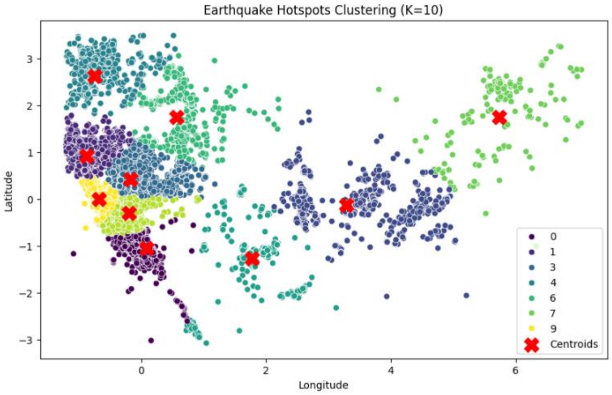


Figure 5: K-Mean Cluster Heatmap

For instance, we anticipated clusters differentiating West Coast quakes from Midwest or Eastern U.S. earthquakes. We determined an appropriate number of clusters K by evaluating the silhouette score, which measures how well-separated the resulting clusters are. The silhouette coefficient compares the distance of each sample to others in its own cluster versus other clusters. We tested K from 2 through 10 and found an optimal clustering at around K=4 (where average silhouette score was maximized). This unsupervised analysis provides an objective way to identify risk-prone areas based purely on historic seismic activity patterns, without any labels.

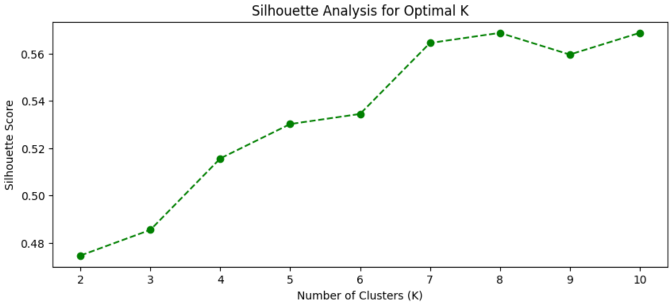


Figure 6: Optimal K Graph for silhouette score

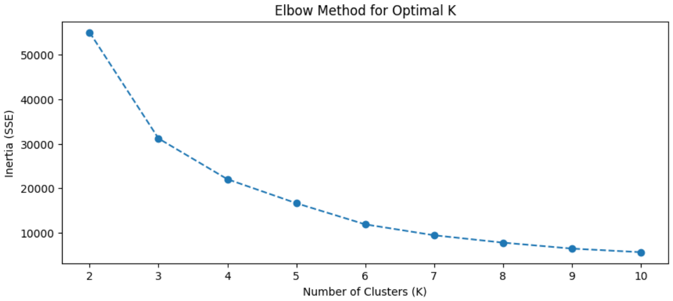


Figure 7: Optimal K Graph for Elbow method

IV. DISCUSSION OF RESULTS

Our results highlight both the strengths and limitations of using machine learning model for earthquake analysis. On the positive side, models effectively learned regional seismic patterns. For example, they could infer from location data whether an earthquake was likely to occur in a high-magnitude zone and adjust predictions accordingly. They also identified key trends, such as the dominance of certain states in seismic activity, and clustered earthquakes into known seismic zone, without any prior supervision. This demonstrates that data-driven approaches can replicate and quantify expert knowledge, such as recognizing that "Alaska and California experience frequent quakes" or that "plate boundary quakes behave differently from interior quakes." Additionally, these models can provide continuously updated assessments as new data becomes available. However, when it comes to predicting specific earthquake magnitude or classifying individual events as large or small the models faced fundamental limitations. Earthquake size is influenced by factors that weren't included in our dataset, such as stress accumulation on faults or precise fault-line positioning.

This aligns with established seismological understanding: while we can estimate probabilities, predicting the exact magnitude of an individual earthquake remains beyond our reach. For example, two earthquakes occurring in the same region at the same depth can still have very different magnitudes due to subtle differences in how the rupture unfold, something no simple dataset can fully capture.

One notable finding was that ensemble models (such as Random Forest, XGBoost, and LightGBM) consistently outperformed linear models, suggesting that non-linear patterns in the data play a vital role in prediction. XGBoost performed slightly better than Random Forest, likely because boosting focuses more on difficult-to-predict cases. LightGBM showed similar performance to XGBoost but trained faster, making it a strong option for scalability. Furthermore, neural network didn't significantly outperform the tree-based models. This may be due to the relatively small dataset or a lack of additional features for it to exploit. With more geological data or a much larger dataset, deep learning might prove more useful, but in this experiment, simpler models were both sufficient and easier to interpret.

Clustering results and trend visualizations provided useful context for the predictive modeling. They reinforced the idea that earthquake activity is highly location-dependent and that our models were essentially learning a mapping from location to typical earthquake magnitudes. This is like traditional seismic hazard maps, which categorize regions based on their seismicity rates. In this sense, machine learning serves as a complement to traditional statistical seismology, automatically identifying patterns such as "earthquakes in region X tend to be smaller" or "region Y frequently experiences large quakes." These insights help validate the model against known science.

In summary, while short-term or precise earthquake prediction remains unattainable (consistent with the scientific consensus), machine learning proves valuable for seismic risk assessment. These models can quickly analyze new data, update seismic activity maps, estimate the likely range of magnitudes in a given area, and potentially flag unusual pattern, such as a sudden swarm of moderate quakes in a typically quiet zone. These insights are valuable for earthquake early warning systems and hazard preparedness efforts.

V. REFERENCES

○ munichre.com

○ usgs.gov

○ kaggle.com

○ scikit-learn.org

○ papers.nips.cc

○ kdd.org

○ en.wikipedia.org

○ scikit-learn.org

○ en.wikipedia.org

○ medium.com

○ worldatlas.com

○ worldatlas.com

○ usgs.gov

○ en.wikipedia.org