**Predictive Modeling of Earthquake Magnitude Using Ensemble Learning: A Comparative Study of Random Forest Regressor, XGBoost, LightGBM, and CatBoost**

Saloni Dembla1, Divya Gautam2\*

STME, SVKM’s NMIMS, Indore, India

1salonidembla57@gmail.com

\**Corresponding Author* - 2divyagautam06@gmail.com

**Abstract**: Using a variety of seismic characteristics, this study attempts to create a predictive model for earthquake magnitude estimation. We used machine learning methods, like XGBoost, LightGBM, Random Forest and CatBoost, to examine the connection between earthquake magnitudes and seismic parameters. The Mean Squared Error (MSE) and R2 score were used to evaluate each model. With an R2 score of 0.7844 and an MSE of 0.1386, Random Forest scored better than the other models, with XGBoost coming in second. The findings show that ensemble learning methods—Random Forest in particular—offer a dependable method for predicting earthquake magnitude, providing important information for risk management and disaster preparedness.

**Keywords :** Earthquake, Random Forest Regressor, XGBoost, LightGBM, CatBoost

**1. Introduction**

Earthquakes are abrupt movements across faults that broadcast seismic waves that travel all across the planet and unleash elastic energy held in rocks. Every day there are about fifty earthquakes worldwide that are strong enough (magnitude > 2.5) to be felt locally, and every few days an earthquake occurs that is capable of damaging structures [1]. Additionally, there are numerous minor earthquakes having magnitude less than 2.5 occurring (Fig. 1) that are easily detected by contemporary instrumentation but are too weak to be felt. These small earthquakes provide valuable information about earthquake processes [2]. In the last few decades, with the rise in urbanization and population density, the effects of earthquakes have become more severe, posing an ever-growing threat to communities around the world. Understanding the issues caused by earthquakes, as well as effective management and prediction, is critical 1.2 for minimizing their impact.[3]

**1.1 Issues Caused by Earthquakes**

The earthquake's most immediate effect is the damage to infrastructure. Buildings, bridges, roads, and other structures are not often able to bear the seismic forces, causing either collapse or severe damage. This puts lives in danger-not just from the earthquake's direct-impact effects, but also from indirect-impact effects such as fires or landslides initiated by the damage to infrastructures. Another of these processes is flooding that comes about after an earthquake due to badly constructed embankments falling apart.

Besides the physical damages inflicted on the area, the earthquake disrupts the communication lines and utilities, including water, electricity, etc., which causes even more complications for rescues and relief. Moreover, the deprivation of clean water and medical facilities in the aftermath will cause another secondary round of health emergencies, such as disease outbreaks. The psychological impacts on survivors, including trauma and anxiety, will remain unattended for long.

The economic consequences of earthquakes can be devastating, with local economies directly suffering and, in extreme cases, grossly affecting the global economy. This is enormous when we include the costs associated with rebuilding improved and damaged infrastructure, homes, and businesses.

**1.2 Earthquake Management**

On this note, the management of earthquakes is very important in mitigating the impacts of earthquakes. Preparedness and planning can do a lot to reduce the dangers arising from earthquakes. Some of the factors include educating the public on safety measures, putting in place strict building codes that ensure structures can withstand seismic activity, and preparing appropriate disaster response mechanisms. Another focus for governments and organizations should be to develop highly equipped emergency management systems with rapid response teams, sufficient medical supplies, and clear communication channels through which the rescue operations can be coordinated.

Earthquake drills that involve evacuation plans are vital means of increasing public preparedness. These initiatives can serve a useful purpose by ensuring that individuals know what to do in an earthquake and where it is safe for them to go and thus will save lives. In an earthquake-prone area, it is crucial that people are provided with information regarding nearby safe zones, shelters, and evacuation routes so that such preparedness could save lives.

**1.3 Benefits of Earthquake Prediction**

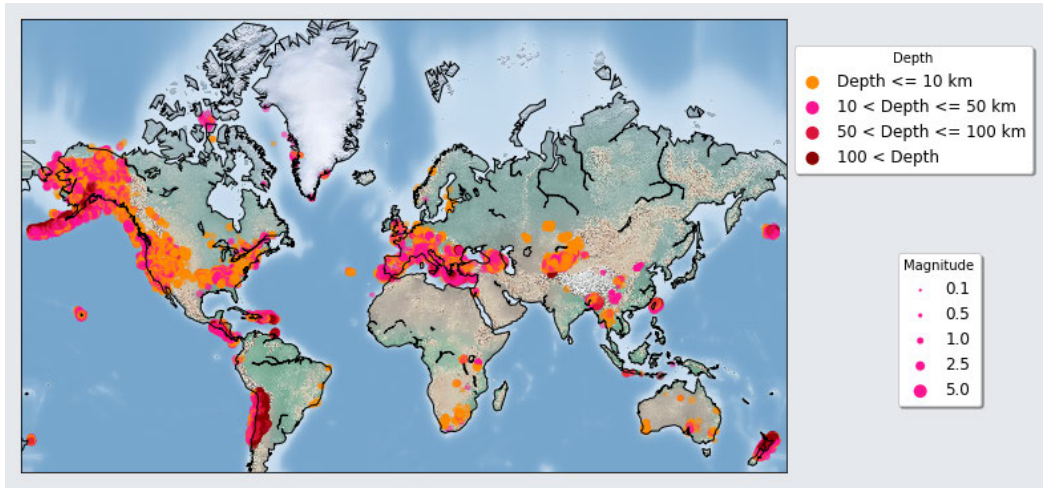
Earthquake early prediction holds tremendous benefits whereby even a few seconds of warning could allow people to take cover, assist with preventing industrial accidents, and halt critical infrastructure like trains or power plants from undergoing catastrophic failures.

Thus, early warning systems allow rapid initiation of emergency response protocols so that resources can be mobilized faster: this saves lives and reduces damage impact. While prediction technology today cannot be guaranteed 100% accurate, additional research and investment in this area shall expect more precise and timely earthquake forecasts in the future.

*Easing Evacuation and Reducing Economic Loss*

In fact, one of the most concrete advantages of an efficient earthquake management and prediction system is the ease and speed in which they would facilitate evacuations. Providing an early warning and tell-tale signs of danger and directing them through a preplanned evacuation course would at times save lives and minimize panic. Evacuation areas will be free of signs, and residents need to be acquainted with emergency procedures by trying to bring to-the-citizen awareness, especially at risk areas.

Even, by attempting to align all these prediction technologies with disaster management strategies, government and private organizations can proactively undertake measures to safeguard critical infrastructure from incurring huge losses due to disaster. Shutting down essential services, reinforcing structures, and organizing post-disaster recovery efforts beforehand can mitigate the damage caused by an earthquake and cut the total financial burden on the affected communities.



**Fig 1.** Location, size, and depth distributions of recorded earthquakes[3]

**2.Literature Review on Earthquake Prediction Using Machine Learning**

Earthquake prediction, therefore, has become the linchpin in ensuring livelihood and infrastructure safety, calling for solid predictive strategies against mustering evidences for its confirmations. Conventional earthquake prediction strategies, based primarily on analysis of historical seismic data, suffer from dependence on high accuracy and timely verification. Recently, machine learning (ML) methods have derived prominence in earthquake prediction for analyzing vast datasets, discovering patterns not easily recognizable by conventional methods.

Different seismic cases were attempted to be correlated to different machine learning models for predicting earthquakes. Ananthabhotla and Liu describe QuakeCast, making use of GNSS data for earthquake prediction, which allows the detection of pre-seismic anomalies in the atmosphere that may herald imminent seismic events[4]. This delineates early warnings ranging from hours to a week for disaster mitigation and response schemes. The model looks at over ten years of total electron content (TEC) data connected with thousands of earthquakes, thus showing the prospects of using signal from the atmosphere for earthquake forecasting.

The Random Forest Classifier for historical earthquakes to predict earthquake magnitudes was employed using the methods outlined in Bhardwaj and Kaushik[5]. Their methodology has proved to reduce false-positive forecasts by increasing prediction precision with ensemble learning techniques. Results indicated that potential seismic event identification by this model can be achieved with several input features based on seismic history.

Long short-term memory (LSTM)-based networks were implemented by Zhang and Wang for time-series seismic data referenced in[6]. The model accounts for temporal dependencies, thus providing a stronger basis for predicting earthquake timing and magnitude. Thus, the adoption of LSTM networks demonstrates that deep learning has truly excelled in enlightening complex temporal relationships inherent in seismic activity.

A hybrid model combining support vector machines with principal component analysis was presented by Singh and Verma to predict earthquake magnitudes from geophysical data sampling[7]. This method realized high accuracy in actual cases, thus indicating the high viability of hybrid techniques in machine learning applications.

The authors Kaur and Gupta applied the K-means clustering algorithm to determine earthquake-vulnerable regions based on a set of seismic and environmental parameters[8]. It underscores the spatial analysis importance relating to risk mitigation and preparedness.

**3. Research Aim**

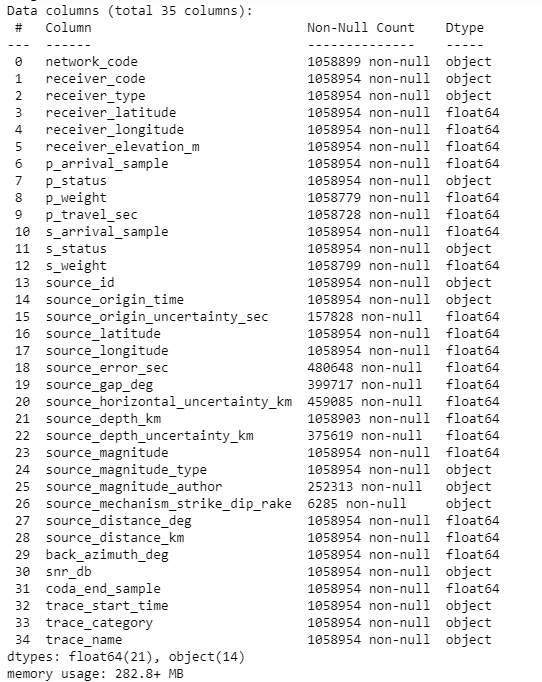
The intended study aims to create a high-accuracy earthquake magnitude prediction model based on AI regression techniques. The comparative performance evaluation of Random Forest Regressor and gradient boosting models (XGBoost, LightGBM, and CatBoost) will be performed using Mean Squared Error (MSE) and R2 score. The purpose is to select the best model for preparedness enhancement and disaster prediction by utilizing seismic data, from 2015 onward.

**4. Dataset Overview**

**4.1 Dataset Description**

The STEAD forms a holistic global data set of seismic waveforms, recorded for several decades. Using various features such as distance to the source, depth, and other conditions regarding the seismic occurrence, we aim to predict the magnitude of the earthquake source. The unique feature of our study is that the traditional seismological analysis uses the events of seismic waves (P-wave, S-wave) in evaluation, while here we use machine learning to predict the magnitude straight away, thereby providing an unprecedented application of ML in the geophysics domain. This study deals with-the earthquake-specific features to predict the source

magnitude.



**Fig.2. Dataset Information**

The collection (fig. 2) offers comprehensive data from receiver stations and seismic events. The station's location is specified by its receiver latitude and receiver longitude, and its height is indicated by its receiver elevation (meters). P-wave and S-wave arrivals are indicated by the P arrival sample and S arrival sample, respectively, with P and S status denoting quality. P travel time (seconds) logs P-wave travel time, and P weight and S weight evaluate dependability. Earthquakes are uniquely identified by their source ID, which specifies their position and depth using source latitude, source longitude, and source depth (km). Source horizontal uncertainty (km) and source error (seconds) show uncertainties, whereas source magnitude indicates intensity. Back azimuth (degrees) provides direction, while source distance (degrees and km) indicates source-receiver distance. Included are the event date, recording start time, coda end sample, and signal-to-noise ratio (dB).

**4.2 STEAD applications**

Earthquake detection, a challenging problem in earthquake seismology, is one of the first steps in data processing. Background noise insensitivity, a low number of false positives (which mistakenly identify non-earthquake signals as earthquakes), a low number of false negatives (which do not overlook small or weak earthquake signals), good generalization (not limited to a specific shape, range, or setting of earthquakes), and the ability to process large amounts of data are all characteristics of a good detection algorithm. The two main types of detection methods are similarity-search-based (e.g. [12]–[14]) and characteristic-function-based (e.g. [11]).

Additionally, STEAD may be used to directly identify the sites of earthquakes by machine learning techniques (e.g. [15]–[17]), which is a difficult topic that hasn't been fully resolved yet. One of the most crucial components included in seismic hazard evaluations are these models [18], [19]. Given a fictitious earthquake source, the strong motion is estimated using ground-motion prediction techniques. Ground-motion prediction equations are frequently developed using linear regression analysis [20], [21]. But ML has proven to be an effective method for creating these kinds of models [22]–[25]. Despite certain variations, STEAD has special features not found in most comparable audio data sets, such as millions of human-picked labels and additional information like known source and receiver locations.

**5. Data Cleaning**

**5.1 Handling Missing Values**

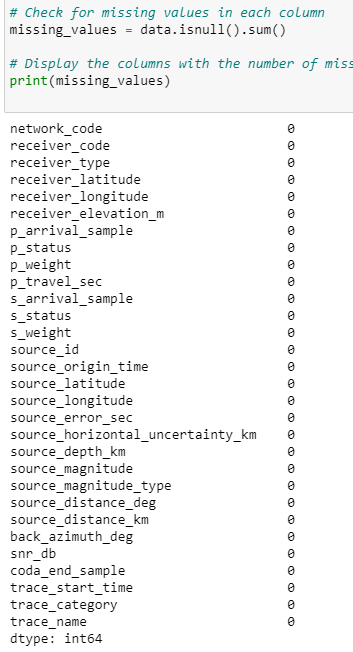
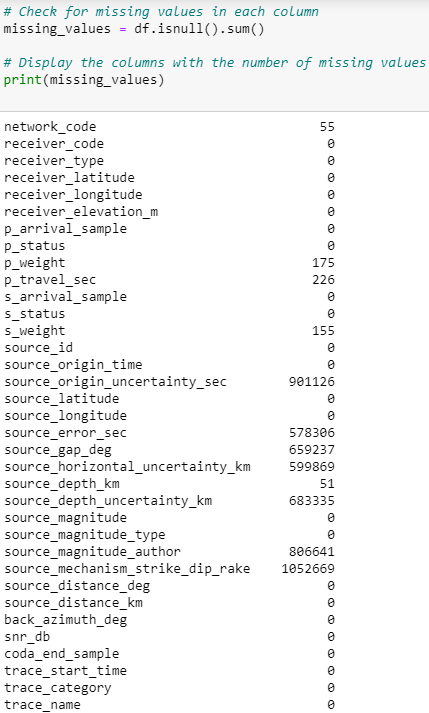
In the raw dataset, many columns contained missing values. To address this:

**Dropping Columns with Missing Data**:  
Columns with more than 40% missing values were deemed too sparse and were removed to avoid introducing skewed or unreliable results into the model.so we decided to remove a few unnecessary columns:

* source\_origin\_uncertainty\_sec: This column had a large proportion of missing values, indicating an uncertainty that couldn’t be reliably filled.
* source\_gap\_deg, source\_horizontal\_uncertainty\_km, and source\_magnitude\_author: These columns similarly contained too much missing data, which would have weakened the model’s effectiveness if included.

**Imputation of Remaining Missing Data**:

* **Numerical Columns**: Missing values in numerical columns were imputed using the median value of each related column (refer fig. 3). The median is a more trustworthy indicator than the mean since it is less vulnerable to extreme outliers. By doing this, the model was shielded from aberrant numbers that could have warped the distribution of the dataset as a whole.
* **Categorical Columns**: The mode, or most frequent value in the column, was used to fill in the missing values for categorical features. This prevented anomalies from being introduced while maintaining the categorical distribution's integrity.



**Fig.3.** Missing Values Information Before and After Cleaning

**5.2 Dropping Irrelevant Columns**

After addressing missing values, the next step involved removing irrelevant columns that were not useful for modeling earthquake magnitude prediction. These columns included metadata and redundant information that did not contribute meaningfully to the model:

* Metadata Columns: Columns such as network\_code, receiver\_code, receiver\_type, and trace\_name were dropped. These fields provided information about the recording process or station codes, but they were not directly related to predicting earthquake magnitudes.
* Redundant or Non-contributive Features: Columns like source\_magnitude\_type, which specifies the magnitude scale, were dropped as they are redundant given the goal to predict the magnitude itself. Additionally, time-related columns such as origin\_month, origin\_day, origin\_hour, and origin\_minute were removed, as the precise timestamp does not have a direct impact on the source magnitude prediction.

**6. Filtering Data for Analysis and Outlier Removal**

Once the dataset was cleaned, the next step involved refining the data by applying specific filters to focus the analysis on more recent earthquake events and remove outliers that could distort the predictive modeling process.

**6.1 Filtering for Recent Earthquakes**

To ensure a focus on a recent seismic activity and current dataset conditions, the rows with the earthquake origin year prior to 2015 are dropped.

* Filtering Data: The filtering for the dataset is done on the lines of:data = data[data[‘origin\_year’] > 2015]

**6.2 Removing Outliers**

The dataset was further refined, with outliers of several key features being removed, in order to avoid skewed predictions by the model.

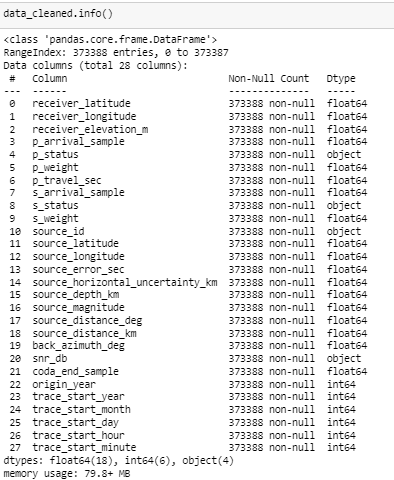
● Source Magnitude: For this analysis, outliers would be earthquakes with magnitudes of over 5. True, larger earthquakes are significant; however, for model applicability for common day-to-day instances, moderate earthquakes that tend to occur frequently are used.

● Source Depth: Earthquakes with depths greater than 200 km were considered extremely outlier-type. With most of the relevant seismic activity occurring in this range, this data set would leave out anything that lies outside the set limit.

● Source Distance: Source distance above 250 km was discarded. This thresholding would ensure the dataset retains meaning by selecting only nearby seismic events for which the source magnitude prediction is fairly relevant.

● P-wave Travel Time (p\_travel\_sec): Any data whereby P-wave travel time is above 1,000,000 seconds was seen as unrealistic and hence removed from the dataset. These huge values indicate the data may be showing anomalies or sensor errors.

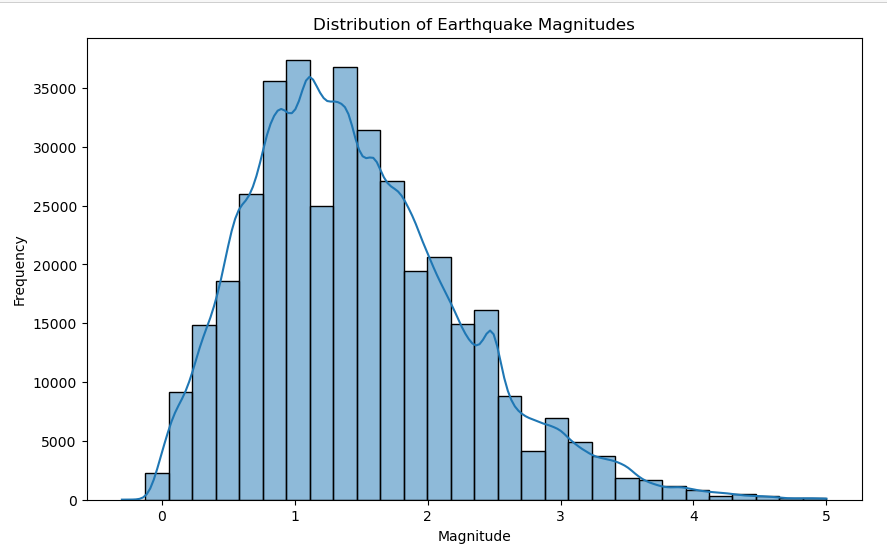
This filtering step would ensure that only relevant and accurate data would be available for modelling(refer fig. 4).



**Fig.4.** Cleaned Data Information

**7. Data Analysis**

**7.1 Distribution of Magnitude**

****

**Fig.5. Earthquake Magnitude Distribution**

The distribution of earthquake magnitudes for dataset values ranging from 0 to 5 is represented in the histogram above(refer fig.5). The x-axis shows earthquake magnitudes, while the y-axis shows frequency of occurrences at each magnitude level. Such data suggest that small to moderate earthquakes are most frequent, with most of the earthquakes occurring between 0.5 and 2.5 in magnitude with the highest frequency occurring between 1.0 and 1.5. The histogram is notably right-skewed, indicating that frequency of occurrence decreases with increasing magnitude. High-magnitude events especially those above 3.5, are comparatively rare, while the long right tail strengthening the observation that higher-magnitude earthquakes are less frequent than their lower-magnitude counterparts follows the global pattern of seismicity.The distribution thus illustrates the abundance of low to moderate magnitude earthquakes.

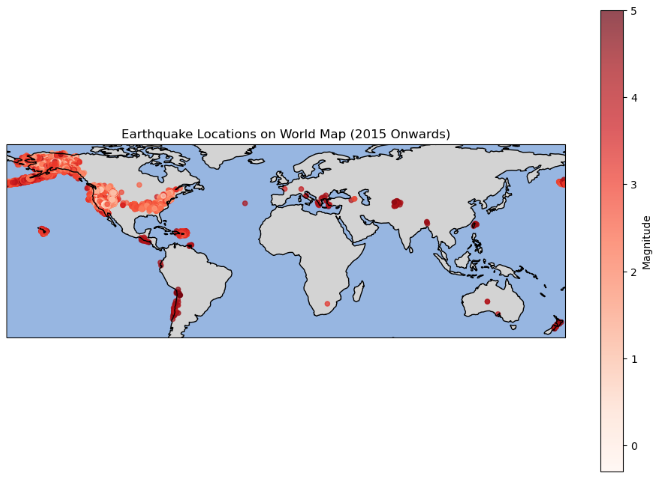
**7.2 Box Plot Across Different Magnitudes**

****

**Fig.6. Source Distance Across Different Magnitude Ranges**

The box plot displayed above examines the relationship between earthquake magnitude and the source distance across different magnitude ranges.(Refer fig. 6) The x-axis categorizes earthquakes into three groups based on magnitude: Small (0-3), Medium (3-6), and Large (6+), while the y-axis shows the source distance in kilometers. For small-magnitude earthquakes, the source distances are more concentrated and exhibit a lower range overall, with most distances falling between 20 km and 75 km, and a few outliers extending beyond 150 km. The medium-magnitude earthquakes, however, display a broader spread in source distances, with values ranging between 50 km and 150 km. The upper whisker reaches close to 200 km, indicating that medium-magnitude earthquakes generally occur further from the observation points compared to small ones. Interestingly, no data points are displayed for the large-magnitude range, possibly due to a lack of sufficient high-magnitude earthquake events in the dataset.

**7.3 Mapping Locations on World Map**



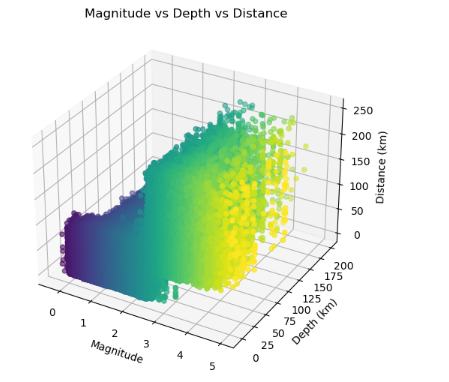
**Fig.7. Earthquake Locations on World Map**

The world map plot visualizes earthquake locations from 2015 onwards, with the earthquake magnitudes represented by varying shades of red. The darker red areas correspond to higher magnitude earthquakes, while the lighter red spots represent lower magnitudes(fig. 7).

The earthquake occurrences are concentrated in specific regions known for seismic activity. These regions include:

* The Pacific Ring of Fire, notably along the coasts of North and South America, Japan, and New Zealand, where a high concentration of earthquakes is evident.
* The Himalayan region, especially near India, Nepal, and Pakistan, is another active zone for seismic activity.
* Parts of Europe, such as Italy and Turkey, show moderate seismic occurrences.

**7.4 3-D Scatter Plot Between Physical Factors**

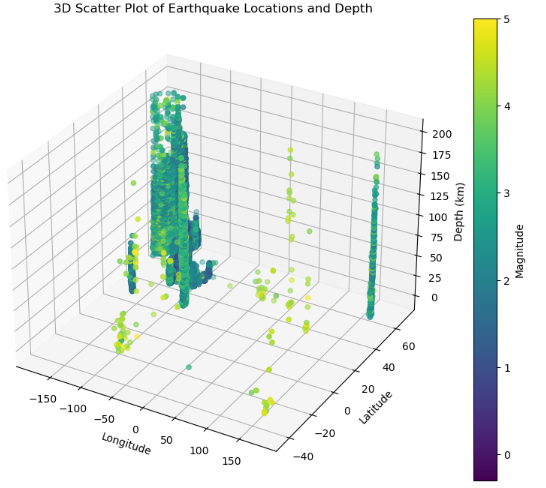
****

**Fig.8. 3D Scatter Plot of Magnitude vs Depth vs Distance**

Earthquake magnitude, depth, and distance relationship has been depicted in 3D scatter graph(Fig.8) deduced from STEAD dataset. While the y-axis and z-axis depict the source depth (in kilometers) and source distance (in kilometers), respectively, the x-axis shows the magnitude of the earthquake.

The plot reveals a concentration of earthquake events with magnitudes predominantly between 1 and 3, depths ranging from 0 to 25 km, and source distances clustered around 50 to 150 km. As the depth increases, earthquakes with greater magnitudes become less frequent, and their distances from the source generally extend further, reflecting the geological behavior of seismic events. The color gradient further emphasizes this relationship, with darker points representing lower-magnitude earthquakes and lighter points indicating higher-magnitude events.

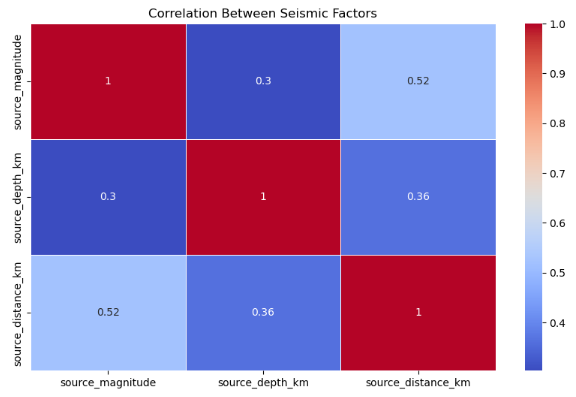
This 3D visualization effectively demonstrates the correlation between earthquake magnitude and its depth and distance from the source. The plot suggests that while shallow earthquakes occur more frequently, larger magnitudes tend to be associated with greater depths and distances, an observation consistent with common seismic patterns. The overall distribution offers valuable insight into the spatial characteristics of earthquakes in the dataset.



**Fig.9. 3D Scatter Plot of Earthquake Locations vs Depth**

The 3D scatter plot presents the distribution of earthquake events as a function of latitude, longitude, and depth, while each point corresponds to a color intensity that denotes the size of the event(refer fig 9). The X and Y axes are the geographical coordinates (longitude and latitude), while the clustering signifies areas of increased seismic activity. On the Z-axis, the depth of the earthquakes is in kilometers, which shows that most events are shallow, with some down to about 200 km. The magnitude is color-coded: yellow for larger earthquakes (almost magnitude 5), while blueish-green signifies the smaller. The plot suggests that larger earthquakes are quite scarce and display a varying depth-range, while smaller earthquakes dominate and are shallow. This visualization gives a good overview of the relationships between geographical location, earthquake depth, and magnitude, enabling the recognition of seismic patterns across different areas.

**7.5 3-D Correlation Plot**



**Fig 10. Correlation between various factors**

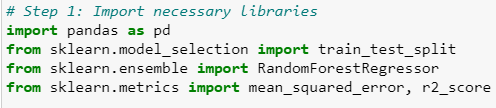
The heatmap(fig 10) shows the interrelation among three different seismic parameters, source\_magnitude, source\_depth\_km, and source\_distance\_km. Moreover, it indicates that the distance between source and observation site becomes longer as the earthquake magnitude increases since source\_magnitude shows a relatively positive correlation (0.52) with source\_distance\_km. There exists a lower correlation (0.3) between source\_magnitude and source\_depth\_km, which suggests that there is a less close relationship between magnitude and depth of sources of earthquake. Another relationship shown is a weak (0.36) relationship between source\_depth\_km and source\_distance\_km. This indicates that those earthquakes are more likely to be deeper and may have a slight connection with the distance from the source, but not as much with the magnitude. Such relations could help us better understand the behavior of these variables during earthquakes.

**8. Model Development for predicting earthquakes**

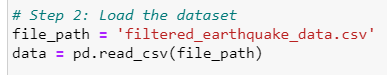
**8.1. Random Forest Regressor**

With the help of several decision trees, one can combine their learning output, which has the advantage of reducing the extent of overfitting and providing more stable and accurate predictions than one can achieve using just one decision tree. A model of this type is built as a whole collection of decision trees or a "forest," and after this, the prediction produced through the voting-deciding process is finally produced. It is because of this ensemble approach that Random Forest makes much accurate results while modeling complex and noisy datasets such as earthquake databases, which contain non-linear relationships in terms of factors like depth, magnitude, and distance from the source itself.

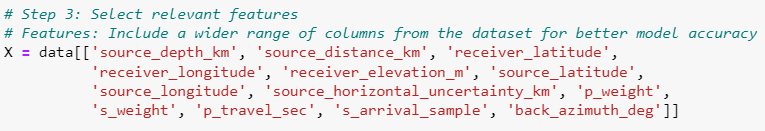
**Step 1**



**Step 2**



**Step 3**



**Step 4**



**y** is the target variable you are trying to predict, which is the **magnitude** of the earthquake (source\_magnitude). This represents the strength of the seismic event.

**Step 5**



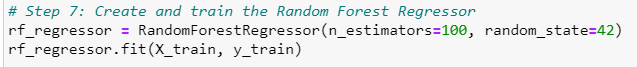
Any missing values in the feature matrix X are replaced with 0. This ensures that the dataset is concise and ready for modeling.

**Step 6**



Training (80%) and testing (20%) sets of data are separated.

**Step 7**



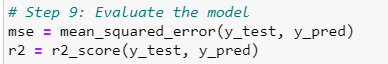
A RandomForestRegressor model is created with 100 trees (n\_estimators=100), meaning the model will train 100 decision trees and average their predictions.

**Step 8**

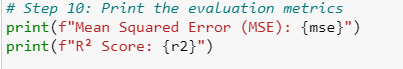


The trained model makes predictions on the test data (X\_test). The predicted values are stored in y\_pred.

**Step 9**



**Step 10**





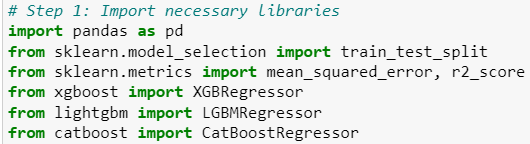
* **Mean Squared Error (MSE)**: 0.13859 — This is the average squared difference between the actual and predicted earthquake magnitudes. A lower MSE indicates that the predictions are good.
* **R² Score**: 0.78438 — This indicates that around **78.4%** of the variance in the target variable (earthquake magnitude) is explained by the model. The closer this value is to 1, the better the model explains the variance in the data.

**8.2. Boosting Models (XGBoost, LightGBM, CatBoost)**

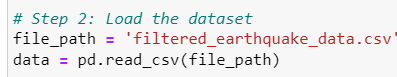
These boosting models- XGBoost, LightGBM, and CatBoost- are better known for refining the prediction accuracy in successive iterations correcting errors. For instance, XGBoost (Extreme Gradient Boosting) is one of the most favored models in machine learning contests and real-world applications due to its efficacious ways in managing missing values and capacity to minimize overfitting through regularization. With its gradient-boosting structure ensuring that every new tree keeps its focus on errors made by the previous trees, the model is quite effective in locating patterns in data that are very "rocky".

Moreover, LightGBM and CatBoost offer additional advantages of boosting in speeding and specific data types. LightGBM is designed with enormous datasets in mind, and trees will build in a leaf-wise rather than level-wise manner, meaning that it will train faster for large datasets where training time is a factor. CatBoost does well in managing categorical variables because it manages automatically noisy data with almost no preprocessing.

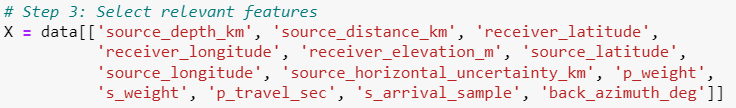
**Step1  
Importing**



**Step 2  
Loading**



**Step 3  
Select relevant features**



**Step 4  
Define the target variable**



**Step 5  
Handle missing values**



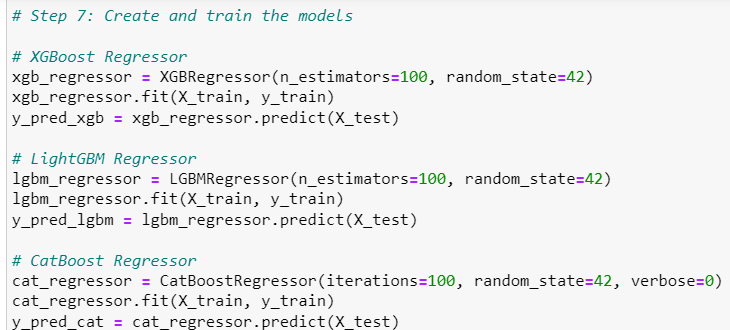
Any missing values in the dataset are filled with 0 using X.fillna(0, inplace=True).

**Step6  
Train-test split**



Spliting the data into training and test sets using an 80-20 ratio (test\_size=0.2).

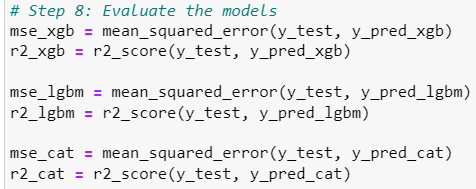
**Step 7  
Create and train the models**



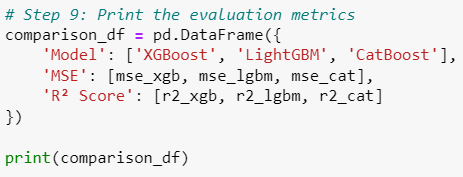
Three separate models (XGBoost, LightGBM, CatBoost) are created and trained:

* **XGBoost**: A popular gradient boosting algorithm known for speed and performance. The model uses 100 estimators (trees) and random\_state=42 for reproducibility.
* **LightGBM**: Another gradient boosting algorithm optimized for speed, especially with large datasets. Like XGBoost, it uses 100 estimators and is trained with the same training data.
* **CatBoost**: A gradient boosting algorithm that’s particularly efficient with categorical data, although here it’s applied to numeric features. It’s set to run for 100 iterations.

**Step 8  
Evaluate the models**



**Step 9  
Print the evaluation metrics**



* The performance of all three models is displayed in a DataFrame for comparison.

**Model Performance Summary**

* **XGBoost** achieved the lowest **MSE** of 0.156662 and an **R² score** of 0.756625, indicating it has performed better in predicting earthquake magnitude compared to LightGBM and CatBoost.
* **LightGBM** had a **MSE** of 0.170578 and an **R² score** of 0.734613.
* **CatBoost** gave a **MSE** of 0.168288 and an **R² score** of 0.738176.

**Key Takeaways**

* XGBoost is the best-performing model among the three in terms of both **MSE** and **R² score**.
* While LightGBM and CatBoost are still strong contenders, XGBoost is preferable if you prioritize accuracy for your dataset.

**9. Results and Discussion**

**9.1. Model Comparisons**

In our study, we employed several machine learning models to predict earthquake magnitude based on various features such as source depth, source distance, receiver latitude, and other seismic attributes. We tested three advanced Gradient Boosting techniques (XGBoost, LightGBM, and CatBoost), along with RandomForestRegressor, to evaluate which algorithm provided the best performance for this task.

**Performance Comparison:**

|  |  |  |
| --- | --- | --- |
| MODEL | MEAN SQUARE ERROR | R2 SCORE |
| Random Boost | 0.138592 | 0.784378 |
| XGBoost | 0.156662 | 0.756625 |
| LightGBM | 0.170578 | 0.734613 |
| CatBoost | 0.168288 | 0.738176 |

**9.2. Analysis of Results**

**Random Forest Regressor**

Random Forest outperformed all of the models in means of the Mean Square Error and R2 score. It has MSE=0.138592 and an R2 score of 0.784378; hence this model has proved its worth in handling high-dimensional data with quite complex relations between the features and target variable.

Another advantage that helped differentiate Random Forest was its ensemble property; the Random Forest model averages over a large number of decision trees to reduce overfitting. The ability to evaluate different features and the robustness against missing values and outliers make it suitable for the earthquake magnitude prediction problem.

**XGBoost**

The second-best was XGBoost, a popular boosting algorithm, with R2 of 0.756625 and an MSE of 0.156662. Optimization benefits for XGBoost, like regularization, parallelism, and its ability to handle missing data effectively, have made it widely accepted.

Since it reduces model bias and variance and yields a relatively high prediction accuracy, XGBoost, therefore, found itself emerging as the main competitor in our study; however, its performance lagged slightly behind Random Forest, and this could have been due to the structure of the dataset and hyperparameter sensitivity of the model.

**LightGBM**

LightGBM recorded an R2 score of 0.734613 and MSE of 0.170578. In spite of this level of performance, it was only ranked third. While LightGBM is an extremely fast, efficient gradient boosting framework, it may not have captured some subtleties in the data as Random Forest or XGBoost did.

The scalability and speed advantage of LightGBM make it highly efficient for big data. However, in our datasets, there occurs a trade-off between speed and accuracy, and therefore, it is able to perform relatively poor than Random Forest and XGBoost.

**CatBoost:**

With respect to MSE=0.168288 and R²=0.738176, CatBoost performed fairly well but did not outperform Random Forest or XGBoost. CatBoost is designed to manage categorical features and is often considered a state-of-art approach in those domains where such types of features are predominant. Since our dataset consisted mostly of numerical features, CatBoost's advantages in this department could not be fully exploited.

In our comparison, CatBoost's automated feature encoding and high-level treatment of categorical data helped, but Random Forest and XGBoost gave somewhat better generalization for this particular problem.

**9.3. Key Takeaways:**

* **Random Forest** was the most appropriate model for earthquake magnitude prediction because it performed the best in terms of both MSE and R². The most accurate forecasts were produced by its capacity to manage intricate patterns in the data.
* **XGBoost**, while performing slightly worse than Random Forest, still provided strong results and would be a good alternative, especially in scenarios where model interpretability and hyperparameter tuning are important.
* **LightGBM** and **CatBoost** showed competitive performance, with **LightGBM** providing faster training times, while **CatBoost** excelled in automatically handling feature types.

Overall, this comparative analysis shows that while all models performed reasonably well, **Random Forest** stood out as the most accurate model for this dataset. By exploring multiple algorithms, we were able to better understand the strengths and limitations of each, helping us refine our approach to earthquake magnitude prediction.

**10. Conclusion**

Although we've been trying different ML models, the Random Forest model has appeared to be the best and most reliable in predicting earthquake magnitudes. XGBoost also scored highly, which is a compelling alternative. The study establishes the point that techniques on ensemble learning like Random Forest and boosting algorithms such as XGBoost are powerful for predictive modeling on complex datasets like seismic events.

Future work from this study opportunity can involve optimizations on hyperparameters tuning of the particular model, new advanced algorithms, or more seismic feature incorporations, thus attaining even better accuracies.

**References**

[1]P. M. Shearer, Introduction to Seismology. Cambridge, U.K.: Cambridge  
Univ. Press, 2009.

[2]E. E. Brodsky, ``The importance of studying small earthquakes,'' Science,  
vol. 364, no. 6442, pp. 736737, 2019.

[3] STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI  
S. Mostafa Mousavi , Yixiao Sheng, Weiqiang Zhu , And Gregory C. Beroza

[4]B. Ananthabhotla and J. Liu, "QuakeCast: Earthquake Forecasting Using Machine Learning to Identify and Classify Preseismic Ionospheric Anomalies," MIT Lincoln Laboratory Project Report TIP-159, 2022.

[5]S. Bhardwaj and A. Kaushik, "Machine Learning Algorithms for Earthquake Prediction Using Seismic Data," 2023.

[6]M. Zhang and Y. Wang, "Deep Learning for Earthquake Prediction Using Seismic Time-Series Data," 2021.

[7]R. Singh and P. Verma, "Hybrid Machine Learning Model for Earthquake Magnitude Prediction," 2020.

[8]H. Kaur and D. Gupta, "Prediction of Earthquake-Prone Regions Using ML Techniques," 2022.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, ‘‘ImageNet large scale visual recognition challenge,’’ Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015. doi: 10.1007/s11263- 015-0816-y.

[10] K. J. Bergen, P. A. Johnson, V. Maarten, and G. C. Beroza, ‘‘Machine learning for data-driven discovery in solid Earth geoscience,’’ Science, vol. 363, no. 6433, p. eaau0323, 2019.

[11] R. V. Allen, ‘‘Automatic earthquake recognition and timing from single traces,’’ Bull. Seismol. Soc. Amer., vol. 68, no. 5, pp. 1521–1532, 1978.

[12] S. J. Gibbons and F. Ringdal, ‘‘The detection of low magnitude seismic events using array-based waveform correlation,’’ Geophys. J. Int., vol. 165, no. 1, pp. 149–166, 2006.

[13] C. E. Yoon, O. O’Reilly, K. J. Bergen, and G. C. Beroza, ‘‘Earthquake detection through computationally efficient similarity search,’’ Sci. Adv., vol. 1, no. 11, 2015, Art. no. e1501057.

[14] K. Rong, C. E. Yoon, K. J. Bergen, H. Elezabi, P. Bailis, P. Levis, and G. C. Beroza, ‘‘Locality-sensitive hashing for earthquake detection: A case study of scaling data-driven science,’’ Proc. VLDB Endowment, vol. 11, no. 11, pp. 1674–1687, 2018.

[15] M. Kriegerowski, G. M. Petersen, H. Vasyura-Bathke, and M. Ohrnberger, ‘‘A deep convolutional neural network for localization of clustered earthquakes based on multistation full waveforms,’’ Seismol. Res. Lett., vol. 90, no. 2, pp. 510–516, 2018.

[16] X. Zhang, J. Zhang, C. Yuan, S. Liu, Z. Chen, and W. Li, ‘‘Locating earthquakes with a network of seismic stations via a deep learning method,’’ 2018, arXiv:1808.09603. [Online]. Available: https://arxiv. org/abs/1808.09603

[17] D. T. Trugman and P. M. Shearer, ‘‘GrowClust: A hierarchical clustering algorithm for relative earthquake relocation, with application to the Spanish Springs and Sheldon, Nevada, earthquake sequences,’’ Seismol. Res., vol. 88, no. 2A, pp. 379–391, 2017.

[18] S. M. Mousavi, G. C. Beroza, and S. M. Hoover, ‘‘Variabilities in probabilistic seismic hazard maps for natural and induced seismicity in the central and eastern United States,’’ Lead. Edge, vol. 37, no. 2, pp. 1–141, 2018. doi: 10.1190/tle37020141a1.1.

[19] S. M. Mousavi and G. C. Beroza, ‘‘Evaluating the 2016 one-year seismic hazard model for the central and eastern United States using instrumental ground-motion data,’’ Seismol. Res. Lett., vol. 89, no. 3, pp. 1185–1196, 2018. doi: 10.1785/0220170226.

[20] Y. Bozorgnia et al., ‘‘NGA-West2 research project,’’ Earthquake Spectra, vol. 30, no. 3, pp. 973–987, 2014.

[21] B. Derras, P. Y. Bard, and F. Cotton, ‘‘Towards fully data driven ground motion prediction models for Europe,’’ Bull. Earthquake Eng., vol. 12, no. 1, pp. 495–516, 2014.

[22] D. T. Trugman and P. M. Shearer, ‘‘Strong correlation between stress drop and peak ground acceleration for recent M 1–4 earthquakes in the San Francisco Bay area,’’ Bull. Seismol. Soc. Amer., vol. 108, no. 2, pp. 929–945, 2018.

[23] B. Derras, P.-Y. Bard, F. Cotton, and A. Bekkouche, ‘‘Adapting the neural network approach to PGA prediction: An example based on the KiKnet data,’’ Bull. Seismol. Soc. Amer., vol. 102, no. 4, pp. 1446–1461, 2012.

[24] A. Alimoradi and J. L. Beck, ‘‘Machine-learning methods for earthquake ground motion analysis and simulation,’’ J. Eng. Mech., vol. 141, no. 4, 2014, Art. no. 04014147.

[25] A. H. Alavi and A. H. Gandomi, ‘‘Prediction of principal groundmotion parameters using a hybrid method coupling artificial neural networks and simulated annealing,’’ Comput. Struct., vol. 89, nos. 23–24, pp. 2176–2194, 2011.