Pilot Research Report: Epoch One

Kazi Sakib Hasan
Most. Afia Anjum Borsha

Epoch One – Independent AI Collective
BRAC University

November 22, 2025

Abstract

This pilot research report documents an initial investigation into data-driven mainshock identification using historical earthquake records. The study leverages a global seismic dataset from 1930 to 2018, applies feature engineering and spatiotemporal labeling of earthquake sequences, and evaluates gradient-boosting models for predictive accuracy. Preliminary results demonstrate meaningful predictive signals for mainshock events, highlighting the potential for machine learning–based early-warning frameworks.

1 Title

Project Name:
Machine Learning–Driven Mainshock Identification Using Seismic Spatiotemporal Patterns and Enhanced Earthquake Feature Engineering

Project ID:
E1-EQ

2 Motivation

In recent months, Bangladesh has experienced a surge in seismic activity, including a 5.7-magnitude earthquake near Narsingdi and Ghorashal, resulting in fatalities, structural damage, and public panic. For a developing country with weaker infrastructures and high urban density, distinguishing whether an initial quake is a *mainshock* or a *foreshock* is critically important. Traditional seismological methods cannot make this distinction before subsequent events unfold, leaving communities vulnerable.

- Challenges in real-time mainshock identification
- Potential impact for national earthquake preparedness and hazard mitigation
- Necessity of data-driven AI approaches to supplement conventional methods

3 Pilot Study Abstract

This study investigates whether machine learning models can identify mainshocks using global historical earthquake data. The dataset spans 1930–2018 and contains over 797,000 earthquake events with latitude, longitude, depth, magnitude, and seismic station measurements. The methodology involves cleaning, magnitude standardization (all converted to Mw), iterative imputation of missing values, and extraction of temporal and spatiotemporal features. Each earthquake was algorithmically labeled as a foreshock, mainshock, or aftershock based on temporal (30-day) and spatial (50 km) windows using KDTree-based nearest-neighbor search. Foreshocks and aftershocks were later removed to focus on mainshock prediction.

Gradient-boosting models (XGBoost and LightGBM) were trained and evaluated. XGBoost achieved 78% accuracy with an AUC of 0.851, while LightGBM achieved 76% accuracy with an AUC of 0.830. SHAP analysis indicated that magnitude, latitude, longitude, depth, and time since previous quake were the most influential features. Results demonstrate that mainshock events can be predicted with meaningful accuracy, laying the groundwork for AI-assisted seismic hazard assessment.

4 Methodology

4.1 Data Collection

- Source: Kaggle dataset "Earthquakes for ML Prediction" by Gustavo Martins; data provided by USGS
- Data type: numerical features of earthquakes (latitude, longitude, depth, magnitude, RMS, station info, etc.)
- Size: 797,046 events covering 1930–2018

4.2 Data Preprocessing

- Conversion of all magnitudes to moment magnitude (`mag_in_mw`)
- Chronological sorting of events and timestamp conversion to datetime
- Iterative imputation of missing values in depth, nst, gap, dmin, rms
- Extraction of temporal features: year, month, day, hour, minute, second, day of week, time since previous quake
- Seismic sequence labeling: foreshock, mainshock, aftershock (using 30-day and 50 km windows with KDTree)
- Retention of only mainshock labels for modeling

4.3 Model/Approach

- Models: XGBoost and LightGBM (gradient boosting classifiers)

- Training procedure: standard train-test split, default hyperparameters with early stopping

- Evaluation metrics: Accuracy, Precision, Recall, F1-score, AUC

4.4 Experimental Setup

- Hardware/Software: Python 3.10, scikit-learn, XGBoost, LightGBM, pandas, NumPy, Jupyter Notebook

- Dataset split: 80% training, 20% testing, stratified on target (class distribution preserved)

5 Results and Discussion

5.1 Quantitative Results

Table 1: Model Performance Metrics

Model	Accuracy (%)	Precision	Recall	AUC
XGBoost	78	0.78	0.86	0.851
LightGBM	76	0.76	0.86	0.830

5.2 Qualitative Discussion

- Magnitude (`mag_in_mw`), latitude, longitude, depth, and temporal recurrence were the most influential features according to SHAP analysis, as shown in Figure 1.

- XGBoost slightly outperformed LightGBM in overall accuracy and AUC.

- Results demonstrate that meaningful signals exist in historical earthquake data for mainshock identification.

- Limitations: only event-level features considered; foreshock/aftershock temporal sequences not explicitly modeled in this pilot.

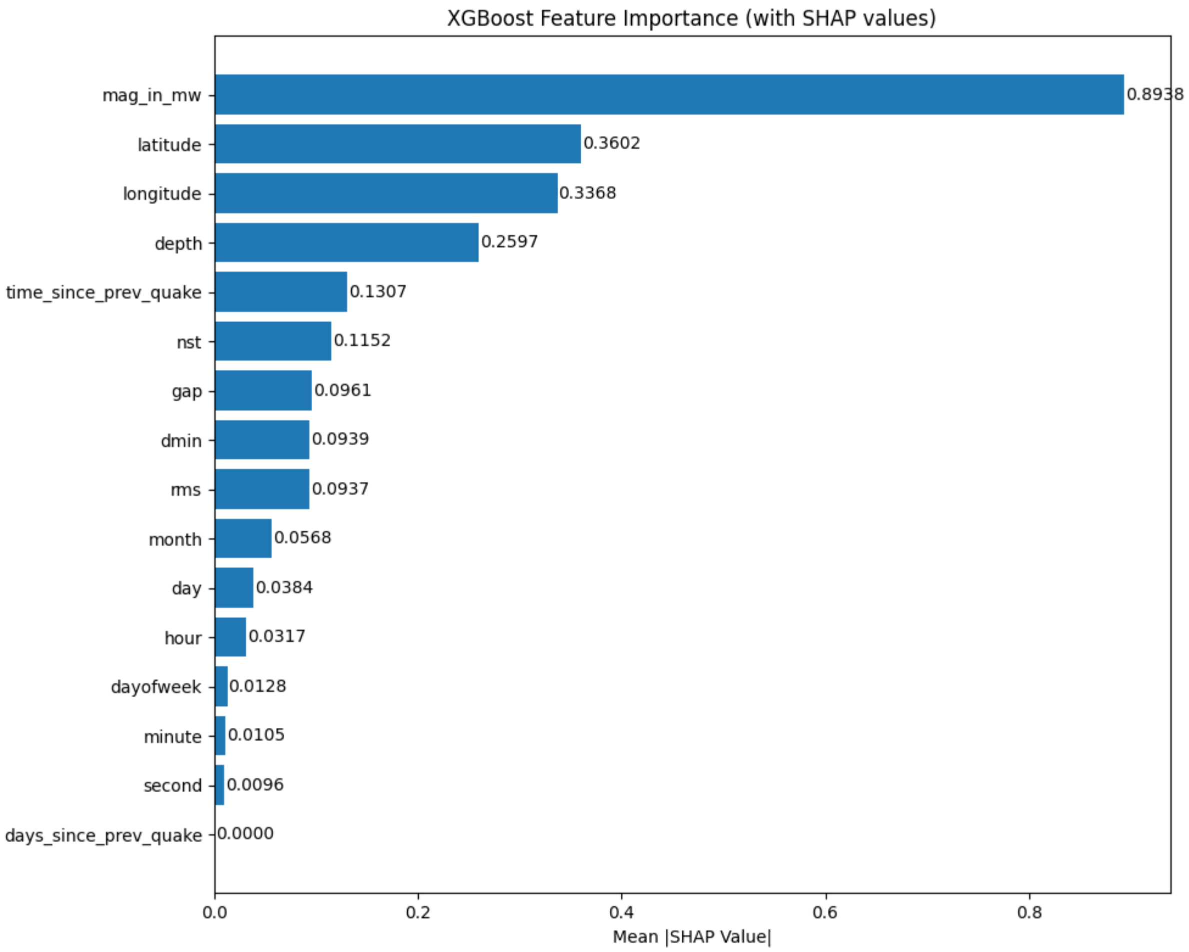

Figure 1: XGBoost Feature Importance (SHAP)

6 Conclusion

- The pilot study demonstrates that mainshock events can be predicted with moderate accuracy using gradient-boosting models on engineered seismic features.
- The study provides a reproducible data preparation pipeline, a labeled mainshock dataset, and baseline models.
- Future work includes integrating more models, tuning hyperparameters, gaining additional insights, and publishing the full research paper.