Analysing and Predicting Major Earthquakes in Pakistan (2025-2030)

Project By: Abdur Rehman Tariq, Ahmad Hassan

Introduction

During our time at Space Summer School, we explored satellite data and learned how to access and utilize it effectively. Drawing on our newly acquired skills and experience, we were inspired to create a project that reflected what we had learned. This led us to develop an earthquake prediction system using data provided by Mr. Abdul Mateen from the Space Summer School team. The data, sourced from the USGS and credited to NCGSA, formed the foundation of our work, specifically to create an Analysis Dashboard and a prediction Model for Earthquakes in Pakistan.

Goals

1. **Create a Model to predict Earthquakes for the next 5 years in Pakistan:**
2. **Additionally, create an analysis dashboard to analyse risk in different districts across Pakistan in the event of a major earthquake.**

Analysis of Earthquake Data and Problems with Model Prediction

We initially had 5 years of data on Earthquakes (2020-2025) (since that was the most we could access for free as students) with the following fields:

Field Name	Information / Description
`date`	The calendar date when the earthquake occurred (YYYY-MM-DD).
`time`	The exact time of the earthquake in UTC (HH:MM: SS).
`latitude`	Geographic coordinate specifying the north-south position of the epicenter.
`longitude`	Geographic coordinate specifying east-west position of the epicenter.
`depth_km`	Depth of the earthquake's focus below Earth's surface, in kilometers.
`magnitude`	Magnitude of the earthquake (Richter scale).
`place`	Description of the location relative to nearby landmarks or regions.
`type`	Type of seismic event, usually labeled as "earthquake".

Our initial premise was to predict the longitude, latitude, and date of future earthquakes. However, a little EDA (exploratory data analysis) revealed that this would be impossible to do due to the spontaneous and random nature of earthquake data.

Regardless, we decided to proceed with the making of a predictive model, testing out quite a few different algorithms such as linear, lasso, ridge, and polynomial regression, as well as ensemble algorithms such as decision trees and Random Forest regressors, adding on XG Boost and KNN too. Only to be met with negative R^2 values. At this point, it was clear that our approach was wrong, and sure enough, there were a ton of factors, with the most significant being:

1) Overfitting on small earthquakes with magnitudes <= 4.0
2) Not enough data: only 5 years of data proved to be insufficient
3) Models not understanding geospatial data (for example: -179 and 179 degrees latitude are only 2 degrees apart spatially)
4) Lack of significant tectonic. geospatial data

The Fix

We decided that if we were to meaningfully predict earthquakes, we would have to significantly alter our approach. So we made the following modifications

1. ### We requested more data

   Mr. Abdul Mateen from the Space Summer School team was kind enough to provide us with 50 years of data on earthquakes from 1975-2025.

2. ### Filtering Data and using Feature Engineering to create more geospatial data

   We decided to use only major earthquakes (i.e, Magnitude >= 5.0) to avoid overfitting on this data. Furthermore, with some basic feature engineering, we would be able to build a better model

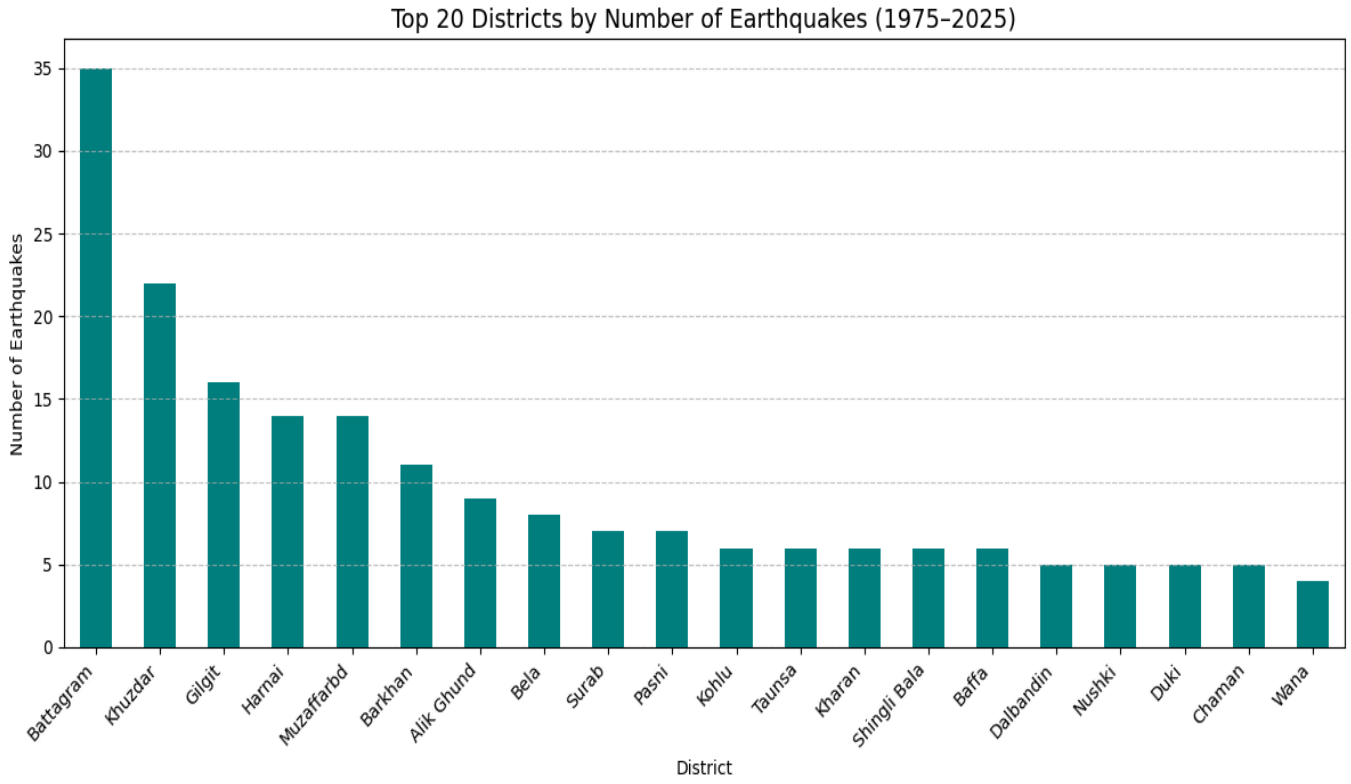
3. ### Changing Our Prediction Target

   We changed our prediction target somewhat, instead of predicting random and spontaneous data, such as the exact epicenter coordinates and date of the earthquakes, We instead decided to measure the frequency of Major Earthquakes by year and predict the expected number of major earthquakes in Pakistan within the next 5 years.

Building a new Model

Now that our prediction target had changed, we first normalized the ‘place’ field in our dataset. The place field had distance, direction, district, and country data within it (e.g, 113 km NW of Bela, Pakistan). We then normalized this into 3 different fields (distance, direction, district), excluding the Country field since we only had data on Pakistan.

With this, we were able to better grasp the distribution of earthquakes by district and analyse areas at high risk. Below is a graph of the top 20 districts by frequency of Major Earthquakes in the past 50 years:



Once we had a grasp on areas at high risk, we Feature Engineered more data, adding the following fields:

Feature Name	Description
`total_earthquakes`	Total number of earthquakes in the year
`major_earthquake_count`	Count of major earthquakes (magnitude ≥ 5.0)
`activity_rate`	Average earthquakes per day in the year
`major_rate`	Average major earthquakes per day
`avg_magnitude`	Average earthquake magnitude that year
`std_magnitude`	Standard deviation of magnitudes
`skew_magnitude`	Skewness (asymmetry) of magnitude distribution
`mag_75th`	75th percentile of magnitudes (upper quartile)
`mag_90th`	90th percentile of magnitudes
`moderate_count`	Count of earthquakes with magnitude between 4.0 and 5.0
`strong_count`	Count of earthquakes with magnitude between 5.0 and 6.0

`energy_release`	Total energy released by earthquakes (approximated by a formula using magnitude)
`energy_per_event`	Average energy released per earthquake
`avg_depth`	Average depth of earthquakes (in km)
`std_depth`	Standard deviation of earthquake depths
`shallow_ratio`	Proportion of earthquakes with depth ≤ 35 km
`intermediate_ratio`	Proportion with depth between 35–70 km
`deep_ratio`	Proportion with depth > 70 km
`spatial_spread`	Spatial variability of quake epicenters (combined lat/lon std)
`spatial_range`	Geographical area covered by earthquake locations
`lat_std`	Standard deviation of latitudes
`lon_std`	Standard deviation of longitudes
`*_3yr_mean` (4 fields)	3-year rolling mean of key indicators (e.g., `total_earthquakes_3yr_mean`)
`*_5yr_mean` (4 fields)	5-year rolling mean of key indicators
`*_lag3` (4 fields)	3-year lag values of select indicators
`*_lag5` (4 fields)	5-year lag values of select indicators

Now with a good amount of data and features for accurate model training, we trained a simple Random Forest Regressor Model and were able to achieve the following results:

Metric	Value
CV Score	0.1416 ± 0.0973
Test R^2	0.6952
Test MSE	0.0274
Test MAE	0.0710

Metric	Value
Total Predicted	2.0
Average per Year	0.3
Historical Average	0.4 ± 0.7

Summary of Results

Our model successfully analyzed 50 years of seismic data (1975-2025) to identify high-risk earthquake zones across Pakistan's districts. The analysis revealed Battagram as the highest-risk district with 35 recorded earthquakes, followed by Khuzdar with 22 events, both classified as "Very High Risk" zones.

The study processed earthquake data with magnitudes of 5.0 or greater, revealing clear geographic patterns, with Northern Pakistan and Balochistan province showing the highest concentrations of seismic activity.

Using Random Forest machine learning algorithms with an R^2 accuracy of 0.70, the system predicts 2.0 major earthquakes (magnitude ≥6.0) for the period 2025-2030, specifically forecasting events in Northern Pakistan (2026, M6.2) and Balochistan (2028, M6.1). The interactive web we developed (Quake Compass) successfully transforms complex seismological data into an accessible visualization tool featuring dynamic markers, color-coded risk zones, and comprehensive statistics dashboards.

This provides critical evidence-based insights for disaster preparedness planning, emergency resource allocation, and public safety awareness across Pakistan's earthquake-prone regions, serving government agencies, researchers, and communities in understanding and preparing for seismic risks. Further Analysis and statistics can be found on our web app Quake Compass, linked in the Bibliography section below.

Future Improvements

Several enhancements could significantly expand the capabilities and impact of the Pakistan Earthquake Risk Map system. Integration of real-time seismic data feeds from multiple monitoring stations would enable continuous updates and immediate alerts for ongoing seismic activity, while incorporation of additional geological factors, such as soil composition, fault line proximity, and topographical variations, could improve prediction accuracy beyond the current $R^2 = 0.70$ threshold. Advanced machine learning approaches, including ensemble methods, deep neural networks etc, could provide more sophisticated uncertainty quantification and multi-model consensus forecasting. The system would benefit from expanded temporal analysis incorporating paleoseismic data to extend the historical baseline beyond 50 years, coupled with climate correlation studies to examine potential relationships between monsoon patterns, temperature variations, and seismic activity.

Bibliography

Project Resources

ARTariqDev. (2025). *Quake-Compass: Pakistan Earthquake Risk Map - Interactive seismic data visualization and analysis tool*. GitHub Repository. https://github.com/ARTariqDev/Quake-Compass/

ARTariqDev. (2025). *Quake-Compass Analysis Dashboard: interactive Pakistan earthquake risk visualization*. Web Application. https://quake-compass.vercel.app/

Data Sources and Credits

National Centre for GIS and Space Applications (NCGSA). (2025).

Space Summer School Team, Institute of Space Technology (IST). (2025).

United States Geological Survey (USGS). (2025). *Global earthquake database and seismic hazard assessment tools*. Earthquake Hazards Program. https://earthquake.usgs.gov/