Earthquake Prediction

DATA/MSML602: Principles of Data Science
Final Project
by
Varun Jain
Srijinesh Alanka
Subha Venkat Milind Manda
Professor: Mohammad T. Hajiaghayi
TA(s): Max Springer
December 12, 2023

Introduction:

An earthquake is what happens when two blocks of the earth suddenly slip past one another. The surface where they slip is called the fault or fault plane. The location below the earth's surface where the earthquake starts is called the hypocenter, and the location directly above it on the surface of the earth is called the epicenter.

Earthquake prediction is a branch of the science of seismology concerned with the specification of the time, location, and magnitude of future earthquakes within stated limits and particularly "the determination of parameters for the next strong earthquake to occur in a region". Earthquake prediction is sometimes distinguished from earthquake forecasting, which can be defined as the probabilistic assessment of general earthquake hazard, including the frequency and magnitude of damaging earthquakes in a given area over years or decades. Not all scientists distinguish "prediction" and "forecast", but the distinction is useful. Prediction can be further distinguished from earthquake warning systems, which, upon detection of an earthquake, provide a real-time warning of seconds to neighboring regions that might be affected.

In the 1970s, scientists were optimistic that a practical method for predicting earthquakes would soon be found, but by the 1990s continuing failure led many to question whether it was even possible. Demonstrably successful predictions of large earthquakes have not occurred, and the few claims of success are controversial. For example, the most famous claim of a successful prediction is that alleged for the 1975 Haicheng earthquake. A later study said that there was no valid short-term prediction. Extensive searches have reported many possible earthquake precursors, but, so far, such precursors have not been reliably identified across significant spatial and temporal scales.

Possibility:

The main question now is can we predict when and where earthquakes happen at the exact place and at the exact time?

It is unlikely they will ever be able to predict them. Scientists have tried many different ways of predicting earthquakes, but none have been successful. On any particular fault, scientists know there will be another earthquake sometime in the future, but they have no way of telling when it will happen.

Need For Prediction:

As the purpose of short-term prediction is to enable emergency measures to reduce death and destruction, failure to give warning of a major earthquake that does occur, or at least an adequate evaluation of the hazard, can result in legal liability, or even political purging. Even if our prediction gives out correct forecasting for the earthquakes it might help in saving lives in huge numbers. The forecasting or prediction need not be perfect but it should at least let the people know about the chances of an earthquake happening. As we all know previously earthquakes have taken many lives and caused destruction of properties, a forecast might save all this from happening. If our prediction model helps people even in the slightest way possible then we would call this a successful model

Data Acquisition

The United States Geological Survey (USGS) serves as a prominent and reliable source for acquiring open-source earthquake data, offering a comprehensive repository of seismic information. USGS operates a sophisticated seismic monitoring network that spans the globe, capturing real-time data on earthquake occurrences, magnitudes, locations, and depths. This valuable resource is made available to the public, researchers, and organizations, fostering collaboration and enabling the development of innovative earthquake prediction models. The open nature of the data aligns with USGS's commitment to transparency and scientific collaboration, allowing scientists, engineers, and enthusiasts worldwide to access and analyze seismic data for a deeper understanding of earthquake patterns and behaviors. Leveraging the USGS earthquake data provides researchers with a foundation to enhance seismic risk assessments, improve early warning systems, and contribute to the broader scientific community's collective efforts in mitigating the impact of seismic events. USGS allows only 20,000 data points to be scrapped at a time.

We used the “ObsPy” library to retrieve data from “ObsPy”. ObsPy is an open-source Python toolbox designed for seismology and seismological observatories. It provides a set of tools for working with seismological data, such as downloading, processing, analyzing, and visualizing seismic waveforms. ObsPy simplifies complex tasks related to seismological research and enables scientists to focus on their analyses. ObsPy supports various data formats commonly used in seismology, allowing users to read and write seismic data effortlessly. It facilitates data retrieval from online sources, such as data centers or real-time data streams, making it easier for researchers to access the information they need.

We use the “Client” function from ObsPy to access data from USGS. We need to allot specific time frames and give it to this client to retrieve information. This information is only received if the parameters are correctly set and the total data points received are lower than 20,000. We compiled the data points received in a dataframe using the pd. DataFrame feature.

The data which we have is for magnitudes 5 and above. The selection of minimum magnitude is kept as 5 because that’s the limit when we start feeling tremors and an area wide emergency notification is given.

We divided the 127 years into 25 segments of 5 years each and fetched data 25 times to receive a dataframe of 100 thousand data points.

Formatting

The data received from USGS is fairly clean. There were some instances of inconsistencies in the data which we had to clean. Many places whose location was not available, we changed it to “NaN”. The “time” column had to be converted into datetime. pd.to_datetime() function call helped us to easily convert the column into datetime. Then we used the call “dt.year” to get the year part of each time to get in which year the earthquake took place. This will help us get valuable insights when we are working with the data. Then we moved onto the place column. We used .split(“,”) to separate the country and the exact location to give us a reference which will be useful for us as we move ahead in the project. The indexing in the data was also wrong so I had to use the df.index to change the index of the dataframe. After this our data is ready for preliminary analysis and model fitting.

Preliminary Analysis

We use our data to get basic insights about the earthquakes which have taken place throughout the world. We start by finding out the exact locations of earthquakes. For this the latitudes and longitudes which we retrieved plays an important role. Just plotting the latitudes and longitudes won't cut it. We will need to find exact locations where these earthquakes took place. We will use geopandas for this. Using geopandas we will call in the world map and plot our latitudes and longitudes on it. The plot is shown in the figure 1 below.

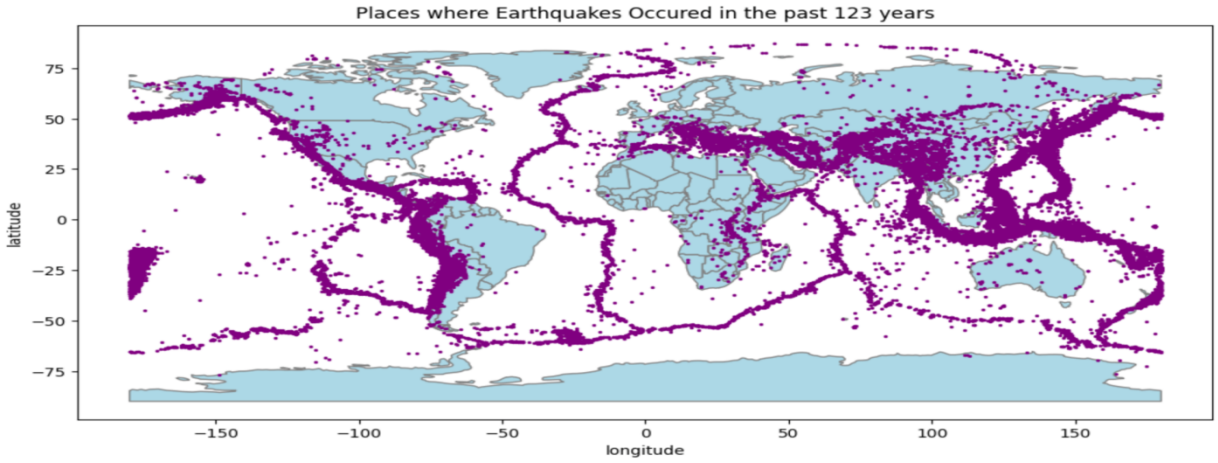

Figure 1: A Visualization of earthquakes for the past 123 years

These are the places where the earthquakes have taken place. From the image we can figure out that most earthquakes have taken place in South Asia and Japan as the points are densely packed in those regions. We will now call “groupby” and sort the countries which have appeared the most in the dataframe. Below is the image of our result.

```
Indonesia                                  11470
Papua New Guinea                            6169
Japan                                       6156
Philippines                                 4658
Russia                                      4315
                                             ...
[Washington]                                   1
[southwest of New Caledonia]                   1
 Uruguay                                       1
[Southern Iran]                                1
[Jamaica region]                               1
```

Figure 2: Most common earthquake occurrences

Our understanding was right as maximum earthquakes have taken place in these regions itself. The places Indonesia, Papua New Guinea and Japan have been affected most. This is because they are small island countries where the tectonic plate movement is the maximum.

Now we have the inference that where the earthquakes take place the most. Now we see that the number of earthquakes in each year is different. We might have to look into the number of earthquakes which have taken place in the past 123 years. There might be a trend which we may have to realize. We now called group-by on the year. The dt.year feature which we did will now come in handy. The average number of earthquakes every year is shown in Figure 3 below.

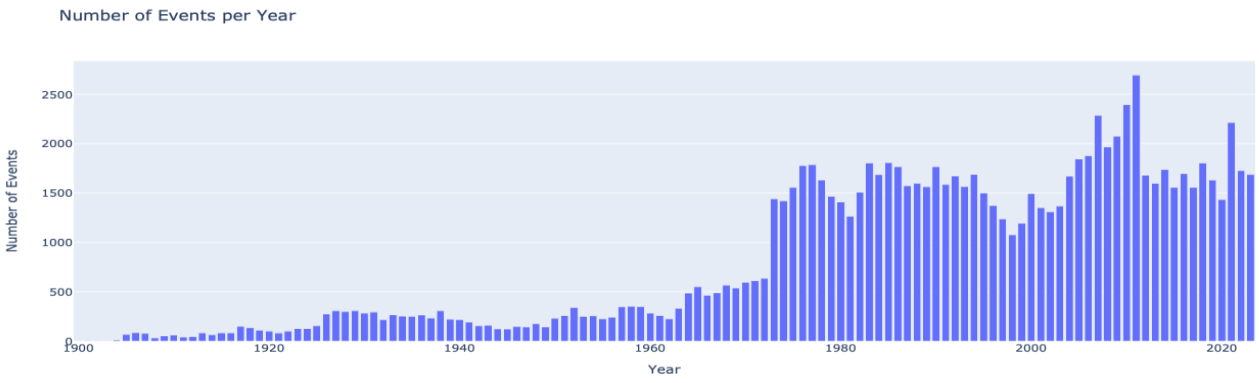

Figure 3: Bar graph of total number of earthquakes per year

The above plot gives us a clear picture of how the number of earthquakes have increased. The number of recorded earthquakes over magnitude 5 in 1900 was 1 and for 2023 we can see that it has significantly increased. The number of earthquakes since 1900 has gone upto a total of 1683. There is over 10,000%

increase in the total number of earthquakes in these past years. Indonesia has suffered the most with around 11470 earthquakes over the magnitude of 5. Figure 4 shows the affected places in Indonesia. The red line outlines the country of Indonesia and the varying levels of colors represented by the colorbar to it's right show the different depths at which the earthquakes with magnitudes over five took place.

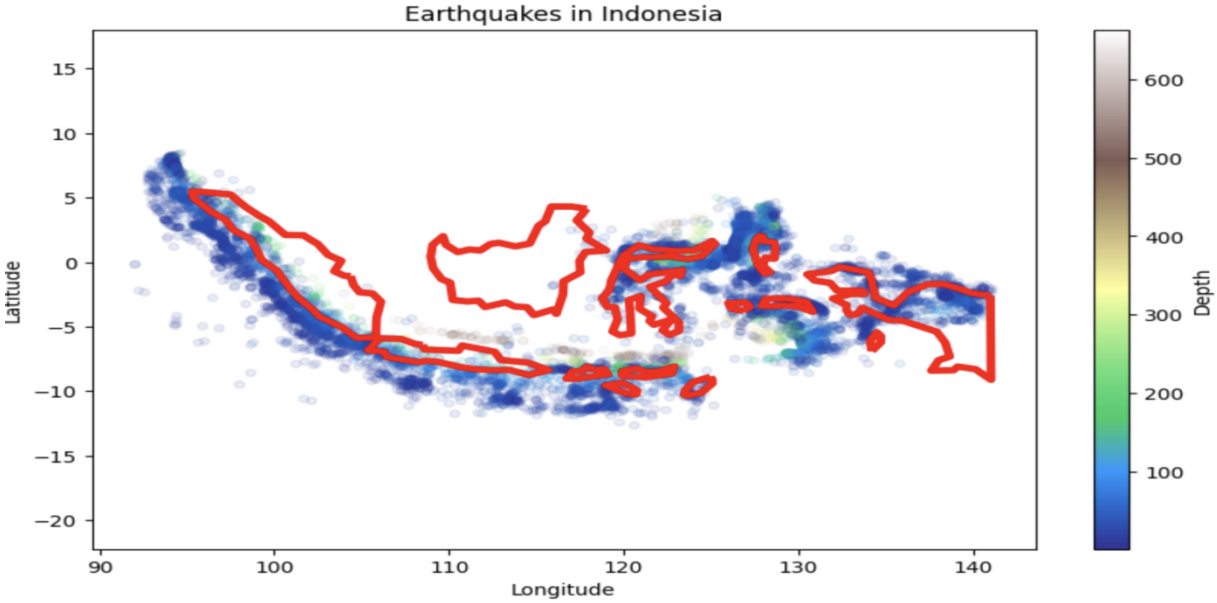

Figure 4: A visualization of earthquakes occurring in Indonesia

We also wanted to fit a regression line for this plot to figure out the trend. We imported the seaborn library to get the regression line for the number of earthquakes. In Figure 5, we can see that the number of recorded earthquakes has increased linearly.

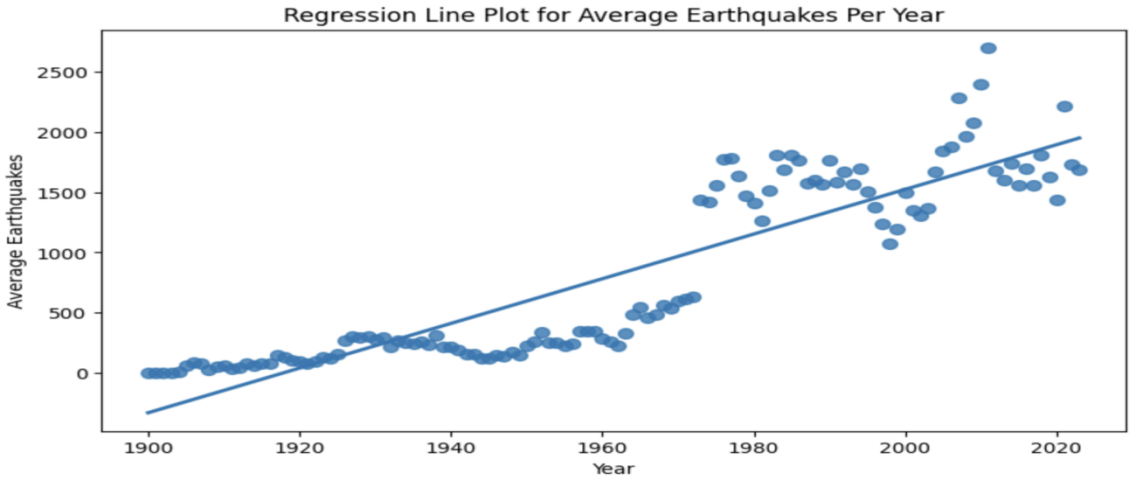

Figure 5: Regression plot

The graph above clearly shows the increasing trend in the average number of earthquakes taking place in each year. We will have to be more careful with our activities as dumping active nuclear waste is one of the biggest reasons for the increased number of earthquakes.

Then another thing we thought of plotting was the depth. There might be certain places on earth that have earthquakes which are deeper than usual. We drew a 3d diagram of earthquakes giving us an indication of the places which have earthquakes deeper than the normal places. The Figure 6 below shows all the required details for us to interpret the necessary information.

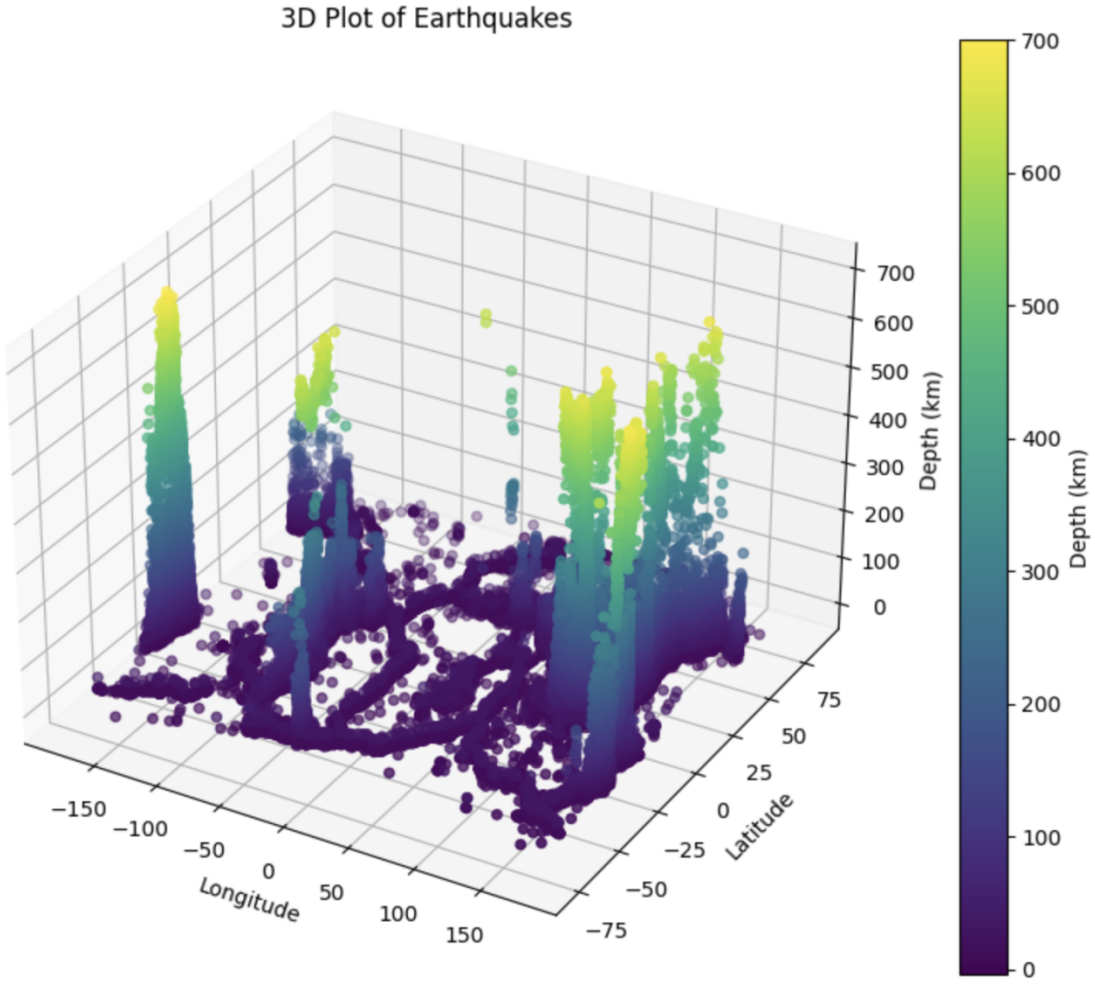

Figure 6: A 3D plot showcasing the depth and the locations

The graph above gives an idea about the places that were affected by earthquakes. We can see that the region between longitude -170 and -180 and latitude -20 and -30 have had deep earthquakes. The earthquake was more than 700 kilometers deep which caused extensive damage to the places affected.
This region again is the Southern Asian region which allows us to interpret that the Southeast Asia region is affected the most in terms of number and depth of earthquakes.

Prediction model

We used Random Forest Regressor, a robust machine learning algorithm, to predict the geographic coordinates where earthquakes are likely to occur and their respective magnitudes. We got the

Mean Squared Error for Location as 0.16182177952915852, Mean Squared Error for Magnitude as 3.124543931497419e-07, which was a better result compared to other models which gave a relatively higher mse and rmse values.

The modeling phase consisted of constructing two distinct models: one to predict location (model_location) and another to forecast magnitude (model_magnitude). The former incorporated a StandardScaler within a pipeline to normalize the features, while the latter was a straightforward application of the RandomForestRegressor. After training both models on their respective training sets, predictions were generated for the test data.

Figure 7 shows a scatter plot with longitude on the x-axis and latitude on the y-axis. The actual locations are marked in blue, and the predicted locations are marked in red. The spread suggests some degree of variance between the predicted and actual locations

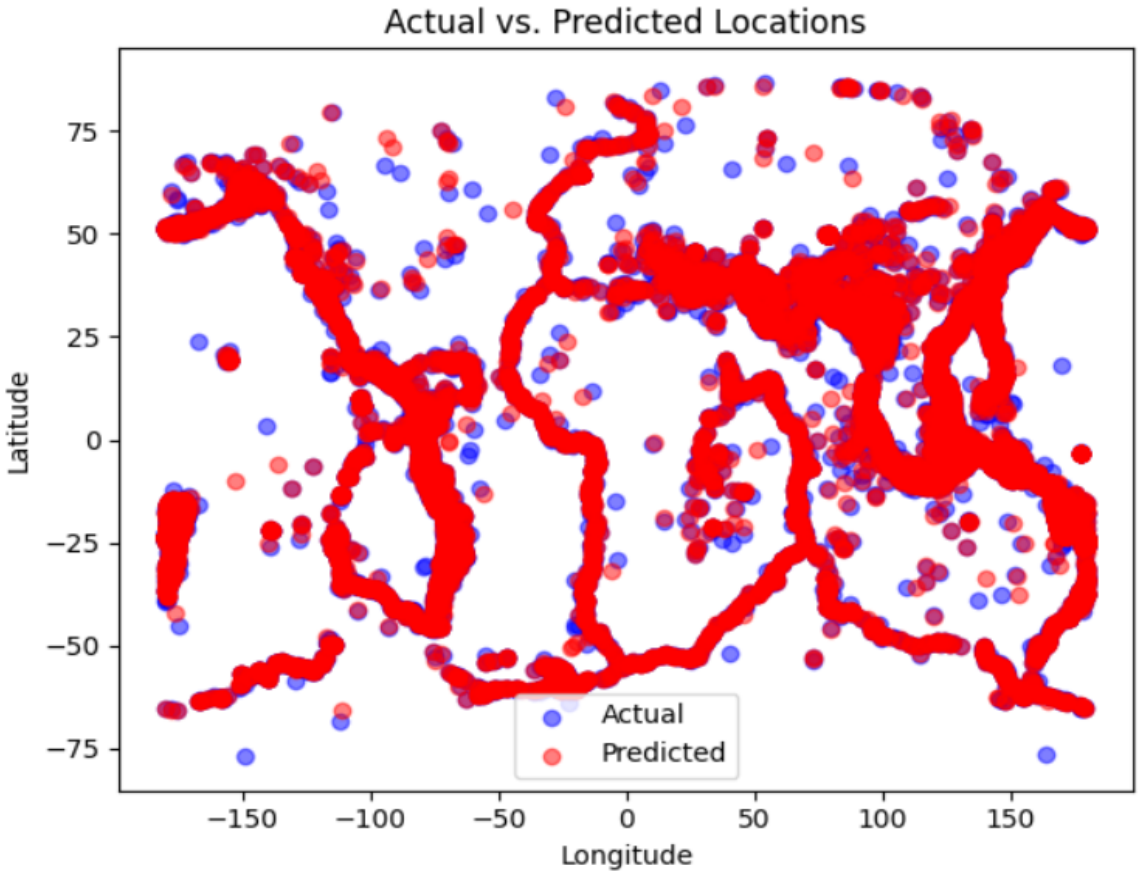

Figure 7: Scatter plot of actual vs predicted locations

Figure 8 displays a scatter plot with the actual magnitudes on the x-axis and the predicted magnitudes on the y-axis. The close alignment of points along a diagonal line suggests that the model predicts magnitudes with a high degree of accuracy.

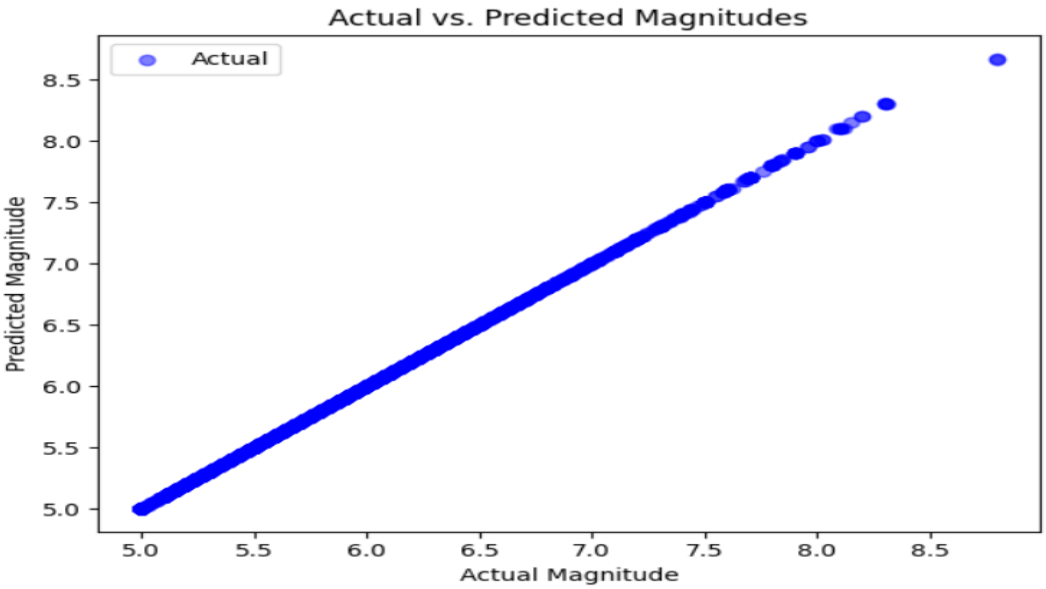

Figure 8: Scatter plot of actual vs predicted magnitudes

These visualizations serve as a testament to the models' predictive performances. While the location predictions exhibit a moderate scatter, indicative of the complex nature of seismic activity prediction, the magnitude model demonstrates a high degree of precision. This suggests that, at least within the scope of our dataset, the Random Forest Regressor holds considerable potential in predicting earthquake magnitudes, if not the exact locations.

User Interface

The Random Forest model is trained on a dataset comprising past earthquake occurrences, considering factors like date, magnitude, latitude, and longitude. The trained model is serialized into a pickle file for integration with the Flask application.

A Flask web server is set up with two primary routes: the home page (/) for input and the prediction page (/predict) for displaying results. The home page (index.html) features a form where users specify the number of days for which they want earthquake predictions.

The prediction route processes this input, uses the model to predict earthquakes for the specified number of days, and renders the results on the result.html page. The predictions include details like the predicted magnitude, latitude, and longitude for each day. The result.html template displays the earthquake predictions. We used Folium, a Python library for visualizing geospatial data, to plot the predicted earthquake locations on an interactive map.
There are two other routes, one for displaying the latest entries of data on earthquakes that occurred in real time from our API - USGS, and another API to visually display our earthquake historical data with respect to time, in a time series plot. The below figures showcase the main parts of our user interface.

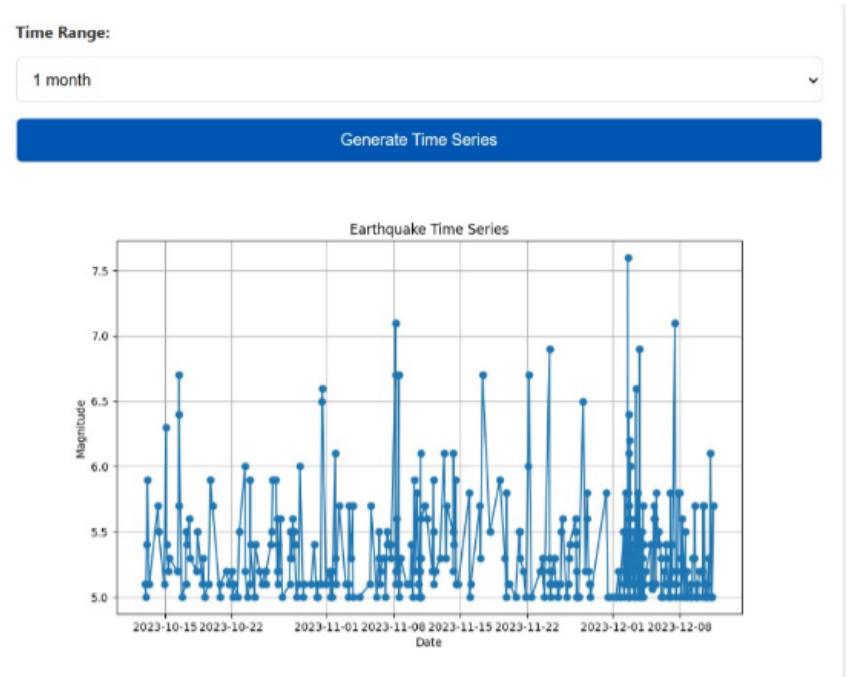

Figure 9: Time series graph of recent earthquakes

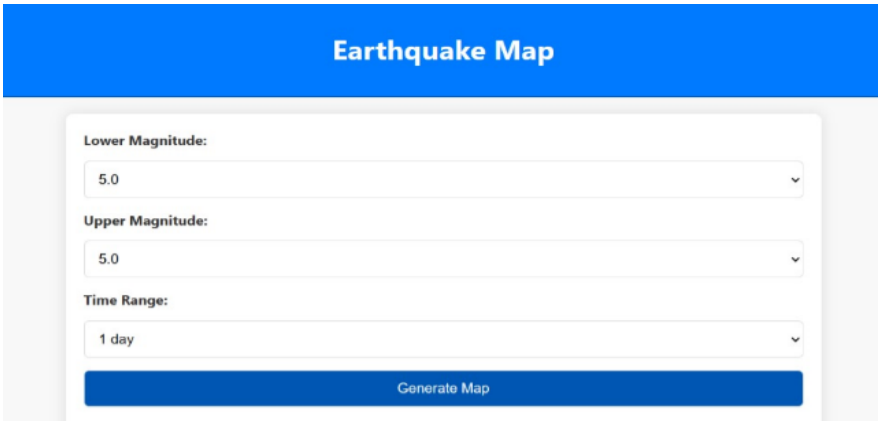

Figure 10: Sample user input for generating earthquake predictions

Figure 11: Displaying earthquake prediction on map

Conclusion

This project successfully demonstrates the integration of real-time seismic data with a machine-learning-based approach to predict earthquakes. Utilizing data from the USGS (United States Geological Survey) API, the system harnesses the most current information to feed into a Random Forest Regressor model, which was chosen for its superior performance in terms of lower root mean square error (RMSE) compared to other models, including LSTM. The Flask-based web application provides a user-friendly platform for accessing these predictions, effectively translating complex data analyses into actionable insights. This innovative approach exemplifies the potential of combining real-time data with machine learning for enhanced natural disaster preparedness and response.

Future scope of study

- While the Random Forest Regressor was chosen for its lower RMSE, continuous exploration and testing of other models or ensemble methods could further improve prediction accuracy.
- Incorporating additional data sources, such as geological surveys or historical seismic activity records, might enhance the model's predictive capabilities
- Developing an alert system within the application to notify users of potential seismic activities in their region.
- Enhancing the UI/UX for a more engaging and informative user experience, including personalized settings and region-specific information would improve the overall experience for the users.

By continuing to refine the model and expanding the application's capabilities, future developments can significantly contribute to earthquake preparedness, risk mitigation, and public safety, harnessing the power of real-time data and advanced analytics