Name:R.Suriya

Reg No:422221104039

College Code:4222

Team: T6

EARTHQUAKE PREDICTION USING MACHINE LEARNING

EARTHQUAKE PREDICTION USING MACHINE LEARNING

Introduction

Different natural phenomena like the fall of meteorites, tsunamis, volcanic eruptions, droughts, ice ages, the reversal of geomagnetic field, forest fires, droughts, earthquakes, and others can pose a significant danger and threat to human life and humanity's economic developments and resource managements (Murray, 2021).

Earthquakes are caused not only by natural seismic and tectonic processes but often time can also be induced by various anthropogenic activities such as nuclear bomb detonations, large dams, and subsurface exploitation of natural resources. The danger and risk posed by usually low intensity earthquakes induced by anthropogenic activities can be indeed mitigated by reducing or completely stopping the human activities that are responsible by these types of minor earthquakes. In a sharp contrast, especially earthquakes of great intensity that are caused by natural processes cannot be avoided but only forewarned with their often catastrophic and damaging impacts minimized.

Different sources and mechanisms have been suggested as triggers and modulators of earthquakes (see, for example Batakrushna et al., 2022, for a full review). For example, even the Sun's activity has been suggested as a significant agent causing earthquakes (Anagnostopoulos et al., 2021). Other proximate causes discussed in the literature include pole tide (Shen et al., 2005), pole wobble (Lambert and Sottili, 2019), surface ice and snow loading (Heki, 2019), glacial isostatic rebound (Hampel et al., 2007), heavy precipitation (Hainzl et al., 2006), atmospheric pressure (Liu et al., 2009), sediment unloading (Calais et al., 2016), seasonal groundwater change (Tiwari et al., 2021), seasonal hydrological loading (Panda et al., 2020). In addition, the Earth's rotation and tidal spinning have also been suggested as driver of plate tectonic activity.

The present geological paradigm about solid Earth is the plate tectonic theory which describes that the lithosphere is segmented into a series of plates that are in constant motions due to mantle mobility or convection. As a result of their interaction, a series of geological, mainly convergent and divergent, processes take place at their plate margins, ranging from seismicity, orogenic processes, and volcanism. The World Stress Map (WSM)[1] compiles the orientation of maximum horizontal stress (σ_{Hmax}) where we delimited our study areas in Figure 1 (Heidbach et al., 2016).

Figure 1

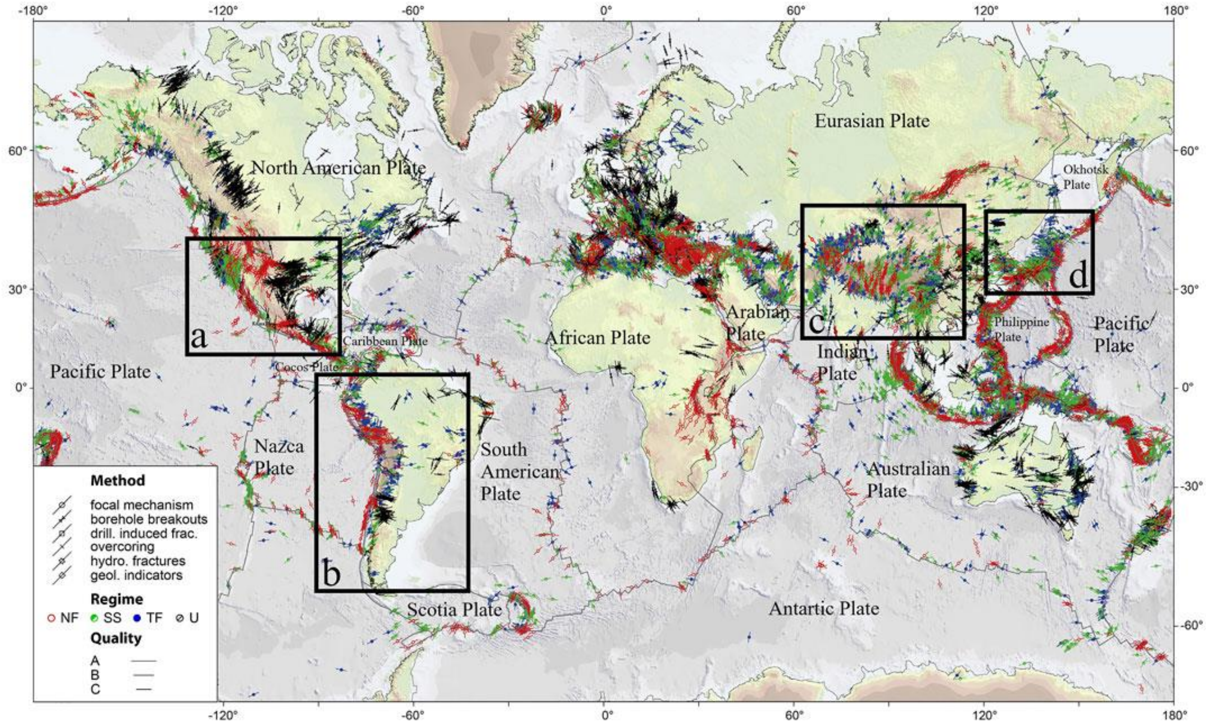


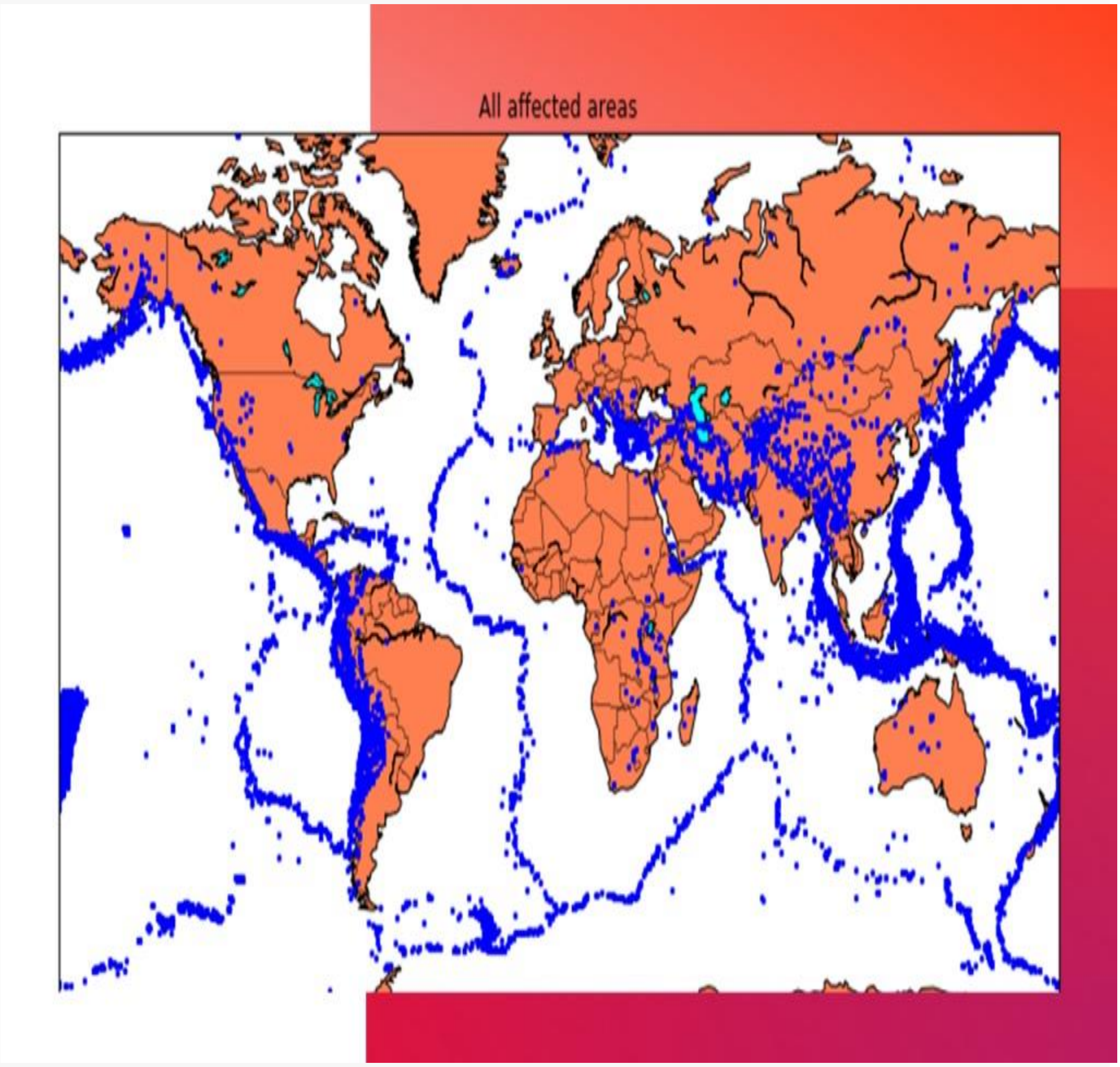
FIGURE 1. World Stress Map from WSM 2016 database release. Lines show the orientation of maximum horizontal stress (σ_{Hmax}) for the 40 km upper crust from different stress indicators displayed by different symbols; line length is proportional to data quality (A–C). Colors indicate the stress regime: I) red = normal faulting (NF), II) green = strike-slip faulting (SS), III) blue = thrust faulting (TF), and IV) black = unknown regime (U). Grey lines give plate boundaries from global model PB2002 of Bird (2003). The seismic zones analyzed are shown in the marked rectangles: (A) United States-Mexico, (B) South America, (C) Southern China and Northern India, and (D) Japan.

The dynamics of the plate tectonics provide a framework to understand the evolutive shape and dynamics of the earth's surface. Plate boundaries involve either divergence, like at oceanic spreading centers and continental rifts, or convergence, such as subduction (ocean to continent or ocean to ocean) and collision zones with different angles of displacement ranging from orthogonal towards subparallel one.

Only minor cases involve transform boundaries that facilitate plate kinematics on the global sphere. These boundaries accommodate plate-parallel relative displacement by strike-slip motion on vertical or steeply dipping faults. Due to these frictional contacts between the different types of plates, seismicity is triggered, producing a succession of earthquakes that progressively decrease in intensity in increasingly distant/remote areas away from the seismic center/zone.

The sliding between tectonic plates is quite varied. Some plates slide without any consequences on Earth's surface, while catastrophic failures punctuate others. Also, after a few hundred meters some earthquakes stop. Nevertheless, others continue to collapse even after thousands of kilometers (Kanamori and Brodsky, 2004).

The driving mechanisms of plate tectonics remain not well unknown or poorly understood. Are they due to internal factors or external astronomical forces? We are hoping that the analysis of seismic patterns could provide some clues and information about the sources and mechanisms that are responsible for both tectonic movements and earthquakes.



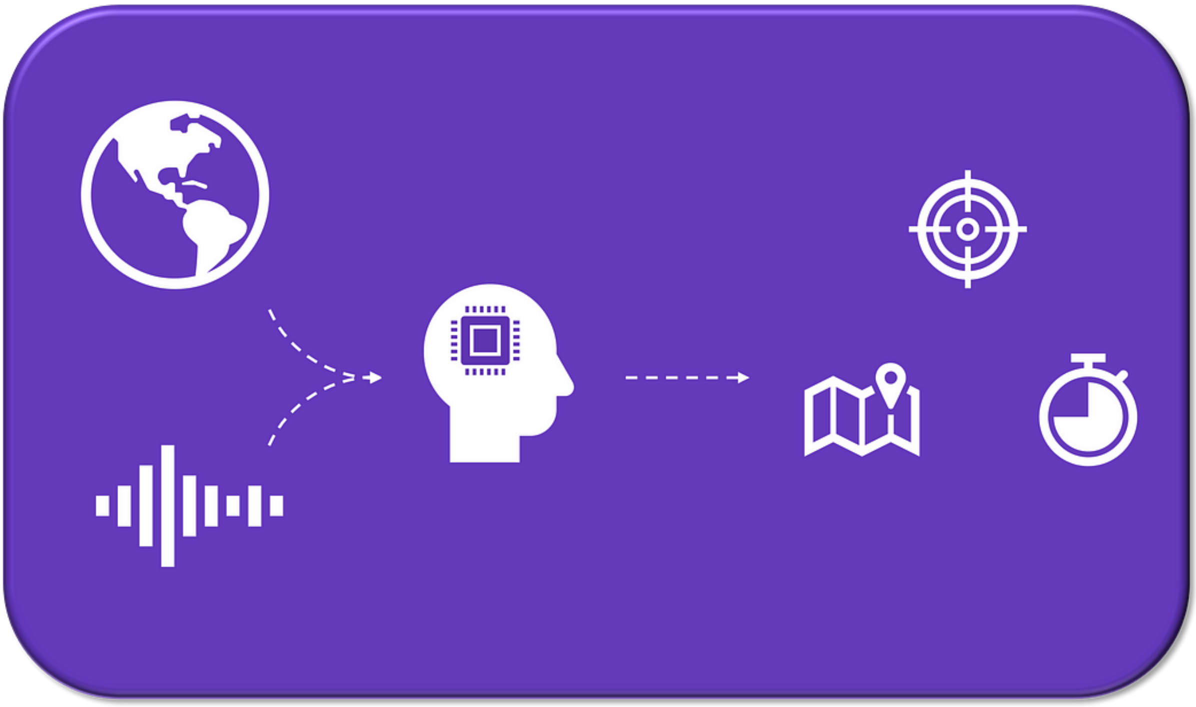
Earthquake forecasting is one of the most difficult areas of research even though it is clear that its early prognosis can save many lives (Jain et al., 2021). Deterministic prediction of the exact coordinates of the epicentre, its depth, magnitude and exact time of one earthquake at the moment remains difficult and possibly impossible (see, for example, Shcherbakov et al., 2019; Beroza et al., 2021). Ogata (1988) suggests that the seismic pattern and temporal variation are usually very complicated. Furthermore, temporal seismic clustering is complex and difficult to discern or anticipate in advance. Different models have been proposed to analyze space-time clusters of seismicity in a region. One example is the Epidemic Type Aftershock Sequence (ETAS) model. This model suggests that the earthquake of a particular magnitude (M) in a region during a period of time can be approximately considered as a Poisson process (Ogata, 1988). In addition, the method of the minimum area of alarm for earthquake magnitude prediction (Gitis and Derendyaev, 2020) and a method for earthquake predictions based on alarms (Zechar and Jordan, 2008) have all been suggested and evaluated.

The studies of earthquake precursors such as observing crustal geochemical fluids and gases, ultra-low frequency magnetic signals, atmospheric effects including ionospheric total electron content measurements, and several recording seismicities in regions experiencing earthquakes in terms of atmospheric, geochemical, and historical information can all help to improve and refine earthquake prediction (see for example, Pulinets and Boyarchuk, 2005; Ouzounov et al., 2018; Pulinets and Ouzounov, 2018). Since 2007, the Collaboratory for the Study of Earthquake Predictability (CSEP) has actively conducted and rigorously evaluated earthquake forecasting experiments as well as the prospective evaluations of earthquake forecast models and prediction algorithms (see, for example, Schorlemmer et al., 2018). CSEP's main targets and focuses are to optimize earthquake forecasting, advance forecast model development, test model hypotheses, and improve seismic hazard assessments.

The medium-term prediction of the strongest earthquakes has been carried out by the *M8* algorithm, which is an algorithm for evaluating times of increased probability (TIPs) for strong earthquakes (Keilis-Borok and Kossobokov, 1990) from intensity of an earthquake flow and rate differential on a specific seismic region of earthquake source concentration and clustering. Also, the prediction of extreme events such as earthquakes demonstrates the efficiency and potential of the algorithms based on a pattern recognition approach as example the *M8* algorithm (Kossobokov and Soloviev, 2008, 2018). In addition, the *M8* algorithm shows that the hypothesis that the largest earthquake events are mere random variations in seismically active regions can be confidently rejected (Kossobokov and Soloviev, 2021). Kossobokov et al. (2015) suggested that "forecasting earthquake information must be reliable, tested, confirmed by evidence, and not necessarily probabilistic". We disagree slightly with this opinion and interpretation of Kossobokov et al. (2015). Probabilistic forecasts in the last century have provided new results to understand natural phenomena (see, for example Wigner, 1967; Landau and Lifshitz, 1988b; Feynman et al., 2011b). In this work, we show the results of a Bayesian model of Machine Learning, which is a probabilistic model. We do agree with Kossobokov et al. (2015) that all forecasts which are either probabilistic or not probabilistic must indeed be confirmed by evidence. We think that only future events will show if our probabilistic Machine Learning predictions are on the right track or not.

In recent years artificial intelligence (AI), deep learning (DL), machine learning (ML) (see, for example, Essama et al., 2021) have been applied to earthquake forecasting. In particular the use of ML in the study of earthquakes has been implemented in the detection, arrival time measurement, phase association, location and characterization (Beroza et al., 2021). In addition, the use of ML has focused on forecasting the exact magnitude of the next strong earthquake in different seismic zones (see, for example, Yousefzadeh et al., 2021).

In this paper, we propose a new method of analysis and algorithm for forecasting strong earthquakes (i.e., magnitude ≥7). We suggest that one promising progress to earthquake forecasting may consist in changing the prediction paradigms from an "exact" approach to probabilistic forecasting of future seismic activity cycles. This work aims to find the temporal seismic patterns of high and low seismicity in four major seismic zones: 1) the United States and Mexico, 2) South America, 3) Japan, and 4) Southern China and Northern India as sketched in Figure 1. We have made a probabilistic long-term earthquake prediction using a Bayesian ML model in these seismic zones based on the seismic patterns deduced from our wavelet analyses.



Earthquake prediction using Machine Learning. Image by author

TYPES OF EARTHQUA

- **Earthquakes basic concepts**
- **Related studies**
- **Data**
- **Modelling**
  - Problem statement
  - Preprocessing and feature engineering
  - Model selection
- **Results and discussion**

Earthquakes basic concepts

Earthquakes are well-studied events, with plenty of academic studies coverage, so only the basic concepts will be described here.

The majority of seismic activity happens between the movement of lithospheric plates (a.k.a. *tectonic* plates). This movement accumulates energy in the form of rock stress, and then it is suddenly released.

After the quake happens, it can be determined the location (longitude, latitude, and depth), time, and magnitude. Magnitude is the physical size of the earthquake, and the energy released can also be roughly estimated by converting the moment magnitude [1].

Earthquakes can cause destruction and loss of lives. Not only by the ground shaking event but also by secondary effects such as landslides, fissures, avalanches, fires and tsunamis [2].

Between 1998–2017, earthquakes caused nearly 750,000 deaths globally, more than half of all deaths related to natural disasters. More than 125 million people were affected by earthquakes during this time period, meaning they were injured, made homeless, displaced or evacuated during the emergency phase of the disaster.

- *World Health Organization*

Building a pre-emptive warning system can greatly increase risk management effectiveness. Being able to prepare for those rare events would help to minimise the harm caused, with actions such as local community alert and government provisioning.

➢ Related studies

To the best of my knowledge, 2 studies try to predict when the next earthquake will happen with machine learning [3][4]. Both conclude that is very difficult to predict the next occurrence, due to its randomness and difficulty to prove that earthquakes follow a specific pattern.

It is important to note that both studies use the table of recorded earthquakes to build the machine learning model. Please refer to the **Problem statement** sub-section for further discussion.

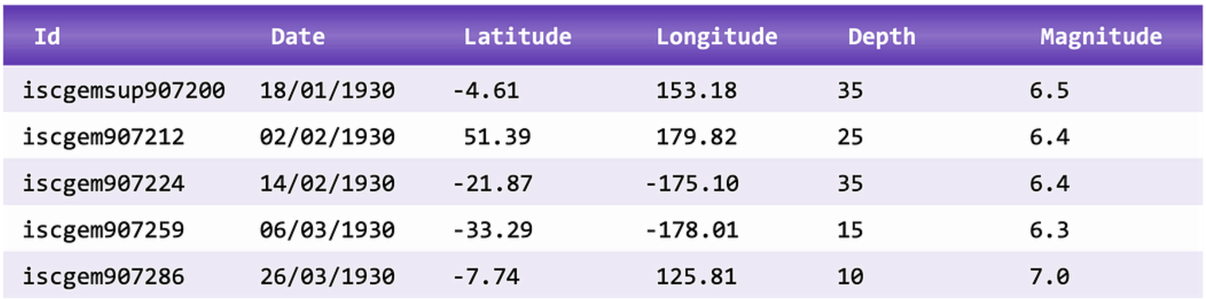
Other ML applications have also been explored:

- A study achieved nice results focusing on predicting aftershock events, which happen after larger ones, and is an important subject since aftershocks cause a lot of damage as well [5]. Some discussions have arisen about the data science methodology used [6][7][8].

- Laboratory earthquakes experiments are studied through the application of ML trying to predict time to failure [9][10].
- Another paper found patterns in energy signals from low-amplitude seismic waves to the timing of slow slip events [11].
- The lateral spreading prediction has been explored [12]. A competition for modelling earthquake damage has also been held [13].
- Earthquake detection and phase picking automation [14].

➢ Data

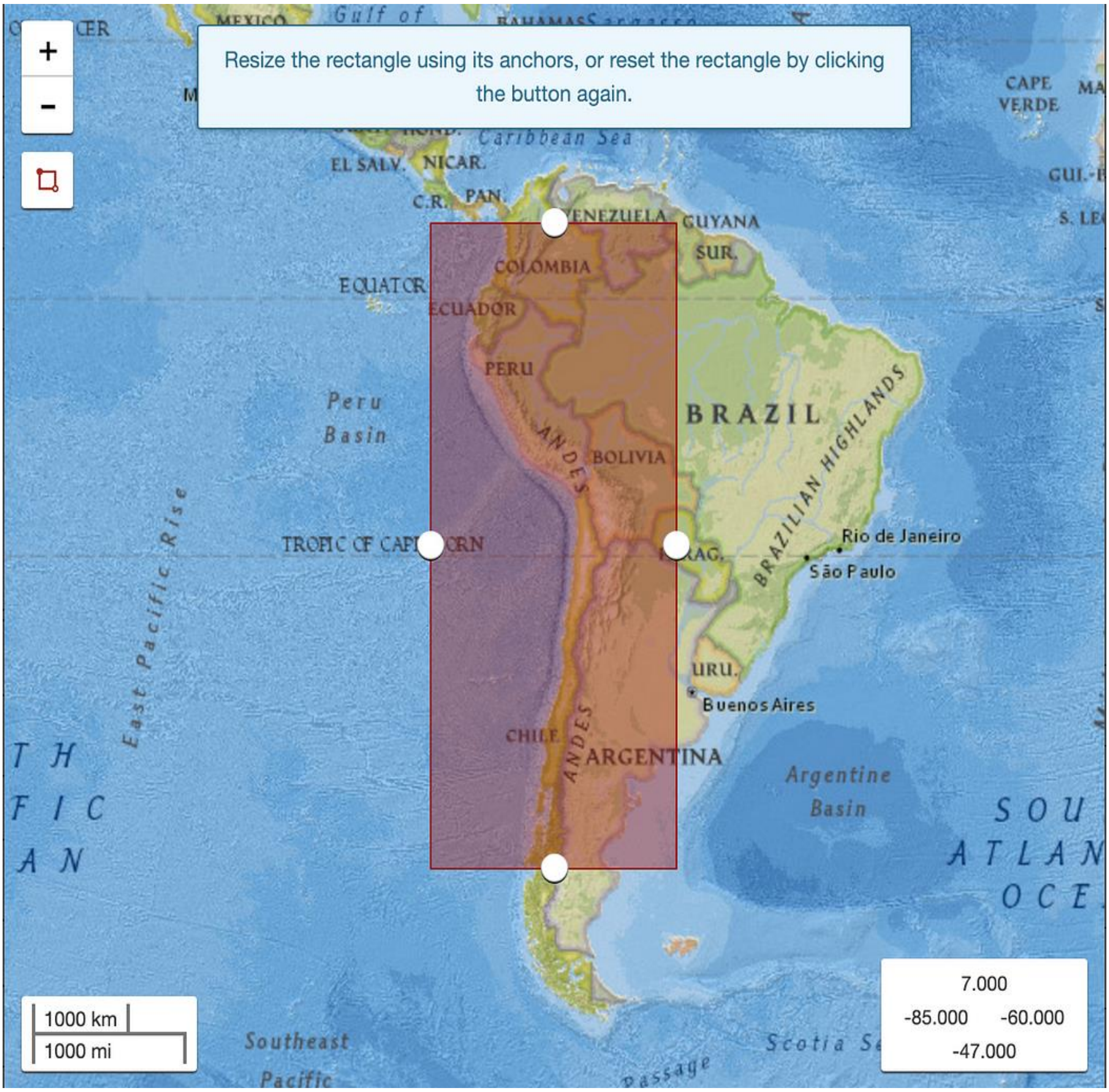
The raw data was sourced from The United States Geological Survey (USGS) earthquake catalog [15]. All worldwide earthquakes from the beginning of records until the end of 2018 were downloaded and later filtered as described below.



Id	Date	Latitude	Longitude	Depth	Magnitude
iscgemsup907200	18/01/1930	-4.61	153.18	35	6.5
iscgem907212	02/02/1930	51.39	179.82	25	6.4
iscgem907224	14/02/1930	-21.87	-175.10	35	6.4
iscgem907259	06/03/1930	-33.29	-178.01	15	6.3
iscgem907286	26/03/1930	-7.74	125.81	10	7.0

Earthquakes recorded data sample. Source [15]. Image by author.

The Nazca-South American plate boundary area was selected (latitude between -47º and 7º, and longitude between -85º and -60º).



Resize the rectangle using its anchors, or reset the rectangle by clicking the button again.

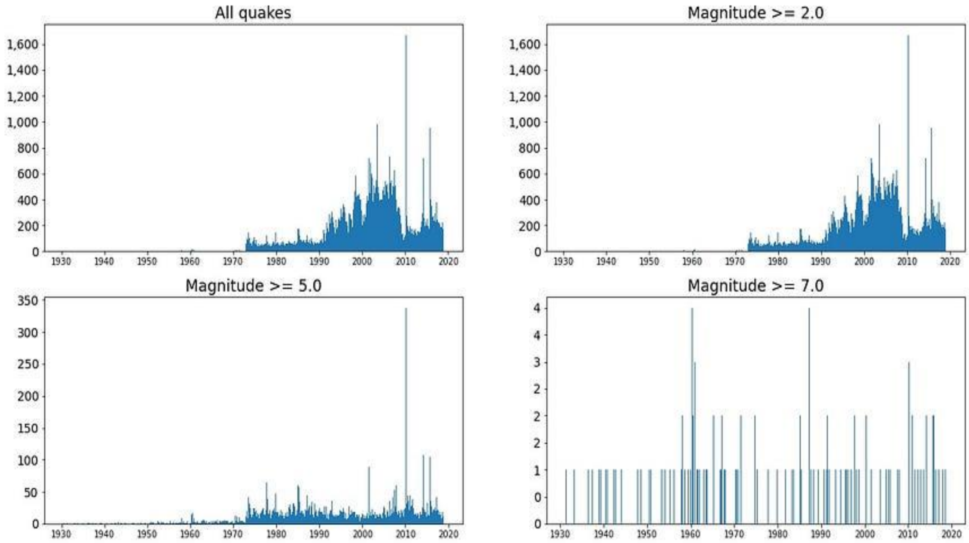
7.000

-85.000 -60.000

-47.000

Area selected. Source [15].

Years between 1973 and 2018 were selected. Comparing the histograms for the number of earthquakes across dates, there is a clear increase in the number of recorded events, mostly for lower magnitudes. This is most likely due to an increase in the number of seismometers, not an actual increase in the number of quakes.



Histograms of earthquakes. Image by author.

Please refer to the **Next steps** section for more information.

All the downloaded data used for this study can be found here.

➢ Modelling

○ Problem statement

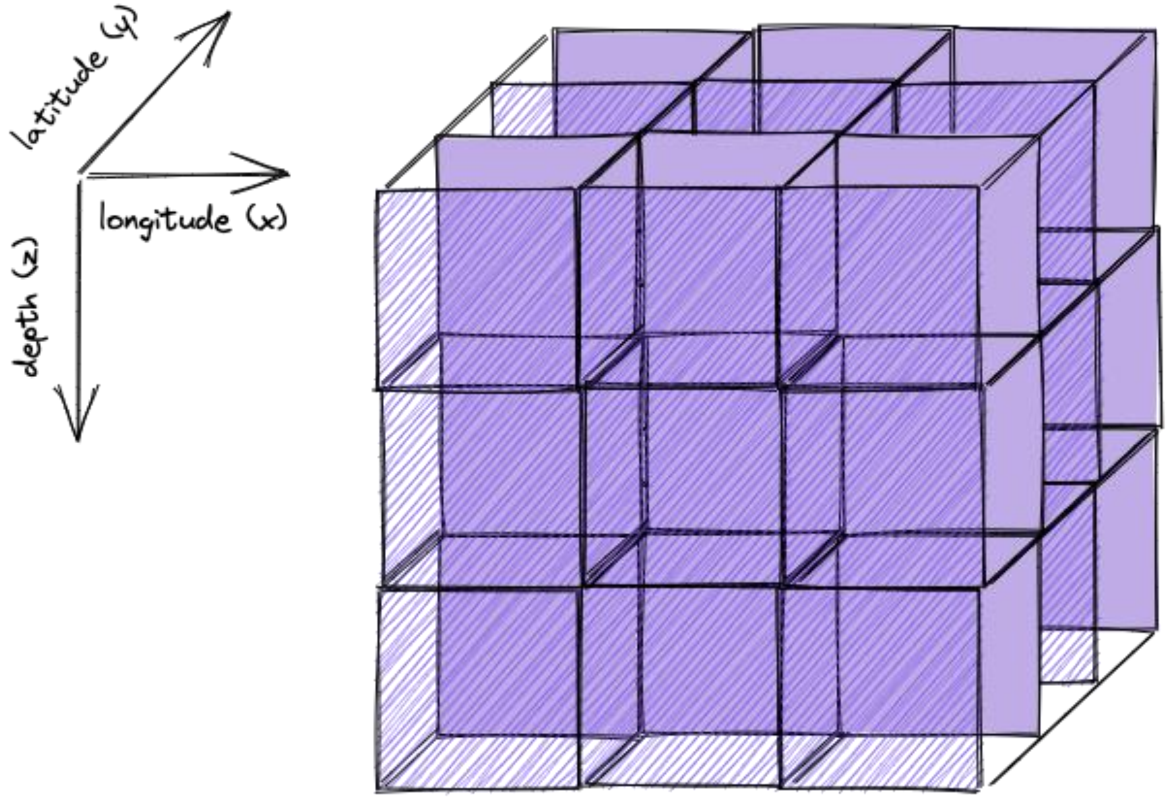
Instead of using the table of recorded earthquakes to appraise the final model, a different approach was selected.

If the goal is to build a warning system capable of predicting the earthquake risk for any time period and specific areas, in my opinion, a fairer assessment of results is more complex. We need to replicate a real scenario in our dataset and **add the time periods**

with no seismic events in our time frame. That ensures that we will evaluate the predictions even when there is no earthquake.

The following steps were done to achieve that.

First, the selected area was divided into a 3D grid. The spatial resolution dimensions selected were 10 degrees latitude, 12 degrees longitude, and 100 km depth.



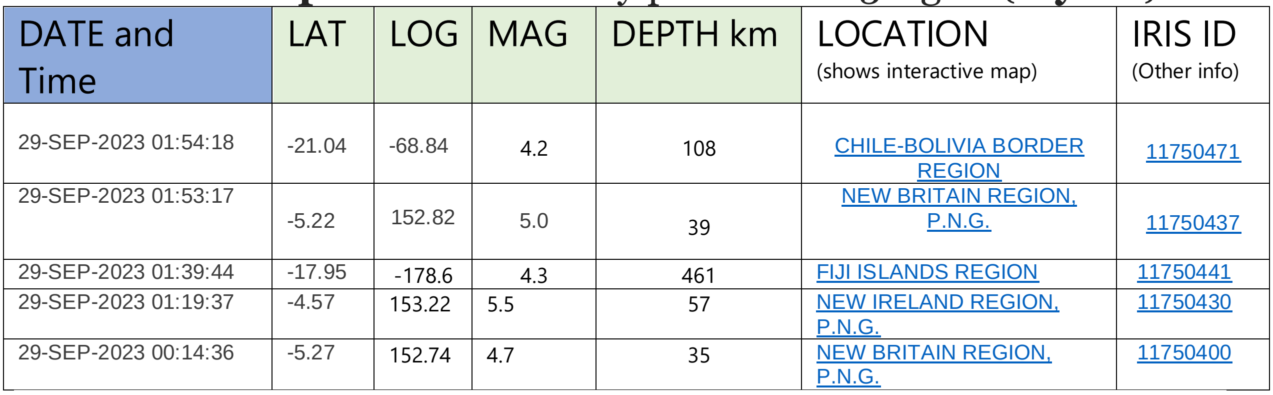
3D grid representation. Made with Excalidraw. Image by author.

Second, the data was grouped time-wise for two periods, thus having two final models. One with 7 days (**weekly model**) and another with 1 day (**daily model**). It was also added a range warning of periods, 2 for the **weekly model** and 3 for the **daily model**. For example, for the **daily model**, if an earthquake will happen on Friday, not only

Thursday evening it should make an alert, but also Wednesday and Tuesday. For the **weekly model** , if an earthquake will happen on the third week of a given month, it should make alerts on the first and second weeks.

Lastly, for the actual type of alert, it was chosen to warn for every quake with a magnitude (M) greater than or equal to **5.0**. This magnitude level was chosen because it can cause damage if is close to a population centre (not only regarding the latitude-longitude plane but also the depth, being or not close to the surface). Even earthquakes with lower magnitudes have already caused deaths, e.g. 4.9 M Afghanistan 1997 killed 15 people [16], although that is uncommon. Starting from that level, they can become even more destructive, e.g. 5.3 M Tajikistan 1989 killed 274 people [17].

In summary, the data was translated into a **time-series binary classification problem** for every point in the 3D grid (**x-y-z-t**).



DATE and Time	LAT	LOG	MAG	DEPTH km	LOCATION (shows interactive map)	IRIS ID (Other info)
29-SEP-2023 01:54:18	-21.04	-68.84	4.2	108	CHILE-BOLIVIA BORDER REGION	11750471
29-SEP-2023 01:53:17	-5.22	152.82	5.0	39	NEW BRITAIN REGION, P.N.G.	11750437
29-SEP-2023 01:39:44	-17.95	-178.6	4.3	461	FIJI ISLANDS REGION	11750441
29-SEP-2023 01:19:37	-4.57	153.22	5.5	57	NEW IRELAND REGION, P.N.G.	11750430
29-SEP-2023 00:14:36	-5.27	152.74	4.7	35	NEW BRITAIN REGION, P.N.G.	11750400

Problem statement for the daily model. Earthquake usp0000yt6. Image by author.

Please refer to the **Next steps** section for a discussion on the parameters and type of problem.

Preprocessing and feature engineering

The core of this modelling is to track energy dispersion. Hence all earthquakes, above or not the selected warning level, were transformed into energy, allowing us to group different events on the same 3D grid point (that cannot be done with magnitude: two events of M = 3 are not the same as one of M = 6).

$$\text{Log}\ Energy = 5.24 + 1.44\ Magnitude$$

Moment magnitude to energy equation. Source [1]. Image by author.

The following step is to reduce the data according to the problem statement. From that is yielded the energy for each point, for every time period, **filling periods with no events with 0 energy**.

With a well-formatted x-y-z-t dataset, the feature engineering process can be done. Features created are energy moving averages (periods of 30, 60, 90, 180, 330, and 360), ratios of those M. A., and also the moving average for the neighbours' 3D data points. Those last features are created to try to capture the relation of energy between close points.

Another feature created is the tracking of days from the last event, an attempt to capture the frequency of events.

The resulting datasets characteristics are displayed here:

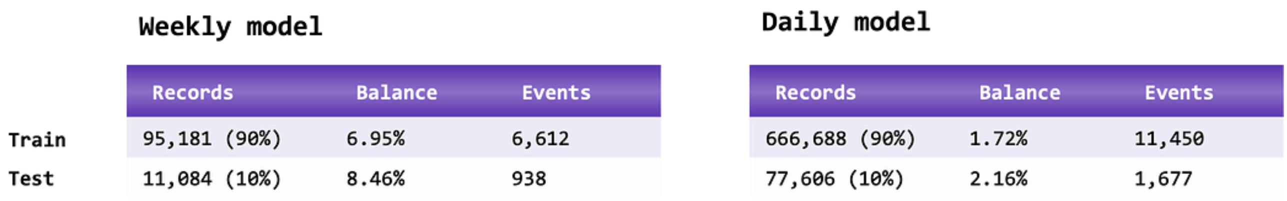
```
* weekly model
Balance: 7.10%
Number of records: 106,265
Number of features: 18
```

```
* daily modelBalance: 1.76%
Number of records: 744,294
Number of features: 18
```

➢ Model selection

Since this is a highly imbalanced problem, a better metric to be used is the F-score, which is the weighted harmonic mean between **precision** and **recall**. In one of my previous articles [18], I explain the difference between those metrics. Precision is penalised from false alarms, and recall is penalised from missed events.

The data split was 90% train and 10% test.



Weekly model

	Records	Balance	Events
Train	95,181 (90%)	6.95%	6,612
Test	11,084 (10%)	8.46%	938

Daily model

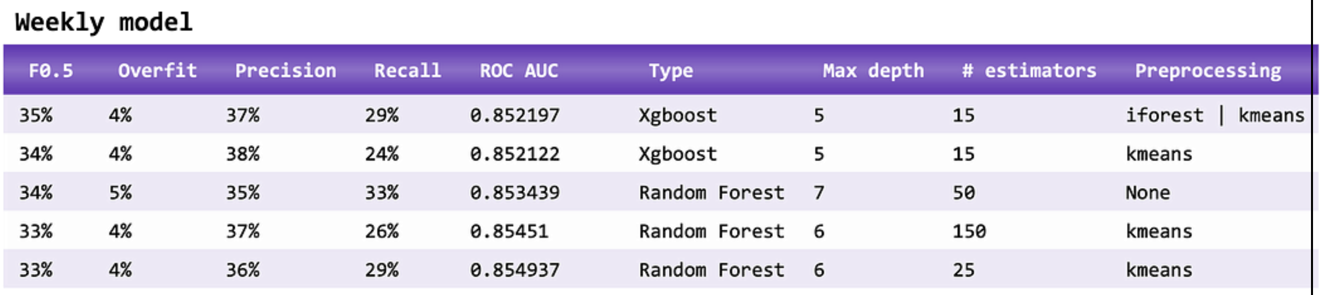
Records	Balance	Events
666,688 (90%)	1.72%	11,450
77,606 (10%)	2.16%	1,677

Datasets split balances. Image by author.

Within the training data, a grid search was performed to find the best models, using a 3-fold cross-validation time-series.

```
###########################
# linear models
###########################lin_params = {
    'C' : [0.01, 0.1, 1.0],
    'solver': ['lbfgs', 'newton-cg']
}###########################
# random forest models
###########################rft_params = {
    'max_depth': [6, 7, 8],
    'n_estimators': [25, 50, 75, 100, 150],
}###########################
# xgboost models
###########################xgb_params = {
    'max_depth': [5, 6, 7],
    'n_estimators': [15, 25, 35],
}###########################
# additional preprocessing
# all features have the NaN filled and inf values clipped
# for linear models, standard scaling is applied
###########################additional = [None, iforest, kmeans]
```

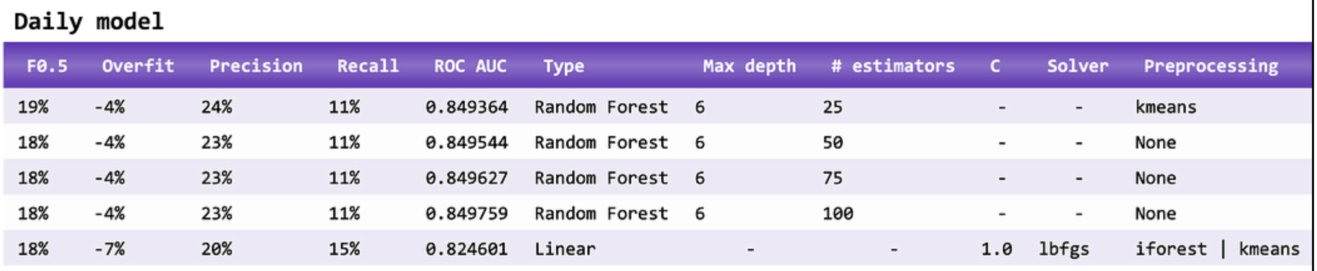
Cross-validation results:



Weekly model

F0.5	Overfit	Precision	Recall	ROC AUC	Type	Max depth	# estimators	Preprocessing
35%	4%	37%	29%	0.852197	Xgboost	5	15	iforest | kmeans
34%	4%	38%	24%	0.852122	Xgboost	5	15	kmeans
34%	5%	35%	33%	0.853439	Random Forest	7	50	None
33%	4%	37%	26%	0.85451	Random Forest	6	150	kmeans
33%	4%	36%	29%	0.854937	Random Forest	6	25	kmeans

Top 5 cross-validation results for the weekly model. Mean values. Overfit from F0.5 score. Image by author.



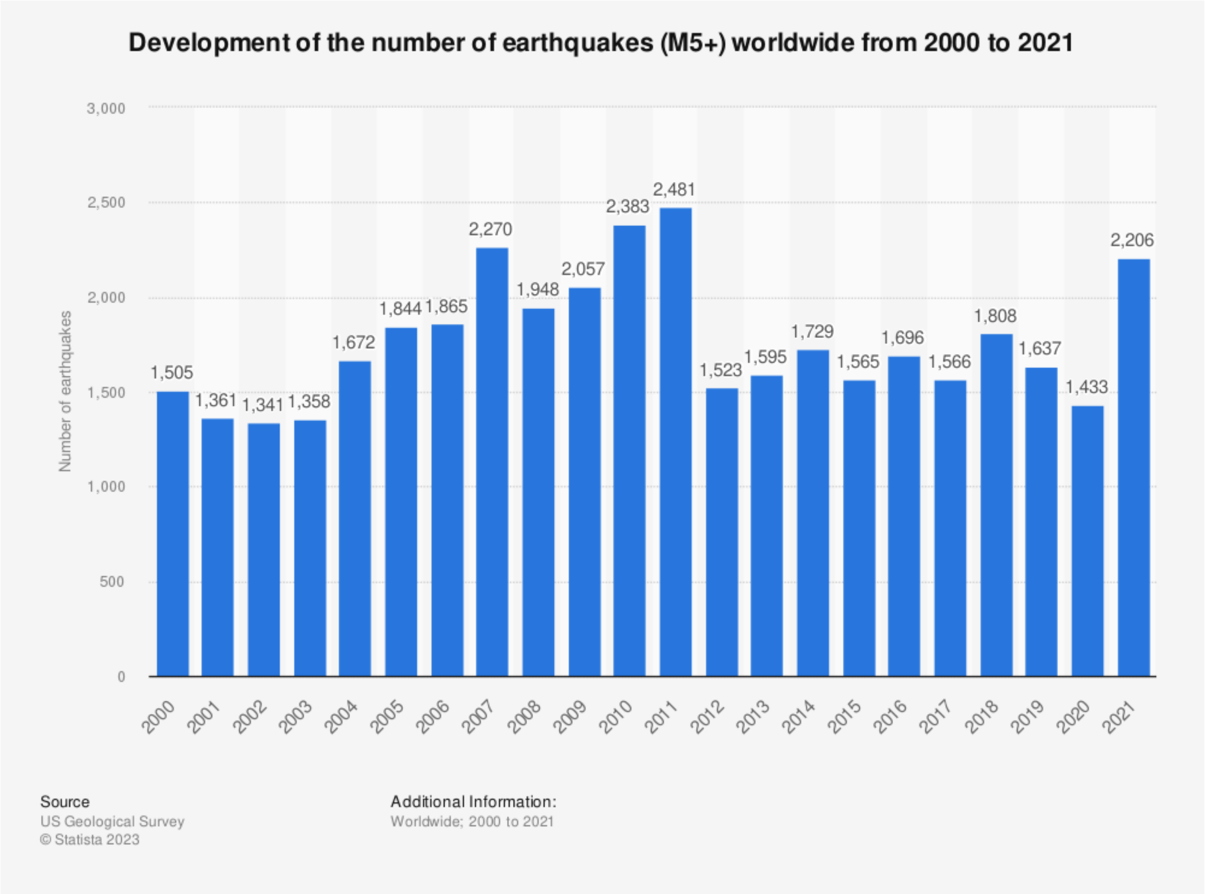
Daily model

F0.5	Overfit	Precision	Recall	ROC AUC	Type	Max depth	# estimators	C	Solver	Preprocessing
19%	-4%	24%	11%	0.849364	Random Forest	6	25	-	-	kmeans
18%	-4%	23%	11%	0.849544	Random Forest	6	50	-	-	None
18%	-4%	23%	11%	0.849627	Random Forest	6	75	-	-	None
18%	-4%	23%	11%	0.849759	Random Forest	6	100	-	-	None
18%	-7%	20%	15%	0.824601	Linear	-	-	1.0	lbfgs	iforest | kmeans

Top 5 cross-validation results for the daily model. Mean values. Overfit from F0.5 score. Image by author.

All the code can be found here.

Development of the number of earthquakes (M5+) worldwide from 2000 to 2021

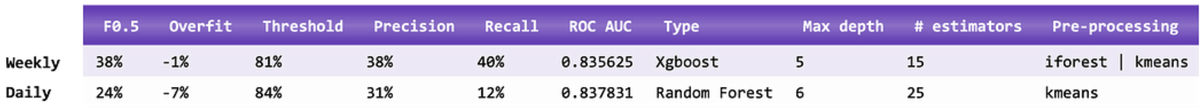


Source
US Geological Survey
© Statista 2023

Additional Information:
Worldwide; 2000 to 2021

➢ Results and discussion

With the best model, the results in the test dataset were determined:



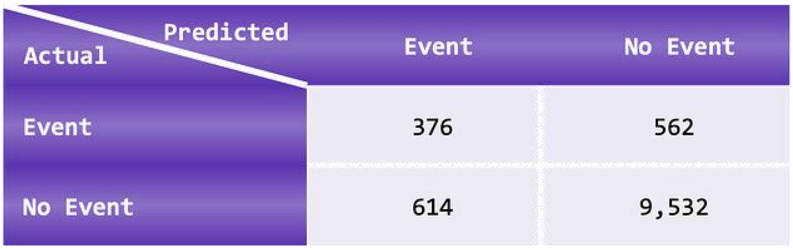
	F0.5	Overfit	Threshold	Precision	Recall	ROC AUC	Type	Max depth	# estimators	Pre-processing
Weekly	38%	-1%	81%	38%	40%	0.835625	Xgboost	5	15	iforest | kmeans
Daily	24%	-7%	84%	31%	12%	0.837831	Random Forest	6	25	kmeans

Final results for each model. Results are better than the cross-validation data because in the latter not all 90% training data is used for training, there is a fold for testing. Image by author.

For a Proof of Concept, if put in perspective of the challenge (highly imbalanced), especially for the precision, results are reasonable.

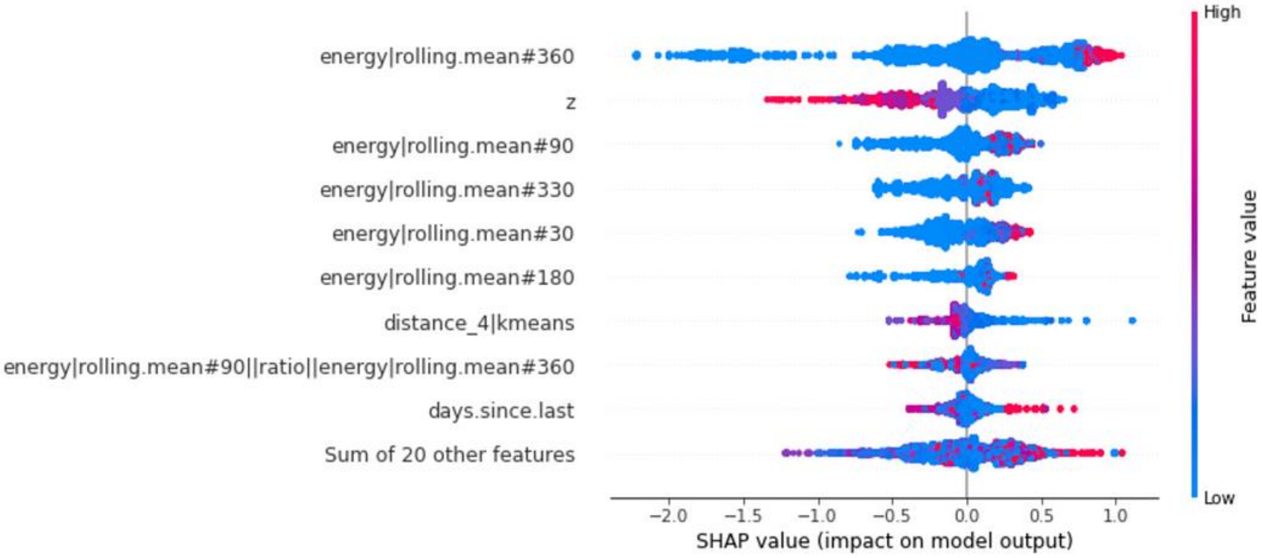
The precision and recall levels achieved for the weekly model are acceptable. Perhaps not suited for alerting the general population of some area. But it can be helpful for either governments or high-risk installations (e.g. nuclear power plants) to plan and be in a better preparedness state.

Confusion matrix and Shapley values are also presented:



Actual \ Predicted	Event	No Event
Event	376	562
No Event	614	9,532

Weekly model confusion matrix. Image by author.



Shapley values for the weekly model. Image by author.

➢ Next steps

Because this is an initial study, there are a lot of areas for potential improvement and exploration.

- **Granularity**

Both space and time can be differently discretised, but if it unbalances the model golden source (for example, increasing the spatial resolution) it will perform most likely worse. That also includes the range warning.

The dataset balance is probably one of the most important factors.

- **Problem statement**

The alert level of 5.0 can be changed, but if set for higher magnitudes, it will decrease the dataset balance, affecting its performance.

An alternative to a binary classification problem is regression, targeting the energy. Then, if the predicted energy reaches a certain level, an alert is produced.

- **Energy**

Currently, the energy calculation is an approximation, but in reality, it differs depending on the type and level of magnitude. This needs to be analysed if impacts the model.

- **Adding features**

Moving standard deviations are the next batch of features to be tested.

- **Geomagnetic field**

Adding more information, like the magnetic field from IMOs, can perhaps improve the results. The data needs to be sourced, transformed and engineered to verify if there is any hidden pattern that could improve the results.

- **Regions**

Other regions can be explored, and transfer learning is a possibility.

- **Models**

Neural Networks and Autoencoders can be tested. 3D Convolutional Neural Networks are also an option, making the location (x-y-z) part of the architecture.

- **Hyperparameter search**

More advanced and modern hyperparameter search strategies should be deployed, like Bayesian optimisation.

- **Explainability**

XAI can be more explored since here only the basics were covered. More conditional features relations can be inspected, like in here [19].

- **Sampling**

Oversampling and undersampling can also be explored.

- **GCP pipeline**

All steps can be automated, from data sourcing to hyperparameter search, and automated retraining.

References

- AI applied for business course

[1] United States Geological Survey, Earthquake Magnitude, Energy Release, and Shaking Intensity, Earthquake Hazards.

[2] World Health Organization, Earthquakes, Health topics.

[3] Sameer, Earthquake History (1965–2016): Data Visualization and Model Development (2019), Medium.

[4] K. Dilbaz, Do Earthquakes Follow A Pattern? (Part 2) (2019), Medium.

[5] DeVries, P.M.R., Viégas, F., Wattenberg, M. et al. Deep learning of aftershock patterns following large earthquakes (2018), Nature 560, 632–634.

[6] Mignan, A., Broccardo, M. One neuron versus deep learning in aftershock prediction (2019), Nature 574, E1–E3.

[7] R. Shah, Stand Up for Best Practices: (2019), Medium.

[8] Synced, Harvard & Google Seismic Paper Hit With Rebuttals: Is Deep Learning Suited to Aftershock Prediction? (2019), Medium.

[9] B. Rouet-Leduc, C. Hulbert, N. Lubbers, K. Barros, C. J. Humphreys, P. A. Johnson, Machine Learning Predicts Laboratory Earthquakes (2017), Geophysical Research Letters 44, 9276–9282.

[10] P. A. Johnson, B. Rouet-Leduc, L. J. Pyrak-Nolte, G. C. Beroza, et al., Laboratory earthquake forecasting: A machine learning competition (2021), Proceedings of the National Academy of Sciences, 118.

[11] C. Hulbert, B. Rouet-Leduc, P. A. Johnson, A Silent Build-up in Seismic Energy Precedes Slow Slip Failure in the Cascadia Subduction Zone (2019).

[12] M. G. Durante, E. M. Rathje, An exploration of the use of machine learning to predict lateral spreading (2021), SAGE Journal.

[13] DriveData, Richter's Predictor: Modeling Earthquake Damage (2021).

[14] Mousavi, S.M., Ellsworth, W.L., Zhu, W. et al. Earthquake transformer — an attentive deep-learning model for simultaneous earthquake detection and phase picking (2020). Nat Commun 11, 3952.

[15] United States Geological Survey, Search Earthquake Catalog.

[16] United States Geological Survey, M 4.9 — Afghanistan.

[17] United States Geological Survey, M 5.3 — Tajikistan.

[18] Gustavo Bighellini Martins, Leveraging Open Banking through Data Science (2020), Medium.

[19] Catarina Freitas, Going beyond churn prediction to support customer retention (2021), Medium.