Anticipating earthquake disasters in Mexico
(Predicting probability of occurrence)

Final Report

De Commons:Images donated by Roberto Esquivel Sánchez family, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=35514033

1. Background:

In the last 40 years, Mexico has suffered 137 earthquakes with a magnitude of M6.0 or higher.

In 1985, with a magnitude of M8.1, one of the biggest earthquakes inside Mexican territory happened in Michoacan state and had a very big impact in Mexico city. This event accounted for more than 10,000 deaths and 30,000+ injured. Additionally, 3,000+ buildings were demolished and 100,000 more were damaged.

Since then, billions of pesos have been lost and thousands of people have suffered the impact of the different earthquakes that have occurred mainly in the south and the Pacific coast (i.e. Guerrero and Oaxaca states) as well as the center (i.e. Mexico city) areas.

Coincidentally, in 2017, exactly 32 years after the event in 1985; on the same date, another earthquake of magnitude M7.1 hit Mexico city, leaving a total of 370 deaths and more than 6,000 people injured.

2. Problem statement:

Due to the recurring earthquakes in Mexico, an improved response program should be developed. An early alert system based on predicting the probability of occurrence of an event should be part of this program.

3. Goal:

The goal is to implement a model that can accurately predict the probability of an event occurring within a time span. This prediction will be based on the features that the dataset has, and the additional features that can be developed during the analysis of the data.

4. Dataset:

A dataset containing almost 40,000 earthquake observations from 1970 to 2022 was obtained from UNAM university data collection in Mexico (http://www2.ssn.unam.mx:8080/catalogo/).

This dataset contains the following features:

- Date – Date when the event occurred
- Time – Time when the event occurred
- Magnitude
- Latitude
- Longitude
- Location reference – Description of the location where the event occurred
- Date UTC – Date in Coordinated Universal Time
- Time UTC - Time in Coordinated Universal Time
- Status – Event was reviewed or not
- State – Name of the state in Mexico were the event occurred

5. Data wrangling:

During the Data wrangling step, I confirmed that there were not null values, but there were a few duplicated rows. These duplicated rows were not easy to find because some values in the rows were not exactly equal. It was necessary to compare each date and time values to find the duplicated rows.

Some of the unnecessary columns were eliminated too. These being: UTC Date, UTC Time, and Status columns.

There were some initial plots which indicated that Oaxaca, Guerrero, and Chiapas states, which are in the southwest part of the pacific coast, were the ones with more events.
Additionally, I confirmed that 2017 was the year with most events since 1970.

6. Exploratory Data Analysis:

During the analysis it was important to have a clear idea of the distribution of the earthquakes in a map from the last 50 years. Most of the stronger earthquakes happen in the pacific coast in the southwest of the country (Fig. 1).

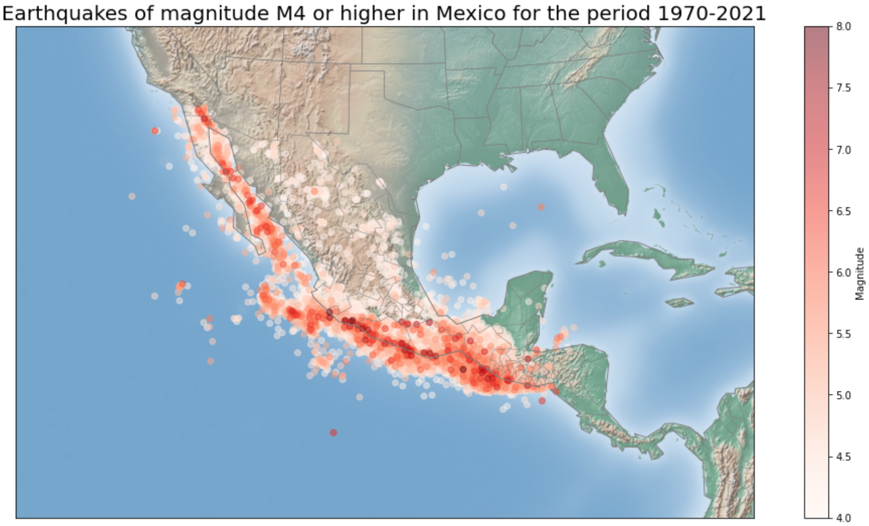

Earthquakes of magnitude M4 or higher in Mexico for the period 1970-2021
Magnitude
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0

Fig. 1

Another important point to check was the location of the tectonic plates which influence majorly in the earthquake phenomenon.

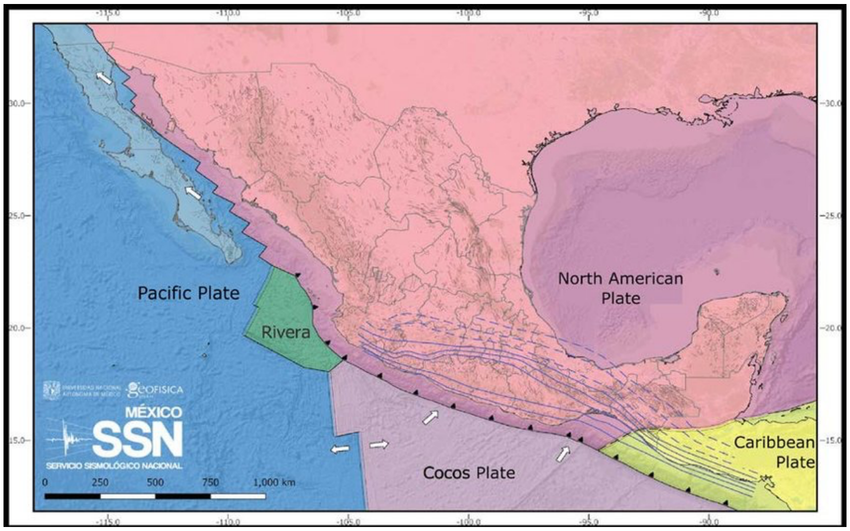


Fig. 2
Reference SSN:
https://www.researchgate.net/figure/Tectonic-plates-that-determine-the-seismicity-in-Mexico-Adapted-from-SSN-2020a_fig1_342570590

There are 3 main plates that apply pressure on the pacific coast through the North American plate. As I saw from the analysis, the main events occur on the southwest, so my main concern was the region around the Cocos plate.

After some analysis on the location of the events was done, it was important to check if there was any correlation on the features of the dataset.

A negative correlation was identified between Magnitude and Depth (Fig. 3). This indicates that stronger events, normally occur closer to the soil surface. If this happens, the impact of the earthquake will be much stronger to buildings and constructions in general.

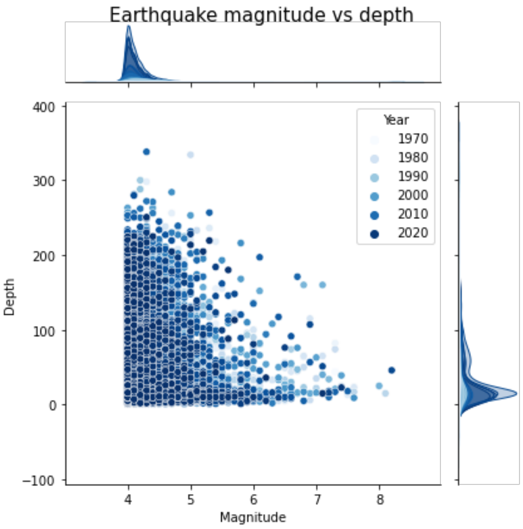


Fig. 3

Plotting a heatmap (Fig. 4) also gave me the information I needed to determine if there was correlation between the remaining features.

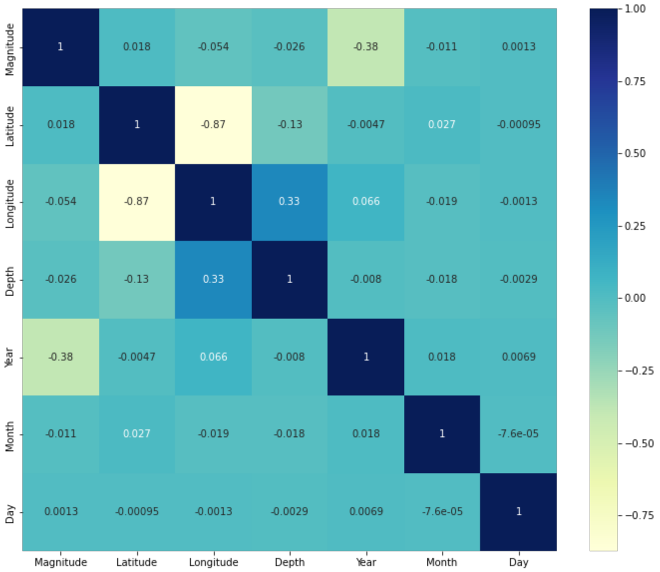


Fig. 4

The latitude and longitude negative correlation shows the distribution of the events like in Fig.1. Additionally, there is some positive correlation between depth and longitude, which indicates that deeper events occur more in the continental surface rather than the ocean.

There also seems to be the case that events greater in magnitude happened more often in previous years than now. It’s more possible that this correlation came from the fact that more events were registered in the last ten years than in the 70’s (Fig. 5). Thus, making the distribution skew to lower values in later years.

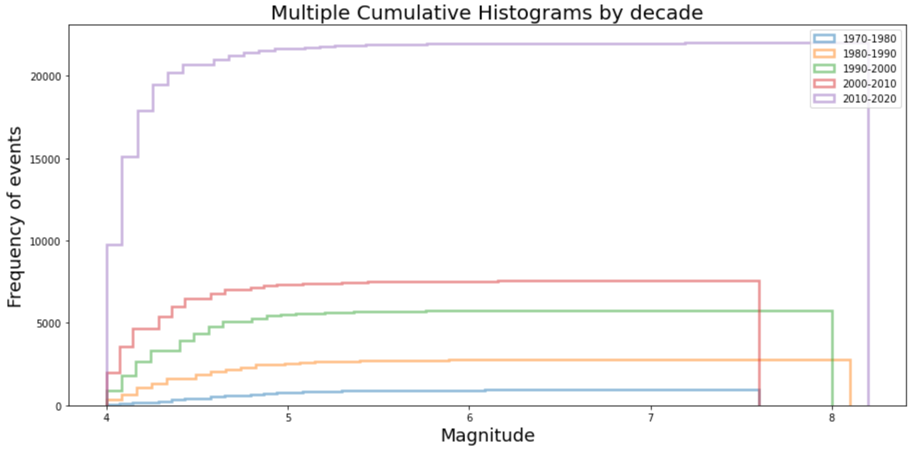


Fig. 5

Lastly, I wanted to check if there was any pattern in different years, that led to some of the biggest earthquakes to occur. The years that I investigated were 1985, 1995 and 2017 (Fig. 6).

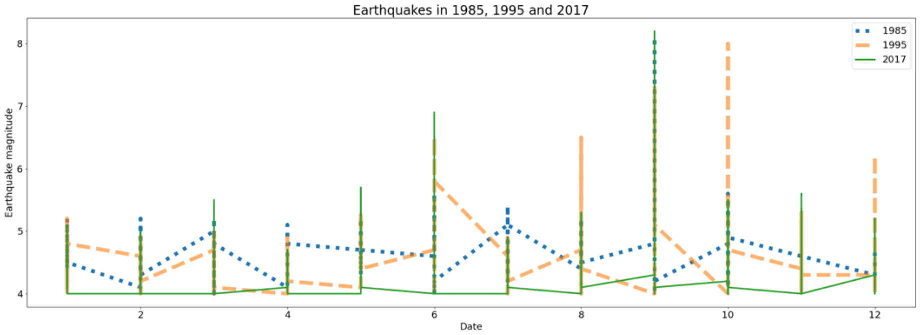


Fig. 6

Figure 6 shows that in the three years, a similar pattern showed. For all three years, the biggest event happened in September or October after increasing the magnitude in the previous months.

I was able to get some insights from the data and confirmed that the focus for events was the Cocos plate area and that there was a pattern that repeated in different years and indicated that some of the biggest events occurred in the last few months of the year.

7. The survival analysis approach:

To be able to achieve my goal, it was important to treat this problem as a “Survival analysis” problem. This type of analysis is normally used in clinical studies or predictive maintenance to predict the time until an event might happen (i.e. death of a patient, failure of a machine). So, in this case, I had to adapt this technique to my problem.

In regular survival analysis we have a number of participants in a study (i.e. number of patients in a clinical study). All the participants join the study at certain point and we record the time that passed since they joined (starting point) until the event of interest happened (i.e. death of the patient) or the patient's data becomes censored. The patient's data becomes censored when we can no longer determine if the event really happened or not. For example, follow up on some patient could be lost by any reason (they moved to another city) or the time of the study could be finished before the event happens, so we don't know what happened after this point. In all of these cases, we handle the data as censored data.

In my case, I considered every earthquake as a participant. Also, I considered every earthquake with a magnitude less than 6.0, a censored participant (class 0) and the rest of the earhquakes as uncensored participants (class 1) where the event happened.

Considering the previous information, I needed to have a ‘y’ label variable with two columns:

a) The ‘Status’ binary column which indicated if the event of interest happened (1) or not (0)
b) The ‘Time to event’ numerical column which indicated the time that passed until an event happened (censored or not censored)

Now, regarding the times to event, let's see Fig. 7 to illustrate it better. In the following image (Fig. 7), on the left side, there are different points when a participant (participants A to H) joined the study. In my case, every starting point of a participant will be the ending point of the last event in the same class. For example, if we consider participants A and F to be on the same class, then, the ending point of A will the starting point of F, just as shown in the image. A similar situation happens with participants C and G. This is how I considered the times to event. After I determined the starting and ending point of each participant (earthquake), I obtained the total time elapsed for each of them, just as shown in the right part of Fig.1.

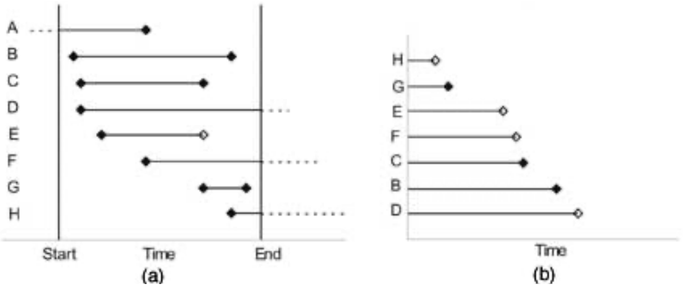


Fig. 7
Reference:
https://www.indianpediatrics.net/sep2010/sep-743-748.htm

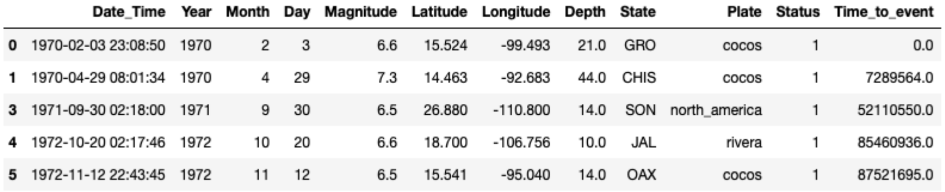
I decided to group earthquakes by state and status. It means that if two or more earthquakes shared the same values on the ‘State’ and ‘Status’ columns, they belonged to the same class, thus they were put together during the calculation of their time to event (value in seconds).

8. Feature engineering and preprocessing:

I knew that I needed to add two more columns to my dataset: ‘Status’, ‘Time to event’. But also, I thought it was important to add a ‘Plate’ column which contained the name of the tectonic plate that was closer to the location of the event. This column gave me a way to classify the states where more events happened in history.

After doing the calculations of time to event, I noticed that many of the values were less than a day. So, I decided to drop all these low values because I considered them as aftershocks.

My final dataset, prior to preprocessing, had 16424 rows and 12 columns (Table 1).

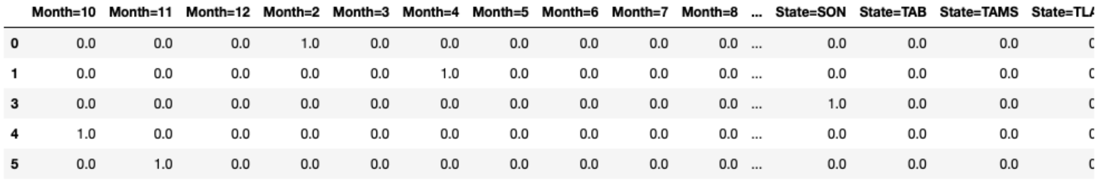


	Date_Time	Year	Month	Day	Magnitude	Latitude	Longitude	Depth	State	Plate	Status	Time_to_event
0	1970-02-03 23:08:50	1970	2	3	6.6	15.524	-99.493	21.0	GRO	cocos	1	0.0
1	1970-04-29 08:01:34	1970	4	29	7.3	14.463	-92.683	44.0	CHIS	cocos	1	7289564.0
3	1971-09-30 02:18:00	1971	9	30	6.5	26.880	-110.800	14.0	SON	north_america	1	52110550.0
4	1972-10-20 02:17:46	1972	10	20	6.6	18.700	-106.756	10.0	JAL	rivera	1	85460936.0
5	1972-11-12 22:43:45	1972	11	12	6.5	15.541	-95.040	14.0	OAX	cocos	1	87521695.0

Table 1

Prior to preprocessing I separated my dataset in ‘X’ and ‘y’. Then, during preprocessing, I decided to drop ‘Date Time’ and ‘Year’ columns. Additionally, I did standardization of numerical columns, except ‘Month’ and ‘Day’ because I wanted to treat these columns as nominal categorical.

After the standardization, I encoded all the categorical columns and my final ‘X’ dataset had 16424 rows and 78 columns (Table 2).



	Month=10	Month=11	Month=12	Month=2	Month=3	Month=4	Month=5	Month=6	Month=7	Month=8	...	State=SON	State=TAB	State=TAMS	State=TL
0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
1	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0
4	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
5	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0

Table 2

9. Modeling:

During modeling I tested 5 different models on 3 different metrics. For every model I used a train set of 80% and a test set of 20%.

5 models:

- Cox’s Proportional Hazard (CPH)
- Out-of-the-box Random Survival Forest (RSF)
- Optimized by randomized search RSF
- Out-of-the-box Gradient Boosting Survival (GBS)
- Optimized by randomized search GBS

3 metrics:

- **Concordance index for right-censored data based on inverse probability of censoring weights (CI_IPCW):** The proportion of all comparable pairs in which the predictions and outcomes are concordant.
- **Cumulative dynamic AUC (AUC):** Area under the ROC curve dependent on time
- **Integrated brier score (IBS):** An extension of the mean squared error for regression

The first model I tried was Cox's Proportional Hazards. This is one of the most used models in survival analysis, but the issue was that it overestimates the concordance index metric upwards when the data is highly right censored. In this case, my data was censored very highly (more than 90%) so I had to use a similar metric for right censored data, the Concordance index for right-censored data based on inverse probability of censoring weights (CI_IPCW). This model had a 76% CI_IPCW and performed a little better than a random model on the AUC with a mean of 57%.

The next model to check was the out-of-the-box RSF which had much better results with 84% CI_IPCW and 89% AUC. I tried to optimize this model using randomized search, but the resulting model didn't perform better. The following plot (Fig. 8) shows the comparison in AUC results between the CPH and OOB RSF models.

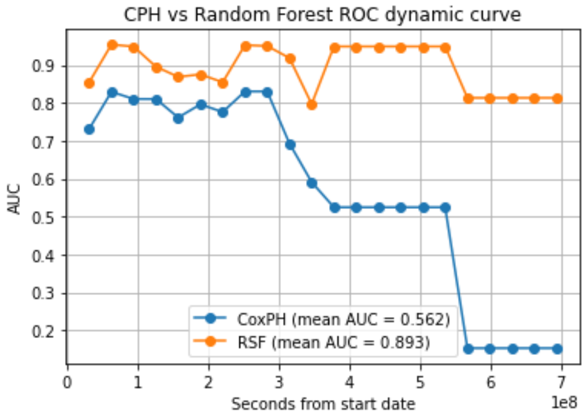


Fig. 8

After this, I trained an out-of-the-box GBS model, which didn't perform better than any of the RSF models. I tried to optimize this model too, but the result was only a bit better than the original GBS model. The following plot (Fig. 9) shows the comparison in AUC results of the best 3 models.

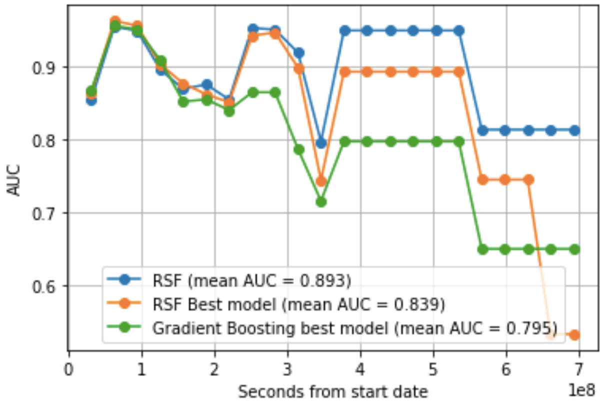


Fig. 9

The comparison on all the metrics showed clearly that the best model was the **out-of-the-box RSF model.** The following table (Table 3) shows the results for every model.

Model	CI_IPCW	AUC	IBS
CPH	0.766	0.562	-
OOB RSF	0.845	0.893	0.106
Optimized RSF	0.829	0.839	0.108
OOB GBS	0.811	0.710	-
Optimized GBS	0.819	0.795	0.120

Table 3

As a final check on the performance of the chosen RSF model, I chose 5 samples from the test set and apply the survival function plot of the model to each sample (Fig. 10).

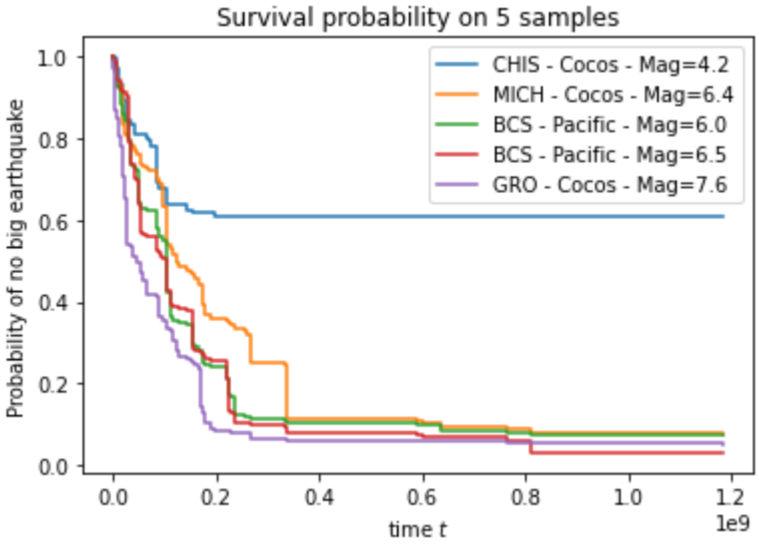


Fig. 10 (time t is in seconds)

The plot shows that there is a probability of ~90% that an earthquake of magnitude 7.6 occur in Guerrero (GRO) within the next ~6.3 years (0.2*10^9 seconds) from the last event that occurred in this state with a magnitude greater than or equal to 6.0.

8. Future work:

I will complement the dataset with more information related to the type of soil and use of it on every location to see if it is possible to improve the model's performance.

Additionally, I would like to try a different approach than survival analysis. Perhaps neural networks.