Source: Images donated by Roberto Esquivel Sánchez family, CC BY-SA 4.0,

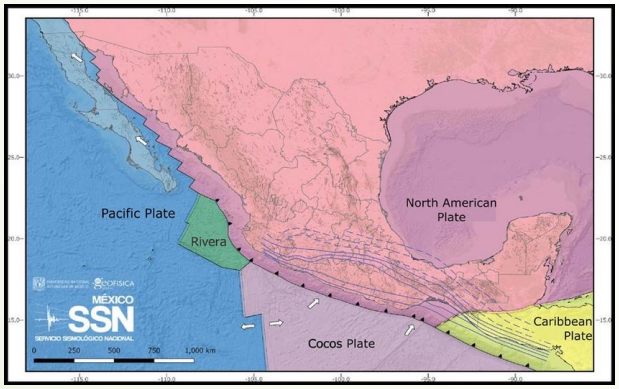
People die in earthquakes!

A system proposal to anticipate earthquake disasters in Mexico

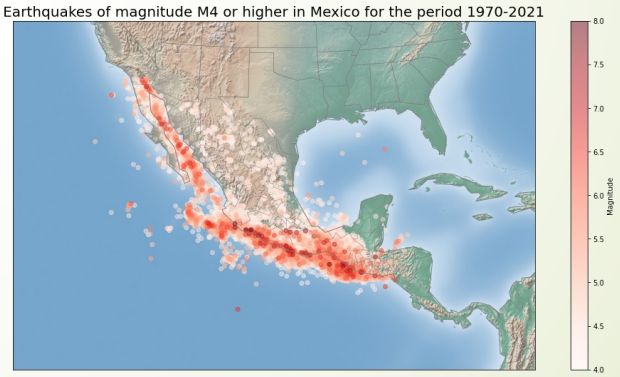
Introduction

Mexico is a country that is prone to have many earthquakes every year.

- Averages 3.4 events / year (last 40 years)
- Located in a subduction zone
- Mexico has 5 tectonic plates (left image)
- The Cocos plate is the most active
- History repeated 32 years later
- More than 10,000 deaths and ~40,000 injured



Source: SSN



Earthquake predictions in Mexico

Problem

Mexico has recurring earthquakes and does not have a way to anticipate these disasters.

- Economical and human losses

Goal

The goal is to implement a model that can accurately predict the probability of an event of M6.0 or higher occurring within a time span.

Is it possible to predict the time until the next big earthquake?

Data

All registered earthquakes in mexican territory from 1970 to 2021

- Data source: SSN (National Seismic Service at UNAM university)
- Observations: 39948 rows
- Features: 11 columns
- Features include:
  - Date
  - Time
  - Magnitude
  - Latitude
  - Longitude
  - Depth
  - State

Data Wrangling

Dataset was clean but needed data type transformation

- Missing or not corresponding values
- Duplicated information: 11 rows
- Data types: 4 numerical, 7 object (4 actual date time)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39948 entries, 0 to 39947
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Date                39948 non-null  object
 1   Time                39948 non-null  object
 2   Magnitude           39948 non-null  float64
 3   Latitude            39948 non-null  float64
 4   Longitude           39948 non-null  float64
 5   Depth               39948 non-null  float64
 6   Location reference  39948 non-null  object
 7   Date UTC            39948 non-null  object
 8   Time UTC            39948 non-null  object
 9   Status              39948 non-null  object
 10  State               39948 non-null  object
dtypes: float64(4), object(7)
memory usage: 3.4+ MB
```

Data Wrangling

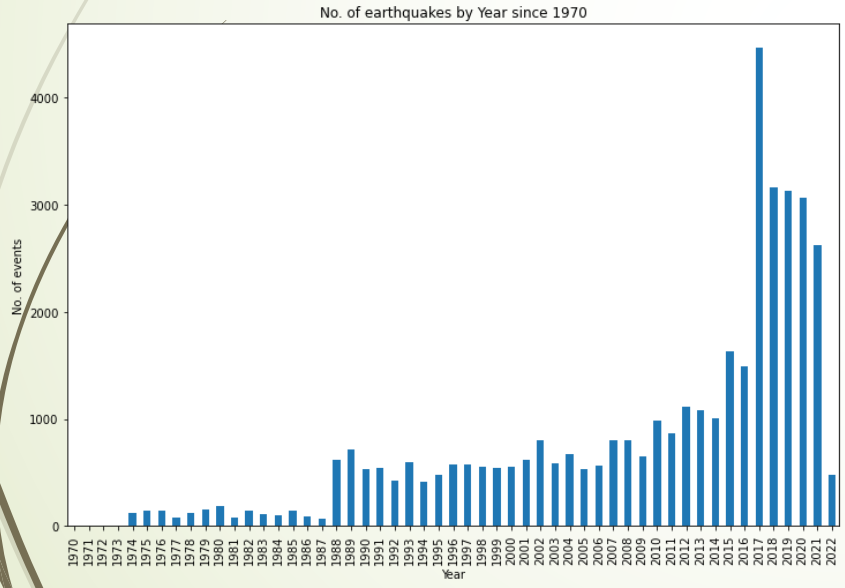
- Date and time duplicated and Status no value
- Date and time as Date time type
- Year, Month, Day added as int type
- Final Dataset: 39937 rows, 10 features

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 39937 entries, 0 to 39947
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype         
---  ------              --------------  -----         
 0   Date_Time           39937 non-null  datetime64[ns]
 1   Year                39937 non-null  int64         
 2   Month               39937 non-null  int64         
 3   Day                 39937 non-null  int64         
 4   Magnitude           39937 non-null  float64       
 5   Latitude            39937 non-null  float64       
 6   Longitude           39937 non-null  float64       
 7   Depth               39937 non-null  float64       
 8   Location reference  39937 non-null  object        
 9   State               39937 non-null  object        
dtypes: datetime64[ns](1), float64(4), int64(3), object(2)
memory usage: 3.4+ MB
```

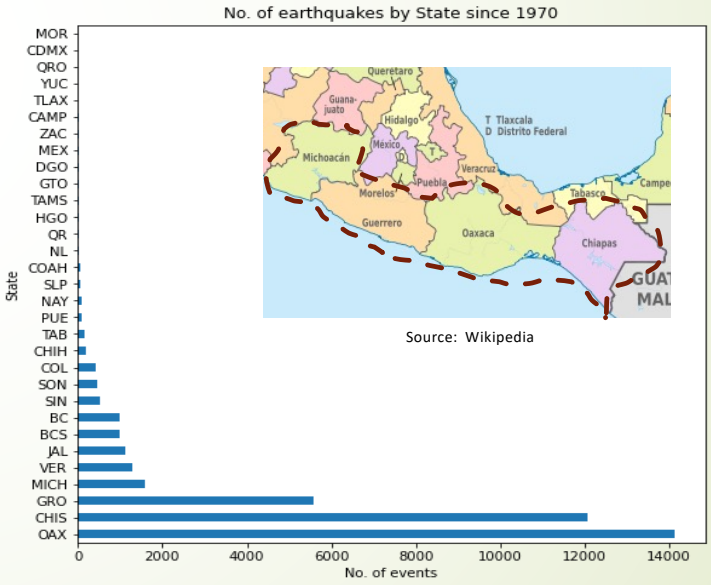
Exploratory Data Analysis

Review of the history and distribution of events

- 2017: The year with most events
- Increasing tendency



- South states in the pacific are most active

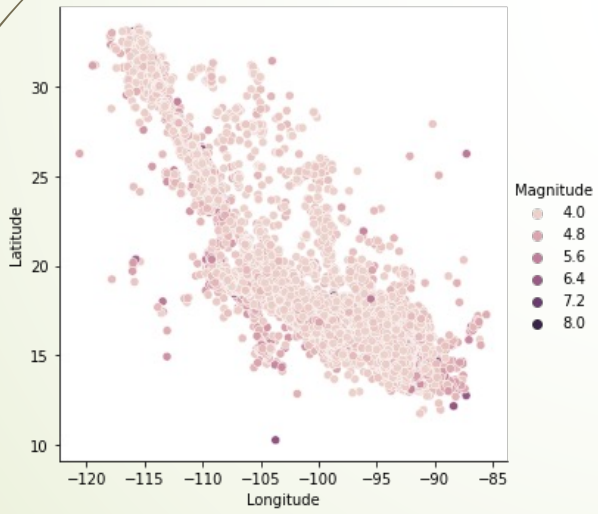


Source: Wikipedia

Exploratory Data Analysis

Correlation between features

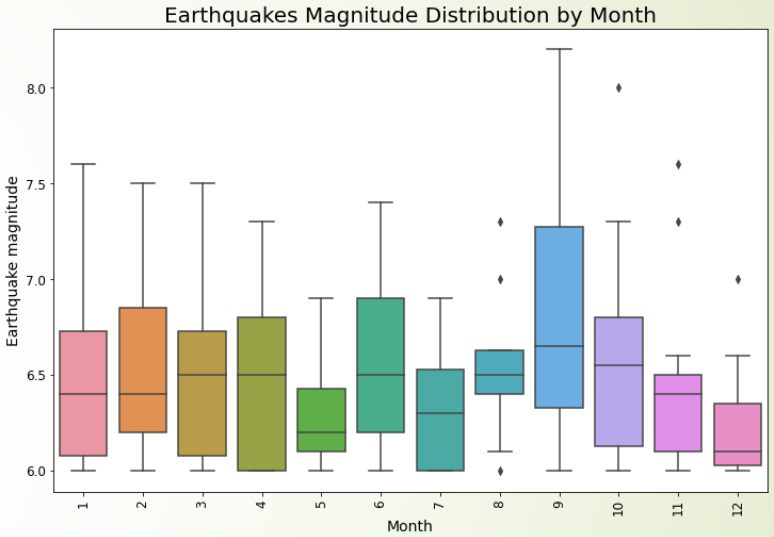
- Depth and longitude
- Magnitude and year
- Latitude and longitude

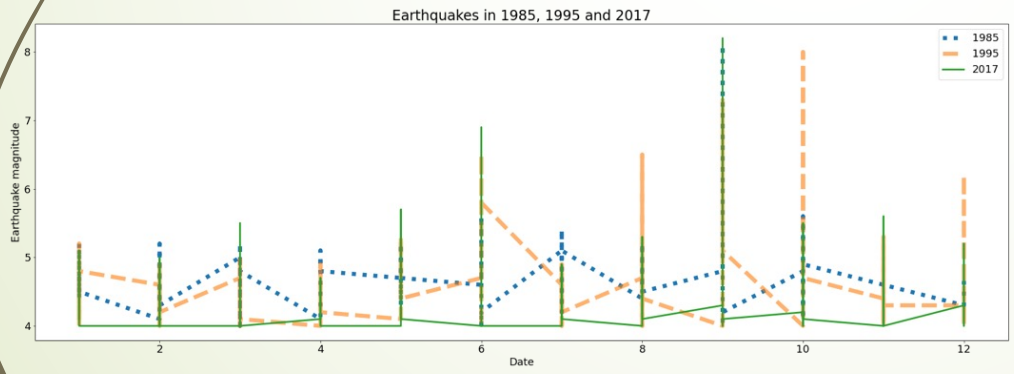


Exploratory Data Analysis

Finding patterns

- Repetitive pattern in different years
- Last few months importance

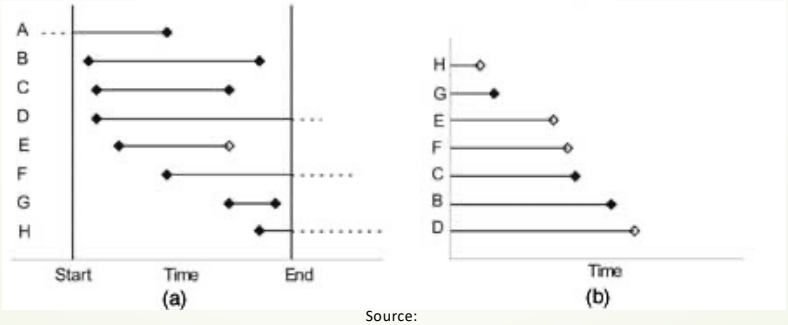




The modeling approach

Survival analysis: Estimate the survival probability within a time span

- Two columns for the 'y' label: 'Status' ('0' or '1') and 'Time to event'
- Define 'Status' values:
  - Status = 0 → earthquakes < 6.0 in magnitude
  - Status = 1 → earthquakes >= 6.0 in magnitude
- Define classes for time to event calculation:
  - Class = 0 → 'Status' = 0 and share same 'State'
  - Class = 1 → 'Status' = 1 and share same 'State'
- Concatenate same class for time calculation (i.e. A and F same class → end of A = start of F)



Source:
https://www.indianpediatrics.net/sep2010/sep-743-748.htm

Feature engineering and Preprocessing

Revise the need for additional features, standardize and encode

- Add 'y' label columns
- Add 'Plate' feature

	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

- Rows dropped in Time to event (Less than one day → Aftershocks)
- Separate in 'X' and 'y'

'X'

'y'

	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

Feature engineering and Preprocessing

- Standardize numerical columns

	Magnitude	Latitude	Longitude	Depth
count	16424.000000	16424.000000	16424.000000	16424.000000
mean	4.296432	17.870507	-99.101216	41.153172
std	0.403712	3.991145	6.099805	45.275427
min	4.000000	10.271000	-120.595000	1.000000
25%	4.000000	15.665600	-102.010000	10.000000
50%	4.200000	16.882100	-98.260000	20.000000
75%	4.400000	18.230000	-94.390000	57.100000
max	8.200000	33.127700	-85.546700	338.000000

- Encode categorical columns ('Month' and 'Day' → nominal categorical)

	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=TLA
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

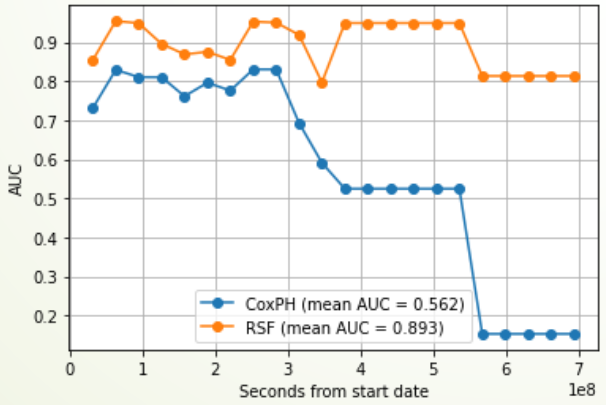
- Final Dataset → 'X' = 16424 rows, 78 columns, 'y' = 16424 rows, 2 columns

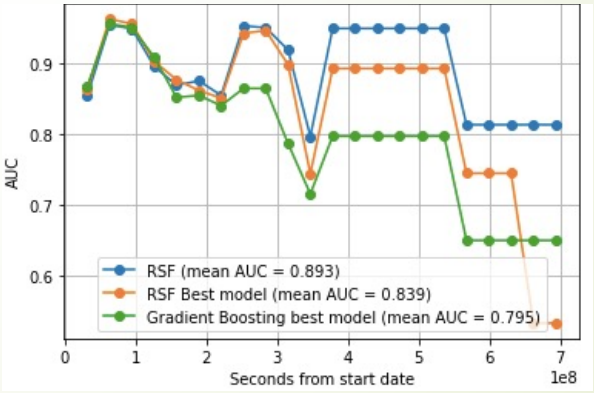
Modeling

3 metrics used to evaluate 5 models

- Models:
  - Cox's Proportional Hazard (CPH)
  - Out-of-the-box Random Survival Forest (RSF) and optimized model (RSF_best)
  - Out-of-the-box Gradient Boosting Survival (GBS) and optimized model (GBS_best)
- Metrics:
  - Cumulative dynamic AUC (AUC)
  - Concordance index for right-censored data (CI_IPCW)
  - Integrated brier score (IBS)

Cumulative Dynamic AUC results



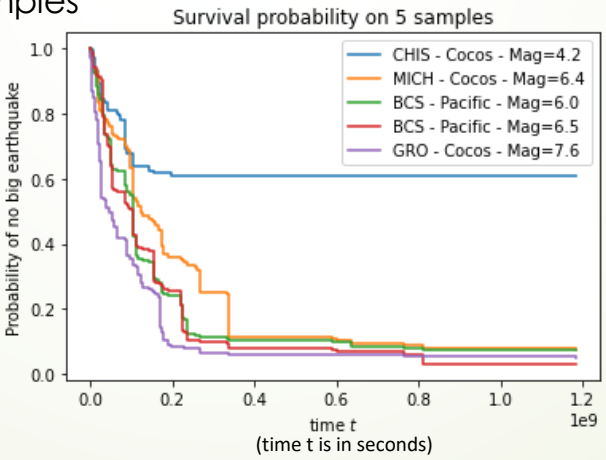


Modeling

Evaluation of 5 models

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

- The best model is out-of-the-box Random Survival Forest
- Trying the model with 5 samples



Conclusions

- Challenging project for a different approach: Survival analysis
- Used 78 features to model
- More information on the type of soil and use of it could help
- In general, the Random Survival Forest model performed the best
- Try other approach (i.e. Neural networks)

Complement the data and share with proper institutions would help to make the model more robust and accurate.