Predicting Tsunami Occurrence from Earthquake Parameters

December 5, 2025
CS 422 Intro to Machine Learning
Professor Jianwen Sun

Glossary

Background [01]

Learning Problem Description [02]

Data Preprocessing [03]

Data [04]

Planned Experiment [05]

Closing Remarks [06]

The problem

Earthquakes that occur off the coast often present an uncertain danger. While the direct damage from the earthquake can be minimal the potential shifting of the sea floor can result in dangerous tsunamis. While visual confirmation can be obtained, evacuation of outlying areas can already take longer than they have before the tsunami makes landfall.

My solution

Using data containing various seismic characteristics and tsunami indicators to build a machine learning model to accurately predict potential tsunami events.

Learning Problem Description

Binary Classification

The model will be trained on several factors to then sort later earthquakes given these same factors into whether a tsunami is likely to occur or not.

Supervised Learning

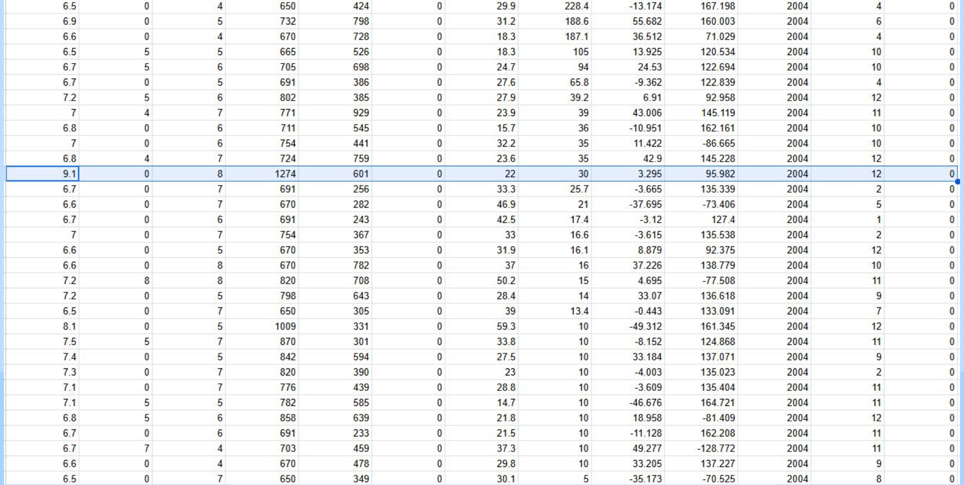
The dataset will be split into a training and test set. The training set will include the information on whether the earthquake resulted in a tsunami. The model will then predict whether the earthquakes in the test data set result in a tsunami without this confirmation.

Gradient Boosting

The dataset contains multiple seismic and tsunami-related factors that all have different correlations to a tsunami outcome both independently and dependently. The plan is to use gradient boosting to combine these individually weak factors into one accurate model.

Data Preprocessing

The dataset that I planned on using in my initial proposal was inherently flawed. All earthquakes before 2012 were marked as having not caused a tsunami.

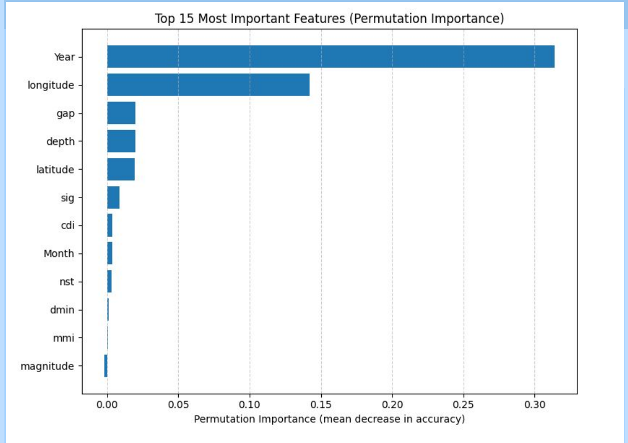


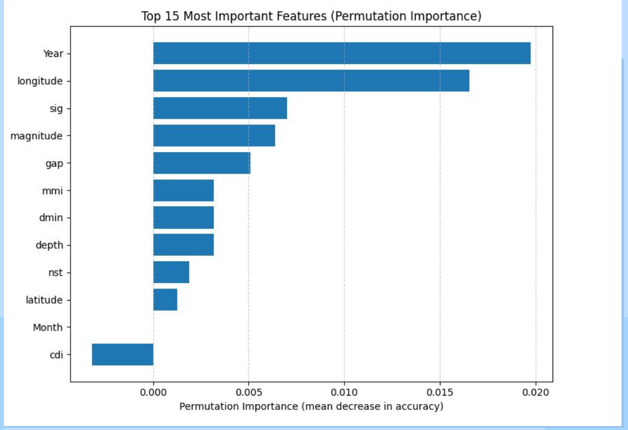
6.5	0	4	650	424	0	29.9	228.4	-13.174	167.198	2004	4	0
6.9	0	5	732	798	0	31.2	188.6	55.682	160.003	2004	6	0
6.6	0	4	670	728	0	18.3	187.1	36.512	71.029	2004	4	0
6.5	5	5	665	526	0	18.3	105	13.925	120.534	2004	10	0
6.7	5	6	705	698	0	24.7	94	24.53	122.694	2004	10	0
6.7	0	5	691	386	0	27.6	65.8	-9.362	122.839	2004	4	0
7.2	5	6	802	385	0	27.9	39.2	6.91	92.958	2004	12	0
7	4	7	771	929	0	23.9	39	43.006	145.119	2004	11	0
6.8	0	6	711	545	0	15.7	36	-10.951	162.161	2004	10	0
7	0	6	754	441	0	32.2	35	11.422	-86.665	2004	10	0
6.8	4	7	724	759	0	23.6	35	42.9	145.228	2004	12	0
9.1	0	8	1274	601	0	22	30	3.295	95.982	2004	12	0
6.7	0	7	691	256	0	33.3	25.7	-3.665	135.339	2004	2	0
6.6	0	7	670	282	0	46.9	21	-37.695	-73.406	2004	5	0
6.7	0	6	691	243	0	42.5	17.4	-3.12	127.4	2004	1	0
7	0	7	754	367	0	33	16.6	-3.615	135.538	2004	2	0
6.6	0	5	670	353	0	31.9	16.1	8.879	92.375	2004	12	0
6.6	0	8	670	782	0	37	16	37.226	138.779	2004	10	0
7.2	8	8	820	708	0	50.2	15	4.695	-77.508	2004	11	0
7.2	0	5	798	643	0	28.4	14	33.07	136.618	2004	9	0
6.5	0	7	650	305	0	39	13.4	-0.443	133.091	2004	7	0
8.1	0	5	1009	331	0	59.3	10	-49.312	161.345	2004	12	0
7.5	5	7	870	301	0	33.8	10	-8.152	124.868	2004	11	0
7.4	0	5	842	594	0	27.5	10	33.184	137.071	2004	9	0
7.3	0	7	820	390	0	23	10	-4.003	135.023	2004	2	0
7.1	0	7	776	439	0	28.8	10	-3.609	135.404	2004	11	0
7.1	5	5	782	585	0	14.7	10	-46.676	164.721	2004	11	0
6.8	5	6	858	639	0	21.8	10	18.958	-81.409	2004	12	0
6.7	0	6	691	233	0	21.5	10	-11.128	162.208	2004	11	0
6.7	7	4	703	459	0	37.3	10	49.277	-128.772	2004	11	0
6.6	0	4	670	478	0	29.8	10	33.205	137.227	2004	9	0
6.5	0	7	650	349	0	30.1	5	-35.173	-70.525	2004	8	0

Data Preprocessing

I attempted to solve this issue cross referencing this dataset with a tsunami dataset encompassing the date range of my existing dataset.

Tsunami Dataset Source: https://www.kaggle.com/datasets/harshalhonde/tsunami-events-dataset-1900-present





Dataset

Dataset Information

Source

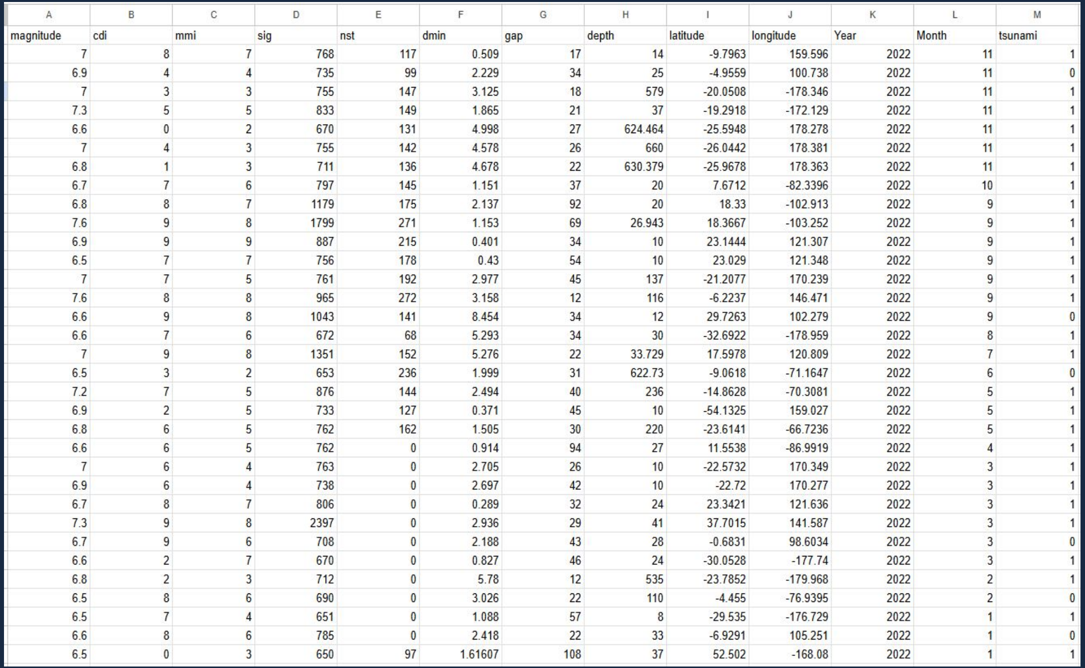
https://www.kaggle.com/datasets/ahmeduzaki/global-earthquake-tsunami-risk-assessment-dataset?resource=download

Records

782 earthquakes between January 1, 2001 and December 31, 2022.

Tsunami Event Percentage

264 or 33.76% of earthquakes resulted in tsunami.



A	B	C	D	E	F	G	H	I	J	K	L	M
magnitude	cdi	mmi	sig	nst	dmin	gap	depth	latitude	longitude	Year	Month	tsunami
7	8	7	768	117	0.509	17	14	-9.7963	159.596	2022	11	1
6.9	4	4	735	99	2.229	34	25	-4.9559	100.738	2022	11	0
7	3	3	755	147	3.125	18	579	-20.0508	-178.346	2022	11	1
7.3	5	5	833	149	1.865	21	37	-19.2918	-172.129	2022	11	1
6.6	0	2	670	131	4.998	27	624.464	-25.5948	178.278	2022	11	1
7	4	3	755	142	4.578	26	660	-26.0442	178.381	2022	11	1
6.8	1	3	711	136	4.678	22	630.379	-25.9678	178.363	2022	11	1
6.7	7	6	797	145	1.151	37	20	7.6712	-82.3396	2022	10	1
6.8	8	7	1179	175	2.137	92	20	18.33	-102.913	2022	9	1
7.6	9	8	1799	271	1.153	69	26.943	18.3667	-103.252	2022	9	1
6.9	9	9	887	215	0.401	34	10	23.1444	121.307	2022	9	1
6.5	7	7	756	178	0.43	54	10	23.029	121.348	2022	9	1
7	7	5	761	192	2.977	45	137	-21.2077	170.239	2022	9	1
7.6	8	8	965	272	3.158	12	116	-6.2237	146.471	2022	9	1
6.6	9	8	1043	141	8.454	34	12	29.7263	102.279	2022	9	0
6.6	7	6	672	68	5.293	34	30	-32.6922	-178.959	2022	8	1
7	9	8	1351	152	5.276	22	33.729	17.5978	120.809	2022	7	1
6.5	3	2	653	236	1.999	31	622.73	-9.0618	-71.1647	2022	6	0
7.2	7	5	876	144	2.494	40	236	-14.8628	-70.3081	2022	5	1
6.9	2	5	733	127	0.371	45	10	-54.1325	159.027	2022	5	1
6.8	6	5	762	162	1.505	30	220	-23.6141	-66.7236	2022	5	1
6.6	6	5	762	0	0.914	94	27	11.5538	-86.9919	2022	4	1
7	6	4	763	0	2.705	26	10	-22.5732	170.349	2022	3	1
6.9	6	4	738	0	2.697	42	10	-22.72	170.277	2022	3	1
6.7	8	7	806	0	0.289	32	24	23.3421	121.636	2022	3	1
7.3	9	8	2397	0	2.936	29	41	37.7015	141.587	2022	3	1
6.7	9	6	708	0	2.188	43	28	-0.6831	98.6034	2022	3	0
6.6	2	7	670	0	0.827	46	24	-30.0528	-177.74	2022	3	1
6.8	2	3	712	0	5.78	12	535	-23.7852	-179.968	2022	2	1
6.5	8	6	690	0	3.026	22	110	-4.455	-76.9395	2022	2	0
6.5	7	4	651	0	1.088	57	8	-29.535	-176.729	2022	1	1
6.6	8	6	785	0	2.418	22	33	-6.9291	105.251	2022	1	0
6.5	0	3	650	97	1.61607	108	37	52.502	-168.08	2022	1	1

Dataset excerpt

Dataset Features

Feature/Indicator	Type	Range	Description
Magnitude	Float	6.5-9.1	Earthquake Magnitude (Richter scale)
Community Decimal Intensity	Integer	0-9	Felt Intensity
Modified Mercalli Intensity	Integer	1-9	Observed Intensity and Structural Damage
Significance	Integer	650-2910	Event Significance
Number Seismic Stations	Integer	0-934	Number of seismic monitoring stations
Distance Minimum	Float	0.0-17.7	Distance to nearest seismic station in degrees
Azimuthal Gap	Float	0.0-239.0	Azimuthal gap between stations in degrees

Dataset Features

Feature/Indicator	Type	Range	Description
Depth	Float	2.7-670.8	Earthquake focal depth in kilometers
Latitude	Float	-61.85-71.63 degrees	Epicenter latitude
Longitude	Float	-179.97-179.77 degrees	Epicenter longitude
Year	Integer	2001-2022	Year of occurrence
Month	Integer	1-12	Month of occurrence
Tsunami	Binary	0,1	Binary representation of tsunami occurrence

Data Preprocessing

Once I had acquired a reliable data set, I split my features into three categories to determine what I wanted to train the machine learning model on.

Earthquake Strength and Characteristics

These features are the main focus of what we want to be training our model on. Magnitude, Intensity, Significance, and Depth

Location and Coordinate Data

Introduces risk of unwanted correlation, but necessary to give more accurate predictions with the machine learning model.

Unnecessary Noise

The year and month data while useful to identify the erroneous data will be removed for the actual training of the model along with the seismic monitoring station data.

Why Gradient Boosting Model?

Accuracy

Gradient Boosting is the method that I predict to be the most accurate in regards to the data being used.

Robustness

Gradient Boosting has a lot of potential parameters and levers through you can refine and compare to design a model that is more accurate to your dataset specifically.

Feature Importance

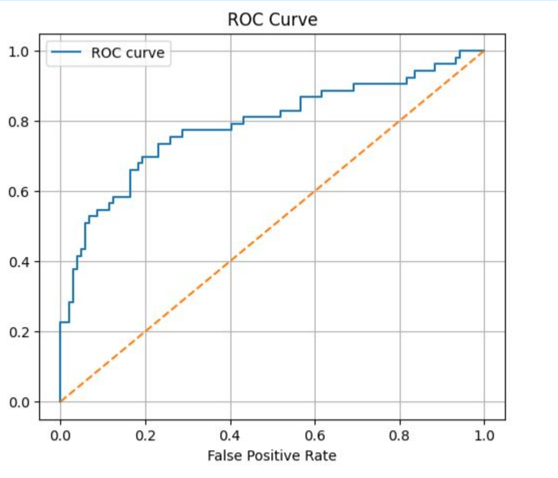
The key focus of my project was to determine the effectiveness of individual features on the predictive outcomes of the model. I found gradient boosting to be the most useful method to compare these features natively.

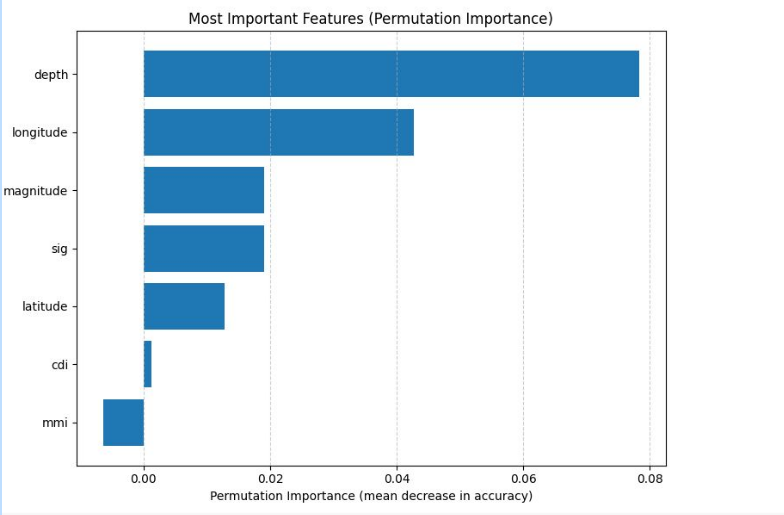
Model Predictive Analysis

After determining the best parameters for our machine learning model we acquired these values for its predictive capabilities.

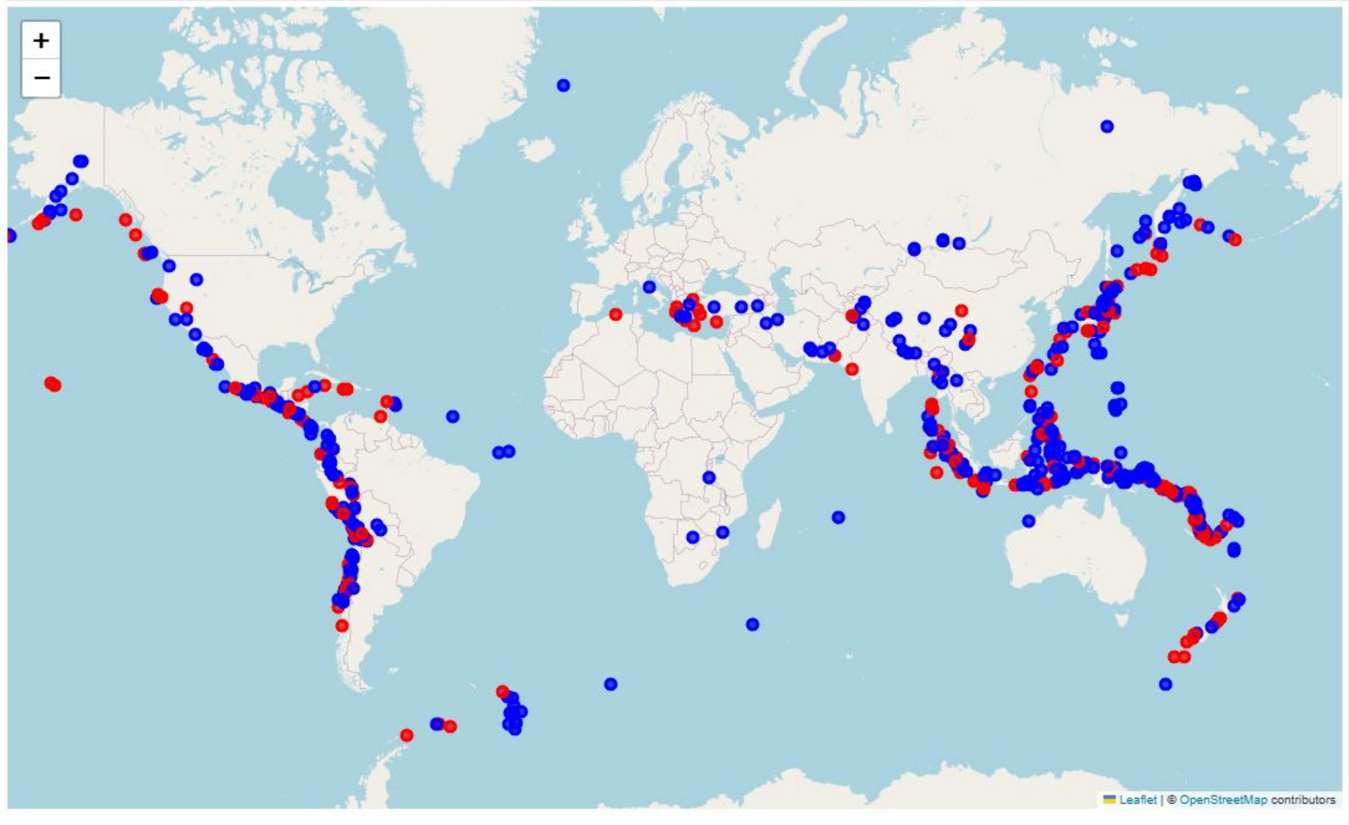
ROC AUC: 78.86%

	Tsunami	No Tsunami
Precision	64%	83%
Recall	68%	81%
F1 Score	66%	82%

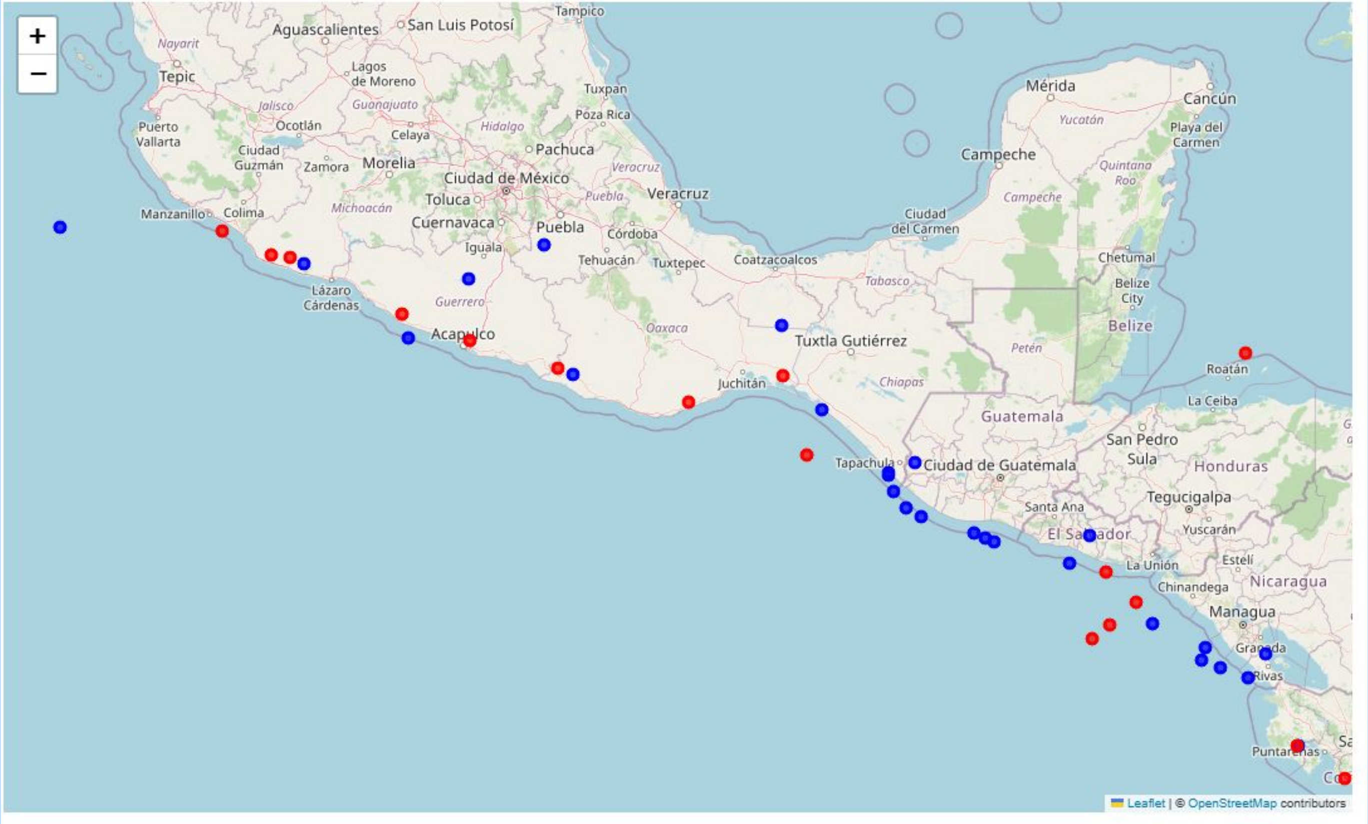




Why include Location Data?



Earthquakes most commonly occur on certain plate boundaries. Which means they naturally occur along the coast in the majority of cases

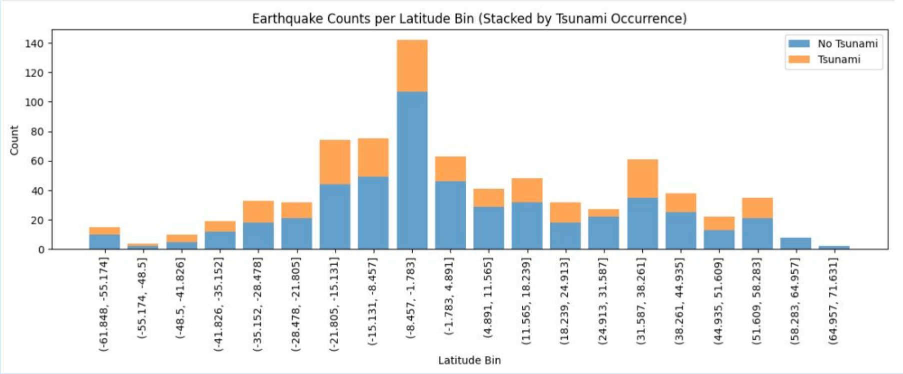


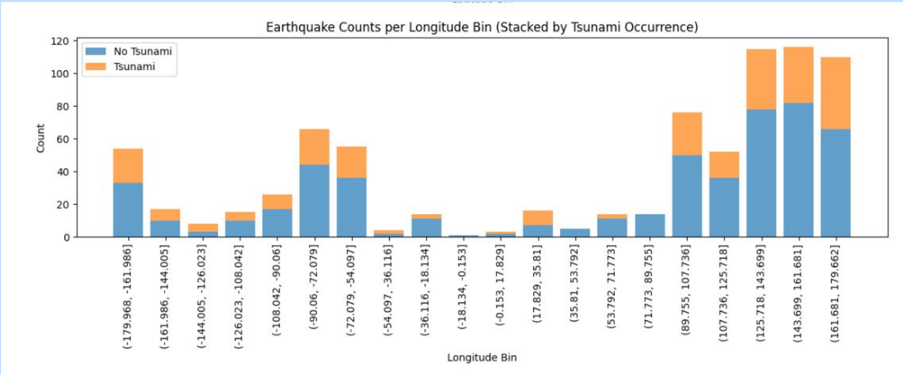
Why include Location Data?

Likewise just because the epicenter of an earthquake occurs over land does that mean that it will not result in a tsunami event.

Location Feature Analysis

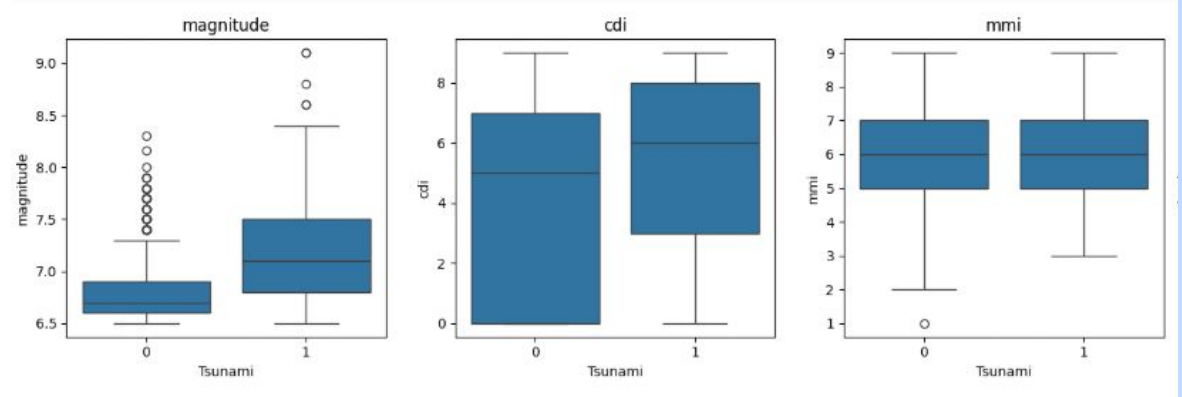
Looking at the graphs we can see that there is indeed some level of correlation between various latitude and longitude coordinates and the likelihood that an earthquake occurring in a specific range would result in an earthquake.

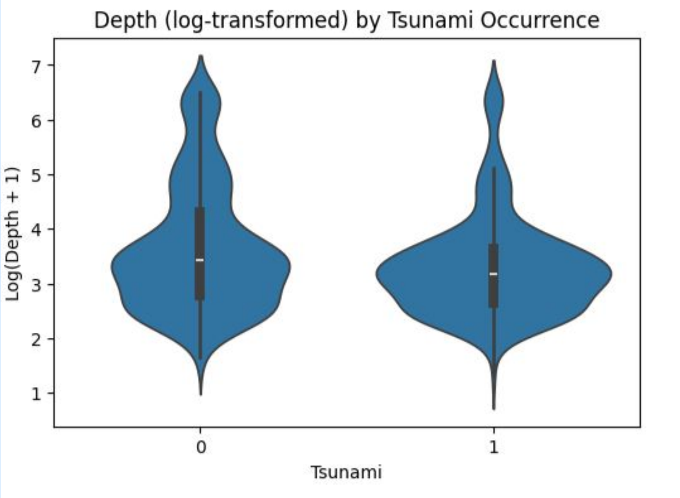




Seismic Factors

Magnitude, Community Decimal Intensity, and Modified Mercalli Intensity are all different ways of measuring the strength of an earthquake. In addition focal depth is how close to the surface the epicenter occurs and we would expect that to have a major effect on tsunami likelihood.





Conclusion

We do not see enough of a correlation between the given features and tsunami outcome to say that a model with these feature can accurately predict the likelihood of a tsunami event. Although the model can predict when a tsunami will not occur with reasonable accuracy, its inability to predict positive tsunami events with that same accuracy prevents it from achieving the goal.

Thank you

I look forward to any replies and questions you have on my project

Stephen Usselman

Email: susse001@odu.edu
Education: Undergraduate
Institution: Old Dominion University