Stephen Usselman
12/05/2025
CS 422

Predicting Tsunami Occurrence from Earthquake Parameters Using Gradient Boosting Models

Introduction

Earthquakes pose a significant tsunami risk to coastal communities. This is particularly dangerous as most earthquakes naturally occur in coastal areas due to plate tectonics. My project intends to take past earthquake's seismic data, coordinate location, and observe tsunami outcomes to predict whether future earthquakes will result in tsunami events.This project will train a gradient boosting model on data encompassing worldwide earthquakes that occurred between January 2001 and December 2022 to predict future tsunami likelihood.

Data Description

The primary dataset contains 782 earthquakes from 2001 to 2022 with the following fields used in this project

- **Magnitude** (Richter magnitude)
- **CDI** — Community Decimal Intensity (reported felt intensity)
- **MMI** — Modified Mercalli Intensity (instrumentally observed intensity)
- **Significance** — a numeric importance value combining magnitude, intensity, and impact measures

- **Depth** — earthquake focal depth (km)
- **Latitude / Longitude** — epicenter coordinates
- **Tsunami** — tsunami occurrence

During preprocessing of the data, the original tsunami field of the dataset was determined to be incorrect in many cases but especially in the 2001 - 2012 date range where every earthquake had a ‘0’ in the tsunami field to indicate no tsunami had occurred. To solve this issue I obtained a separate data of tsunami data from 1900 to 2024, which happened to encompass the original dataset’s date range. We then cross-referenced year, month, magnitude and latitude/longitude to yield a final dataset with accurate tsunami event occurrence. The final class distribution ended at 264 earthquakes resulting in a tsunami and 518 earthquakes that did not.

Machine Learning Methodology

I selected schkit-learn’s HistGradientBoostingClassifier due to its robustness of parameters, accuracy in model predictive outcomes, and feature importance analysis capabilities. I used RandomizedSearchCV with stratified 5-fold cross validation to perform hyperparemeter tuning for my model. Data was then split into an 80% training and 20% test set with balance maintained through stratified sampling.

Experimental Results

The model had an overall accuracy of 76% with a precision of 64% for the earthquake resulting in tsunami class. This shows that while the model can predict a non-tsunami event with reasonable accuracy it begins to struggle with predicting the tsunami-caused earthquake events. Given the stakes of an inaccurate prediction we could accept a reasonable amount of false positives but our threshold for false negatives must remain much higher. The ROC curve however did indicate a clear separation from the random-guess baseline suggesting that there does exist a partial correlation between the given factors and a tsunami-event occurrence.

Feature Importance

Coordinate data indicates that tsunami causing earthquakes are more likely to occur in certain geographic areas. Given our current dataset it is difficult to ascertain the root cause of this correlation. Various unknown factors including the type of plate boundary, sea conditions, distance from coastal areas in each direction, and more could all be needed to properly ascertain exactly why these areas are more likely to produce these earthquake tsunami events.

Of the seismic factors of magnitude, community felt intensity, mercalli intensity, significance, and depth we see that magnitude and depth had the most significance on predicting tsunami outcome but the overall impact on the predictive quality of the model was small. The depth field accurately displays that tsunamis are more likely to occur at lower focal depth, but it is not a hard rule and the vast majority of earthquakes occur at lower depths regardless making it less useful a parameter. Magnitude was revealed to be a far superior predictive factor than the intensity values, which aligns with my understanding that magnitude represents the true strength of an earthquake and intensity measures its effects.

Discussion

The results of this model encourage the idea that there could exist a model to accurately predict tsunami outcomes if it was just given the right parameters and dataset. Between the limited features and the flawed dataset, many of the outcomes of this project are brought into question. The fact that those same outcomes fall short of a conclusive result just exemplifies the issue at hand. Limiting the project to a single form of machine learning model could also have had a negative impact on the outcomes, but I maintain that gradient boosting was most suitable with every aspect taken into account. A future project with a similar goal would do best to acquire a superior dataset with more location-based data.

Conclusion

This project successfully developed a machine learning model to predict tsunami occurrence through earthquake-related data and a gradient boosting algorithm. Although the predictive accuracy of the model fell short of the desired parameters, the model itself performed as intended. Feature importance insights were gained from analysis between the various seismic factors as well as whether specific geological areas were more prone to tsunami-related earthquakes. Future work would best be directed towards creating a more expansive and accurate dataset that includes several additional geophysical features and is initially accurate without requiring cross-referencing.