Using Logistic Regression in the Early Detection of Tsunamis caused by Earthquakes

Ian Chen

Abstract

Tsunamis have forced large financial investments devoted to their preparation and repair. One primary cause of such tsunamis are earthquakes, often those occurring nearby. By using processed NCEI/WDS datasets for earthquakes and tsunamis, we built a logistic regression model to predict the probability a tsunami would form following a given earthquake using features such as magnitude and location with an accuracy of around 80%. In addition, a linear regression model was used to predict the resulting tsunami's size, although it had low efficacy.

1. Introduction

There have been many thousands of documented tsunamis over the last several millennia (NCEI 2022). Many of these have enough energy to cause a large amount of human death, as well as trillions of dollars in damage repair. The problem arises on how to better prepare for these natural disasters and limit its damage, especially in countries such as Japan that are frequented by tsunamis.

Of course, tsunami detection systems already exist and are implemented throughout the globe, but they still have their limitations. In 2011, the Tōhoku Tsunami and Earthquake killed over 15,000 people, and it was largely because of the surprise. Although Japan has an existing early detection system, it failed to accurately warn the public of its scale (Oskin 2022). This project seeks to use logistic regression to predict whether a tsunami would form following an earthquake, as well as provide accurate statistics on the magnitude of such an earthquake (if it would occur) using linear regression.

2. Background

The earth's lithosphere comprises tectonic plates, in addition to the crust and part of the mantle (National Geographic). According to National Geographic, the boundaries of these tectonic plates, called fault lines, are the locations where earthquakes are most commonly found. The plates are constantly shifting along with the mantle, and when two plates move into another. At these convergent boundary sites, one plate moves under the other, causing mountains, or trenches in oceans. The energy released from this can cause earthquakes, and ocean earthquakes are a main cause of tsunamis.

In assessing the risk of tsunami formation, we must look at how data is collected. Earthquake locations can be determined through GPS, and the energy of them can be measured with seismographs. Besides the seismic event, on-sea data is measured with the Deep-ocean Assessment and Reporting of Tsunamis (DART) system. According to NOAA (2006), the DART system uses an underwater sensor that connects to a surface buoy, which relays data such as pressure and water depth to satellites.

There have been other attempts at using logistic regression. Chang et al (2006) used a similar model to predict landslides. They trained a separate model for each cause factor, including rainfall-induced, typhoons, and earthquakes. This allowed for specificity and breadth at the same time. They collected data on dozens of variables. Their data was determined to be statistically significant with a p-value of 0.01 for both the typhoon and earthquake models, showing preliminary promise for this method. A similar procedure was taken for this study.

3. Dataset

Our data was obtained from the National Centers for Environmental Information (NCEI) and the World Data System (WDS). In this study, we used two datasets, the NCEI/WDS Global Historical Tsunami Database and Global Historical Earthquake Database which both have data ranging from the past 4000+ years (2100BC to present).

The Global Historical Tsunami Database had data on 2795 tsunamis from 2150 BCE to 2022 CE. This includes many features, including the date, the magnitude (using the Iida scale), the focal depth, the water height, the location (in latitude/longitude, as well as city/country name), as well as the deaths and injury counts and the housing damage. In this study, all but the following variables were removed: water height, tsunami magnitude, latitude, longitude, earthquake magnitude (that caused), date. All of these variables are numerical and no additional preprocessing was added. After removing all tsunamis with null values for our desired variables, 336 tsunamis were left. This is largely due to missing tsunami magnitude data, as well as a minimum datetime (described in section 4.1).

The Global Historical Earthquake Database had data on 6323 earthquakes from 2150 BCE to 2022 CE. This has many features, and the following were kept: date, magnitude, longitude, latitude, focal depth. In addition, several features were created: *on sea, distance to land* (described in section 4.1). On sea is a categorical boolean variable, with 0 representing not on sea, and 1 representing on sea. After removing all entries with a null value, 1317 data points remained.

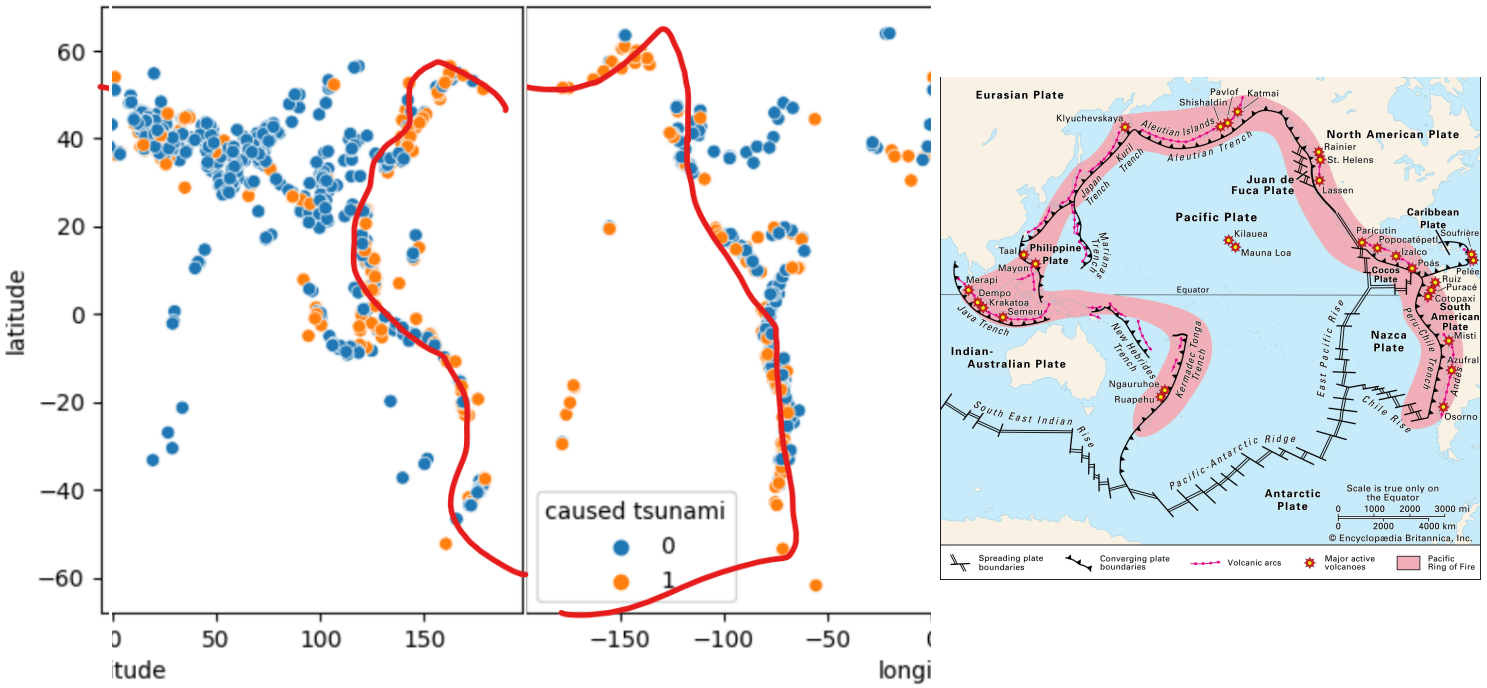

Figure 1. Left: Representation of Earthquake Data, grouped by whether a tsunami was caused. Each point is an earthquake datapoint. Right: Ring of fire tectonic plate boundary (Britannica 2022).

In addition, our study created a combined data frame that links a tsunami to its causal earthquake, which NCEI provides. This has 115 data points, with all the variables from the tsunami and earthquake datasets. All our data was originally obtained in a tsv (or tab separated values) text file directly from the NCEI/WDS website.

4. Methods

Our study uses Python 3.10 and its associated libraries. Namely, the following libraries (in addition to the standard library) are used: *numpy, matplotlib, webscrape, pandas, and sklearn*.

4.1 Preprocessing

Using Python 3.10's *pandas* library, dataframes can be used to read and edit tabular data. Once being read into a dataframe, all unwanted columns are filtered out. We would also remove all rows with null values. Additional filtering was made to specific columns; for example, only tsunamis and earthquakes with an event validity of 1 and 4 respectively (definite occurrence) were kept.

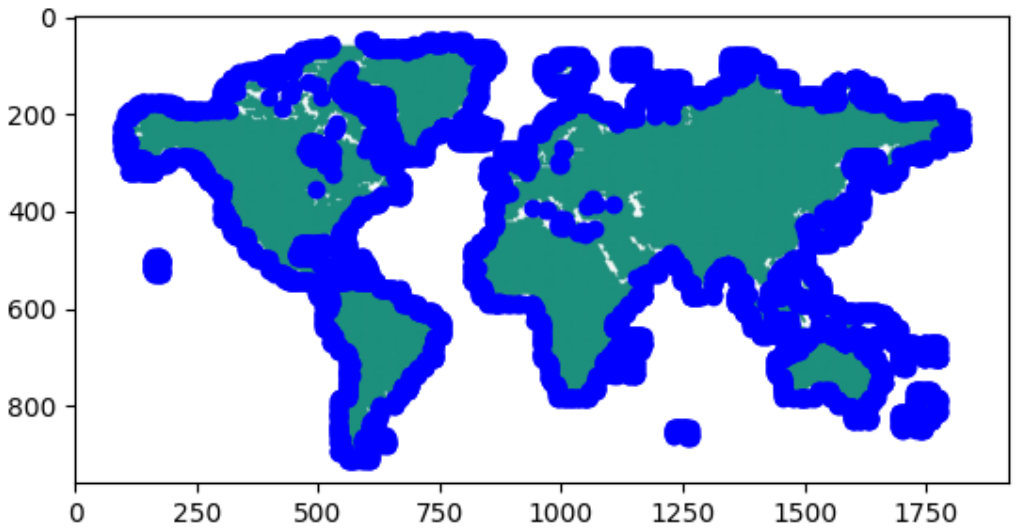

Figure 2: Geographic Map with Outline Detection (blue)

In addition to the variables given from the NCEI/WDS dataset, we used a geographic map with longitude and latitude axis to add the following variables to our earthquake dataframe: on sea and distance to land (shown in figure 2). The amount of pixels that a degree of longitude/latitude took up was created by dividing 360/180 by the width/height respectively. An earthquake's latitude and longitude coordinates were translated into an x/y coordinate on a map by multiplying the latitude/longitude by that scale. The map image had a binary color, one for sea and land. By taking the color value at the x/y coordinate, the *on_sea* variable was determined. To find the distance to land, the outline of each land mass was determined by creating a uniform distribution of random points and filtering by whether it had both sea and land within a degree of latitude/longitude. Thus, we can find the distance to land in degrees of longitude/latitude by returning 0 if the point is not on the sea, or the distance to the nearest boundary point. The distance in kilometers was determined using the Haversine Formula, which accounts for the spherical nature of the globe. The Haversine Formula was calculated as follows where r is the radius of the earth and (lat1, long1) and (lat2, long2) were the latitudes and longitudes of the endpoints:

$$r*\arccos\left(\sin(lat_1) * \sin\left(lat_2\right) + \cos\left(lat_1\right) * \cos\left(lat_2\right) * \cos\left(long_2 - long_1\right)\right)$$

Besides working with the tsunami and earthquake data sets independently, a new dataset was created with variables in both datasets. This had much less data points, but provided more data than working with them individually. The first attempt at merging these datasets attempted to match similar key features, such as location and magnitude. However, this likely introduced bias in our data set, since our dataset had known information (that the earthquake and tsunami was within a certain distance and magnitude away). The second attempt used the reference link in the tsunami data set. This linked the tsunami to its causal earthquake when applicable. By visiting the reference page provided by NCEI and using the *webscrape* library, data from both the tsunami and earthquake could be merged in a dataframe.

After manipulating the dataframes, the training, validation, and testing dataframes were created. We used a 63-30-7 split of training, testing, and validation data respectively. From reading the data, we split 30-70 between testing and training, and further split the training into 10-90 validation and training. Because 7% of our data consisted of only 47 data points, a cross validation approach could be later used.

4.3 Logistic Regression

We used a logistic regression model to use numerical data (and one discrete) to predict whether a tsunami would occur following an earthquake. This model was from *sklearn.linear_model.LogisticRegression*, which took in multi-dimensional inputs to output a singular probability. We had a probability threshold, and probabilities below that threshold output to a 0 (or no tsunami), with probabilities above the threshold outputting a 1 (yes tsunami). This threshold would be a hyperparameter that was tuned using according to its performance in the validation data set. This was done by iteration, with finding the accuracy, precision, and accuracy for each threshold percentage and finding the largest score, which was calculated by the average of its accuracy, precision, and recall. Using *earthquake magnitude, intensity, focal depth, distance to land,* and *on sea* as inputs, we predicted whether a tsunami was caused. We found that a threshold of 0.65 was optimal for our data set, with an accuracy, precision, and recall of (0.75, 0.57, 0.71) for the training and (0.76, 0.59, 0.69) for the testing.

4.4 Linear Regression

We used a linear regression model to use numerical data (and one discrete) to predict specific attributes of the resulting tsunami if it was predicted to occur. This model was from *sklearn.linear_model.LinearRegression*, which took in multi-dimensional inputs to output one or more numerical outputs. By fitting a linear line to our multi-dimensional inputs, it allowed us to predict new values. We used *earthquake magnitude, intensity, focal depth, distance to land,* and *on sea* as inputs to predict the magnitude and water height of the resulting tsunami. This was not particularly effective, and had a coefficient of determination of 0.22.

5. Results and Discussion

5.1 Logistic Regression

The logistic regression model had an accuracy of 0.75 for the training dataset, and 0.76 for the testing. 30.65% and 38.14% of the training and testing earthquakes caused a tsunami respectively, so this proved better than blindly guessing. However, there is lots of room for improvement.

We note the low relative precision, with 0.57 and 0.59 for the training and testing respectively. This can be from the imbalance in earthquakes that caused and didn't cause tsunamis, which can be improved by balancing the data through removing those that didn't cause tsunamis, or interpolating new earthquakes that did. The former would lead to even less data points, while the latter might lead to a decrease in accuracy.

Figure 3 shows the confusion matrix produced by our logistic regression model in the testing data, with the true label on the vertical axis and predicted on the horizontal axis. The top left and lower right show the true positives, and they encompass the majority of the predictions. However, 14.5% of the predictions were false positives, and 10.0% were false negatives. This means that 1/10 tsunamis remain undetected.

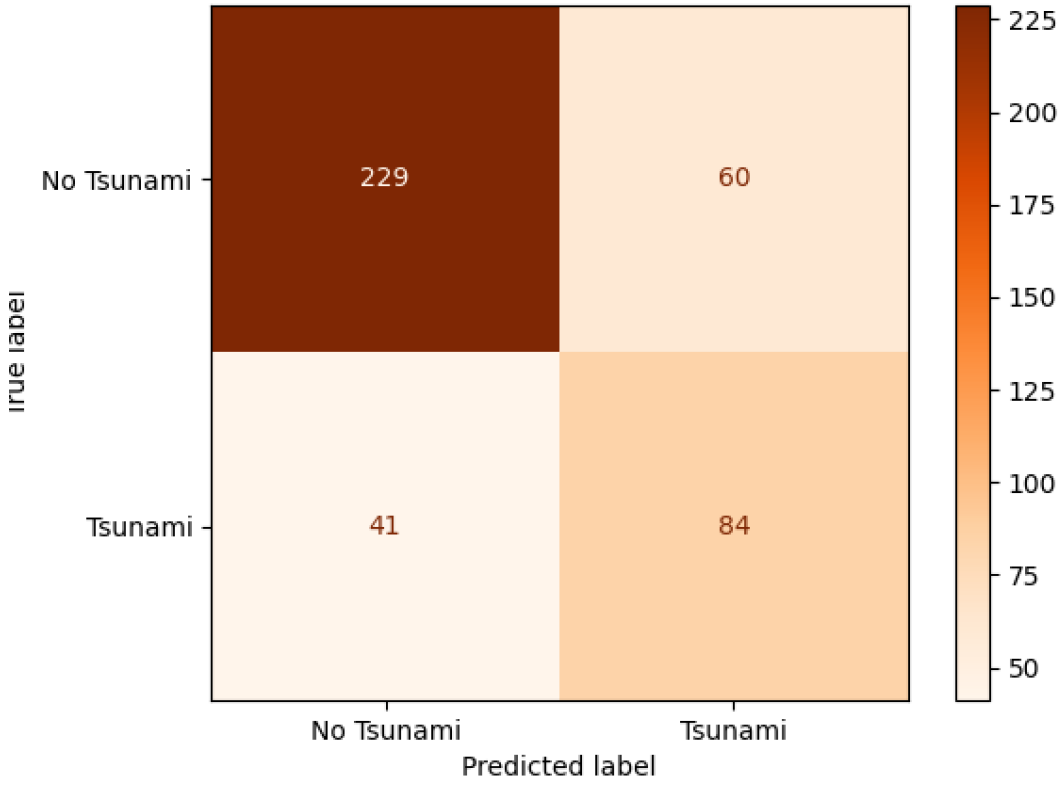

Figure: 3 Confusion Matrix for Logistic Regression Tsunami Predictions

5.2 Linear Regression

The linear regression model had a coefficient of determination of 0.22. This was a very weak linear correlation, which is expected. It is unlikely that these many variables are all linear, and we know that is not the case as magnitude is logarithmic. Using some transformation would likely increase the coefficient of determination and increase the reliability of a linear regression model to predict the tsunami maximum water height and magnitude. It is also important to note that because magnitude is on a logarithmic scale, being off by 1 actually leads to a 10x error in tsunami energy and effect.

In summary, this linear regression model is not viable for use yet, though transforming the data might lead to a more reliable model.

6. Conclusion and Future Work

With the large monetary costs and human casualties associated with tsunamis, a model to predict the occurrence of them has been developed using logistic regression. This paper explored the process of processing a dataset and applying a regression model to it. This model may be used in various countries that are frequented by Japan, or be built upon in future work. There are a variety of different ways to expand upon and improve these models within the preprocessing, logistic regression, and linear regression parts of our process.

6.1 Preprocessing

One issue with working with pandas dataframe is the requirement of working with pandas data types. Specifically, pandas.Timestamp.min = Timestamp('1677-09-21 00:12:43.1452'), which means that no entries could use a date before the 17th century. This means a large portion of our data points could not be used. This can be solved by simply not using the date as an input for the regression models, or alternatively discretizing the date (such as using centuries). However, climate change has greatly influenced the conditions in which tsunamis and earthquakes are formed (Zhu and Fan 2021). By only keeping our data points within a few centuries rather than two millennia, we predict our model may perform better. This can and should be tested, and this is left for future work.

In addition, additional variables categorizing the earthquake's relation to fault lines could be created by instead using a fault line map. This can include absolute distance or angle.

6.2 Logistic Regression

In evaluating and hypertuning the performance of a logistic regression model, we want false negatives to be punished more heavily than false positives as they have more detrimental consequences, although an excess of false positives can lead to a distrust in the system. In future models, taking into account the difference in consequences between different regions of the confusion matrix may lead to a more practical model. This can be done by changing the loss function when deciding on a threshold.

6.3 Linear Regression

The linear regression model proved to be unreliable with a coefficient of determination of 0.22. Our methodology did not include scaling or linearizing the data, which may be employed to create a more linear dataset and increase its efficacy. In addition to this, principal component analysis (PCA) can be used.

6.4 Final takeaways

Logistic regression has been used in past studies to predict landslides in Taiwan, and we have used it to predict earthquake-caused tsunamis. This will allow citizens in areas frequented by natural disasters to be safer. This model also paves the path for future work to be done.

7. References

Britannica, *Ring of Fire*, https://www.britannica.com/place/Ring-of-Fire
Chang et al., *Modeling typhoon- and earthquake-induced landslides in a mountainous watershed using logistic regression*, https://doi.org/10.1016/j.geomorph.2006.12.011
National Geographic, *Plate Tectonics*, https://www.nationalgeographic.com/science/article/plate-tectonics
NCEI, *NGDC/WDS Global Historical Tsunami Database*, https://www.ngdc.noaa.gov/hazard/tsu_db.shtml, doi:10.7289/V5PN93H7
NOAA, *Deep-ocean Assessment and Reporting of Tsunamis (DART)*, https://nctr.pmel.noaa.gov/Dart/dart_home.html
Oskin, Feb 2022, *Japan earthquake & tsunami of 2011: Facts and information*, https://www.livescience.com/39110-japan-2011-earthquake-tsunami-facts.html
Zhu and Fan, Jan 16, *Exploring the Relationship between Rising Temperatures and the Number of Climate-Related Natural Disasters in China,* https://doi.org/10.3390%2Fijerph18020745

8. Acknowledgements

I thank the InspiritAI high school program and their mentors for their mentorship and providing me the opportunity to research this topic.