Machine Learning Final Report:
Earthquake-Triggered Tsunami Prediction

Hsiang Wang

Abstract

Tsunamis are rare but highly destructive natural disasters that can cause severe loss of life and extensive damage to coastal infrastructure. Because tsunami waves can reach shorelines within minutes after an earthquake, early warning systems must operate under extremely tight time constraints. As a result, there is a strong need for fast screening methods that can assess tsunami risk immediately after an earthquake occurs.

In this project, we study a binary classification problem that aims to predict tsunami occurrence using basic seismic features available after an earthquake. We use a global earthquake dataset covering events from 2001 to 2022 and evaluate several classical machine learning models, including Logistic Regression, Support Vector Machines (SVM), Random Forests, and Gradient Boosting. Since false negatives (predicted no-tsunami but actual tsunami) are more costly than false positives, we view recall as a priority over precision and accuracy.

A key challenge we identify is a strong temporal label distribution shift caused by missing tsunami records in early years. By adopting data truncation and randomly stratified splitting, we obtain more stable and reliable results.

Among the evaluated models, Random Forest achieves the best generalization overall, followed by Gradient Boosting. Our results highlight the importance of exploratory data analysis, data preprocessing and metric selection when applying machine learning methods to disaster prediction tasks.

1. Dataset Description

We used the **Global Earthquake–Tsunami Risk Assessment Dataset**, which is available on Kaggle. The dataset contains earthquake events recorded worldwide between **2001 and 2022**, with a total of 782 samples. Each sample corresponds to one earthquake event.

1.1. Target Variable

- 1: a tsunami was triggered by the earthquake
- 0: no tsunami occurred

1.2. Input Features

The original dataset contains 12 features, all of which are numerical. They can be grouped into four categories, as stated below.

1. Magnitude and intensity-related features
   - **magnitude**: 6.5-9.1, Earthquake magnitude (Richter scale)
   - **mmi**: 1-9, Modified Mercalli Intensity (instrumental)
   - **cdi**: 0-9, Community Decimal Intensity (felt intensity) human experience of shaking

- **sig**: 650 - 2910, Event significance score, a calculated score by the USGS that combines magnitude, felt reports, and estimated impact

2. Geometry and distance-related features:
   - **depth**: Earthquake focal depth (km) Shallow earthquakes (less than 70km) are more likely to cause tsunami because they can displace the seafloor more directly.
   - **latitude**: Epicenter latitude (WGS84)
   - **longitude**: Epicenter longitude (WGS84), an epicenter location close to the coastline or under the ocean is more likely to trigger tsunami
3. Data Quality and Reliability features:
   - **nst**: Number of seismic monitoring stations, accuracy of data
   - **dmin**: 0.0 - 17.7, Distance to nearest seismic station (degrees), smaller values means calculated depth is more accurate
   - **gap**: 0.0 - 239.0, Azimuthal gap between stations (degrees), Location reliability, coverage of stations around epicenter
4. Temporal features: **Year, Month**, which is used to find seasonal patterns in data.

1.3. Excluding Data Before Year 2013

Initially, the model was trained and tested on the whole dataset. However, the model was observed to rely almost exclusively data quality and reliability features —specifically, sensor density (`nst`) and distance to the nearest seismic station (`dmin`).

Initial analysis of the data revealed that there were no tsunami records prior to 2013, despite historical occurrences (Figure 1); this suggested a significant reporting bias. Figure 2 and 3 showed a structural break in the reported metadata around 2013. This indicated that the tsunami presence was strongly correlated with recording features. Consequently, the model was predicting the target based on historical data collection shifts rather than earthquake-related features.

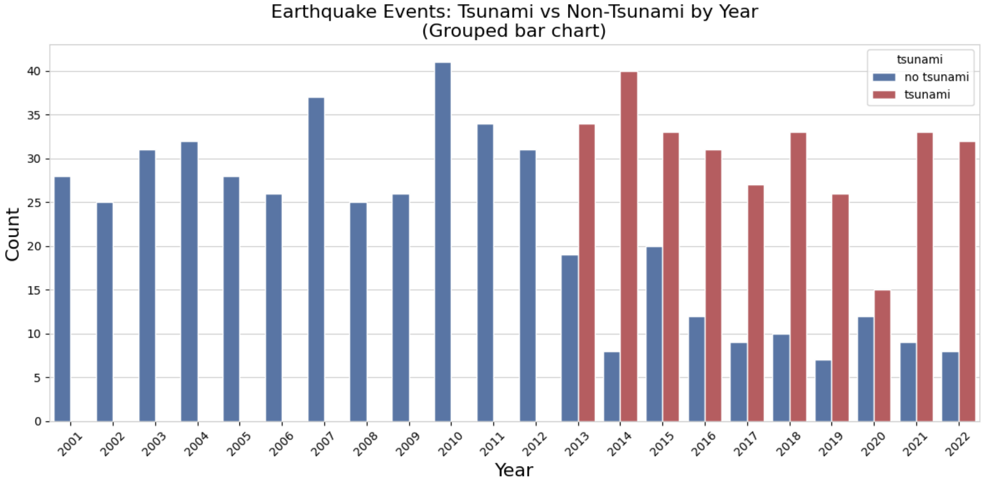

Figure 1. Earthquake Events: Tsunami vs Non-Tsunami by Year

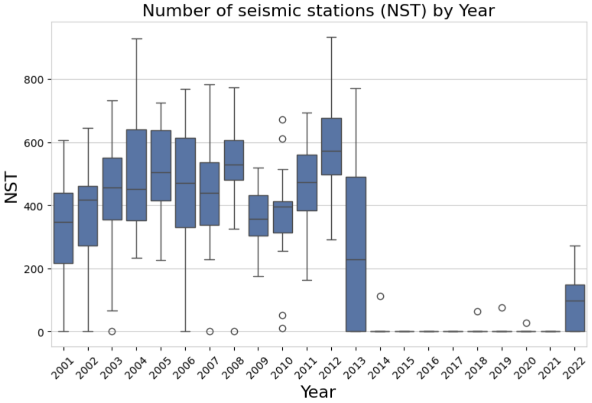

Figure 2. NST by Year

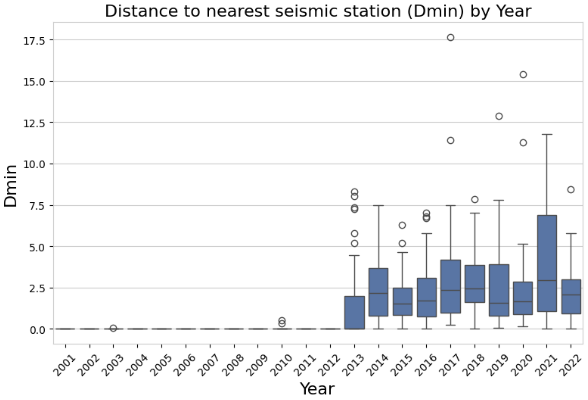

Figure 3. dmin by Year

To prevent the model from learning historical bias, **the data prior to the year 2013 was truncated, leaving with 418 earthquake records from 2013 to 2022.**

2. Exploratory Data Analysis (EDA)

2.1. Missing Values

We inspected the dataset using summary statistics and confirmed that no missing values are present in any column. Therefore, no imputation or row removal was required, and all models are trained on the full set of 418 events.

2.2. Target Distribution

Out of the total 418 earthquake samples, 304 earthquakes triggered a tsunami, accounting for 72.73 percent. Since the dataset is imbalanced, stratification is required during training to ensure constant tsunami proportions across splits.

2.3. Features Correlation

Figure 4 illustrates the correlation matrix among the input features and the target tsunami variable. It is noticeable that the magnitude and intensity related features are positively associated with each other, with ”magnitude” and ”significance” exhibiting the strongest positive correlation of all.

We can also see that after the truncation of data before year 2013, no features show a relatively high correlation with year, reducing the risk of temporal bias.

2.4. T-test of Tsunami and No-tsunami on Features

The data were partitioned into those with labels tsunami and no-tsunami, and two tailed independent sample T-tests were performed separately on features except ”Year” and ”Month”. Since the two groups are assumed to be independent and sampled from a normally distributed population of earthquake data, the ”ttest_ind” function from scipy stats is suitable for the context.

With the significance level set to 0.05, the features that pass the T-test and have a significant difference in mean for the two groups were **”magnitude”, ”depth”, and ”latitude”**.

2.5. Magnitude and Intensity-related Features

In Figure 5, the box plot for magnitude in groups tsunami and non-tsunami depicts the relationship of higher chance of tsunami occurence and larger magnitude. In addition, earthquakes with magnitude 8 or higher account for only a small proportion, suggesting that they are less likely to occur. The distributions of CDI, MMI, and event significance score does not present a notable difference between the two groups.

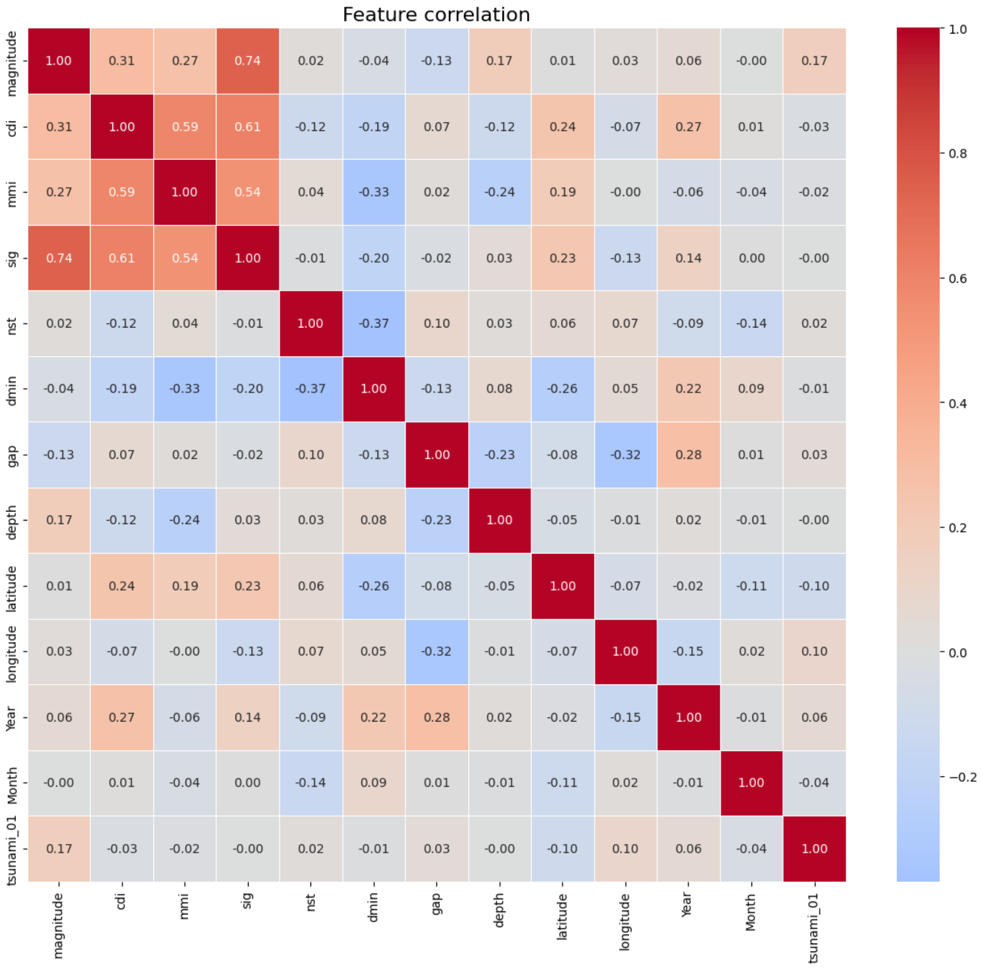

Figure 4. Feature correlation matrix

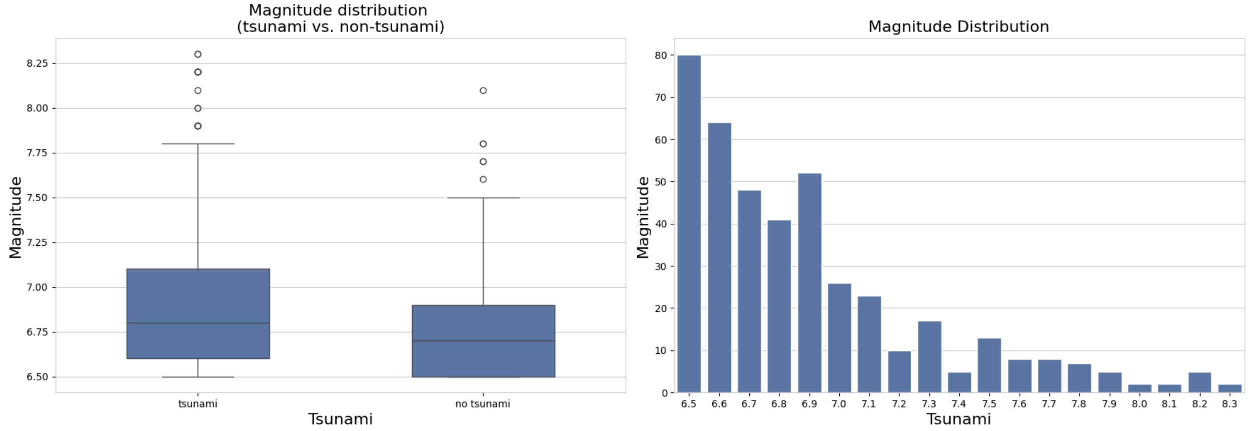

Figure 5. Magnitude

2.6. Location and Geometry-related Features

We divided the earthquakes into three depth categories according to USGS. In both tsunami and no-tsunami groups, shallow earthquakes of 0-70 km dominate the samples, reflecting a skewed depth distribution. Furthermore, the tsunami-triggered earthquakes have higher median of magnitudes in all three depth categories, providing evidence of the positive correlation of magnitude and tsunami while controlling the depth.

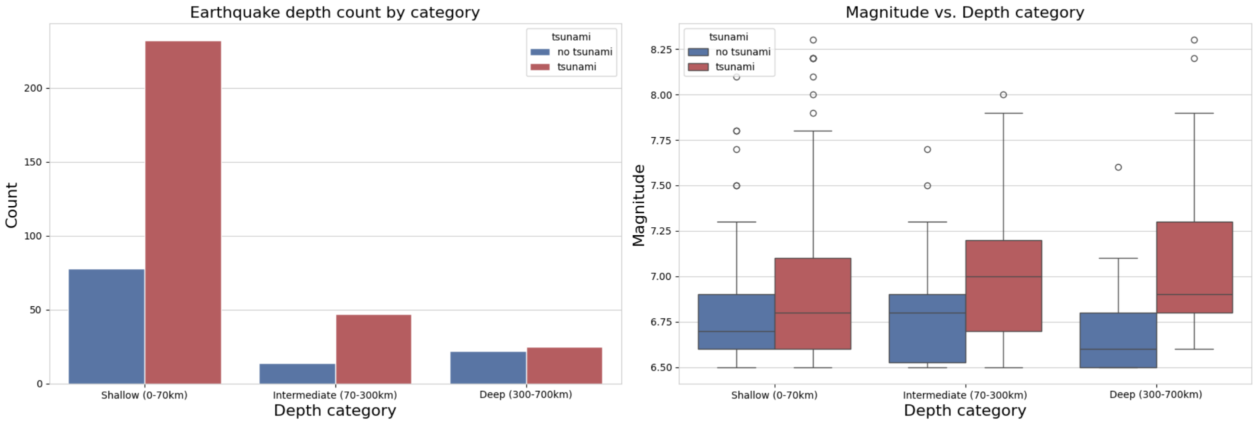

Figure 6. Depth

From Figure 7, we can see that the majority of earthquakes associated with a tsunami occurred near coastal regions, especially along the pacific ring of fire.

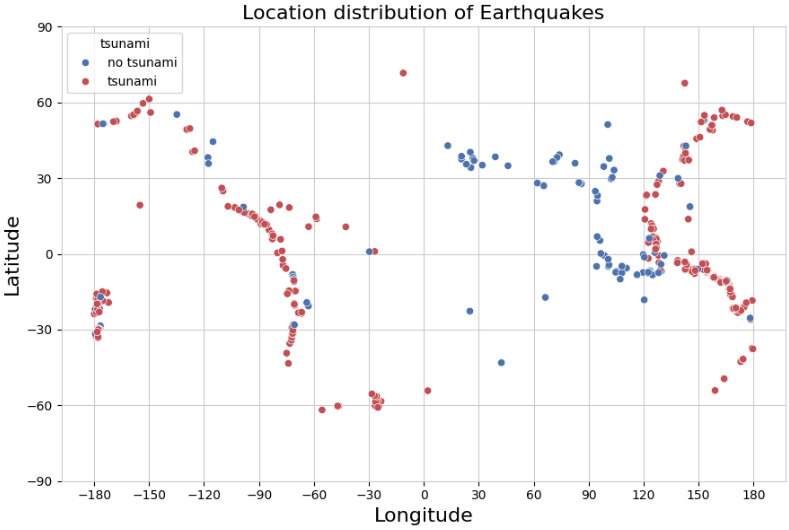

Figure 7. Location distribution of Earthquakes

2.7. Reliablity-related Features

From Figure 8 and Figure 9, we can see that the distribution of ”NST” and ”gap” for both tsunami and non-tsunami groups were highly left skewed. Similarly, the spread of ”dmin” concentrates at the lower values, though it is not as skewed as the other two reliability-related features.

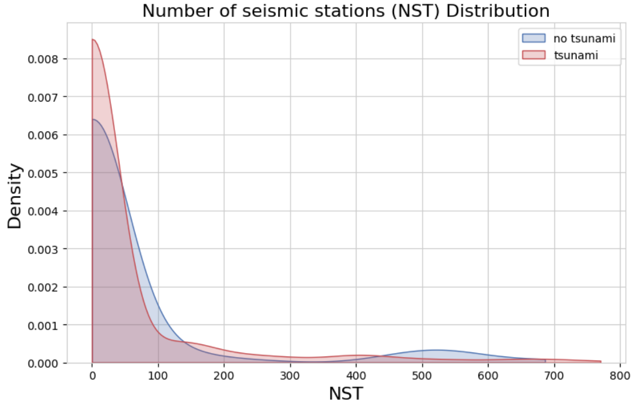

Figure 8. Number of seismic stations

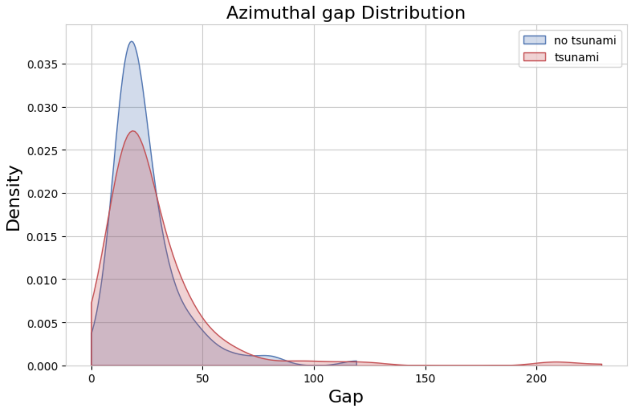

Figure 9. Azimuthal gap

2.8. Temporal Patterns

According to the line graph illustrated in Figure 10, the earthquakes in tsunami and non-tsunami groups both followed similar trends across different months. There are no noticeable seasonal patterns in the happening of tsunami-triggered earthquakes, thus the feature can be eliminated for model training.

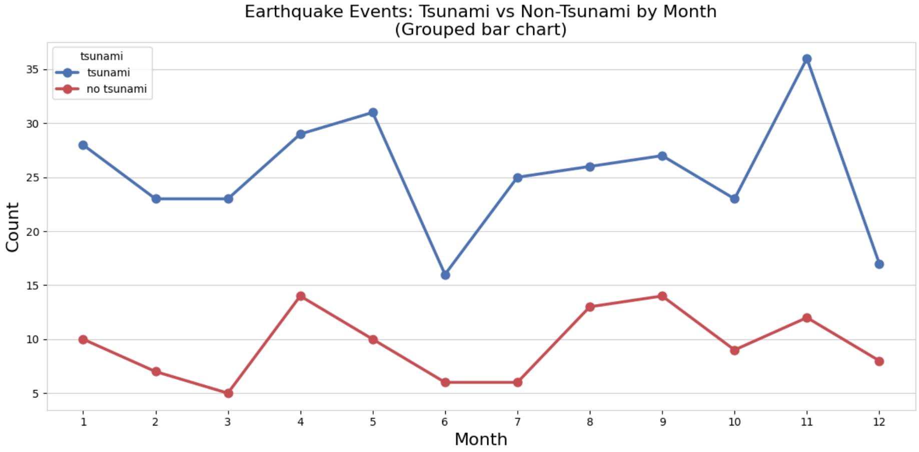

Figure 10. Tsunami vs. Non-tsunami by Month

3. Data Preprocessing and Splitting

3.1. Feature Engineering and Feature Selection

According to the results obtained from EDA, we included a new feature **”mag_depth_ratio”** that divides magnitude by depth. In addition, log transform was applied to the highly skewed features, including **”depth”, ”nst”, and ”gap”**, while their original columns were dropped.

Finally, other unwanted columns ”depth_category”, ”tsunami”, ”Year”, ”Month” were dropped for the final input features. The temporal features were excluded because they may introduce temporal bias which is irrelevant for tsunami prediction.

After feature engineering and feature selection, there are **11 final input features (magnitude, mmi, cdi, sig, depth_log, latitude, longitude, nst_log, dmin, gap_log, mag_depth_ratio)** that are used for training and testing the models.

3.2. Feature Scaling

Since all 10 input features are numeric and have different scales, we applied Z-score standardization using StandardScaler. The scaler is fitted only on the training data and then applied to the validation and test sets to avoid data leakage. This preprocessing step is especially important for distance-based and margin-based models such as Logistic Regression and SVM.

3.3. Randomly Stratified Split

A randomly stratified 80/20 training and testing set split was applied, with 334 training samples and 84 testing samples. We did not include a validation set because cross-validation was already adopted when selecting the best hyperparameters in all models. Stratification ensures consistent tsunami proportions across all splits, enabling fair model comparison and more stable evaluation.

4. Model Implementation

We evaluated four classical machine learning models: Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting. For all models, **a grid search with 5-fold stratified cross-validation** was applied to find the best combination of the specified parameters. We chose **F1-score** as the scoring metric, since it provides a balanced measure of precision and recall, which is more indicative than accuracy for this problem. Furthermore, we used the ”class weight=’balanced’” parameter in Scikit-Learn available for Logistic Regression, SVM, and Random Forest to address the class imbalance problem by assigning larger weights to the minority class.

4.1. Logistic Regression

Logistic Regression serves as a linear baseline. A 5-fold cross-validation was applied on different values of the parameter C with L1 and L2 penalty, respectively. The parameter C in logistic regression is the inverse of the regularization strength; therefore, smaller values of C correspond to stronger regularization.

As depicted in Figure 11, smaller C values does not improve model performance, suggesting that regularization has limited effect to the model. Regarding to penalty type, L2 performs better than L1, which is likely due to the more evenly shrinking of model coefficients, resulting in a more stable model. The best hyperparameter combination identified by the grid search was **L2 penalty with C=1** with the mean F1-score of 0.7046.

On the test set, though most non-tsunami events were identified correctly, the model produced a considerable number of false negatives, which was unacceptable for the problem. This is likely due to the constraint of logistic regression to linear decision boundaries, causing it to struggle with more complex data patterns.

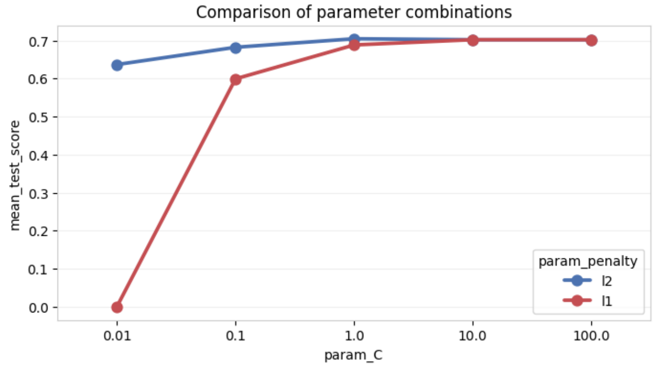

Figure 11. Logistic Regression: Cross-Validation F1-score for different parameters

4.2. Support Vector Machine (SVM)

We employed Support Vector Machine (SVM) classifier using the default **radial basis function (RBF) kernel**, which is a widely used kernel when little prior knowledge about the data distribution is available. The RBF kernel maps the input data into a higher-dimensional feature space, enabling the model to learn complex, non-linear decision boundaries.

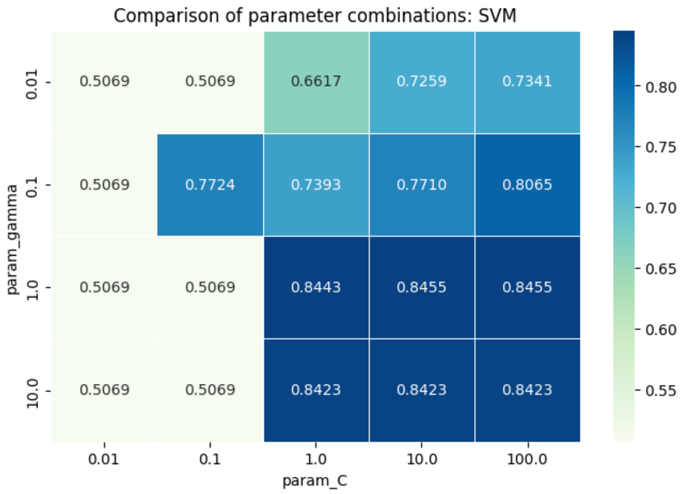

Figure 12. SVM: Cross-Validation F1-Score for different parameters

The key parameters of the SVM are γ **and C**. The parameter γ controls the radius of influence of individual training samples; smaller values lead to smoother and simpler decision boundaries. C is the regularization parameter that balances the margin width and classification errors; lower values of C correspond to a wider margin with stronger regularization. The default L2 penalty was used for this experiment.

Similar as results from logistic regression, increasing regularization (lowering the C parameter) does not improve model performance. This might be because the training sample size is not large enough to cause overfitting issues, therefore regularization is not necessary. On the other hand, the gamma value 1.0 performs the best while values lower than 1.0 worsens the performance.

The optimal configuration was **C=10 and** γ **=1** with the mean F1-score of 0.8455. Under this setting, the model achieved a test accuracy of 0.7738 and test recall of 0.9508. Figure 15 shows that the confusion matrix of the tuned SVM has significantly fewer false positives than logistic regression, albeit at the cost of more false negatives.

4.3. Random Forest

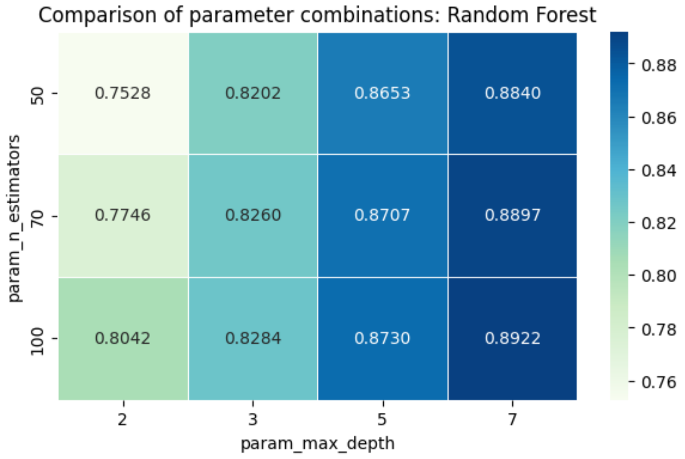

Figure 13. Random Forest: Cross-Validation F1-Score for different parameters

Random Forest is an ensemble learning technique that integrates several decision trees trained independently with a random selection of data points and features. Compared to a single decision tree, random forests are more robust to overfitting as each individual tree captures different perspectives of the data. However, it can also result in overly complex models for relatively small datasets.

We investigated how the number of estimators (trees) and the maximum depth of each tree can affect performance. Figure 13 illustrates that increasing the maximum depth has a greater impact on the F1-score, whereas the number of estimators only improve the model performance slightly. The optimal settings found for the random forest model were **100 estimators with a maximum tree depth of 7**. On the test set, this model achieved an accuracy of 0.8929 and recall of 0.9508.

4.4. Gradient Boosting

Gradient Boosting is a tree-based ensemble technique that combines models sequentially. Each newly added tree corrects the mistakes made by the previous ones, allowing the model to learn complex patterns. Nevertheless, careful hyperparameter tuning and regularization are required to achieve optimal performance.

From the result of cross validation in Figure 14, increasing the maximum tree depth and the number of estimators does not improve the cross-validation recall score, as it can lead to overfitting by making the model overly complex. In the optimal configuration, the number of boosting stages (n_estimators) was reduced to 50, and the maximum depth was reduced to 2. Given the relatively small size of the dataset, these constraints help control model complexity and prevent overfitting.

As illustrated in Table 1, Gradient Boosting attains the highest ROC-AUC score of 0.8895 among all the models evaluated. The ROC curve for gradient boosting is closest to the top-left compared to the other three models. A further implementation is to adjust the decision threshold value with respect to the ROC curve.

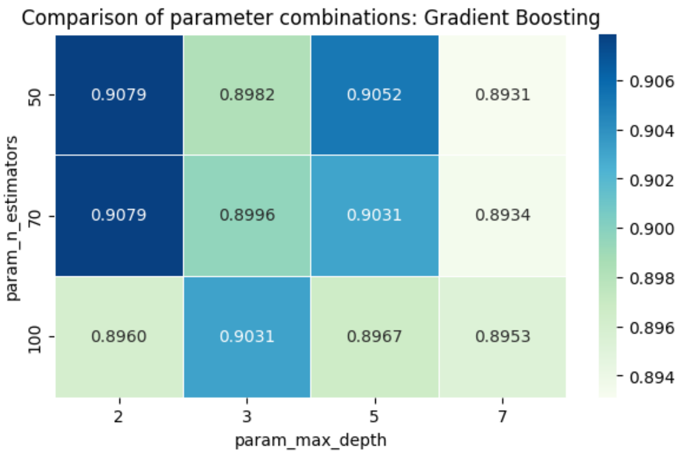

Figure 14. Gradient Boosting: Cross-Validation F1-Score for different parameters

5. Results and Discussion

In the context of earthquake-triggered tsunami prediction, false negatives (predicted no-tsunami but an actual tsunami occurred) are more costly than false positives in tsunami prediction. Thus, **we prioritize recall and F1-score over precision and accuracy** during model evaluation. Table 1 and Figure 15 illustrates the performance of models on the test set.

	Logistic Regression	SVM	Random Forest	Gradient Boosting
Accuracy	0.714286	0.773810	**0.892857**	0.857143
Precision	**0.911111**	0.783784	0.906250	0.865672
Recall	0.672131	**0.950820**	**0.950820**	**0.950820**
F1-score	0.773585	0.859259	**0.928000**	0.906250
ROC-AUC	0.752673	0.714897	0.873842	**0.889522**

Table 1. Evaluation metrics results of all models on testing set

Across all models, we observed that:

- **Random Forest** generalizes best overall, achieving the highest evaluation scores except for precision and ROC-AUC, in which is slightly lower than Logistic Regression and Gradient Boosting, respectively. This is likely due to the sequential nature of Random Forest, allowing it to capture non-linear and complex relationships, producing accurate results that outperform the other three models.
- **Gradient Boosting** has the highest AUC score of 0.8895, providing the best discrimination between tsunami and non-tsunami events. The ROC-Curve was shown in Figure 16. It is the second most effective model after Random Forest in terms of accuracy, precision and F1-score.
- **SVM** excelled at identifying actual tsunami events, achieving the highest recall of 0.95, along with both tree-based models. On the contrary, it has the lowest precision score and an even lower ROC-AUC score than logistic regression.
- **Logistic Regression** achieved the highest precision of 0.91, but produced considerably more false negatives that is intolerable for the problem.

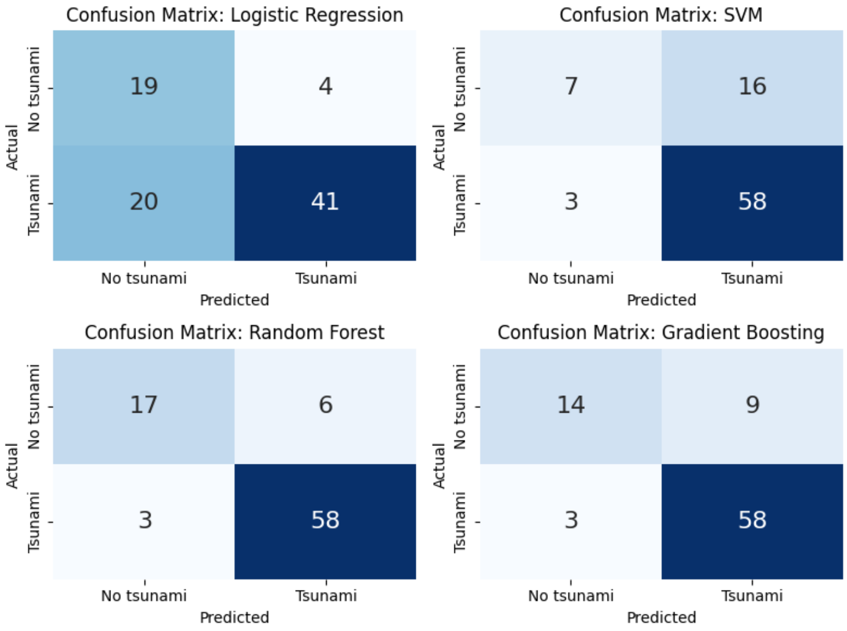

Figure 15. Confusion matrices of all models on testing set

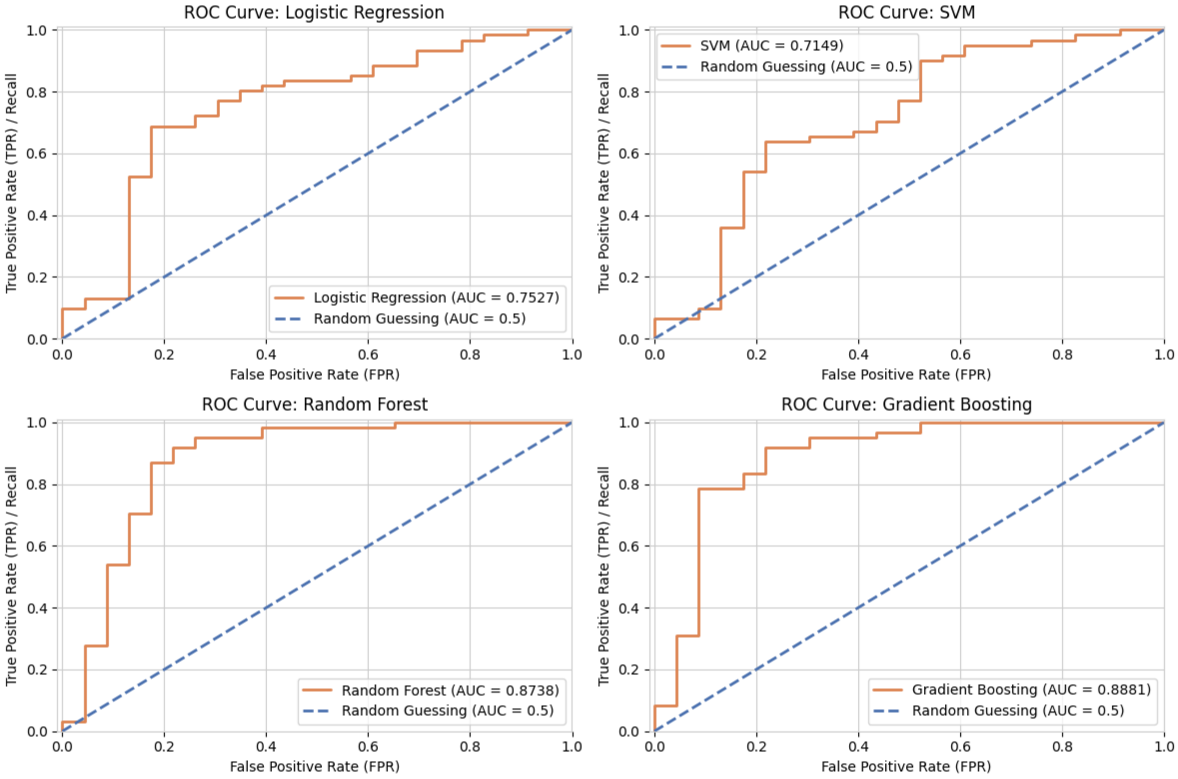

Figure 16. ROC Curve of all models

5.1. Feature Importance

To identify crucial indicators of tsunami occurrence, the importance of features was analyzed.

For tree-based models, the scikit-learn library provides Mean Decrease in Impurity (MDI). This metric calculates importance by combining the fraction of samples a feature contributes to the final prediction with its corresponding decrease in impurity.

However, this approach has two main limitations. First, it is biased toward features with high cardinality (e.g., numerical features or categorical features with many levels). Second, it is calculated solely on the training data, which can provide misleading results if the model is overfitting.

In contrast, Permutation Feature Importance measures the decrease in a model's evaluation score when a specific feature's values are randomly shuffled. This method is unbiased and can be applied to both training and testing sets using any evaluation metric or model type.

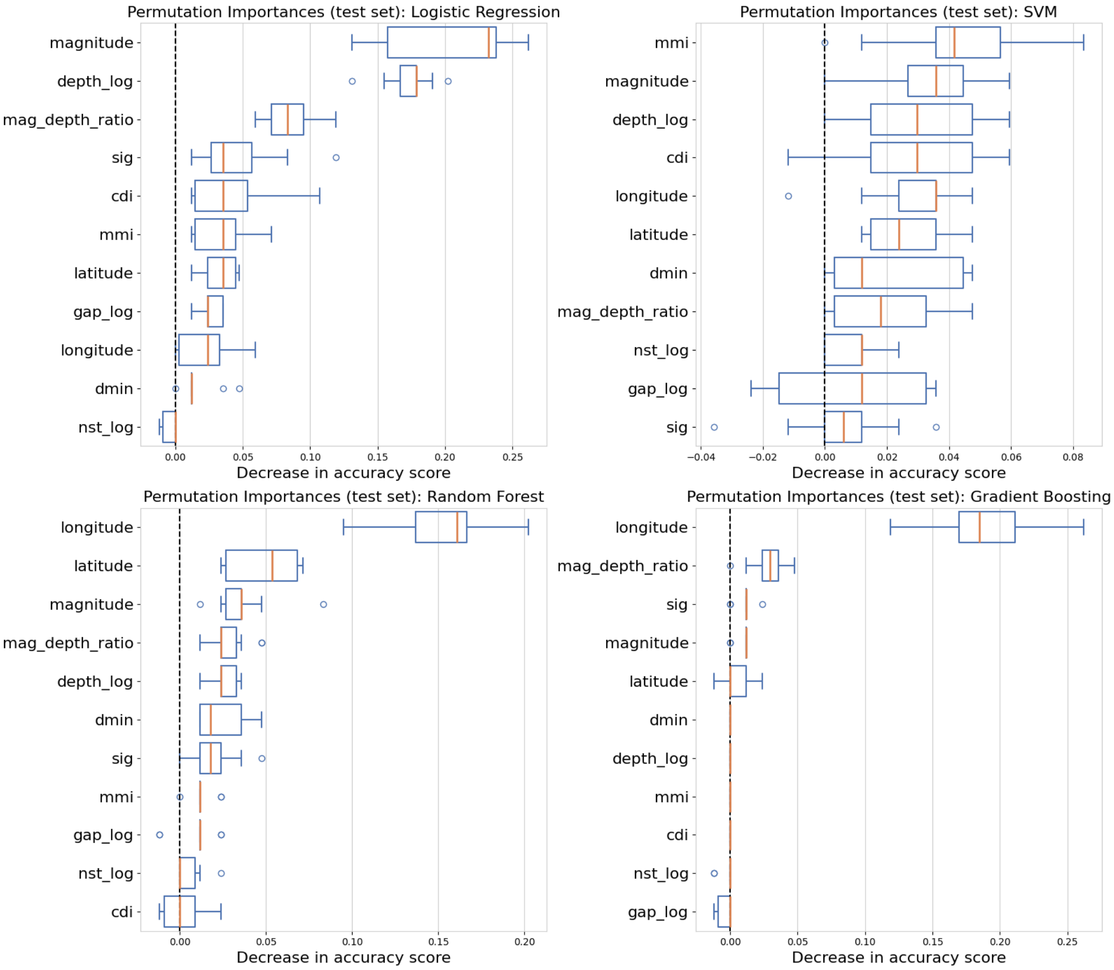

Figure 17. Permutation Importances on Test Set

According to Figure 17, we discovered that longitude serves as the most critical feature for the high performing tree based models, which has the ability to capture non-linear patterns. The new engineered feature ”mag_depth_ratio” also greatly influenced the predictive ability of these models. Magnitude related features were fairly important among all four models, particularly Logistic Regression and SVM, where it topped the feature importances charts.

Different from other models where the predictions are dependable on a few critical features, SVM has relatively equal contributions from all features.

Finally, the models is no longer learning the reliability-related features, after the data prior to 2013 is truncated. The resulting performance metrics are more representative of the model's ability to generalize to future seismic events recorded under current global monitoring standards

6. Conclusion and Future Work

In this project, we studied the tsunami prediction binary classification problem with classic machine learning models. We demonstrated that initial data visualization and processing plays a crucial role. Removing the data that induced historical bias greatly improved the model performance, and also lead to more reliable results.

In addition, we investigated how various parameters in machine learning models can affect the overall performance. Among the four models implemented, random forest generalized the best, while gradient boosting can also be a reasonable choice with higher AUC score. This project highlights the importance of using multiple evaluation metrics and prioritizing recall in disaster prediction tasks.

For future work, we plan to:

- Experiment with more sophisticated models such as XGBoost or neural networks.
- Perform feature engineering of the location features, such as identifying the distances from the coastline.
- Adjust the decision threshold according to ROC-AUC curves.
- Collect more non-tsunami events to reduce the imbalance of dataset.

7. References

- https://www.kaggle.com/datasets/ahmeduzaki/global-earthquake-tsunami-risk-assessment-dataset/data
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
- https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html
- https://www.usgs.gov/programs/earthquake-hazards/determining-depth-earthquake