1.Project Introduction

Our project is centered on training machine learning models to predict the magnitudes of the very next earthquake aftershock within a predefined time and spatial boundary following a main seismic event. Aftershocks can have profound implications for urban centers and post-earthquake recovery operations, often inflicting damage can exceed the initial main shock. The ability to accurately forecast the magnitude of the immediate subsequent aftershock is because of the paramount importance for emergency preparedness and response.

By employing machine learning techniques, we selected a range of features to predict the magnitude of what we define as the next aftershock occurring within a custom time frame after a mainshock. This focused approach allows a more targeted prediction, which is critical for informing timely and effective disaster response measures.

2.Dataset Introduction

2.1 Overview

This dataset encompasses a 30-year record of seismic events from 1993 to 2023 in the South Japan Sea and North Philippine Sea Region, provided by the United States Geological Survey (USGS). The USGS is a leading agency in delivering reliable geophysical data, and this dataset meticulously documents each earthquake in this seismically active area. It includes essential details such as the date and time of occurrence, geographical coordinates, focal depths, and magnitudes, which are crucial for understanding the region's seismic activity.

Each record in the dataset not only quantifies the earthquake's magnitude but also details the type of magnitude measured, the number of seismic stations reporting, the angular gap between stations, and the distance to the nearest station from the epicenter. Additionally, the dataset includes the root mean square (RMS) of the seismic signal, offering insights into the spread of vibrations. Unique identifiers and update timestamps ensure the traceability and currency of the information.

Error metrics such as horizontal error, depth error, and magnitude error provide vital information on the precision of the recorded data, essential for the integrity of any analysis. The review status of the data along with the sources of location and magnitude information, further enhances the dataset's credibility.

2.2 Data Analysis

To better understand and visualize the dataset, we are going to conduct some exploratory data analysis. The earthquake data has been sorted from in a descending order from 1983 to 2023. We can view the magnitude of the past 500 earthquakes using a time series plot.

A screen shot of a graph

Description automatically generated

Figure 1.Plot of recent 500 Earthquakes

The earthquake data in total has about 60000 rows. It may be more suitable to look at the heatmap of all the earthquakes in the South Japan Sea and North Philippine Sea region, based on their latitude, longitude and magnitude. The earthquakes are concentrated along a narrow band of latitude, roughly between 120-160 degrees longitude. This suggests the earthquakes are occurring along a defined fault line or plate boundary in this region. The pattern shows an arc-like shape to the earthquake occurrences. This could align with a curved subduction zone or other known geological feature in this part of the world.

A map of earthquakes from years of earthquake

Description automatically generated with medium confidence

Figure 2.Heatmap of Earthquakes from 1983-2023

3.Feature Engineering

In our earthquake aftershock prediction project, we began with 5 features from the original dataset, including earthquakes, including time, latitude, longitude, depth, and magnitude. Through a series of feature engineering steps, we aimed to extract and construct features that would be helpful in predicting the magnitude of aftershocks. The following are the detailed steps we took in our feature engineering process:

(1) First, we calculated the difference in latitude (dif\_lat) and longitude (dif\_long) between each earthquake and the one that preceded it, taking the absolute values of these differences as new features. We then used these differences to calculate the straight-line distance (distance) between two earthquakes, assuming that each degree on the Earth's surface is approximately equivalent to 111 kilometers, using the following formula:

distance =

(2) If the distance between two earthquakes was under 40 kilometers, we classified the latter as an aftershock (cat\_afs = 1) of the former, otherwise cat\_afs = 0.

We set an initial sequence value (seq\_afs) of 0 for all earthquakes and updated this in reverse order for aftershocks based on their classification (cat\_afs).

(3) For each earthquake classified as an aftershock, we identified the corresponding main shock using the aftershock sequence (seq\_afs) and recorded its magnitude (mag\_main) and depth (depth\_main).

(4) We calculated the magnitude (mag\_pre) and depth (depth\_pre) for the preceding earthquake of each event, setting these to 0 for main shocks.

(5) The target for prediction was set as the magnitude of the next aftershock (target), shifting each earthquake's magnitude forward by one position. Main shocks were assigned a target value of 0.

(6) We removed the first and last rows from the dataset to avoid null values, as they lacked a “previous” or “next” earthquake.

(7) Latitude and longitude were converted into labels (longitude\_label and latitude\_label), with a label assigned for every 3 degrees, ranging from 1 to 20, to avoid treating these as continuous features.

4.Performance Measure

4.1 Mean Squared Error

Mean Squared Error (MSE) is a commonly used metric for evaluating the performance of regression models. It measures the average squared difference between the predicted values and the actual values.

4.2 Mean Absolute Error

Mean Absolute Error (MAE) is another widely used metric for evaluating the performance of regression models. It measures the average absolute difference between the predicted values and the actual values.

The distribution of earthquake magnitudes in the dataset shows a long tail, indicating the presence of some high-magnitude earthquakes(outliers). The magnitudes are highly concentrated within 3.3-6.2 Mw, there are few earthquakes with magnitude higher than 7.0.

MSE is more sensitive to outliers because it squares the errors, giving more weight to larger deviations. MAE is less sensitive to outliers because it uses the absolute values of the errors, treating all deviations proportionally.

Considering the characteristics of the dataset and the importance of accurately predicting high-magnitude aftershocks, MSE might be a more suitable performance measure for this specific task. Its sensitivity to outliers can help identify and penalize models that make larger errors on high-magnitude earthquakes, which is crucial for earthquake preparedness and risk assessment.

5.Model Analysis

Machine learning, one of the fastest-growing branches of artificial intelligence, utilizes computational methods to improve the performance of systems by leveraging data. Various machine learning algorithms have been widely applied in the analysis and prediction of consumer purchasing behavior, demonstrating impressive predictive capabilities.

This section will introduce the algorithms employed, primarily including commonly used supervised learning methods such as Linear Regression, Random Forest algorithm, Gradient Boosting and MLP(Multiple Layer Perceptron) Neural Network algorithm. We will apply these models to historical earthquake data of Japan, employing techniques such as cross-validation to execute the validation and grid search to select the optimal model parameters and compare the predictive performance of each model to find the best parameter for the model.

5.1 Linear Regression

Modeling

**Data Splitting:** The dataset was divided into training and test sets using a standard 80/20 split, ensuring the model's evaluation was performed on unseen data.

**Model Initialization**: A Ridge regression model, a form of linear regression that includes regularization to prevent overfitting was chosen for this task. The model was initialized with an alpha parameter of 1.0 which controls the strength of the regularization.

**Cross-Validation:** The model was subjected to 5-fold cross-validation on the training data. This technique helps in assessing the model's effectiveness and robustness, as it splits the training data into 5 subsets, trains the model on 4 of these subsets, and evaluates it on the 5th subset, repeating this process five times.

**Performance Metrics Calculation**: For each cross-validation fold, the Mean Squared Error (MSE) was calculated and then converted to positive values since cross-validation scores are negated for maximization purposes. The average MSE across all folds was computed to provide an overall performance estimate on the training data.

**Model Fitting:** The model was fitted to the entire training dataset, allowing it to learn the relationship between the independent variables and the target variable.

**Prediction and Normalization:** Predictions were made on the test set and then rounded to one decimal place to maintain consistency with the format of the target data.

Result

**Cross-Validation Results:** The average MSE from the 5-fold cross-validation was approximately 0.436 suggesting that the model on average had a squared error of that magnitude when making predictions on the validation folds.

**Test Set Evaluation:** On the test data, the model achieved a MSE of approximately 0.483, which is slightly higher than the cross-validation MSE indicating some differences in model performance between the training phase and the test phase.

**Additional Metrics:** The Mean Absolute Error (MAE) on the test set was approximately 0.518, and the R² score was 0.219. The R² score, which measures the proportion of variance explained by the model, suggests that the model has a relatively low predictive power in this context.

**Residual Plot Analysis:** The residual plot shows the residuals on the vertical axis and the predicted values on the horizontal axis. The distribution of the residuals does not show a clear pattern, which is good as it indicates that the model's errors are randomly distributed. However, there is some indication of heteroscedasticity, as the variance of residuals appears to be increasing with the predicted values, which might suggest that the model fits differently across the range of predictions.

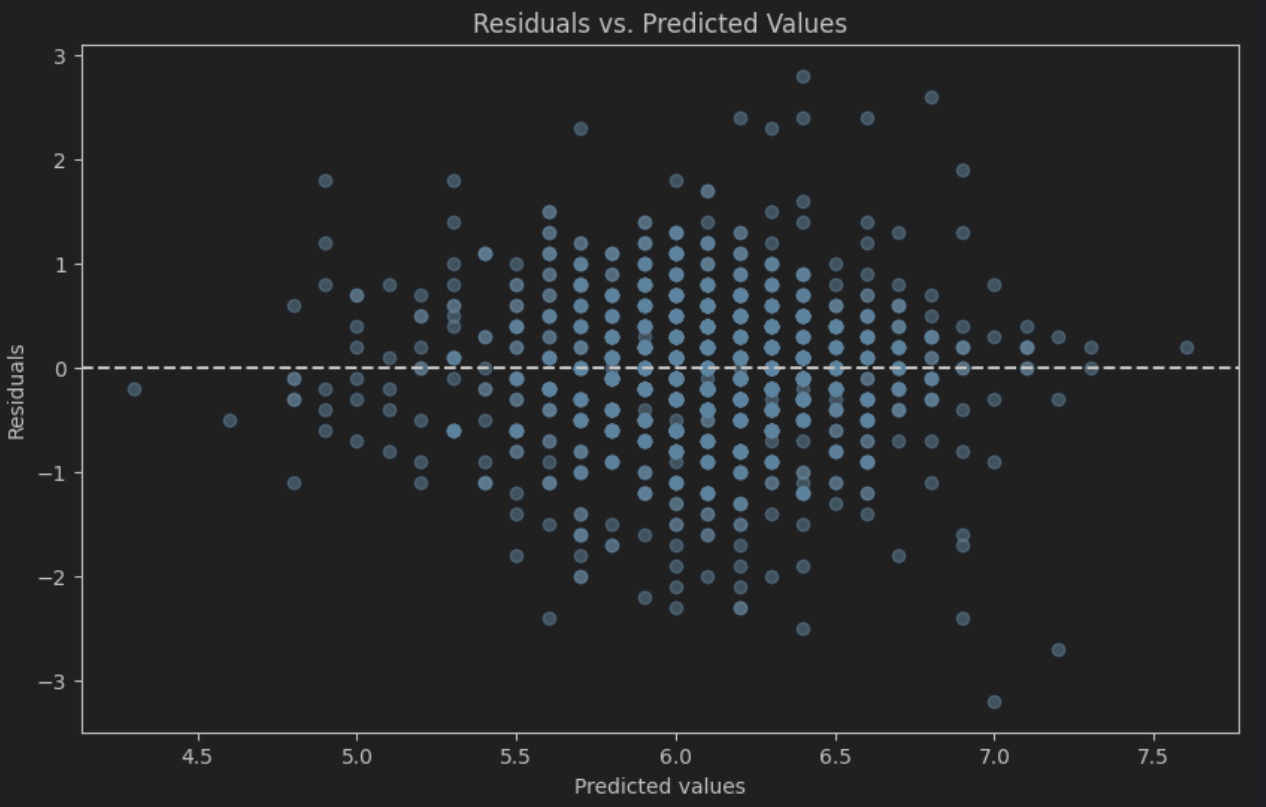


Figure 3.Residual plot of Linear Regression

Conclusion

The Linear Regression Model with Ridge regularization has shown modest performance in predicting the target variable, with a fair degree of error as indicated by the MSE and R² metrics. The residual plot suggests that there may be non-linearity or other factors affecting the model's predictions that are not captured by the linear model. Further model diagnostics and possibly exploring more complex models or feature engineering might provide improvements.

5.2 Random Forest Model

Modeling

**Model Initialization:** A Random Forest Regressor was initiated with 100 estimators (trees) and a maximum depth of 10. This configuration allows the model to learn complex patterns without becoming too deep which can lead to overfitting.

**Cross-Validation:** The model underwent 5-fold cross-validation on the training data, providing a robust estimate of its performance. This technique ensures that the model's ability to generalize is not based on a single split of the data.

**Hyperparameter Tuning:** Hyperparameter tuning was conducted using GridSearchCV, searching over a grid of different numbers of estimators and maximum depths. This exhaustive search across the specified parameter grid aims to find the most optimal settings for the model.

Results

**Cross-Validation Results:** The average cross-validation MSE for the initial Random Forest model was approximately 0.410 suggesting that the model performs consistently across different subsets of the training data.

**Test Set Evaluation:** After fitting the model to the training data, the MSE on the test set was approximately 0.452, which is slightly higher than the cross-validation MSE. The Mean Absolute Error (MAE) was about 0.467, and the R² score was 0.269. The R² value indicates that the model explains approximately 26.9% of the variance in the test set which is a moderate level of predictive power.

**Hyperparameter Tuning Outcome:** The best parameters identified by GridSearchCV were the same as the initial model's parameters, with 100 estimators and a maximum depth of 10. The best CV MSE found was consistent with the initial model's performance which suggests that the chosen hyperparameters were already optimal within the searched range.

**Best Model Test Performance:** Using the best parameters, the Random Forest model's MSE on the test set was about 0.455 which is comparable to the initial model's performance, indicating stability in the model's predictions after hyperparameter tuning.

**Residual Plot Analysis:** The residual plot displays the difference between the actual and predicted values against the predicted values. The residuals appear to be randomly distributed without any apparent pattern or trend. This randomness is desirable as it suggests the model's errors do not follow a systematic bias. Like the Linear Regression model, there seems to be some heteroscedasticity as the spread of residuals varies across the range of predictions which could be an area to investigate further.

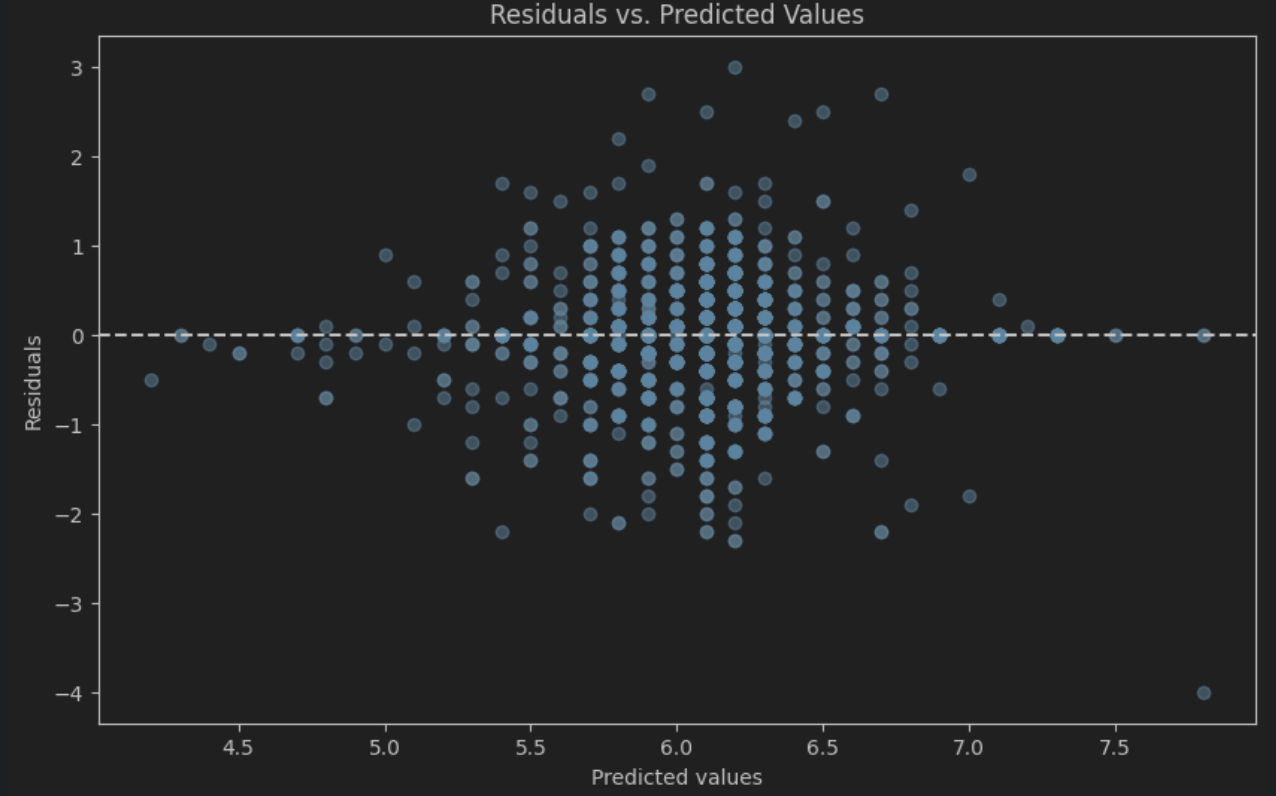


Figure 4.Residual plot of RandomForest

Conclusion

The Random Forest model demonstrated an acceptable level of performance on the training and test datasets with the hyperparameter tuning confirming the initial parameter choices. The residual plot indicates a good distribution of errors, although some heteroscedasticity is present. While the R² score shows that the model has a moderate predictive ability exploring additional feature engineering, trying other ensemble methods, or expanding the hyperparameter search space might lead to improved performance.

5.3 Gradient Boosting

Modeling

**Model Initialization:** A Gradient Boosting Regressor was initialized with 100 estimators and a learning rate of 0.1 providing a balance between model complexity and training speed. The random state was set to 42 to ensure reproducibility.

**Cross-Validation:** To estimate the model's performance, 5-fold cross-validation was performed. This method evaluates the model's generalization ability by training it on different subsets of the training data.

**Hyperparameter Tuning:** GridSearchCV was utilized to explore a range of hyperparameters including the number of estimators and learning rates. This approach seeks to optimize the model's settings for better performance.

Results  
Cross-Validation Results: The Gradient Boosting model achieved an average cross-validation MSE of approximately 0.410. This result suggests that the model's predictions are quite consistent across different parts of the training data.

**Test Set Evaluation:** The model was then fitted to the entire training dataset and used to make predictions on the test set. The MSE on the test set was approximately 0.463, with an MAE of about 0.487, and an R² score of 0.251 which indicates that around 25% of the variance in the test data is explained by the model.

**Hyperparameter Tuning Outcome:** The best hyperparameters identified were the same as those initially set with a learning rate of 0.1 and 100 estimators. The consistent MSE between the initial and tuned models suggests the initial parameters were well-chosen.

**Best Model Performance:** With the best parameters, the MSE on the test set was very slightly lower at approximately 0.463. This marginal difference indicates the model's robustness and that the chosen parameters were close to optimal.

**Residual Plot Analysis:** The residual plot shows the residuals distributed across the range of predicted values. Like the previous models, the residuals display no clear pattern, which implies no systematic errors in the predictions. There is some evidence of heteroscedasticity, as the variance of residuals seems to change with the magnitude of the predicted value, which could imply that the model's performance is not consistent across all levels of predicted values.

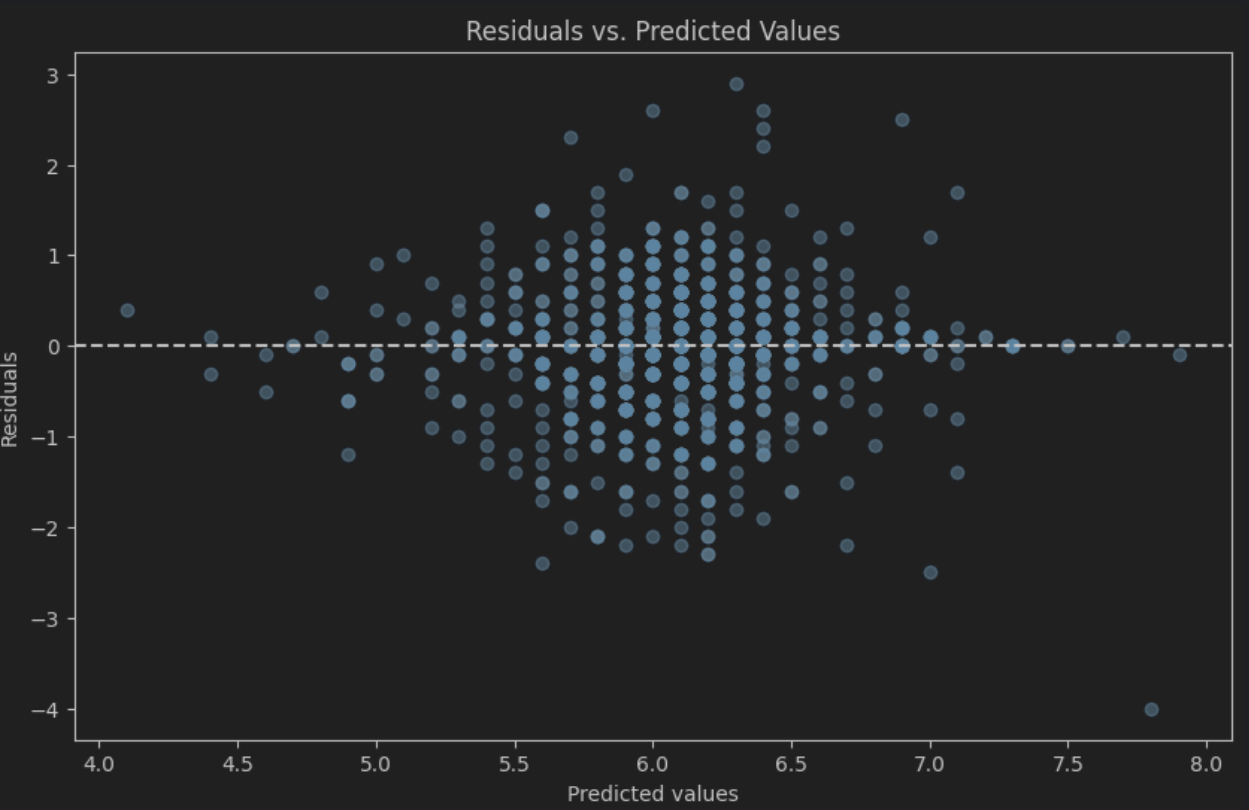


Figure 5.Residual plot of GradientBoosting

Conclusion

The Gradient Boosting model shows a good degree of consistency in its predictive performance as indicated by the cross-validation and test MSEs. The hyperparameter tuning confirmed that the initial model configuration was quite effective. While the R² score shows the model captures some of the variability in the target variable, further improvements might be possible with additional feature engineering, alternative model configurations, or more extensive hyperparameter tuning.

5.4 Neural Network

Modeling

**Data Preparation:** The data was split into training and testing sets with an 80/20 ratio and feature scaling was applied to standardize the features, an essential step for many machine learning algorithms especially neural networks.

**Model Architecture:** The model is a multi-layer perceptron (MLP) with three hidden layers. The first hidden layer has 64 neurons, the second 32 and the third 16, all using ReLU activation functions. The output layer has one neuron with a linear activation function suitable for regression tasks.

**Model Compilation:** The model was compiled with Stochastic Gradient Descent (SGD) as the optimizer and Mean Squared Error (MSE) as the loss function. These choices are standard for regression problems and are likely to converge well for this kind of dataset.

**Model Training:** The MLP was trained over 50 epochs with a batch size of 100. Validation data was used to monitor the performance on the test set during training, providing insights into the generalization of the model.

Results

**Test Set Evaluation:** The neural network achieved an MSE of approximately 0.527 on the test set which is higher compared to the previous models. The Mean Absolute Error (MAE) was around 0.545, and the R² score was 0.148 indicating that the model explains a relatively small portion of the variance in the test data.

**Residual Plot Analysis:** The residual plot displays the residuals versus the predicted values. The spread of residuals across the range of predictions indicates some degree of variance in the model's performance which isn't captured uniformly. As with the previous models, there seems to be heteroscedasticity where the spread of residuals increases with the magnitude of predicted values.

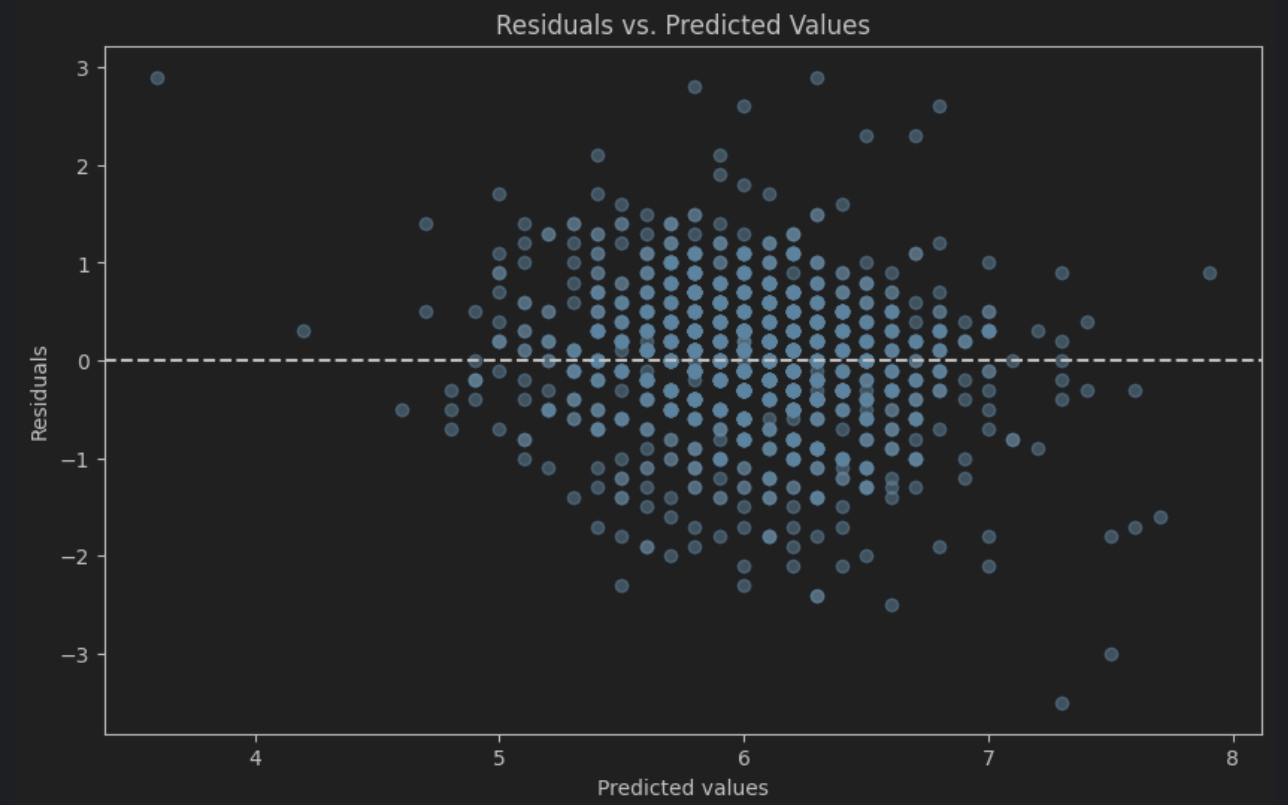
****

Figure 6.Residual plot of Neuro Network

Conclusion  
The Neural Network model shows a moderate to low level of predictive performance based on the MSE and R² values. The higher test MSE suggests that the model may not be capturing the underlying patterns in the data as effectively as the other models. Potential improvements could include experimenting with different architectures, activation functions or more advanced techniques such as regularization and dropout to prevent overfitting. Additionally, further hyperparameter tuning of the learning rate and the optimizer might yield better results.

5.5 Models comparing

|  |  |  |  |
| --- | --- | --- | --- |
| Model | MSE | MAE | R2 |
| Linear Regression | 0.48322908522283 | 0.51782642689601 | 0.218646401470113 |
| RandomForest | 0.45197028928850 | 0.46669272869429 | 0.269190074100523 |
| GradientBoosting | 0.46338545738858 | 0.48670836591086 | 0.250732404755751 |
| Neuro Network | 0.52705237979905 | 0.54511337414488 | 0.147786658206149 |

Overall Conclusion

**Linear Regression Model (Ridge Regression):** Exhibited a Mean Squared Error (MSE) of 0.483 and an R² score of 0.219. This model is straightforward and interpretable, but its predictive power was relatively limited.

**Random Forest Model:** Showed an MSE of 0.452 and an R² score of 0.269. This model performed better, suggesting a greater complexity and a stronger ability to capture nonlinear relationships in the data.

**Gradient Boosting Model:** Had an MSE of 0.463 and an R² score of 0.251. Its performance was comparable to the Random Forest indicating it as another viable option.

**Neural Network Model:** Recorded an MSE of 0.527 and an R² score of 0.148. Although neural networks often excel at capturing complex nonlinear relationships, this instance did not perform as well as the others.

Among these four, the Random Forest and Gradient Boosting models showed the best performance with similar MSE and R² scores. They offer better predictive capabilities and lower errors than both the linear regression and neural network models. The residual plots for these models indicate random distributions of errors without evident systematic biases, a sign of good model fit.

While the Gradient Boosting model's test MSE is marginally higher than that of the Random Forest, the difference is quite minimal. Gradient Boosting is typically robust at handling various types of non-linear problems and can often be improved upon tuning. However, Random Forest models are usually easier to train and less prone to overfitting. Considering these factors, we will use the Gradient Boosting model to do the prediction since our model is aiming at predicting the aftershock to protect the safety of the rescuer, training time and computational resources should not be the limiting factor affecting the accuracy of our prediction.

Therefore, we will use either the Gradient Boosting Model for further tuning or practical prediction.

**Appendix**

|  |  |
| --- | --- |
| **Raw data (combined\_df)** | |
| **Data Field** | **Definition** |
| time | Time of the earthquake occurrence |
| latitude | coordinate of earthquake latitude |
| longitude | coordinate of earthquake longitude |
| depth | earthquake focal depth |
| mag | earthquake magnitude |
| magType | measurement unit of magnitude |
| nst | number of seismic stations reporting |
| gap | angular gap between the stations |
| dmin | distance to the nearest station from the epicenter |
| rms | seismic signal |
| id | Unique identifiers of earthquake |
| updated | update timestamps |

|  |  |
| --- | --- |
| **Processed Data (df\_pure)** | |
| **Data Field** | **Definition** |
| time | Time of the earthquake occurrence |
| latitude | coordinate of earthquake latitude |
| longitude | coordinate of earthquake longitude |
| depth | earthquake focal depth |
| mag | earthquake magnitude |
| dif\_lat | difference in latitude |
| dif\_long | difference in longtitude |
| distance | distance between first shock and the next(aftershock) |
| cat\_afs | label indicate whether it is an aftershock |
| mag main | magnitude of the main shock; if it is ‘0’, meaning this earthquake is the main shock itself |
| depth\_main | depth of the main shock |
| mag\_pre | magnitude of the previous shock |
| depth\_pre | depth of the previous shock |
| target | magnitude of aftershock |
| longitude\_label | binned label of longitude |
| latitude\_label | binned label of latitude |