Depth Prediction Modeling for Earthquakes Near Puerto Rico

Kailande Cassamajor (kc3336)

December 2022

Abstract

Among many of the natural disasters, earthquakes are one of the most damaging and devastating. Determining earthquake source depth still exists as one of the problematic areas within earthquake source seismology work. This project aims to explore the construction of earthquake depth prediction models for earthquakes that occurred around Puerto Rico. Data was sourced from the United States Geological Survey Earthquake Catalog, and included observations from January 1st, 2009 through January 1st, 2019. Input predictor variables included latitude, longitude, normalized magnitude values, log energy released, and the distance from the closest city that the earthquake ruptured. This project used a regression approach; models included Random Forest, and gradient boosting algorithms LightGBM and XGBoost. Results indicated that better modeling approach, additional data, and understanding of the complex relationships between variables is necessary to construct a well developed predictive model for earthquake depth.

1 Project Objectives and Introduction

1.1 Background

Earthquakes, considered among the most dangerous natural disasters, have the ability to inflict devastating loss and destruction of communities, and structural buildings. Earthquakes are caused by slips on a tectonic fault line. A magnitude (Mw) 7.0 earthquake hit the Republic of Haiti on January 12th, 2021 around 4:43pm local time (DesRoches et al. 2011). While there was much uncertainty about the depth at which the earthquake occurred, the estimated depth of the earthquake was recorded to have occurred at a depth of 13km. On January 7th, 2020 a magnitude 6.4 earthquake occurred on at around 4:24pm local time at the southwest coast of Puerto Rico (Li and Pu 2022). Interestingly, researchers state that before 2019, the area whose movement led to the earthquake (state location), was "not very seismically intensive" with the last catastrophic earthquake having occurred in 1918 (Li and Pu 2022). Determining earthquake source depths still stands as one of the most difficult problems in earthquake source seismology (Craig, 2019). Solving such problems would greatly contribute towards earthquake hazard assessment, understanding the Earth's structure, including improving nuclear security (Craig, 2019).

1.2 Objective

It is important for researchers to understand the underlying mechanisms and drivers of earthquake depth, to get a better sense of how earthquakes are occurring at specific locations. In this project, we explore earthquake data from 2009 through 2019 specifically that occurred in proximity to Puerto Rico. The aim of this project is to construct a prediction model for earthquake depth using regression models such as Random Forest, XGBoost, LightGBM, in hopes to better understand the most important drivers of reported earthquake depth.

2 Data and Exploratory Data Analysis

The data used for this project is derived from the United States Geological Survey (USGS) earthquake catalog. The USGS provides relevant earthquake data, reports, and information in order to aid in the effort to reduce deaths, injuries, and property damage from earthquakes (USGS). The information from the earthquake catalog contains the time, latitude, longitude, magnitude, depth, place, along with additional data on the number of stations to detect magnitude, and variable error measurements.

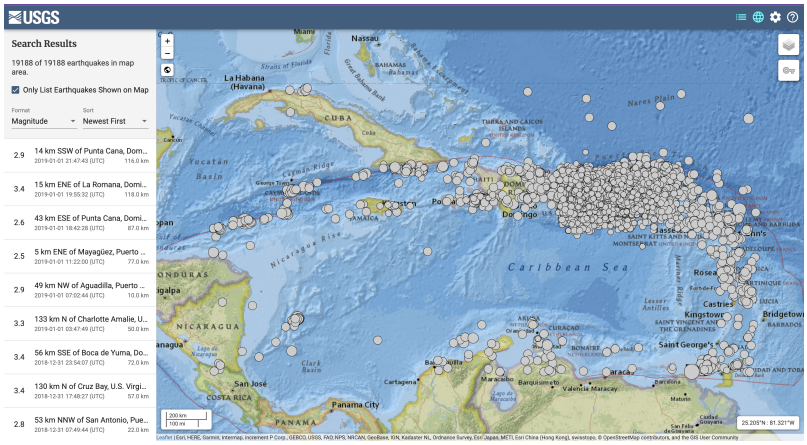

Figure 1: Earthquakes across selected area (USGS) earthquake catalog

For this project, 10 year span of earthquake data from January 1st, 2009 through January 1st, 2019 was selected. The selected area of interest was within the area of the Caribbean Sea with the following coordinates: [10.747, 25.562] Latitude and [-84.99, -59.441] Longitude. Including of all earthquakes that occurred near the multiple Caribbean nations, the resulting dataframe was 19188x22. During the cleaning process, the 'place' variable which included the distance in km and the city, county/territory, was split into 3 additional columns: 'dist_se': (float) the distance in km from the city, country/territory that the earthquake occurred, 'city': (string) the city, and 'country' (string) with the nation/territory name. Two additional variables were calculated and created: the normalized moment magnitude, and log energy released from the earthquake event.

2.1 Variable Distributions

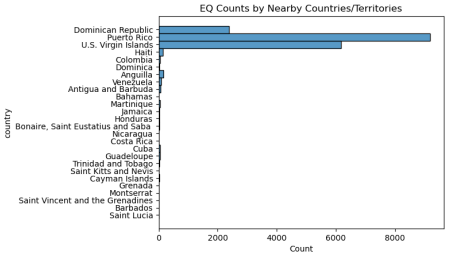

Figure 2: Earthquake Counts by Country

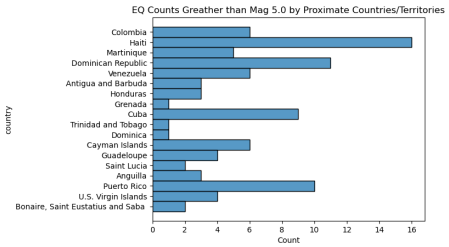

Figure 3: Earthquakes with Mw >=5 by Nearest Country

Figure 2 shows us that the majority of the earthquakes that occurred from 2009 through 2019, within the specified coordinated area, ruptured on or nearest to Puerto Rico making up 47% (9183 out of the 19188) earthquakes, with the U.S. Virgin Islands following behind. It made sense for this project to focus on predicting the depth of earthquakes that occurred at or nearest

to Puerto Rico for simplicity and to control for the location. Figure 3 shows that the majority of earthquakes that occurred between 2009 and 2019, with a magnitude of 5 or above, ruptured nearest to or in Haiti.

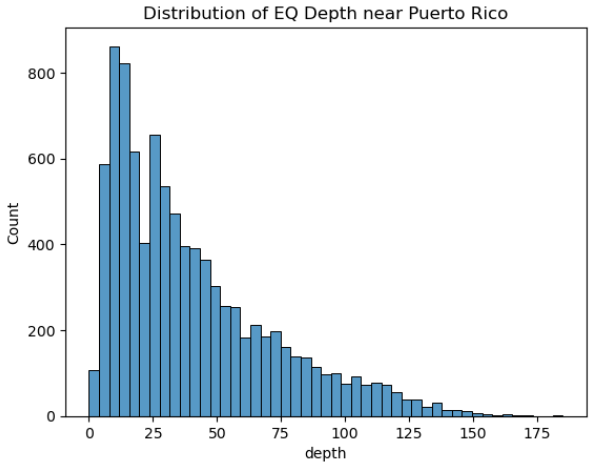

Figure 4: Distribution of Earthquake Depth nearest to Puerto Rico

Figure 4 demonstrates the distribution of the depth of earthquakes that occurred nearest to Puerto Rico, from 2009 through 2019. We notice a positively skewed distribution.

2.2 Correlation

Figure 5 shows the Pearson's-r correlations between the variables in the dataset, specifically where earthquakes occurred near Puerto Rico. We notice that dist_se, the distance from the nearest city in Puerto that the earthquake ruptured, and the latitude are strongly positively correlated. Since log energy released is calculated using the normalized moment magnitude, they are strongly correlated, and only one of these would be used in modeling. In addition, we notice that there are no strong correlations between earthquake depth and the other variables of interest.

3 Created Variables

3.1 Magnitude Normalization to Moment Magnitude (Mw)

The earthquake catalog includes the following magnitude types: moment magnitude (Mw), local magnitude (Ml), body-wave magnitude (Mb), and duration magnitude (Md). For consistency, the magnitude types were first normalized to moment magnitude according to the following investigated empirical relationships in Figure 7. This variable was named mag_norm.

3.2 Log Energy Released

Another way to calculate the size of an earthquake, is to calculate how much energy is released (USGS). Another column for this measure was created based off of the normalized moment magnitude. The energy released is a measure of the earthquakes damage potential (USGS). Energy released is estimated by the equation in Figure 8; E is energy and Mw is moment magnitude.

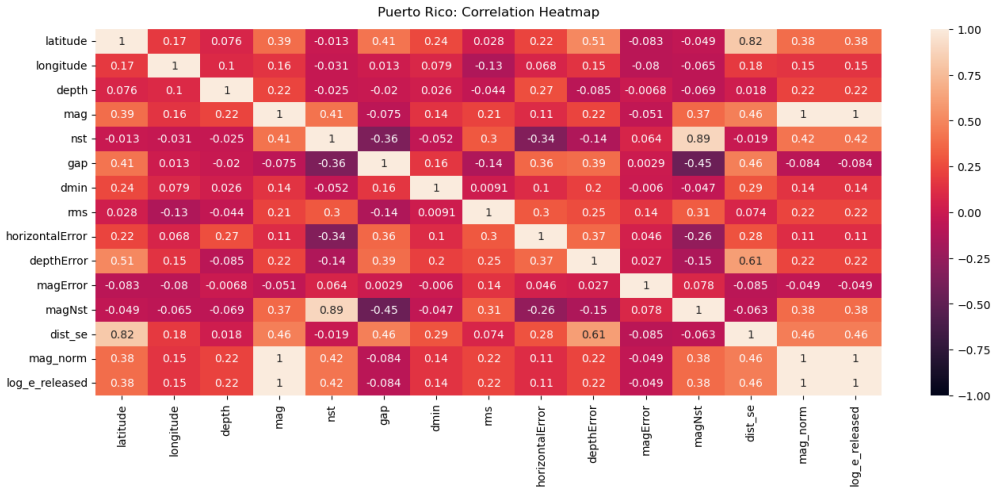

Figure 5: Variable Correlations

4 Methodology

A Random Forest Regressor, and two boosting methods: LightGBM, and XGBoost models were constructed in order to predict earthquake depth from the following input variables: latitude, longitude, mag_norm, dist_se, and log_e_released.

4.1 Models

4.1.1 Random Forest

The Random Forest is a popular and widely used machine learning algorithm and modeling technique that combines the performance of multiple decision trees in order to predict or classify a value. When the Random Forest takes in input variables (predictors), it builds a certain number of K regression trees and averages the results (Galiano et al. 2015). Random Forests are able to handle large and complex data. They are also able to provide a measure of the importance of each input variable, which can be useful for identifying the most relevant features in a dataset. This feature is useful for researchers to gain insight into underlying relationships between variables, and improving upon model development.

4.1.2 Boosting

Originally, boosting algorithms were designed sand implemented for classification problems. The concept behind boosting algorithms, are that they assemble or iteratively combine multiple simple models called 'weak learners' such as trees to construct a better performing learner with improved accuracy (Touzani et al. 2018). This concept was extended by Friedman to address regression problems, this was introduced as the gradient boosting method (GBM) (Touzani et al. 2018). Gradient boosting method arranges the weak learners sequentially - at each step, GBM adds a new decision tree that reduces the loss function to learn the residual between the target and the predicted value (Touzani et al. 2018). This method learns from the prior trees and updates the residual error. The result is the sum of the weighted the corresponding leaf value on each tree.

XGBoost, also known as eXtreme Gradient Boosting, is a gradient boosting algorithm based on weak learners like regression trees. XGBoost is considered an extension of gradient boosted decision (GBM) trees. XGBoost is considered an effective and powerful algorithm for supervised learning tasks.

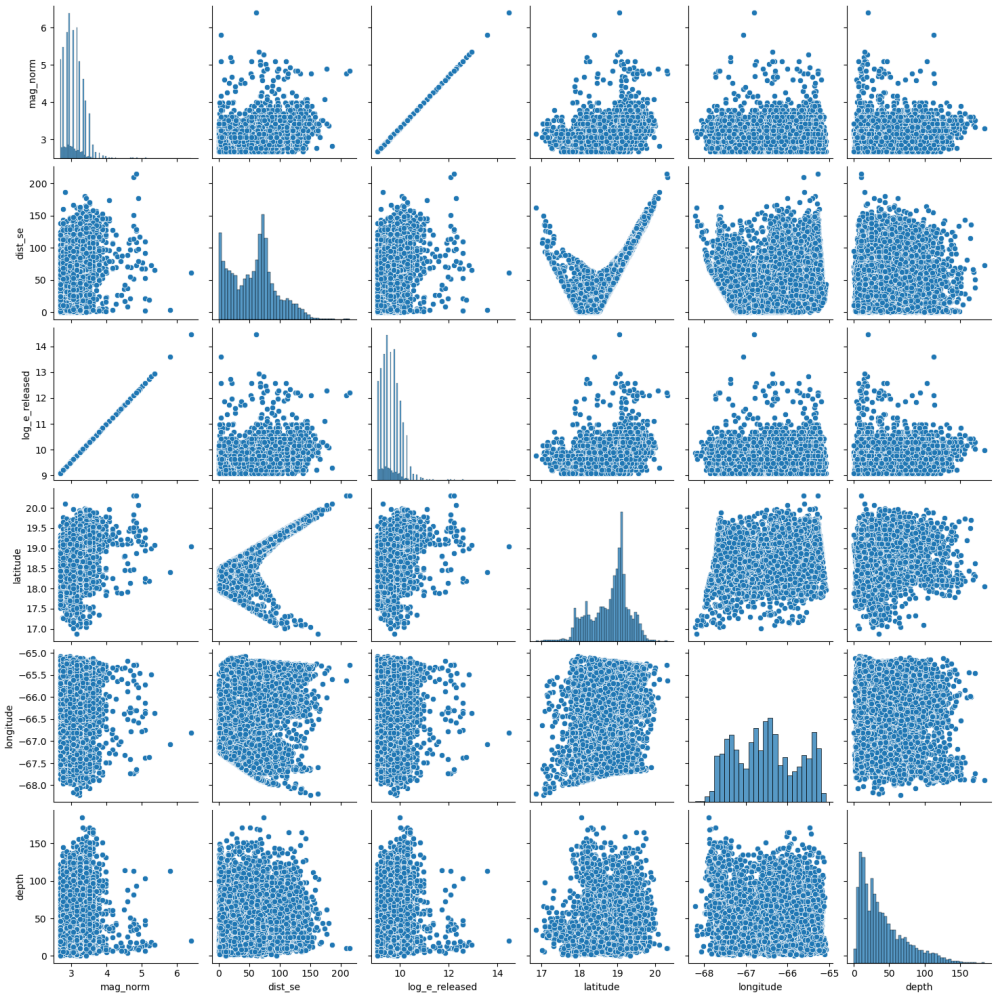

Figure 6: Variable Pairplot. There are no clear relationships between the 5 input variables and earthquake depth.

LightGBM works similarly to XGBoost, however, instead of developing the decision trees level-wise, this is done leaf-wise. LightGBM makes sure that one leaf is split into the next level according to the gain.

4.2 Approach

Gradient boosting regressors were chosen due to the complexity and nature of earthquake phenomena. Before splitting the data, the data was sorted by the date time variable, in ascending order. Due to the fact that the data is sequential (attached to a date-time sequence), the data was not randomized and train-test splitting was completed by completing structured splitting. The first 80% of the data was selected for training, while the rest (20%) were allocated as the test data. A cross-validation method was applied on all the models. Baseline models with a single parameter entered, were ran prior to model improvement efforts. Randomized Search was used for hyper-parameter tuning for each model. Feature importance's were also plotted to visually examine variables the models identified to be the most important drivers of earthquake depth prediction values.

1. M_L to M_w : $M_w = \frac{2}{3}M_L + 1.15(0 \leq M_w \leq 3.8)$
2. $M_w = 0.98M_L + 0.19(M_w > 3.8)$
3. M_b to M_w : $M_w = 0.85M_b + 1.02(3.5 \leq M_b \leq 6.2)$
4. M_D to M_w : $M_w = 0.93M_D + 0.35(3.5 \leq M_b \leq 5)$

Figure 7: Magnitude types and their relationship to moment magnitude (Li and Pu, 2022)

$$\log E = 5.24 + 1.44M_W$$

Figure 8: Log Energy Released (USGS)

4.3 Results

4.3.1 Random Forest Regressor

A simple random forest model with n_estimators = 100 was run and it's r-squared score for the training set was 0.9116, and 0.3038 for the test set. I then tried to improve the performance of the model by tuning several hyper-parameters: n_estimators, max_features, max_depth, min_sample_split, and min_samples_leaf. The best parameters were n_estimators=400, max_features = log2, max_depth = 100, min_sample_split = 10, and min_samples_leaf =4 which resulted in a r-squared of 0.6598 for the training set and 0.3345 for the test set. This resulted in a 10.1% improvement. Figure 9 shows that the most important features were latitude and longitude.

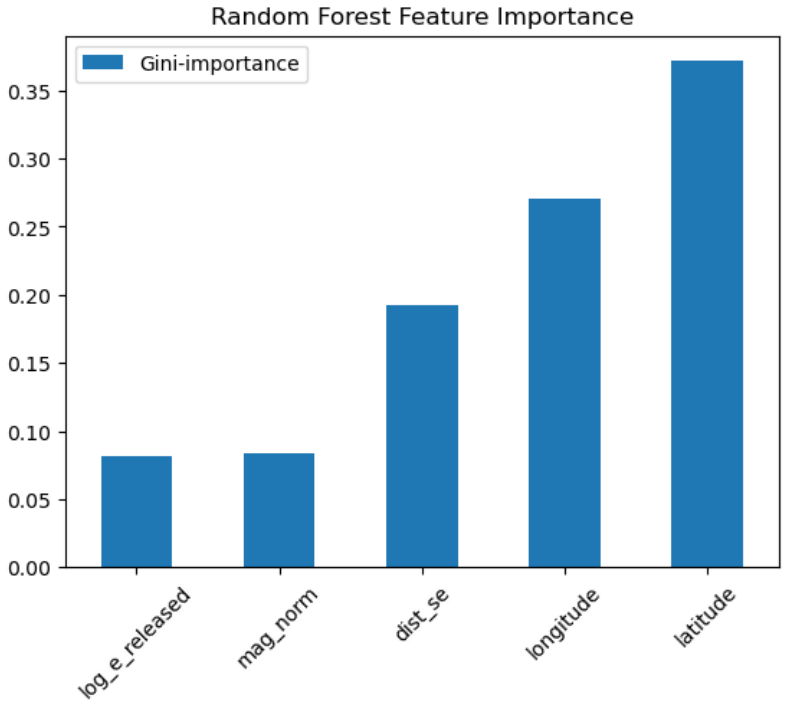

Figure 9: Random Forest Feature Importance

```
r_f r-sq 0.9116316212909321
r_f test r-sq 0.30388463718795244
```

Figure 10: Random Forest Baseline Performance

```
rf r-sq 0.6598533516318388
rf test r-sq 0.3345810906798695
```

Figure 11: Random Forest Performance After Hyper-parameter Tuning

4.3.2 LightGBM

The baseline LightGBM model with n_estimators=100, returned a r-squared of 0.5715 for the training set, and 0.3467 for the test set. After randomized search, the best parameters for the LightGBM model included the following: max_depth = -1, n_estimators=400, and learning_rate=0.01 with a best score of 0.3311. After fitting with the best hyper-parameters, the r-squared for the training set was 0.4957, and 0.3501 for the test set - a less than 1% improvement from the baseline. Similar to the random forest model, latitude and longitude were the most influential features in predicting earthquake depth.

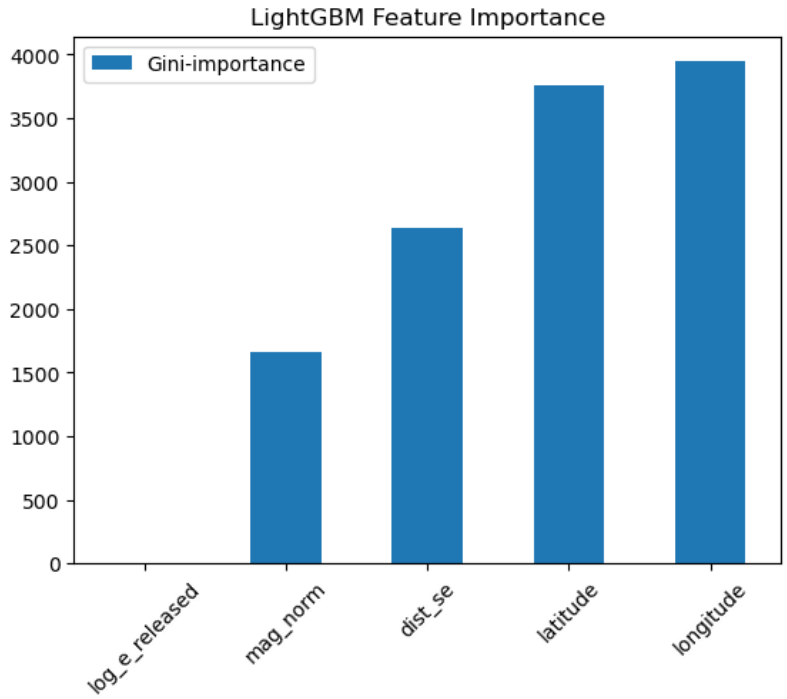

Figure 12: LightGBM Feature Importance

4.3.3 XGBoost

The baseline xgboost model with n_estimators=100, returned an r-squared of 0.7610 for the training set and 0.2672 for the test set. The hyper-parameter combinations were set as follows:max_depth: $[2, 10, 32]$, n_estimators: $[100, 200, 300, 400]$, learning_rate: $[0.01, 0.1, 1]$. The best hyper-parameters after randomized search were learning rate$= 0.1$, maximum depth

```
lgbm r-sq 0.5715258437902883
lgbm test r-sq 0.34675557126414536
```

Figure 13: LightGBM Baseline Performance

```
lgb train r-sq 0.4957067293614592
lgb test r-sq 0.3501162313616121
```

Figure 14: LightGBM Performance After Hyper-parameter Tuning

= 2 and the number of estimators = 200. The r-squared for the training set was 0.3994, and 0.3101 for the test set. The feature importance result was very different from the random forest and LightGBM model. Figure 15 illustrates that instead of longitude being the second most important feature in the model, it is the normalized magnitude, mag_norm.

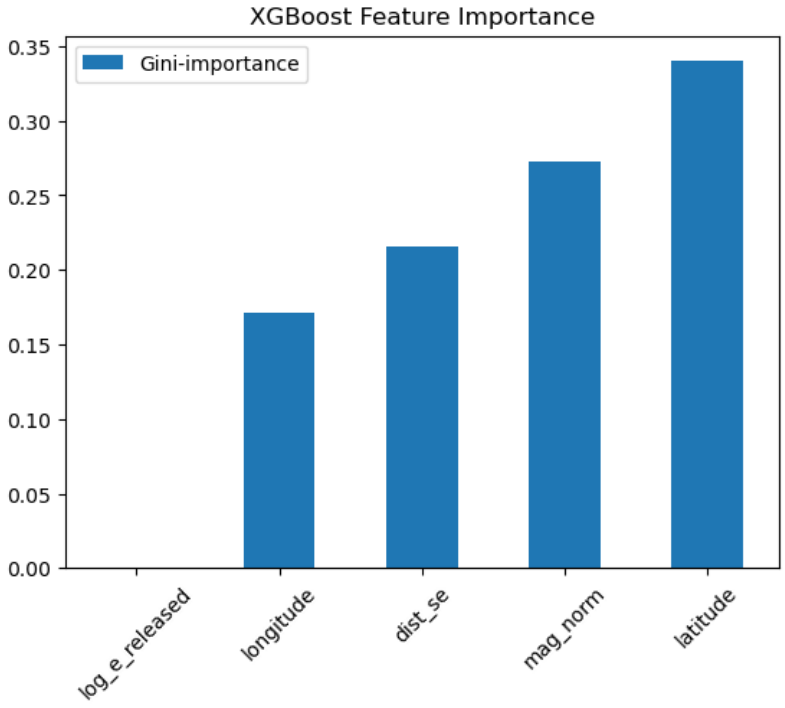

Figure 15: XGBoost Feature Importance

5 Conclusion and Limitations

The aim of this project was to develop a model to predict earthquake depth using the following variables: latitude, longitude, normalized moment magnitude, log energy released, and distance from nearest city. There are multiple issues with the approach and models above. There is a lack of understanding of the intricate relationships between spatio-temporal variables, seismic, and geological variables. Time was not a factor considered in this project (other than in train-test splitting), as it was reasoned that the depth of an earthquake is not a time dependent measure, however, this assumption was not approached with sufficient domain knowledge. There are details here that require domain knowledge, and are lacking here. While LightGBM appeared to

```
xgb r-sq 0.7610529606524495
xgb test r-sq 0.2672022681681203
```

Figure 16: XGBoost Baseline Performance

```
xgb r2 0.3994976506779625
xgb test r2 0.3101180447605484
```

Figure 17: XGBoost Performance After Hyper-parameter Tuning

be the best performing model, all of the models returned low r-squared scores indicating poor performance. There could be a collinearity issue. As demonstrated from the correlation matrix above, latitude and dist_se are highly correlated. log_e_released is also calculated from the normalized magnitude which is also included in all of the models, which could have negatively impacted model learning and performance. Another major point of error here is the lack of log-transformation prior to model training and testing. the variables dist_se, mag_norm, and log_e_released, including the target variable are positively skewed and this absolutely impacted learning and model results. This approach is a direct numerical prediction of earthquake depth, however, a better approach could have been a classification approach, and including some of the other categorical variables included in the dataset. Since it seems that latitude and longitude are most important in predicting depth, access to and understanding of spatio-temporal location data could have not only improved the models but changed the overall approach of this project.

6 References

J., Craig T. (2019). Accurate depth determination for moderate-magnitude earthquakes using global teleseismic data. Journal of Geophysical Research: Solid Earth, 124, 1759– 1780. https://doi.org/10.1029/2018JB016902

V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, M.J.O.G.R. Chica-Rivas Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines Ore Geol. Rev., 71 (2015), pp. 804-818, 10.1016/j.oregeorev.2015.01.001

Yuanhui Li, Wuchuan Pu; Analyzing the 2020 Mw 6.4 Puerto Rico Earthquake Sequence Based on the Epidemic-Type Aftershock Sequence Model. Seismological Research Letters 2022;; 93 (2A): 609–619. doi: https://doi.org/10.1785/0220210217

Megan Torpey Zimmerman, Bingming Shen-Tu, Khosrow Shabestari, Mehrdad Mahdyiar; A Comprehensive Hazard Assessment of the Caribbean Region. Bulletin of the Seismological Society of America 2022;; 112 (2): 1120–1148. doi: https://doi.org/10.1785/0120210157

Touzani, S., Granderson, J., Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy and Buildings, 158, 1533-1543.

USGS. (n.d.). Earthquake magnitude, energy release, and shaking intensity. Earthquake Magnitude, Energy Release, and Shaking Intensity | U.S. Geological Survey. Retrieved from https://www.usgs.gov/programs/earthquake-hazards/earthquake-magnitude-energy-release-and-shaking-intensity