Prediction of Hypocenter Depth With Ground Motion Variables For Seismic Hazard Analysis

MSc Research Project
Data Analytics

Sanket Dayama
Student ID: x18143652

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sanket Dayama
Student ID:	x18143652
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Supervisor:	Noel Cosgrave
Submission Due Date:	20/12/2018
Project Title:	Prediction of Hypocenter Depth With Ground Motion Variables For Seismic Hazard Analysis
Word Count:	XXX
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	9th December 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Hypocenter Depth With Ground Motion Variables For Seismic Hazard Analysis

Sanket Dayama
x18143652

Abstract

Abstract goes here. You should provide a high-level (approx. 150 – 250 words) overview of your paper, its motivation, and the core findings. This is the teaser of your work – it'll probably be best to write it last.

1 Introduction

Seismic hazard analysis has gained its importance over the years into the domain of earthquake engineering to tackle the problem natural disasters like earthquakes which are caused by the movement of tectonic plates and strong ground motions. This leads to destructive consequences resulting in loss of economy, human lives and increases the risk associated with infrastructure development (Huang et al. (2019)). The impact of the earthquakes can be limited by predicting the necessary variables so that the preventive measures can be improved. The ground motions of the areas showing high seismicity are studied periodically to provide useful information regarding the active seismic conditions of the sites and the intensity of the earthquakes which categorises them according to their hypocenter depth into shallow and deep earthquakes (Lin Thu Aung et al. (2019)).

The ground motions produced by the earthquakes are monitored with various intensity measures like Peak Ground Acceleration (PGA), Peak Ground Velocity (PGV) and accelerated response which are mostly taken into consideration for seismic hazard analysis for building earthquake resistant construction sites or structures like bridges, dams and skyscrapers (Xu et al. (2019)). The depth of an earthquake which is termed as the hypocenter is considered as an important variable in measuring intensity, the rupture caused in the area and to analyse the effects caused due to an event of an earthquake. It is also considered to study the deformation process which are caused due to tectonic movement and to analyse the shallow earthquakes which makes it an important variable to consider for seismic hazard analysis (Yu et al. (2019)). The seismic hazard analysis can be divided into two categories which are deterministic seismic hazard analysis (DHSA) and probabilistic seismic hazard analysis (PHSA). The DHSA is a method which performs analysis on the variables in relation with each other. On the other hand PHSA is a method where the analysis performed is time dependent.(Feng et al. (2020))

The motivation for this research came from an area of Nile valley in Egypt which is densely populated and with having a high risk of seismicity as many earthquakes have been recorded in that region and the structures constructed were not earthquake resistant from the strong ground motions. This increases the risk of consequences which need to be faced on the event of an natural calamity like an earthquake (Mostafa et al.

(2019)). The peak ground acceleration (PGA) was estimated in the Nile valley of Egypt in relation with various ground motion variables by performing deterministic seismic hazard analysis.Another aspect of motivation was taken from the Iran where number of earthquakes have been occurred of high intensity by making it one of the most active region in terms of seismicity. Evaluating the loss generated, a seismic hazard analysis was carried out by predicting the PGA and PGV in relation with the ground motion variables to restrict the consequences suffered (Darzi et al. (2019)).

There are some questions which arise in relating with the ground motion variables such as, are the variables PGA and PGV sufficient for performing seismic hazard analysis? The seismic hazard analysis deals with ground motion variables based on intensity of an earthquake. This gives rise to another question which is, can the depth of an earthquake be considered for analysing the seismicity of an area in relation with PGV and PGA and other ground motion variables ? These type questions need to be resolved, as the depth of an earthquake which is termed as hypocenter plays an important role with the intensity of an earthquake occurred. This can be resolved by identifying the relation of ground motion variables with hypocenter depth to differentiate between the shallow and deep earthquakes.

The inspiration to implement this research was gathered from Vaez Shoushtari et al. (2018), Podili and Raghukanth (2019) Zanini et al. (2019), Sabermahani and Ashjanas (2019) and Hamze-Ziabari and Bakhshpoori (2018) by performing supervised machine learning techniques for predicting ground motion variables like PGA, PGV and cumulative absolute velocity. The data used in this research consists of ground motion variables which covers the intensity measures and other variables relating to rupture and compressed principal stresses in fault plane. Couple of authors like Derakhshani and Foruzan (2019) and Raghucharan et al. (2019) also used artificial neural network for predicting various ground motion variables.

The data used in this research was acquired from Pacific Engineering Earthquake Research which was used by some studies for predicting ground motion variables relating to velocity, acceleration and displacement with selected features which are mentioned in related work. This data had a limitation of many missing records as it was polluted with signal noise, and the information were collected in a poor manner (Du et al. (2019)). This research was carried out by implementing regression models with bagging and boosting techniques like random forest regression, light gradient boosting regression and ridge regression. some other techniques implemented were cross validated with k-fold, hyper parameter tuning and transforming the data by using square root. The models were evaluated using different evaluation techniques such as R squared value, RMSE and MAE. Considering all these circumstances a research question was formed.

1.1 Research Question

Can hypocenter depth be predicted using ground motion variables with regression techniques for the purpose of deterministic seismic hazard analysis?

1.2 Research Objective

This research will be focusing on predicting the hypocenter depth of earthquakes in relation with other ground motion variables for seismic hazard analysis.

1. To identify the variables which are best suitable to predict the hypocenter depth for deterministic seismic hazard analysis.

2. To implement algorithms with different experiments and to improve the efficiency of the implemented models.

2 Related Work

This section consists of the previous research work carried out for predicting different variables associated with ground motion data using different techniques. Some past studies relating with depth prediction from different domains have also been included in order to get a broader aspect for prediction of hypocenter depth. In this research, an effort has been made to choose a novel approach and study if those models can be applied for hypocenter depth prediction.

2.1 Past Studies Performed on Different Variables of Ground Motion Data

2.1.1 Analysis on Ground Motion Using Regression Techniques

Different kind of earthquakes occur which are categorized as shallow and deep earthquakes. The seismic nature of the earthquake depends on the depth of an earthquake. The seismic layer is separated into two types by a seismic boundary specifically into an upper layer and lower layer of seismicity in order to categorise the seismic hazard analysis. One such study performed by Chung et al. (2018) shows that the depth decreases or gets vanished due to the cracks formed above and below the seismic layer boundary. The earthquake data consisted of south korea of 53 earthquakes. 7 velocity models were used to determine the hypocenter of the earthquakes along with HYPO71 which is a location capturing software for earthquakes. The RMSE value obtained was 0.5 at 0-30Km. Scattering attenuation analysis was performed by using Multiple Lapse time window analysis for ensuring the seismic boundary between upper and lower layers which was 12Km. It was also analysed that 60 percent of earthquakes happened in the upper seismic boundary layer.

Along with the seismic layers and seismic boundary Chung et al. (2018), it is important to study the ground motions nearby the epicenter which can give a clear idea of the shift of seismic boundary and thereby causing depth of the layers to change. This limitation was brought under the control by Vaez Shoushtari et al. (2018). The regions of Sumatra,peninsula region of Malay and Japan were analysed for seismic hazard analysis as megathrust earthquakes were experienced implying to rough ground motion which can impact negatively on the society. The seismic hazard analysis was performed using ground motion data which was acquired from Pacific Engineering Earthquake Research. The peak ground acceleration (PGA) was predicted by considering the variables like peak ground velocity (PGV), a categorical variable consisting of four site classes based on reducing the earthquake hazards program, pseudo-spectral acceleration (PSA) and long range of hypocenters (Rhyp) which ranged specifically between 120 and 1300 km. As the nature of data set was studied was a non-linear data hence the least-squares regression method was chosen for predicting PGA. The result parameters such as RMSE and MAE were not mentioned.

Amongst the variables chosen for seismic hazard analysis in previous studies of Vaez Shoushtari et al. (2018) and Podili and Raghukanth (2019) completely focus on the PGA parameter, this study consists of analysing the ground motion data with a different variable named cumulative absolute velocity (CAV) which quantifies the risk of potential damage by earthquakes to structures.(Xu et al. (2019)) The data chosen for this study was particular to Taiwan region acquired from TMIP with the magnitude above 4.8 and the epicenter area in the range of 200 Km. Regression models were executed on the data set consisting of number of records of 24667 resulting in the R squared value of 0.74 for deep earthquakes and 0.67 for shallow earthquakes. The models were validated by assessing the distribution of inter-event residuals and intra-event residuals. It was concluded with the analysis performed that Taiwan would experience a CAV greater than 0.97 g-sec per year and having a 10 percent chance of experiencing a seismic hazard in 50 years.

A study related to similar region of Taiwan Xu et al. (2019) of eastern coast was explored on the ground motion data obtained from the Broadband Array in Taiwan for Seismology (BATS) and the Central Weather Bureau 24-bit Seismic Monitoring Network (CWB 24-bit). The captured data consisted of the epicenter of less than 80 Km. (Lee et al. (2019)) This study was performed on an single earthquake of 6.2 magnitude in order to study the strong ground motion and the rupture process to identify the involvement of the structure of rupture in the particular event. A parallel non-negative least squares method was used to perform the analysis. Unlike the studies of Vaez Shoushtari et al. (2018), Podili and Raghukanth (2019) and Chung et al. (2018) were performed on multiple earthquakes and with specific range of the regions and epicentral area which provide more efficient results of the ground motion than one earthquake.

Mostly the ground motion variables are analysed based on 2 to 3 variables or just taking the major features into consideration (Vaez Shoushtari et al. (2018)). But it is important to consider other factors which are related and can influence on the ground motion. This was experimented in the study of Podili and Raghukanth (2019) the ground motions are analysed with 21 features having the properties of ground motion acceleration. For the analysis the Japanese region was studied and analysed having 96880 number of records which were taken from Knet database with the earthquakes having the magnitudes between 5 to 9. Non-linear least square regression analysis was performed and it was concluded that the events were not contributing to the ground motion equation. The inter-event and intra-event residual plots were plotted to analyse the errors. unlike other studies Vaez Shoushtari et al. (2018), Xu et al. (2019), Lee et al. (2019) which only focused on ground motion variables Podili and Raghukanth (2019) took the factors which could promote the rough ground motion like volcanic arc motions and tectonic activities and behaviour of model was studied.

Inspired from the studies of Vaez Shoushtari et al. (2018) and Podili and Raghukanth (2019) which consists of predicting the ground motion variable PGA, another study from Zanini et al. (2019) aims to contribute by covering the gap of EMS-98 intensities and ground motion variables like PGA,PGV and peak ground displacement (PGD) by establishing regression relationship. The ground motion/EMS-98 data was collected from Italian accelerometric archive and for checking the normality of the data, Kolmogorov-Smirnov normal test was performed. The orthogonal distance regression technique was applied in order to overcome the problem of uncertainty between the dependent and independent when chosen the ordinary least squared method. The R squared value obtained with the parameters chosen of ground motion were 0.94. This research gradually performed better than the others by checking the normality of data and implementing appropriate methods which resulted in high R squared value.

Similar studies performed by Vaez Shoushtari et al. (2018), Podili and Raghukanth (2019) and Zanini et al. (2019) for prediction of the ground motion variables like PGA, PGV and PGD was performed but the intent was to improve the prediction using bagging techniques used for regression.(Hamze-Ziabari and Bakhshpoori (2018)) The dataset chosen consisted of 15521 records and 322 earthquakes upon which different regression techniques like CART,M5 and Ensemble CART+M5 algorithms were applied. Parameter tuning by decreasing the size of the leaf and depth of the trees was done on all applied algorithms and the results were compared. The results obtained for PGA,PGD and PGV using CART, M5 and Ensemble CART+M5 were compared using MAE and RMSE parameters. The lowest result obtained was from the ensemble approach from CART+M5 having MAE of 0.027 and RMSE of 0.38 for PGA. For PGV the MAE and RMSE were 0.27 and 0.37 and for PGD was 0.35 and 0.50. Amongst the studies for predicting these variables, this was the best approach taken to achieve results as the parameters were hyper parameter tuned unlike in other studies.

On the similar grounds by using the same database in studies of Vaez Shoushtari et al. (2018) and Hamze-Ziabari and Bakhshpoori (2018) but with different features, A regression technique named ridge regression was applied to predict and analyse the parameters like PGA,PGD and PGV with having different features such as earthquake magnitude, hypocenter distance and fault type were considered. The results were studied on the metrics of RMSE of 0.038 for PGA and R squared value obtained was 90 percent. This technique resulted the best R squared value with lowest RMSE than other studies, with probability of selecting the best features having the problem of multi co-linearity as it is handled by ridge regression (Sabermahani and Ashjanas (2019)).

Amongst many properties listed to predict the ground motion relating with seismicity, ground motion cycles was predicted in this study Du et al. (2019) with choosing the same database used by Vaez Shoushtari et al. (2018) and Hamze-Ziabari and Bakhshpoori (2018).Number of records were removed because of poorly recorded data thereby choosing four variables for analysis by using rainflow range counting method. Mixed effects regression analysis and a library named ”nlme” from R were used analyse the data. The validation process for done by plotting the residuals and checking their distribution.

Another region of south east of the Europe which is the Vrancea region was explored by Borleanu et al. (2017) for analysing the seismic activity associated with intermediate-depth which was monitored by stations of Romanian seismic network.The prediction of peak delay time for S waves, consisting of earthquakes having the hypocentral depth of 100 to 250 km was carried out by a least-squares regression technique. The results were calculated with RMSE values on different frequencies which were obtained in the following pattern, for 2-4 Hz RMSE ontained was 0.087, for 4-8 Hz RMSE ontained was 0.12,RMSE ontained was for 8-16 Hz 0.15and 12-24 hz RMSE ontained was 0.17. It was concluded that the peak delay time and scattering of seismicity was decreasing with increase in depth hypocentres.

2.1.2 Analysis on Ground Motion Using Artificial Neural Network

Another analysis was performed on ground motion data which was used by Vaez Shoushtari et al. (2018), Hamze-Ziabari and Bakhshpoori (2018) and Sabermahani and Ashjanas (2019) that is pacific earthquake engineering center from NGA-West2 database. This analysis used a deep learning approach which was Artificial neural network for predicting the variables like PGA, PGV and PGD on the basis of earthquake magnitude, rake angle,source to site distance and soil shear wave velocity. The analysis was performed on 12556 records after pre-processing. The weights were normally distributed for the network and Rectified linear units(ReLU) was the activation function used. The results of the model were measured with RMSE and MAE metrics. The PGV predicted with ANN had RMSE of 0.503, MAE of 0.397; for PGD the results came out to be RMSE of 0.735 and MAE of 0.575 for PGA the RMSE was 0.029 and MAE of 0.024.(Derakhshani and Foruzan (2019)) The RMSE was relatively low as compared to the study of Sabermahani and Ashjanas (2019) using ridge regression. In this case with the selected features ANN performed better.

The same technique used by Derakhshani and Foruzan (2019) of ANN was used on the data of ground motions of Indo-Gangetic region.The data used was obtained from two databases which were program for excellence in strong motion studies (PESMOS) and a network of central Indo-Gangetic plains. Two new variables were introduced which were focal depth and average shear wave velocity from surface to depth of V30s along with the independent variables which are commonly taken to predict the PGA. The optimal number of nodes for ANN were searched which turned out to be 9 at which the standard deviation for PGA was 0.287.(Raghucharan et al. (2019)). But Thomas et al. (2016) used a randomized adaptive neuro fuzzy inference system (RANFIS) model with 100 epochs was executed which gradually after trial and error increased the performance with requiring less computational time. Amongst the neural networks and regression techniques this model performed the best with obtaining lowest error values. The results obtained after predicting the PGA, PGV and PGD were as follows, for PGA the RMSE obtained was 0.052, MAE was 0.026. For PGV the RMSE obtained was 0.045, MAE was 0.025 and PGD had RMSE of 0.035 with MAE being 0.016. This model was executed on 2815 records after pre-processing.

2.2 Motivation for Prediction of Depth in Different Domains

The hypocenter of an earthquake is always below the ground, to predict that a similar approach of predicting the soil depth is referred. The independent variables which are referred to as the environmental factors were selected by using three methods which are Pearson correlation, general additive models and random forest to predict the soil depth (yuan LU et al. (2019)). The use of pearson co- relation was also used in the study of Han et al. (2018) for testing the multi col-linearity and the variables having value greater than 0.9 were removed. In order to predict the soil depth, random forest regression and Support vector regression models were used. Random forest performed better by having the R squared value of 0.62 and RMSE of 10.62 against the Support vector machine which had R squared value of 0.49 and RMSE of 12.67 (yuan LU et al. (2019)). The study of Han et al. (2018) implemented linear and non linear algorithms. The multiple linear regression was performed on two hill slopes of H1 and H2. The best R squared value of 0.63 was obtained for H1 was having RMSE of 23.34 and MAE of 15.68 where as for H2 the best R squared value 0.64 having on H2 having the lowest RMSE of 24.12 and MAE of 17.60. But the ANN1 for H1 gave relatively better results in terms of MAE and RMSE being 20.47 and 26.37 which was used with same variables as MLR than ANN2 which was used with all variables with MAE being 21.09 and RMSE 27 for the Rsquared value 0.63.For H2, the ANN2 performed better than ANN1 having less RMSE and MAE score of 23.93 and 15.03 with greater R square of 0.75. This study further applied direction algorithms D8 and D-infinity to calculate the surrounding area.

The motivation was also gathered from prediction of groundwater depth. This study aims to predict the groundwater depth on the maximum potential height of the tree .(Yang et al. (2019)) A regression analysis was performed to choose the most important and best predictor variables upon which maximum tree height and volume were taken into consideration. A classical measurement error (CME) model was used with obtaining the R2 of 0.82 and RMSE of 0.33. Leave one out cross-validation technique was used upon which the optained R squared value was 0.86, which showed that the prediction of ground water depth was correlated in a significant manner.

The water table depth was predicted for managing the groundwater availability. For feature selection, a similar approach carried out by Yang et al. (2019) was carried out where the Lasso regression technique was used and important features were taken into consideration. The values were scaled and then the LSTM model was applied in order to predict the ground water depth. (Zhang et al. (2018)) As deep learning techniques are prone to overfitting with less number of data, a regularization method named dropout was used to fix the issue.The Rsquared value obtained particularly ranged between 0.78 to 0.95 and the RMSE from 0.070 to 0.184 for five areas.

A different approach of predicting the water table was carried out by using vertical 1-D numerical model by considering four soil variables to identify the soil properties by using the rainfall data. Amatya et al. (2019) The R squared value ranged between 0.80 to 0.87 and another metric named NSE(Nash sutcliffe efficency) was considered which ranged from 0.77 to 0.87. Comparing with the studies of Yang et al. (2019) and Zhang et al. (2018) the deep learning technique LSTM tends to perform better than CME model and the 1-D numerical model with appropriate techniques like lasso and dropout applied to implement the model.

2.3 Conclusion

After going through the past studies of ground motion data, there consists of many variables from which the seismic hazard of a particular area can be analysed and predicted with machine learning techniques. The techniques used to analyse the ground motion data improved gradually with regression techniques being used by Vaez Shoushtari et al. (2018), Xu et al. (2019) and Podili and Raghukanth (2019) to study the variables like peak ground acceleration and peak ground velocity. Bagging methods Hamze-Ziabari and Bakhshpoori (2018) were also used to improve the prediction of these variables. Advanced machine learning techniques like ANN were used Derakhshani and Foruzan (2019) and Raghucharan et al. (2019) which showed the improvement in prediction of ground motion variables. It was also explored that the research was not only limited to PGA and PGV but other variables were explored like cumulative absolute velocity (CAV) by Xu et al. (2019), which shows that there are other variables which can be explored and used for seismic hazard analysis. The motivation was created after studying the related work to adopt a machine learning approach to predict the depth of the hypocenter with the ground motion variables for seismic hazard analysis.

3 Methodology

The methodology selected for this research was CRISP-DM. This method was selected as some of the phases like business understanding, data understanding and preparation of data out of others were crucial for this research (Bošnjak et al. (2009)). An execution plan was created for the methods which were implemented in this research. The Plan is was carried out in number of steps.

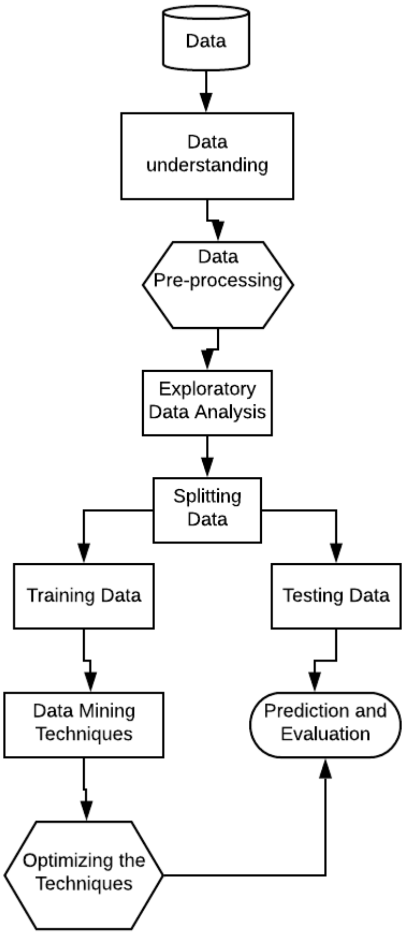

Figure 1: Research Flow Diagram

- **Step 1:** The data consisting of multiple earthquakes was gathered from Pacific Earthquake Engineering Research into the CSV format consisting of 21541 records and 273 columns.

- **Step 2:** With a large number of columns present in the data, the data dictionary was referred to understand and gain importance of existing columns .

- **Step 3:** The data set consisted of missing values at a large extent, it was necessary to perform a missing data analysis and identify the pattern of missing data. It plays a major role in deciding of which columns to be removed and retained. As this research focuses on implementing regression models which is discussed in the related work[2], the data should be in numeric format. The columns consisting of names like Earthquake names, Station names, File Name Horizontal, File Name Vertical were removed and some of the columns like Instrument Model, PEA Processing Flag, Type of Filter, HP-H1 (Hz), HP-H2 (Hz), LP-H1 (Hz), LP-H2 (Hz), Lowest Usable Freq - H1 (Hz) , Lowest Usable Freq - H2 (H2) , Lowest Usable Freq Ave. Component (Hz) were removed on the basis of poor recording of data and uncorrected time series present in the data as per the knowledge gained from data dictionary.

- **Step 4:** In order to get the better understanding of the data exploratory data analysis is performed on a clean set of data.Scatter plots were plotted on against different attributed with the dependent variable were interpreted. The normal distribution of the data set was also checked and it was found that the data was skewed towards the right which means positively skewed data. To handle this issue, firstly the data was transformed by converting the values into a logarithmic form. This resulted in not much effect which was overcome by applying square root transformation on data. Dealing with such large dimensional data, there was a need of reducing the dimensions for making the model work efficiently and getting rid of columns which are causing the problems of multicollinearity. The most common dimensional reduction technique is principal component analysis (PCA) which has its own limitations that is, it depends on linear relationships of amongst the features. The features which are not contributing or gives a very little contribution to the variance in data are hidden and at times it removes those important features which contribute less to the variance but generate an impact. This reduces the performance of the model. Hence, a feature selection technique named boruta was chosen in this research.

  Boruta works on creating the shadows of all the features present in the data and then shuffles the shadow features to check the correlations. A random forest is executed to calculate the Z score of the features. The Z scores of the features are compared with the shadow features having maximum score and the features having significant low score are termed as not important and are removed. The features having higher score are retained by terming an important feature. The shadow features are removed and this process is iterated a number of times until the important features are gathered (Kursa and Rudnicki (2010)).

- **Step 5:** After exploring the data and performing pre-processing on data, different machine learning techniques were applied. The machine learning models were applied with two experiments. The first experiment was performed by simply executing with simple split which is termed as hold-out method and the second experiment was performed by hyper-parameter tuning with k-fold by using randomized search algorithm.

  Random search works by generating random combinations for tuning the parameters to find the optimal solution for the algorithm. Unlike the grid search which creates a predefined list of values and then tries on creating different combinations. The limitation of grid search technique is it cannot deal with high dimensional data as the tuning parameters are increased. On the other hand random search has an advantage of extracting the best parameters with a range of values present in the data with number of iterations and combinations according to dimensions present in the data set. In this research, the models were chosen according to the behaviour of the data which was identified by plotting scatter plots. As the data showed no linearity in its behaviour, the bagging and boosting techniques were used. Random Forest was used as a bagging technique. The training data is divided into several parts and the decision trees are trained. Each tree gives its own predicted value and its mean is used for the purpose of the accuracy. Another technique into the category of boosting which is Light gradient boosting was used in this research. It is an ensemble technique where the focus is on minimising the residual errors by fitting a new independent variable in place of the previous variable contributed more error. With the CART algorithms used for predicting the PGA and PGV as seen in the previous study of Hamze-Ziabari and Bakhshpoori (2018), these algorithms were used on predicting the hypocenter depth of an earthquake. As the data was suffering from multicollinearity a regularization technique which reduces the value of coefficients was used which is Ridge regression. The Ridge regression was chosen over Lasso regression because Lasso reduces the value of coefficients to zero which makes the features of no use. Hence, these were the methods chosen in this research according to behaviour of the data.

- **Step 6:** At last, the models used in this research were evaluated using R squared value, RMSE and MAE as discussed in the section 2.

4 Implementation

The discussed plan in the methodology section [3] was implemented with different techniques are tried which are discussed in this section.

In terms of data, it was downloaded from the Pacific Earthquake Engineering Research into a CSV format which was loaded into the google drive and Google Colab was linked with the data to perform the operations using Python 3 using the libraries ”google.colab” and ”drive”. The ”Pandas” library was used to read the data with ”read-csv()” function and perform various operations. The data consisted of -999 values which as per the data dictionary were denoted as the values which were not recorded properly. These values were replaced to NA with a replace function. As the research focuses on the regression problem, the columns with names and some unnecessary columns mentioned in the step 3 of the section [3] were removed with referring the data dictionary.

4.1 Identifying the Missing Data Pattern

The data consisted of large number of missing values therefore, it was a need of missing data pattern to be identified. A Nullity matrix was created using ”Matplotlib” and ”Missingno” libraries.A ’Monotonous’ missing data pattern was identified which is mostly present in the periodic data. The missing data is said to be monotonous when the features in the data set are interdependent on each other or are auto correlated. If the variables such as X1 to Xn present in the data set has a feature Xk which is has missing values then the columns which depend on Xk will have missing values.(Leke and Marwala (2019)) Below is the matrix visualised to identify the missing data pattern.

From the above graphs we can see that there were huge number of missing values in the data set. The columns which had missing data greater than 50 percent were removed by using a for loop by first getting the per column percent missing values. The relations between the features were also studied by plotting a dendrogram which helps in showing the hierarchical relationships. It was plotting by using the ”Missingno” library in python.

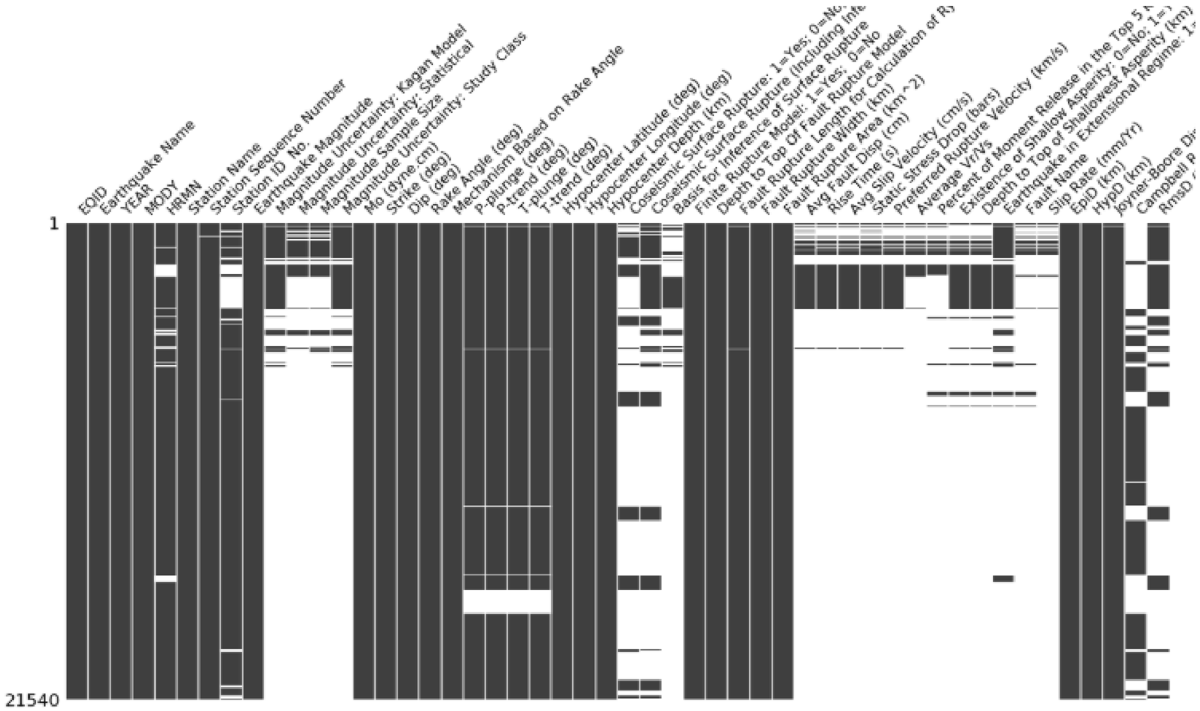

Figure 2: Monotonous Missing Data pattern

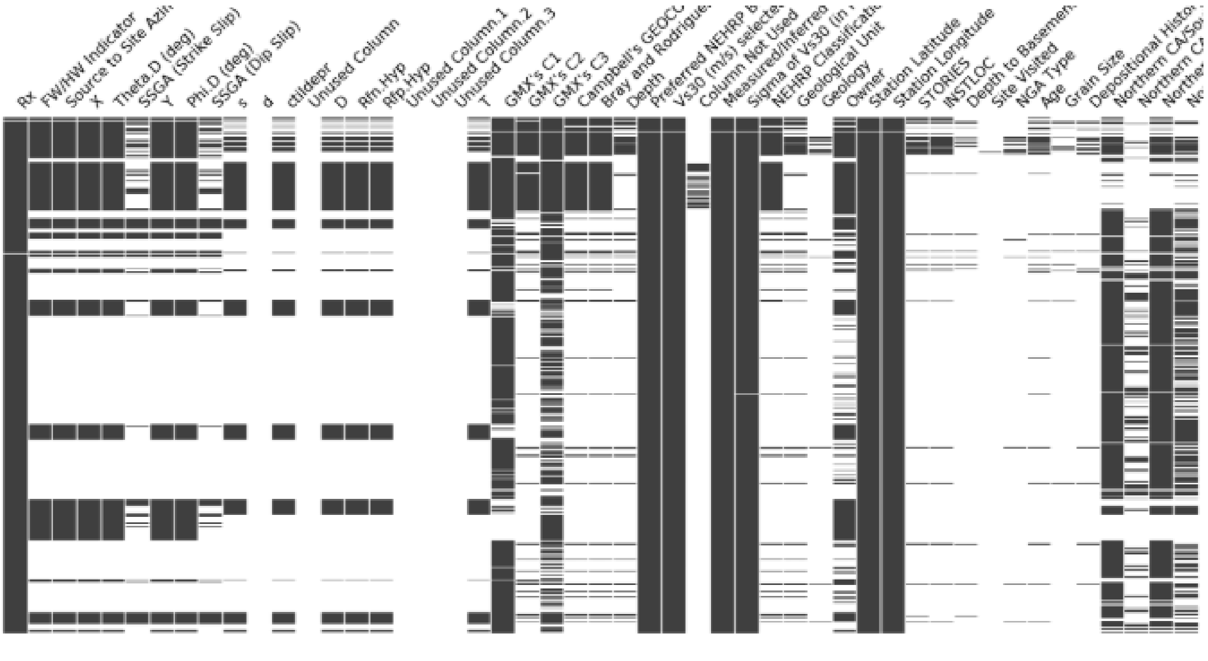

Figure 3: Monotonous Missing Data pattern

4.2 Feature Selection with Boruta

After removing the unnecessary columns and the columns having the large number of missing values there were 165 features left in the data set. The correlation matrix would be difficult to interpret, hence a feature selection technique named ”Boruta” was used as mentioned in the section [3]. As Boruta needs numeric data, the categorical features were converted to numeric by using label encoding. The label encoding was chosen because there consists some information in the dataset which are labelled as per the terms in data dictionary. The label encoding was implemented by using the ”sklearn” library and creating the LabelEncoder object. For imputing the missing data the standard deviation of mean imputed data and median imputed data was checked and as the mean consisted of less standard deviation. The data imputed with mean was chosen. The ”mean()” and ”fillna()” functions were used to implement the imputation of data.

After implementing feature selection, 19 features were chosen by Boruta on 50 iterations based on Z scores. This algorithm was executed by splitting the data into X which consisted of independent variables Y with dependent variable. The ”BorutaPy()” function was used from the boruta library. Below is the output of confirmed list from boruta.

Features	Status
EQID	Confirmed
Earthquake Magnitude	Confirmed
Mo (dyne.cm)	Confirmed
Strike (deg)	Confirmed
Rake Angle (deg)	Confirmed
Mechanism based on Rake Angle	Confirmed
P-plunge (deg)	Confirmed
P-trend (deg)	Confirmed
T-Plunge (deg)	Confirmed
T-trend (deg)	Confirmed
Depth to Top of Fault Rupture Model	Confirmed
Fault Rupture Length for calculation or Ry (km)	Confirmed
Fault Rupture Width (km)	Confirmed
Fault Rupture Area (km^2)	Confirmed
T0.035S	Confirmed
T0.036S	Confirmed
CRjb.4	Confirmed
CRjb.5	Confirmed

Table 1: Features obtained from Boruta

The correlation matrix was plotted by using ”corr()” function in python and ”seaborn” library was used to apply the ”coolwarm” feature where the highly correlated columns can be identified easily. The multicollinearity observed between the features obtained from Boruta were removed. Below is the correlation matrix plotted with keeping couple of important variables although showing a bit of high correlation.

The features were scaled using min max scaling technique for making the values in all features at same scale. This was implemented by using library ”sklearn” and importing ”preprocessing” package from it. The MinMaxScaler() function was used. The data was right skewed at a high level, an effort was taken by taking square root of the data by using the ”numpy” library and ”np.sqrt” package. The data was split into train and test with naming the variables ”Xtrain”, ”Xtest”, ”ytrain” and ”ytest”. This was done by using a library called ”sklearn.model-selection” and importing the ”train-test-split” package into 80:20 ratio.

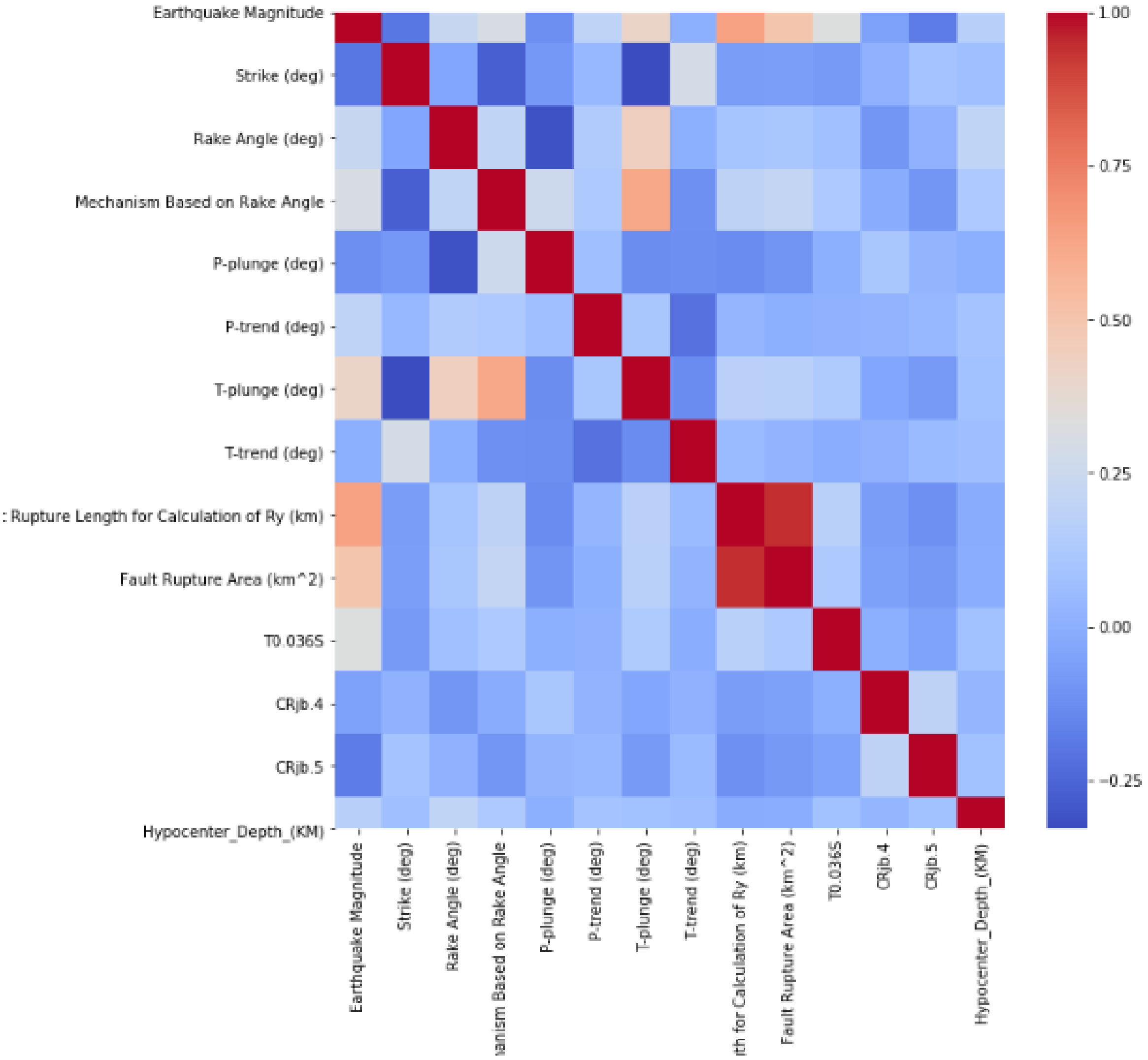

Figure 4: Correlation Matrix

4.3 Implementing with Various Models and Methods

The data was ready to be experimented with the different models as discussed in section [3]. Various experiments were performed by using three models. The first one implemented was random forest regression by using the library named ”sklearn.ensemble” and ”RandomForestRegressor” package was imported. The second experiment was implemented by using light gradient boosting technique. This was implemented by importing the ”lightgbm” package and ”LGBMRegressor()” function was used. The third experiment was conducted using Ridge regression which was implemented using a library named ”sklearn.linear-model” and ”Ridge” package was imported. The function used was Ridge() and alpha parameter was passed in it. These experiments were done using a simple split which is called the holdout method.

The second part of experiments was to observe the models performing under hyperparameter tuning. They were performed using random search and implemented by using the package ”RandomizedSearchCV” and using different functions like ”RandomizedSearchCV()”, ”RandomForestRegressor()”, ”LGBMRegressor()” and ”Ridge()” were used. K-fold was also applied by importing the ”Kfold” package from ”sklearn.model-selection" library and "cross-val-score" library with "cross-val-score()" function and cv is the argument where kfold parameter was passed with 5 folds. The aplha value parameter for tuning was made to choose from "sp.rand()" function by importing the "scipy.stats" library. The parameters passed for tuning the random forest model were n-estimators, max-features, max-depth, min-samples-split and min-samples-leaf. The best parameters upon which the model was executed were listed by a property called "random.best-params-".

4.4 Evaluating the Models

For evaluation purpose, the R squared value, RMSE and MAE were taken into consideration. These were calculated by importing the "metrics" package and functions used to get the values were "mean-absolute-error()" , "r2-score" and the square root was taken of "mean-squared-error()" for RMSE.

$$\text{RMSE} = \sqrt{\tfrac{1}{n}\textstyle\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|$$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2}$$

Where,

y_j = Predicted Hypocenter Depth

$\hat{y}_j$ = Actual Hypocenter Depth

n = Sample size

5 Evaluation

As discussed in section [1], various models were implemented with different experiments like simple split which is hold-out method and hyper parameter tuning. The metrics used for evaluating the models are listed and explained below.

- **Random Forest Regression with Simple Split.**
  The first experiment was performed with splitting the data in the ratio of 80:20. The metrics used for evaluation such as R squared value, RMSE and MAE were obtained as follows,
  R squared value : 0.9632
  RMSE : 1.046

MAE : 0.05

- **Random Forest Regression with Hyper-Parameter Tuning**
  In order to observe the results by considering the best parameters, hyper-parameter tuning by using random search method was performed with k-fold using 5 folds. By using the parameters mentioned in section [4.3] the results obtained were as follows:
  R squared value : 0.99
  RMSE : 0.42
  MAE : 0.046

  The best parameters obtained on which the model was executed upon which the results were obtained were,
  n-estimators : 150, min-samples-split: 2, min-samples-leaf: 1, max-features : log2, max-depth: 50

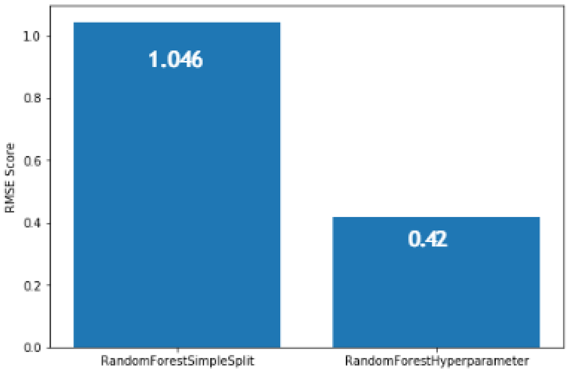

Figure 5: Random Forest RMSE Comparison with Simple split and Hyperparameter tuning

- **Light Gradient Boosting Regression with Simple Split**
  The second experiment performed was with light gradient boosting model by simple slit dividing the data in a ratio of 80:20. The results obtained with used metrics were,
  R squared value : 0.60
  RMSE : 3.62
  MAE : 2.62

- **Light Gradient Boosting Regression with Hyper-Parameter Tuning**
  In order to improve the results by considering the best parameters, as done with the previous model, hyper-parameter tuning by using random search method was performed with k-fold using 5 folds. By using the parameters mentioned in section [4.3] the results obtained were as follows:
  R squared value : 0.99
  RMSE : 0.41
  MAE : 0.087

The best parameters obtained on which the model was executed upon which the results were obtained were,
colsample-bytree : 0.7427013306774357, max-depth: -1, n-estimators: 200, num-leaves : 90, subsample: 0.7103996728090174.

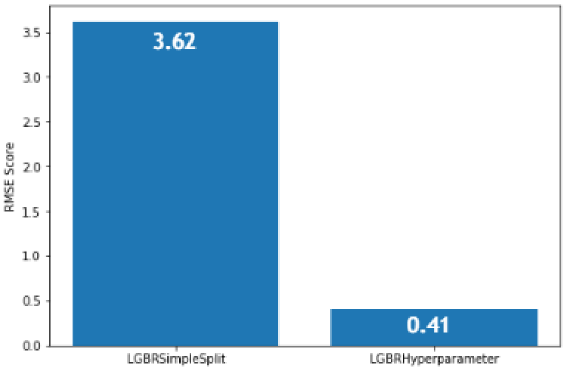

Figure 6: Light Gradient Boosting RMSE Comparison with Simple split and Hyperparameter tuning

- **Ridge Regression with Simple Split.**
In third experiment, the procedure of splitting the data was done in the same way with the ratio of 80:20 and alpha value was taken as 10. The metrics used for evaluating the model were obtained as follows,
R squared value : 0.13
RMSE : 4.24
MAE : 5.43

- **Ridge Regression with Hyper-Parameter Tuning**
In order to improve the results by considering the best parameters, hyper-parameter tuning by using random search method was performed with k-fold using 5 folds. By using the parameters mentioned in section [4.3] the results obtained were as follows:
R squared value : 0.14
RMSE : 5.40
MAE : 4.20
The alpha value obtained by random search was 0.0093

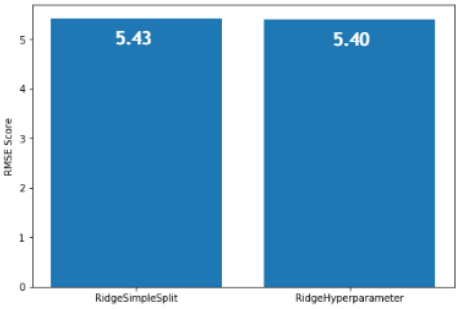

Figure 7: Ridge Regression RMSE Comparison with Simple split and Hyperparameter tuning

6 Discussion

Various experiments were evaluated along with different method as seen in section5. Initially the research was experimented with random forest as the data was not fulfilling the assumptions of linearity and in comparison with the study of Hamze-Ziabari and Bakhshpoori (2018). The RMSE and MAE obtained on the ground motion variables for PGA and PGV were relatively less than one which are obtained in this research for Hypocenter depth. But it is observed that performance of the model was improved drastically after executing with tuned parameters. The R squared value of the model drastically increased to 99 percent, which is the variation in the target variable Hypocenter depth explained by its predictors. The RMSE also decreased from 1.046 to 0.42 when the parameters were tuned using random search on the range of 0.02 to 45, which says that it has a very low error between the actual values and predicted values. In the second experiment light gradient boosting was used as per the behaviour of data and the R squared value obtained with simple split was 60 percent but increased drastically to 99 percent after tuning the parameters. The RMSE also decreased from 3.62 to 0.41 resulting in low residual error. The third experiment was carried out using ridge regression which performed poorly in comparison with Sabermahani and Ashjanas (2019) by giving R squared value of 14 percent after tuning the parameter alpha against 90 percent for predicting the ground motion parameters. The RMSE obtained after tuning the alpha parameter was 5.40 which was very high against the literature.

Comparing the RMSE of the models implemented in this research, it can be said that Random forest performed the best and can be considered to predict the hypocenter depth and plan a deterministic seismic hazard analysis.

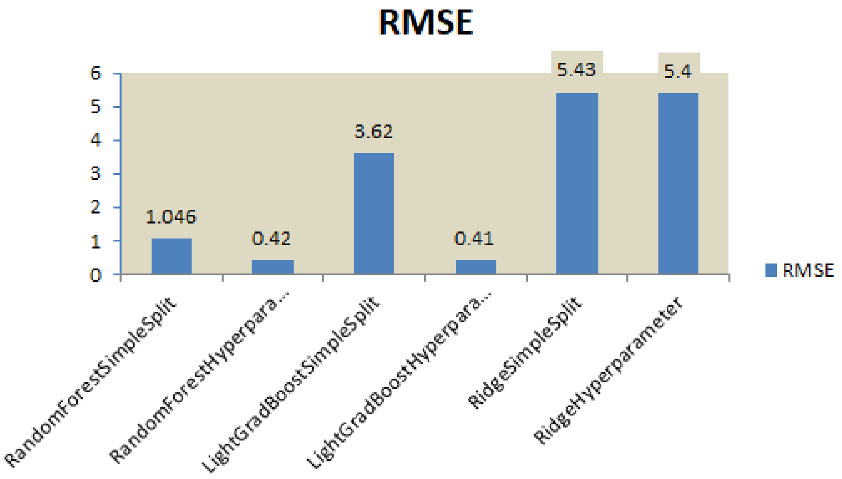

Figure 8: Comparing the models implemented in this research.

7 Conclusion and Future Work

Relating to the question of this research which is framed in section [1], the first objective of this research was to select best features for predicting the hypocenter depth which was done by a technique named Boruta. Before that, a missing data pattern was identified relating with the other variables which turned out to be a monotonous missing data pattern. Regression models were used to predict and find the relation between the ground motion variables and Hypocenter depth, in order to perform a deterministic seismic hazard analysis. The second objective of this research was to improve the efficiency of models by transforming the data, scaling and by tuning the parameters using random search. The results shown by the experiments in this research performed well against the state of the art models Hamze-Ziabari and Bakhshpoori (2018), Sabermahani and Ashjanas (2019) with Random Forest regression and Light Gradient Boosting showing the highest variance. The Ridge regression performed poorly. In the research implemented, the data was poorly recorded with having a high level of noise present which was a limitation observed. Regardless of different techniques applied to prepare the data and train the models to achieve maximum results, some things cannot be overshadowed that a lot of columns were unavailable to use because of the incorrect data as per the data dictionary.

As this research focuses on the the Deterministic Seismic Hazard Analysis (DHSA), in future a probabilistic approach can be taken for performing Probabilistic Seismic Hazard Analysis (PHSA) with corrected time series data by using the Vector Auto Regression model to predict the depth periodically.

8 Acknowledgment

I am thankful for the support and guidance achieved from professor Noel Cosgrave to perform this research.

References

Amatya, D. M., Fialkowski, M. and Bitner, A. (2019). A Daily Water Table Depth Computing Model for Poorly Drained Soils, *Wetlands* **39**(1): 39–54.

Borleanu, F., De Siena, L., Thomas, C., Popa, M. and Radulian, M. (2017). Seismic scattering and absorption mapping from intermediate-depth earthquakes reveals complex tectonic interactions acting in the Vrancea region and surroundings (Romania), *Tectonophysics* **706-707**: 129–142.

Bošnjak, Z., Grljević, O. and Bošnjak, S. (2009). CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data, *Proceedings - 2009 5th International Symposium on Applied Computational Intelligence and Informatics, SACI 2009* (114): 509–514.

Chung, T. W., Iqbal, M. Z., Lee, Y., Yoshimoto, K. and Jeong, J. (2018). Depth-dependent seismicity and crustal heterogeneity in South Korea, *Tectonophysics* **749**(April): 12–20.

Darzi, A., Zolfaghari, M. R., Cauzzi, C. and Fäh, D. (2019). An empirical ground-motion model for horizontal PGV, PGA, and 5% damped elastic response spectra (0.0110 s) in Iran, *Bulletin of the Seismological Society of America* **109**(3): 1041–1057.

Derakhshani, A. and Foruzan, A. H. (2019). Predicting the principal strong ground motion parameters: A deep learning approach, *Applied Soft Computing Journal* **80**: 192–201.
URL: *https://doi.org/10.1016/j.asoc.2019.03.029*

Du, W., Yu, X. and Wang, G. (2019). Prediction equations for the effective number of cycles of ground motions for shallow crustal earthquakes, *Soil Dynamics and Earthquake Engineering* **125**(September 2018): 105759.
URL: *https://doi.org/10.1016/j.soildyn.2019.105759*

Feng, J., Wang, E., Ding, H., Huang, Q. and Chen, X. (2020). Deterministic seismic hazard assessment of coal fractures in underground coal mine: A case study, *Soil Dynamics and Earthquake Engineering* **129**(May 2019): 105921.
URL: *https://doi.org/10.1016/j.soildyn.2019.105921*

Hamze-Ziabari, S. M. and Bakhshpoori, T. (2018). Improving the prediction of ground motion parameters based on an efficient bagging ensemble model of M5 and CART algorithms, *Applied Soft Computing Journal* **68**: 147–161.

Han, X., Liu, J., Mitra, S., Li, X., Srivastava, P., Guzman, S. M. and Chen, X. (2018). Selection of optimal scales for soil depth prediction on headwater hillslopes: A modeling approach, *Catena* **163**(December 2017): 257–275.
URL: *https://doi.org/10.1016/j.catena.2017.12.026*

Huang, Q., Meng, S., He, C. and Dou, Y. (2019). Rapid Urban Land Expansion in Earthquake-Prone Areas of China, *International Journal of Disaster Risk Science* **10**(1): 43–56.
URL: *https://doi.org/10.1007/s13753-018-0207-4*

Kursa, M. B. and Rudnicki, W. R. (2010). Feature selection with the boruta package, *Journal of Statistical Software* **36**(11): 1–13.

Lee, S.-J., Wong, T.-P., Liu, T.-Y., Lin, T.-C. and Chen, C.-T. (2019). Strong ground motion over a large area in northern Taiwan caused by the northward rupture directivity of the 2019 Hualien earthquake, *Journal of Asian Earth Sciences* p. 104095.
URL: *https://doi.org/10.1016/j.jseaes.2019.104095*

Leke, C. A. and Marwala, T. (2019). Missing Data Estimation Using Firefly Algorithm, (i): 73–89.

Lin Thu Aung, Martin, S. S., Wang, Y., Wei, S., Myo Thant, Khaing Nyein Htay, Hla Myo Aung, Tay Zar Kyaw, Soe Min, Kaung Sithu, Tun Naing, Saw Ngwe Khaing, Kyaw Moe Oo, Suresh, G., Chen, W., Phyo Maung Maung and Gahalaut, V. (2019). A comprehensive assessment of ground motions from two 2016 intra-slab earthquakes in Myanmar, *Tectonophysics* **765**(1): 146–160.
URL: *https://doi.org/10.1016/j.tecto.2019.04.016*

Mostafa, S. I., Abdelhafiez, H. E. and Abd el aal, A. e. a. K. (2019). Deterministic scenarios for seismic hazard assessment in Egypt, *Journal of African Earth Sciences* **160**(September).

Podili, B. and Raghukanth, S. T. (2019). Ground motion prediction equations for higher order parameters, *Soil Dynamics and Earthquake Engineering* **118**(November 2018): 98–110.
URL: *https://doi.org/10.1016/j.soildyn.2018.11.027*

Raghucharan, M. C., Somala, S. N. and Rodina, S. (2019). Seismic attenuation model using artificial neural networks, *Soil Dynamics and Earthquake Engineering* **126**(August): 105828.
URL: *https://doi.org/10.1016/j.soildyn.2019.105828*

Sabermahani, S. and Ashjanas, P. (2019). Geodesy and Geodynamics Sensitivity analysis of ground motion prediction equation using next generation attenuation dataset, *Geodesy and Geodynamics* (November): 1–6.
URL: *https://doi.org/10.1016/j.geog.2019.09.004*

Thomas, S., Pillai, G. N., Pal, K. and Jagtap, P. (2016). Prediction of ground motion parameters using randomized ANFIS (RANFIS), *Applied Soft Computing Journal* **40**: 624–634.
URL: *http://dx.doi.org/10.1016/j.asoc.2015.12.013*

Vaez Shoushtari, A., Adnan, A. B. and Zare, M. (2018). Ground motion prediction equations for distant subduction interface earthquakes based on empirical data in the Malay Peninsula and Japan, *Soil Dynamics and Earthquake Engineering* **109**(June 2017): 339–353.

Xu, Y., Wang, J., Wu, Y.-M. and Kuo-Chen, H. (2019). Prediction models and seismic hazard assessment: A case study from Taiwan, *Soil Dynamics and Earthquake Engineering* **122**(March): 94–106.
URL: *https://doi.org/10.1016/j.soildyn.2019.03.038*

Yang, X. D., Qie, Y. D., Teng, D. X., Ali, A., Xu, Y., Bolan, N., Liu, W. G., Lv, G. H., Ma, L. G., Yang, S. T. and Zibibula, S. (2019). *Prediction of groundwater depth in an arid region based on maximum tree height*, Vol. 574, Elsevier B.V.
URL: *https://doi.org/10.1016/j.jhydrol.2019.04.022*

Yu, C., Hauksson, E., Zhan, Z., Cochran, E. S. and Helmberger, D. V. (2019). Depth Determination of the 2010 El Mayor-Cucapah Earthquake Sequence (M 4.0), *Journal of Geophysical Research: Solid Earth* **124**(7): 6801–6814.

yuan LU, Y., LIU, F., guo ZHAO, Y., dong SONG, X. and lin ZHANG, G. (2019). An integrated method of selecting environmental covariates for predictive soil depth mapping, *Journal of Integrative Agriculture* **18**(2): 301–315.
URL: *http://dx.doi.org/10.1016/S2095-3119(18)61936-7*

Zanini, M. A., Hofer, L. and Faleschini, F. (2019). Reversible ground motion-to-intensity conversion equations based on the EMS-98 scale, *Engineering Structures* **180**(November 2018): 310–320.
URL: *https://doi.org/10.1016/j.engstruct.2018.11.032*

Zhang, J., Zhu, Y., Zhang, X., Ye, M. and Yang, J. (2018). Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas, *Journal of Hydrology* **561**: 918–929.
URL: *https://doi.org/10.1016/j.jhydrol.2018.04.065*