A picture containing shape

Description automatically generated

**Module Name** Data Analytics

**Module Lecturer** David Leonard

**Date Submitted** 6th December 2022.

|  |  |  |  |
| --- | --- | --- | --- |
| **Student Name** | **Student Number** | **% Input** | **Signature** |
| Abusomwan  Amadevboho  Idowu Omoniyi James | K00266081  K00266211 |  | AMADE      JAMES |

1. **INTRODUCTION**

Earthquakes are an old natural phenomenon that has existed in the world and has continued to become a growing problem. Earthquakes are tremendous vibrations which can span over an entire geographical area caused from sudden movements of the earth’s crust. Earthquakes happen due to the movement of tectonic plates. As the world has moved into the technological age, it has found new ways to monitor tectonic activities beneath the earth surface, and even advanced to been able to predict future possible earthquakes, its “magnitude”, “depth”, “date” and “location”.

This data we will be working with was gotten from the National Earthquake Information Centre (NEIC), this data deals with the information gotten form various Earthquakes activities from around the world which deals with its location, size etc and distributes this important information to agencies, facilities, scientists and the general public.

The data includes a record of the “date”, “time”, “location”, “depth”, “magnitude” and source of every earthquake with a reported magnitude 5.5 or higher since 1965. This report will be investigating the significance of the earthquakes provided by the data published by Kaggle grandmaster Gabriel Preda (2019) which he used the dataset to illustrate earthquakes using bubbly. Nevertheless, we would also work on the Earthquakes, 1965-2016-EDA and TS predication done by Pierre Jeanne, parsing date and earthquake visualization by Pierre-Louis Danieau for different years.

**2.0 PROBLEM DEFINITION**

Upon first glance of the data, which consist of 21 attributes and 23412 rows. The data included some missing values. Hence, these was addressed by using the average values of these attributes to replace the missing ones. It was also noticeably discovered due to the high volume of rows; our point map graph couldn’t visualize trends across the various locations and dates. So, based on this data provided we couldn’t work with the chosen attributes needed to make suitable predictions and to generate a model to get a good performance.

Furthermore, making predictions using a linear regression model, which consist of a dependent and independent variable, rapid miner provided an operator “select attributes to get the relevant column useful for the prediction. this attribute will be used to build a model that predicts the likelihood of occurrences of earthquakes in different locations over the years.

A linear regression will be used to make predications on the depth having a time series data. However, our data was also built on a classification using a cross validation to test the accuracy of the model.

**3.0 Data Exploration**

Based on the working data which needed to undergo cleaning, replacing all missing values, selecting the necessary attributes required to provide accurate predictions, selecting the required attribute to create a label role and splitting the data into the required testing and training data. This data consist of different attributes such as “magnitude”, “longitude”, “latitude”,” depth”,” time”,”types”,”date”, ”source”, “location”, “status” etc.

**3.1 Magnitude Types**

Chart, histogram

Description automatically generated

**Fig 1.0 *Magnitude Histogram Chart***

As shown in the fig 1, this is a chart showing magnitude and its frequency. It clearly shows the highest frequency coming in with a magnitude of 5.65. It also shows the different magnitude types in various colours with the ML representing Local Magnitude, MW representing Moment Magnitude, MS representing Surface-wave magnitude, and MB representing Body-wave Magnitude.

This chart shows a trend of an inversely proportional relationship between the frequency and the magnitude and also the magnitude types.

**4.0 Identification of Data Details from Previous Steps**

Previous work was done using this same data from the Kaggle website. The context of the project would refer using work done by Kaggle grandmaster Gabriel Preda (2019) which he used the dataset to illustrate earthquakes using bubbly. Also, we would refer to the code, which was generated by Pierre Jeanne, which he talked about the Earthquakes, 1965-2016 – EDA and TS prediction.

Chart, scatter chart

Description automatically generated

**Fig 2.0 *Earthquake Magnitude Prediction Map***

This map is being compared with the work done by Kaggle expert Pierre-Louis Danieal from his work which he used the earthquake visualization using plotly. Here he was able to address where the earthquakes are located, what are the main characteristics (magnitude, recurrence, status etc). Nevertheless, Using Rapid Miner, we used the point map to make visualization of the likelihood of occurrence of earthquakes across various locations. Using the Rapid Miner point map features which we were able to go get the magnitude across different locations in respect of our latitude and longitude.

Chart, scatter chart

Description automatically generated

**Fig 3.0 *Earthquake Position (longitude, latitude, and magnitude)***

The bubble plot is showing the data in (Longitude, Latitude) coordinates, animated, each animation frame showing one year. Each bubble shows a major earthquake. The size of the bubbles is correlated with the amplitude (calculated from the magnitude). The bubble displays the magnitude (on Richter scale). Kaggle Grand master Gabriel Preda wrote a report investigating the significance of the earthquakes provided by the dataset published by Kaggle website in 2019 he used the dataset to illustrate earthquakes using bubbly. Nevertheless, we were also able to replicate the same visualization using rapid miner. This process was done with the bubble plot feature whereby we placed the latitude on the Y axis and longitude on the X axis in respect of the magnitude, which is displayed in colour bubble, hence, the blue magnitude shows a lower dense area while the red magnitude shows a high dense area.

**5.0 Descriptive Analytics**

Table 1a*:  Linear regression of summary statistics*

Graphical user interface

Description automatically generated

Y= B0 + B1X1 + B2X2+ B3X3

Y= 37.901 + (- 0.291). Latitude +(-0.075). Longitude + (6.092). Magnitude

using simple linear regression to make prediction we used the model

Where Y= “depth”, B0 = ‘Intercept”, X1 = Latitude, X2 = Longitude, X3= Magnitude

With this using the dataset we used the explanatory or dependent variables which include magnitude, longitude, latitude as factors which was used in predicting the depth.

Where, variable Y is the dependent variable, and the X variables are known as the explanatory variables, predictors, covariates, or regressors variable.

Given the illustration in Table 1a, Table 1b and Table 1c can be used for forecasting or prediction in the context of time series analysis, we could see interestingly that “depth” earthquake is likely to happen using the multiple linear regression increases with time; therefore, we can conclusively say that the depth will have an impact of earthquake happening over time. Predictions were made possible using different operators that can be seen in the attached process file. The significant earthquake dataset was imported before selecting a subset of five (3) variables from the dataset. The variables include “latitude”, “longitude” and “magnitude” which is the explanatory variables, predictors, covariates, or regressors variable and thereby making the “depth” as label to make prediction of the likelihood of earthquake to occur in respect to the depth. 70% of the dataset was used for training in our multiple linear regression model and the remaining 30% of the dataset was used for testing before Importing of linear regression model using the rapid miner.

Secondly, we noticed longitude and latitude has a negative or inverse relationship with the depth and on the other hand the magnitude has a positive relationship with the depth, therefore the higher or increase in the depth level it would result to an increase in the magnitude of earthquake and vice versa.  Furthermore, having a negative coefficient shows it is inversely related.

Above all, the p value from our findings, having a unit of 0 means it is a good predictor (less than 0.5).

**6.0 Evaluation Strategy and Performance Measure**

Our evaluation and performance strategy for this project work was Regression and Classification, respectively. As previously performed, we were able to predict our Depth values and with these values, we were able to get our confusion matrix.

**Confusion Matrix:**This confusion matrix gave out a tabulation on how well the model performed on a specific data (in this case between the year 1965-1975). In this case the labels are “Hi” and “Lo”, these labels were created after creating a new attribute called ‘Class’ and the from the average predicted Depth table, we got the median value of 33 and then created a classification of ‘HI’ and ‘Lo’ from this.

The True Predicted High value gotten was 960 and the True Predicted Lo is 1559. The model also predicted a True high but was wrong 361 times and predicted Lo but was 868 times. If all predicted values are summed up together, we will get the total sum of rows used for this model (3748).

The accuracy indicated in the top left corner of the table indicates a helpful single measure of performance, as it is the correct predictions divided by all the predictions, giving us the accuracy of 67.21%

Accuracy = (TP + TN)

(TP+FP+TN+FN)

This table also shows the Standard Deviation of 2.58%, the class precision shows the percentage of how correct the model predicted the outcome and the class recall shows how well the True ‘Hi’ and True ‘Lo’ was predicted.

Table

Description automatically generated with medium confidence

**ROC Curve: Receiver Operator Characteristics** is a means of distinguish between classes across different thresholds. This curve is gotten by rank ordering the confidences (descending order) giving of the curved nature of the graph with x-axis being the ‘False Positive Rate’ and y-axis being the ‘True Positive Rate’.  The Area under the Curve (AUC) which gets an average measure of performance difference between two classes in this case is given as 0.750.

Chart, line chart

Description automatically generated

**CONCLUSION**   
Based on the data used, we have been able to learn more about earthquakes, we have also been able to work with and determine new values from this dataset like the Histogram chart explaining the Magnitude frequency, the earth location prediction on the world map and give a more precise location with its Longitude and Latitude.

After wrangling this dataset, we went further to using Linear Regression to create a Prediction Attribute where a Depth column was predicted from the year 1965-1975. Moving with this depth prediction, we further analysed and classified the data to get a performance accuracy and an AUC value, and a confusion matrix with an accuracy value

Appendix

Table 1b*:  Depth Prediction*

Table

Description automatically generated

Table 1c*:  root mean squared error*

Graphical user interface, text, Word

Description automatically generated

Table 1d*:  Mininum,maxmum and average prediction of depth*

A screenshot of a computer

Description automatically generatedText

Description automatically generated

**Referencing list**

En.wikipedia.org. 2021. *Earthquake - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Earthquake> [Accessed 29 December 2021].

Kaggle.com. 2021. *Significant Earthquakes, 1965-2016*. [online] Available at: <https://www.kaggle.com/usgs/earthquake-database> [Accessed 31 December 2021].

Usgs.gov. 2021. *National Earthquake Information Center (NEIC) | U.S. Geological Survey*. [online] Available at: <https://www.usgs.gov/programs/earthquake-hazards/national-earthquake-information-center-neic> [Accessed 31 December 2021].

Kaggle.com. 2022. *Significant Earthquakes, 1965-2016*. [online] Available at: <https://www.kaggle.com/usgs/earthquake-database> [Accessed 6 January 2022].

Kaggle.com. 2022. *Run Data Science & Machine Learning Code Online | Kaggle*. [online] Available at: <https://www.kaggle.com/code> [Accessed 6 January 2022].

Kaggle.com. 2022. *Earthquakes animation using bubbly*. [online] Available at: <https://www.kaggle.com/gpreda/earthquakes-animation-using-bubbly> [Accessed 6 January 2022].

Kaggle.com. 2022. *💥 Earthquakes visualization - Plotly Intro 👁️*. [online] Available at: <https://www.kaggle.com/pierrelouisdanieau/earthquakes-visualization-plotly-intro> [Accessed 6 January 2022].