Study of Various Natural Calamities

Shaikh Mohammed Nauman
Student id: 202218010
202218010@daiict.ac.in

Asish Kumar Sahoo
Student id: 202218022
202218022@daiict.ac.in

Swayista T. Ahmed
Student id: 202218035
202218035@daiict.ac.in

***Abstract*—This project aims to study and understand natural calamities. our target is to develop a machine learning model which accurately predicts the occurrence of the natural disaster. This involves the study and analyzing the past data of the calamity occurred to identify the pattern and trend that may indicate increase in the disaster. Project may use various techniques such as data preprocessing, feature selection, feature engineering and ensemble learning to improve accuracy of the model. Our ultimate objective is to provide effective and reliable means of predicting such calamities that can potentially save lives.**

I. Introduction

Due to the rise in natural disasters in this decade, people started to study these Disasters with the hope of getting a pre-hint about any natural disaster Which would help save lives and would help the country reducing much economic loss for the country .

Earthquake means the shaking of the Earth's surface. It is a sudden trembling of the surface of the Earth. Earthquakes are the result of sudden movement along faults within the Earth. Some Earthquakes are weak in nature and probably go unnoticed.

The effects of earthquakes include ground shaking, surface faulting, ground failure, and less commonly, tsunamis. Earthquakes can range in severity from minor tremors to catastrophic events that cause extensive damage and loss of life. The severity of an earthquake is measured on the Richter scale, which ranges from 0 to 10, with each number representing a tenfold increase in seismic activity.

Earthquake forecasting is one of the most significant issues in Earth science because of its devastating consequences. Our current project is about earthquake prediction which focuses on three key points: when the disaster will occur, where it will occur, and how big it will be.

Phase 1- We are applying various data analysis techniques to gain insights about our data

Phase 2- We plan to implement some existing papers and heuristics to solve the problems

Motivation- earthquake, any sudden shaking of the ground caused by the passage of waves through Earth's rocks. Seismic waves are produced when some form of energy stored in Earth's crust is suddenly released, usually when masses of rock straining against one another suddenly fracture and "slip." Earthquakes occur most often along geologic faults, narrow zones where rock masses move in relation to one another. The world's major fault lines are located at the fringes of the huge tectonic plates that make up Earth's crust.

II. Exploratory Data Analysis

(1) Plotting graph of null values
(2) 5 value summary
(3) Total number of earthquake for particular year(UDF)
(4) Number of earthquake happened for desired magnitude(UDF)
(5) Plotting position of earthquake on map based on longitude and latitude

III. Data Description

The data set we used here is collected from 'https://earthquake.usgs.gov/earthquakes/search/' consist 21 features of 32000+ rows for Japan from year 1900 (January) - 2023 (April)

No.	Name	Description of feature
1	time	Time when the earthquake occurred in the yyyy-mm-dd HH:MM:SS format
2	latitude	Latitude of the place
3	longitude	Longitude of the place
4	depth	Depth of the earthquake in kilometres
5	nst	number of Seismic stations, which is used to determine the earthquake location
6	gap	Seismic Gap in degree (0 to 180 degrees)
7	dmin	Horizontal distance between epicentre and nearest station in degrees
8	rms	The root mean square of the travel time residual.
9	Net	Data contributor ID
10	id	Database id of record
11	updated	Most recently updated time of earthquake
12	Place	Description of Geographical position
13	Type	Type of Seismic Event ("Earthquake", "Quarry")
14	location	Name of the network that reported the location of the earthquake
15	magSource	Name of the network that reported the magnitude of the earthquake
16	Horizontal error	Horizontal error of the location in kilometres
17	Depth error	Depth error of the location in kilometres
18	Mag error	Standard error of the magnitude
19	Mag nst	The earthquake magnitude, which is determined using the number of Seismic stations
20	status	Indicates that the earthquake was reviewed by humans
21	class	Target Variable (Fatal Earthquake, Moderate Earthquake and Mild Earthquake)

IV. Data Pre-processing

Firstly we divided time column in two parts first consisting the time (hh:mm:ss) format, and other column was for Date part which was in (DD:MM:YYYY) format. Then we further divided date column in 3 different columns namely Date,Month and Year.

Adding a new column name 'earthquake type', which will have values 'Fatal','Moderate' and 'Mild' according to magnitude.

- 'Fatal' if $magnitude \geq 6$
- 'Moderate' if $6 > magnitude \geq 4$
- 'Mild' if $magnitude < 4$

Then converting that column to numerical data I,e Label Encoding. For fatal = 2, Moderate = 1, Mild = 0.

V. K means Clustering:

K-means is a type of clustering algorithm that groups together similar data points based on their characteristics. It randomly selects points as the centroid of clusters, where a number of clusters are given by the user beforehand. It checks and assigns the point to the cluster nearest to the point and updates the cluster's centroid. This loop repeats for all points.

Inertia: The sum of squared distances between each data point and its centroid is called inertia. The lower the inertia is, the better the algorithm has performed. Result: **inertia** = 38647240.7906963 Conclusion: After the analysis of the inertia of the data given, it is determined that the application of the clustering algorithm may not yield a satisfactory result.

Conclusion: After the analysis of inertia of the data given, it is determined that the application of clustering algorithm may not yield satisfactory result.

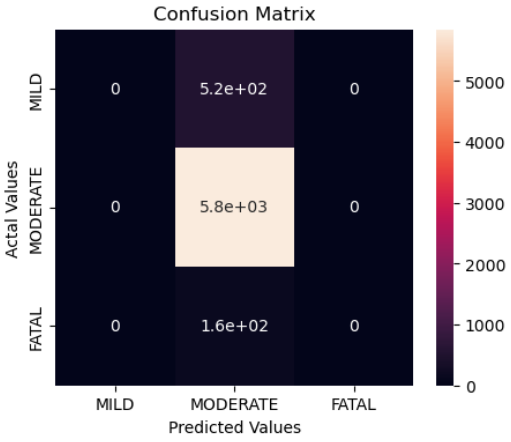
VI. Logistic Regression

Logistic regression is a statistical method used to analyse and model the relationship between binary (0 or 1) dependent variables and one or more independent variables. If a dependent variable has more than two outcomes then, the model is tuned in such a way that the sum of the probability of occurring of every event is equal to 1.

Result:
Accuracy: 89.66733098267669 %
Precision: 29.889110327558893 %
Recall: 33.33333333333 %
F1 score: 31.517404892768617 %

A. *Explaination:*

From the above graph we can see that every value that has been predicted by our model is '1'. It is happening because classes are unbalanced. It is also showing that our trained



model is highly biased.

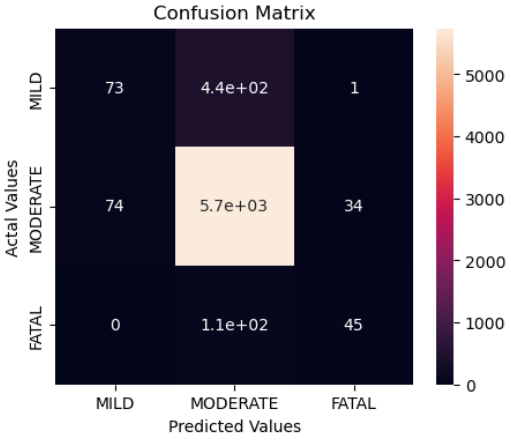
Even when Accuracy is around 90% we are unable to make any conclusion from it.

VII. Random Forest Classifier

Random Forest Classifier: Random forest classifier is a machine learning algorithm that predicts the class of decision variables based on features. It works by combining multiple decision trees, where each tree is trained with a random subset of input features. Each decision tree used 'If-then' rules to predict the outcome.

The Random Forest algorithm works better when data is in higher dimensions.

Result:
Accuracy: 89.82063467729571 %
Precision: 65.69824774467178 %
Recall: 47.03078281895173 %
F1 score: 51.54381459926376 %



A. *Explaination:*

Above graph is showing us that only 45 of the fatal earthquakes are predicted successfully. Again class

unbalancedness causes the problem of biased results.

Random forest is giving better results than logistic regression in term of predicting the fatal earthquakes, but there is need to improve the machine learning model

VIII. Problem occurring in applied algorithms

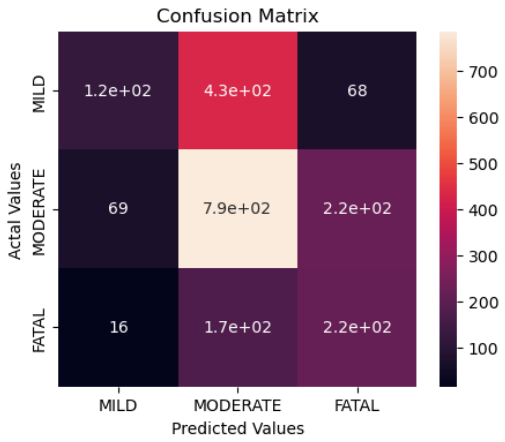
The data we have is unbalanced. Here nearly 90% of values in decision variable is 1, 8% values are 2 and only 2% values are 0. Result in higher accuracy but less precision. Outcomes of the models are highly biased.

A. *Solution Applied:*

Making data balanced, by oversampling of the rows which have less entries. initially the data was of rows 32611 and 80% of the data was used as train randomly and 20% was used as test.In 32611 rows, 29078 rows are 'Moderate', 2624 are 'Mild' and only 909 are 'Fatal' earthquakes .After applying method of oversampling we have created new dataset, in which train set has 60000 rows. 28000 are 'Moderate', 24000 are 'Fatal' and 28000 are 'Mild' earthquakes. And applied the algorithms again.

IX. Applying Oversampling method

Improved Logistic Classification Result:

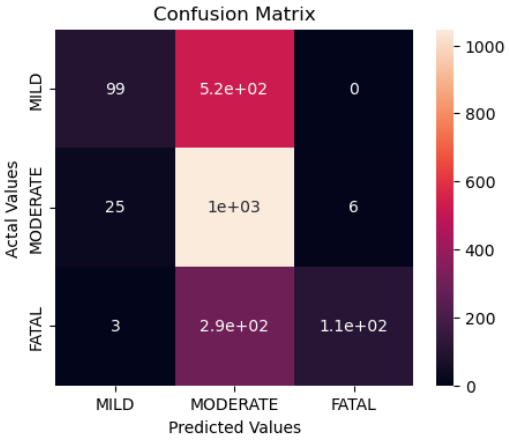


Accuracy: 52.34486025580294 %
Precision: 50.17179878094954 %
Recall: 51.76314303540989 %
F1 score: 50.61209988486222 %

A. *Explaination:*

Above graph is showing improved results of logistic regression. We can see the accuracy decreasing, but there is significant increase in precision. In other words we can say that our model is predicting fatal earthquakes better than before.

Improved Random forest Classification



Result:
Accuracy: 59.59261013737566 %
Precision: 76.32577830054747 %
Recall: 46.79118398774354 %
F1 score: 46.665769365053436 %

B. *Explaination:*

Above graph is showing improved results of Random Forest Classificatiob. We can see the accuracy decreasing, but there is significant increase in precision. It's accuracy and precision is better than logistic regression. But in terms of predicting Fatal earthquake, Logistic regression works better.

X. Time Series forcasting

Time series forecasting occurs when you make scientific predictions based on historical time-stamped data. It involves building models through historical analysis and using them to make observations and drive future strategic decision-making. An important distinction in forecasting is that at the time of the work, the future outcome is completely unavailable and can only be estimated through careful analysis and evidence-based priors.

There are different methods to forecast the values. while Forecasting time series values, 3 important terms need to be taken care of and the main task of time series forecasting is to forecast these three terms.

A. *Seasonality*

Seasonality is a simple term that means while predicting a time series data there are some months in a particular domain where the output value is at a peak as compared to other months.

B. *Trend*

The trend is also one of the important factors which describe that there is certainly an increasing or decreasing trend time

series, which actually means the value of organization or sales over a period of time and seasonality is increasing or decreasing.

C. *Unexpected Events*

Unexpected events mean some dynamic changes occur in an organization, or in the market which cannot be captured. for example a current pandemic we are suffering from, and if you observe the Sensex or nifty chart there is a huge decrease in stock price which is an unexpected event that occurs in the surrounding. Methods and algorithms are using which we can capture seasonality and trend But the unexpected event occurs dynamically so capturing this becomes very difficult.

XI. PROPHET

Prophet is a forecasting tool developed by Facebook that is used to predict future values of a time series data. It is designed to handle large datasets with several seasonality factors, such as daily, weekly, and yearly patterns, as well as holiday effects.

Prophet uses a generalized additive model (GAM) that combines various time-series components, including trend, seasonality, and holiday effects, to model the data. The model is then trained on historical data to capture the patterns and variations in the data. Once the model is trained, it can be used to make future predictions with a high level of accuracy.

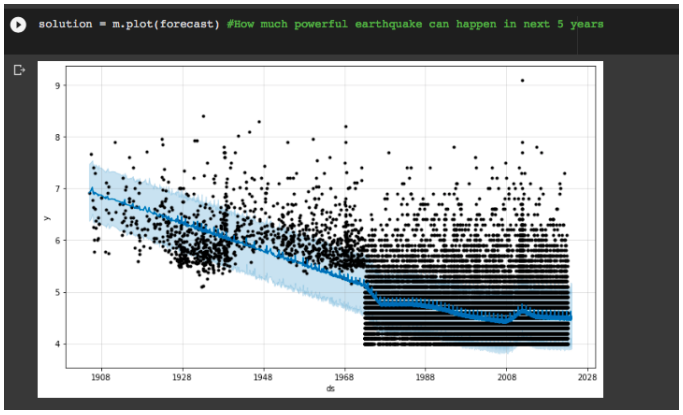
Prophet also provides an interactive tool that allows users to easily visualize the forecasts and adjust the parameters to optimize the model's performance. This makes it a useful tool for data analysts and business users who need to make accurate predictions for planning and decision-making purposes.

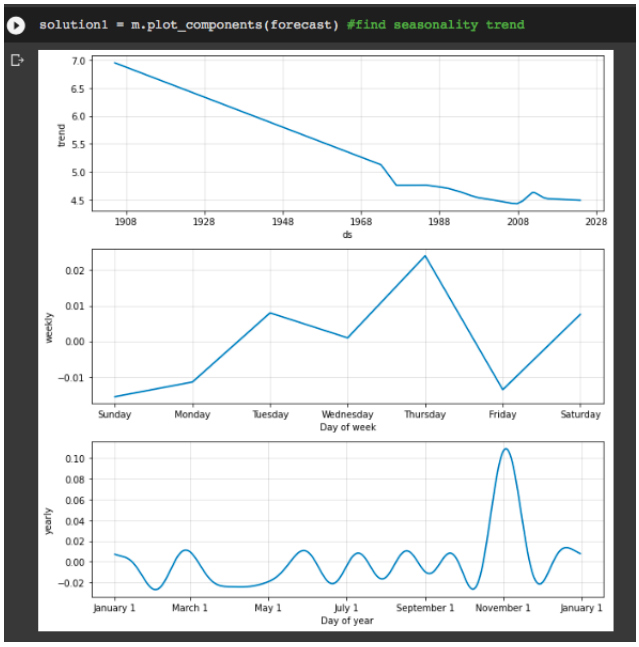
A. *How Prophet works?*

The procedure makes use of a decomposable time series model with three main model components: trend, seasonality, and holidays. Similar to a generalized additive model (GAM), with time as a regressor, Prophet fits several linear and non-linear functions of time as components. In its simplest form;

$$y(t) = g(t) + s(t) + h(t) + e(t)$$

Prophet is essentially "framing the forecasting problem as a curve-fitting exercise" rather than looking explicitly at the time-based dependence of each observation. Ultimately, Prophet was engineered to help analysts with a variety of backgrounds produce more forecasts with less time invested in doing so. This was achieved by sticking to a relatively plain model.





XII. CONCLUSION:

It is determined that after applying oversampling, precision has increased and algorithms can now predict Fatal earthquakes better than before.
It is also determined that Logistic classifier is predicting the Fatal earthquake better than Random Forest

XIII. FUTURE WORK:

Till now we have done some classification algorithm for prediction. In next phase we will move to time series analysis. We may apply time series algorithms for predictions like ARIMA, Vector Auto regression, Moving Average, LSTM, and classification algorithms like Support Vector Classifier, Naive Bayes.

XIV. REFERENCES

REFERENCES

[1] Dataset: "https://earthquake.usgs.gov/earthquakes/search"
[2] paper we referred: "https://mdpi-res.com/data/data-clustermdpi/dms/sustainability/sustainability-13-00971"
[3] K means clustering: "https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1"
[4] Logistic Classification: "https://www.analyticsvidhya.com/blog/2021/05/logistic-regression-supervised-learning-algorithm-for-classification/"
[5] "https://towardsdatascience.com/the-perfect-recipe-for-classification-using-logistic-regression-f8648e267592"
[6] "https://towardsdatascience.com/understanding-random-forest-58381e0602d2"
[7] "http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html"
[8] "https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575"
[9] "https://www.w3schools.com/python/python$_{m}l_{c}onfusion_{m}atrix.asp$"
[10] "https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/"
[11] "https://ieeexplore.ieee.org/abstract/document/10073687"
[12] "https://www.researchgate.net/publication/347119111$_{T}ime_{S}eries_{A}nalysis_{o}n_{E}arthquakes_{U}sing_{E}DA_{a}nd_{M}achine_{L}earning$"
[13] "https://www.mdpi.com"