Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Seismic Activity Analysis in California: Patterns, Trends, and Predictive Modeling

Pravakar Debnath[1]✉, Mitu Karmakar[2], MD Tushar Khan[3], MD Azam Khan[4], Abdullah Al Sayeed[5], Arifur Rahman[6], and Md Fakhrul Islam Sumon[7]

[1]School of Business, Westcliff University Irvine, California, USA
[2476]School of Business, International American University, Los Angeles, California, USA
[3]Masters of Science in Business Analytics, Trine University
[5]Masters of Business Administration in Project Management, Central Michigan University

Corresponding Author: Pravakar Debnath, **E-mail**: p.debnath.259@westcliff.edu

| ABSTRACT

For decades, California has been considered one of the most seismically active areas in the United States, if not the most, due to the frequent occurrence of earthquakes from tectonic activity, notably along the San Andreas Fault. Understanding seismic activity is a matter not only of safety but also of major importance for urban planning. In the recent past, Machine learning algorithms (ML) have emerged as a promising invention for advancing earthquake prediction by facilitating the pinpointing of patterns in large datasets that may be hard to detect through traditional techniques. The primary objective of this research project is to assess historical seismic data to identify trends and patterns in California's seismic activity. Equally important, the goals of this research focused on developing predictive models that can provide insight into the possibility of seismic events in the future. To assess seismic activity in California, data were gathered from a credible and reputable source, most notably, the United States Geological Survey (USGS) Earthquake Database, which provides detailed records of earthquakes spanning over six decades. This dataset included all recorded seismic events in California, capturing the key details about the place of occurrence in latitude and longitude, the magnitude of the seismic event, the depth at which it happened, and the time when the seismic activity took place. To devise and curate a reliable earthquake prediction algorithm for California, distinct machine learning models were considered, comprising, Random Forest, XG-Boost, and Logistic Regression. Overall, the Random Forest algorithm exemplified high accuracy. Optimizing this algorithm will lead to more reliable predictions, potentially aiding disaster preparedness and risk mitigation in California's earthquake-prone areas. The findings from seismic activity analysis have deep implications for urban planning and disaster preparedness in California. Knowledge of the pattern, place, and magnitude of earthquakes over time will help urban planners and policymakers structure communities that can remain resilient during such calamities. For instance, higher earthquake frequencies would automatically call for increased stringency in building codes, especially along fault lines, to ensure that houses situated on such lines are designed to withstand major seismic activity. This operation may even limit or reallocate urban development out of high-risk zones into less risky zones.

| KEYWORDS

Seismic activity, earthquake prediction, California, earthquake trends, urban planning, disaster preparedness, machine learning

| ARTICLE INFORMATION

ACCEPTED: 20 October 2024 **PUBLISHED:** 06 November 2024 **DOI:** 10.32996/jcsts.2024.6.5.5

1. Introduction

Motivation and Background

Perez-Oregon et al. (2021), state that for decades, California has been considered one of the most seismically active areas in the United States, if not the most, due to the frequent occurrence of earthquakes from tectonic activity, notably along the San

Andreas Fault. The number of fault lines that run through the state maintains the constant rate of seismic events that range from small tremors in nature to very significant earthquakes, ones that affect a wide scope. The converse of this, however, has been the realization that there is a dire need to revisit and enhance the building codes to understand seismic patterns guiding future preparedness. Al banna et al. (2020), argue that while urbanization does take place in earthquake-prone areas with an increase in population, the need to understand seismic trends has increased to an extent where it either needs to be kept at bay or minimize damage to key infrastructures for public safety.

Muhammad et al. (2021), posit that understanding seismic activity is a matter not only of safety but also of major importance for urban planning. Indeed, the design of earthquake-resistant structures and infrastructure systems requires city planners and policymakers alike to have a sound basis on which to conduct risk assessments. Equally important, Asim et al. (2020), assert that the reduced seismicity knowledge cannot be divided from any disaster preparedness program as it will ensure emergency response plans are timely and effective in reducing any life, human, and economic loss to the minimum extent possible. This research project explores how seismic prediction techniques are devised and the growing possibility of integrating forecasting into urban planning and policy development using data-driven decision-making and proactive risk management strategies.

Research Objectives

The primary objective of this research project is to assess historical seismic data to identify trends and patterns in California's seismic activity. By assessing these data, we aim to depict the frequency, magnitude, and spatial distribution of earthquakes within the state, contributing to a clearer comprehension of the underlying trends that define California's seismic landscape. The goals of this research can be extended toward developing predictive models that can provide insight into the possibility of seismic events in the future. In this regard, applying machine learning and statistical techniques could develop a model that could predict the probability and possible magnitude of future earthquakes in various parts of the state. It finally discusses the implications of these findings for urban planning and disaster preparedness, studying the ways that predictive modeling can enhance existing risk mitigation strategies and aid in making better-informed long-term decisions about how to create resilient communities.

Significance of the Study

This research project holds substantial value for seismic risk management and disaster response planning in California. Our research contributes to the management and reduction of seismic risks by allowing a deeper understanding of historical seismic activity and its possible future manifestations. Predictive model development will enhance preparedness for earthquakes, thus enabling communities and response agencies to better one day anticipate and mitigate the impact of future earthquakes. Insights developed from such an analysis provide the basis for policies and construction practices protecting human lives and infrastructure. Finally, it is expected that the findings from this study will contribute toward significant disaster response frameworks that support evidence-based approaches toward the enhancement of resilience to seismic events with implications for guiding urban planners, engineers, and government officials on the necessary insights to make appropriate, proactive decisions.

2. Literature Review

Earthquake Activity in California

Rundel et al. (2021), indicated that California is also one of the most seismically active places on Earth because of its peculiar tectonic setting-it lies over a complex network of tectonic plates, notably the Pacific and North American plates, which meet at the San Andreas Fault, one of the more famous fault lines over 800 miles across the state. These include the Hayward and San Jacinto faults as subsidiary faults running together with the overall system of faulting, which explains the frequency of earthquakes in California. Laccarino et al. (2020), argue that the integrated interaction of the tectonic plates creates ongoing stress, and then strain, in the Earth's crust, which manifests itself either as fault slips or earthquakes. That would mean the state experiences several thousands of seismic events every year, but only a few are strong enough to be felt by its residents.

Rundel et al. (2021), added that California has had its fair share of major earthquakes throughout history, events that seem to have always left a permanent mark on the community and infrastructure. Among the major events, the earthquake experienced in San Francisco in 1906 was of a magnitude of 7.9, destroying much of the city and further causing several fires and destruction. It brought into perspective the vulnerability of urban centers to seismic hazards and set the tone for modern earthquake-resistant building codes. More recently, the devastating 1989 Loma Prieta and 1994 Northridge earthquakes underlined that earthquake preparedness and response measures had to be continually improved. Morell et al. (2020), pointed out that these events caused billions of dollars in damages, displaced thousands of residents, and pointed toward further research in earthquake-resistant infrastructure and early-warning systems. Although engineers and policymakers have made refinements, the frequent seismic violence that grips California requires constant study in mitigation of the risks that will accompany future earthquakes.

Existing Methods for Earthquake Prediction

The great majority of conventional earthquake prediction methodologies have positioned their emphases on respective seismological and statistical methods in analyses of seismic activity patterns. In seismic methods, the movement of tectonic plates, the activity of fault lines, and the deforming of the ground are studied to assess the probability of a future earthquake. These often employ data on changes in tectonic stress from seismographs, GPS measurements, and other monitoring technologies (Pinilla-Ramos et al., 2024). By contrast, statistical analyses look at the pattern of earthquake history for possible cycles or clusters in seismic activity that could determine whether and where an earthquake would occur. Traditional approaches in these regions have already allowed researchers to identify those areas that are prone to earthquakes, thus increasing general preparedness.

However, traditional approaches that support such forecasts have serious setbacks. The intrinsic complexity of tectonic phenomena, along with deficiencies in the definite precursors that would signal the near occurrence of an earthquake, further aggravates this challenge. While it is possible to assess risks over long periods on an appropriate scale and identify areas at high risk, the exact times, places, and magnitudes of such an event are rather difficult to complete. Current models equally fail to explain the variation in earthquake patterns, such as foreshocks, aftershocks, and earthquake swarms, that make reliable prediction even more daunting (Pinilla-Ramos et al., 2024). There is, therefore, an increasing demand for sophisticated techniques that will enhance the accuracy of earthquake forecasting for better planning in disaster response.

Machine Learning in Seismic Analysis

Kourehpaz & Molina (2022), holds that in the recent past, Machine learning algorithms (ML) have emerged as a promising invention for advancing earthquake prediction by facilitating the pinpointing of patterns in large datasets that may be hard to detect through traditional techniques. Application of ML algorithms such as neural networks, support vector machines, and decision trees to seismic data leverages recognition of patterns that precede earthquakes. These are especially helpful in high-dimensional data analysis. For researchers, Machine Learning algorithms methods mean processing huge volumes of information obtained from seismic records, satellite imagery, and measurements over fault lines. Hu et al. (2021), assert that applying machine learning models using such data allows for the detection of slight changes in tectonic stress and deformation that could be indicative of seismic events.

Reyes (2021), articulated that advances in data-powered predictive modeling have illustrated the capability of machine learning to enhance earthquake forecasting. For instance, machine learning algorithms have successfully identified possible foreshock sequences that might precede larger events and forecasted the locations of aftershocks following a major seismic event. Additionally, models are currently being trained to make predictions based on the probability of subsequent earthquakes from real-time seismic data and historical trends. Martinelli (2020), argues that these deals are a sea change in seismic analysis, considering machine learning offers adaptive and subtle predictions. While machine learning cannot predict the exact time an earthquake could occur, it has the potential to process complex data and detect patterns to further improve strategies for earthquake preparedness and response.

3. Data Collection and Preprocessing

To assess seismic activity in California, data were gathered from a credible and reputable source, most notably, the United States Geological Survey (USGS) Earthquake Database, which provides detailed records of earthquakes spanning over six decades. This dataset included all recorded seismic events in California, capturing the key details about the place of occurrence in latitude and longitude, the magnitude of the seismic event, the depth at which it happened, and the time when the seismic activity took place (Pro-AI-Robikul, 2024). These are the fundamental variables for analyzing spatial and temporal patterns in earthquake activities, while magnitude points toward the severity of the event experienced, depth shows tectonic behavior, and time stamps provide trend and frequency analyses. The current study uses a robust and extensive dataset to establish an accurate and detailed overview of historic seismic patterns in California, forming a reliable basis for developing predictive models.

Data Cleaning and Preparation

Cleaning and preparation of data are the most important steps toward the authenticity and operability of earthquake data for meaningful analysis and model development. Handling missing or incomplete records is one of the things that needs to be carefully judged in this initial phase. Many seismic datasets have always contained some missing values due to equipment limitations, environmental factors, or a lapse in recording. These omissions lead to biases if left unprocessed. First, missing records were identified in key fields related to location, latitude and longitude, magnitude and depth, and time of occurrence. When records had important data missing either on location or time, the record was discarded, since such information is crucial for carrying out spatial and temporal analyses (Pro-AI-Robikul, 2024). In cases where actual data was not available, such as depth or magnitude, imputation techniques were performed with mean or median imputation. Only such cases were considered where the value represented less than a small percentage of the dataset. The goal of such a selective approach toward data imputation is to minimize data losses and avoid compromising the integrity of the dataset.

Feature Engineering

S/No	Key Feature	Description
1.	**ID**	Unique identifier for each entry.
2.	**STATE_CODE**	Code representing the state (California in this case).
3.	**STATE_NAME**	Name of the state (California).
4.	**CITY**	Name of the city where the earthquake occurred.
5.	**County**	County of the occurrence
6.	**LATITUDE & LONGITUDE**	Geographical coordinates of the location

Table 1: Depicts the Key Features

The above-derived features are included due to their relevance to earthquake prediction. For example, the identification of clusters of magnitudes and inter-event times may elucidate patterns that are precursors to large earthquakes for early warning capabilities. Another critical feature was the depth-to-magnitude ratio, which captures the relationship between earthquake depth and its potential surface impact. This is particularly relevant for urban planning, as shallow earthquakes tend to have more surface damage while deeper ones do (Pro-AI-Robikul, 2024). Also, the time of day and day of the week were added to check if any temporal features could be discerned that might reinforce predictive capabilities, though these features are less directly impactful. Features it chooses can either describe the physical aspects and magnitude, for example-or temporal ones-for example, frequency of events, and time interval-seismic activities that, by enriching the pattern recognition capability of the predictive model, may lead to more accurate, informed forecasts of seismic events.

Exploratory Data Analysis

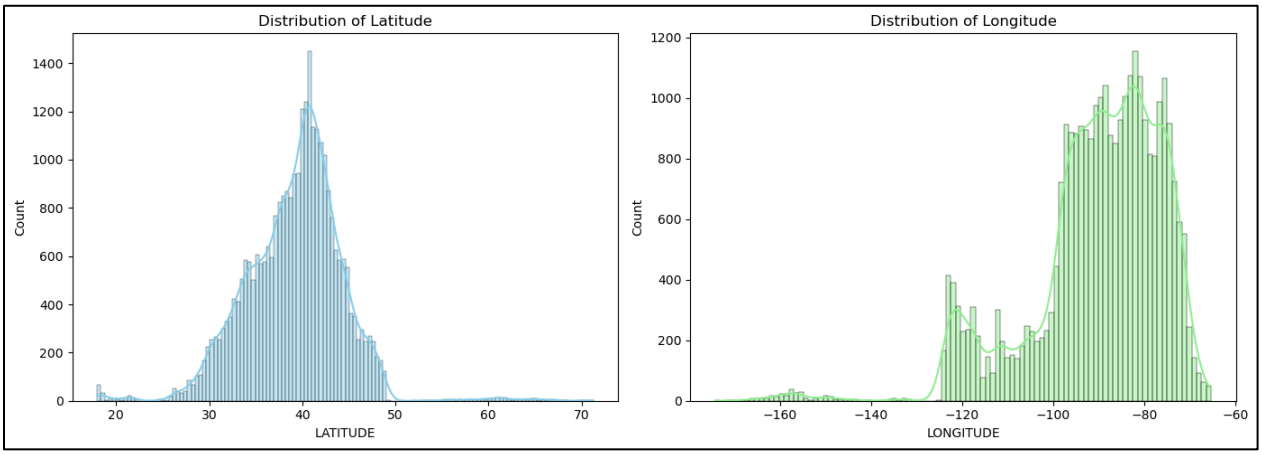

Figure 1: Portrays the Distribution of Latitude and Longitude

The above plots show the distribution of the occurrence of earthquakes by latitude and longitude, which in return provides insight into the spatial distribution of seismic activity. The **latitude distribution** indicates that most earthquakes occur between 30° and 45° latitude, with a peak at 37°–39°. That range of latitude corresponds to central and northern California and might hint toward high seismic activity nearby, including major lines of faults such as the San Andreas Fault. From the **longitude distribution**, it can be seen that the earthquakes within this longitude range are mostly concentrated between -120° and -90°, with a high concentration at around -120°. This corresponds to the western United States, specifically the coastal and inland fault zones of California. The result shown above reinforces that seismic activity in California is concentrated along certain latitudinal and longitudinal bands-which is consistent with the state's tectonic plate boundary and fault system layout.

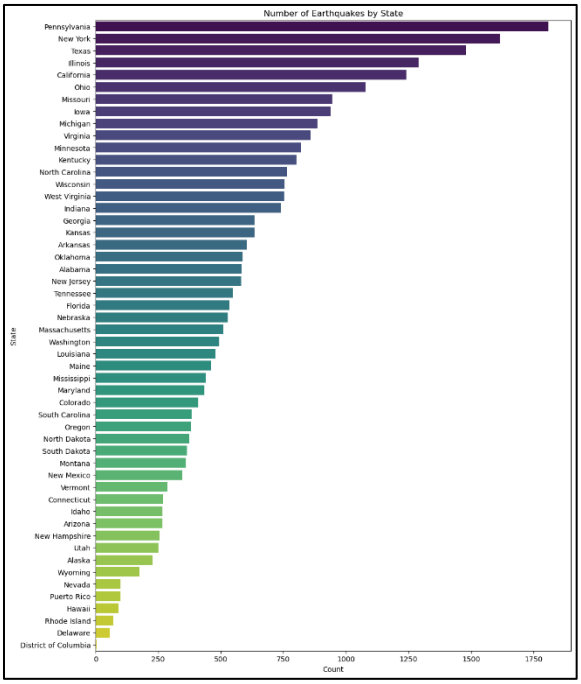

Figure 2 Displays the Number of Earthquakes by State

The above graph shows the number of recorded earthquakes, by state: Pennsylvania, having the highest number, followed by New York, Texas, Illinois, and California. Shockingly enough, Pennsylvania and New York rank higher than California, which tends to be thought about as very seismically active due to its positioning along significant fault lines, such as the San Andreas Fault. This might mean that more small earthquakes occur in those states without any damage, or simply reporting thresholds or practices in these regions are different. States such as Hawaii, Alaska, Nevada, and Puerto Rico are known to be seismic; it could place them lower on this list either because of the differences in data samplings or simply different magnitudes of earthquakes. On the whole, the chart underlines a wide scattering of seismic activity in the U.S. due to some unexpected states appearing as leaders according to high counts of earthquakes, probably by dint of regional geological and reporting factors that invite further investigation.

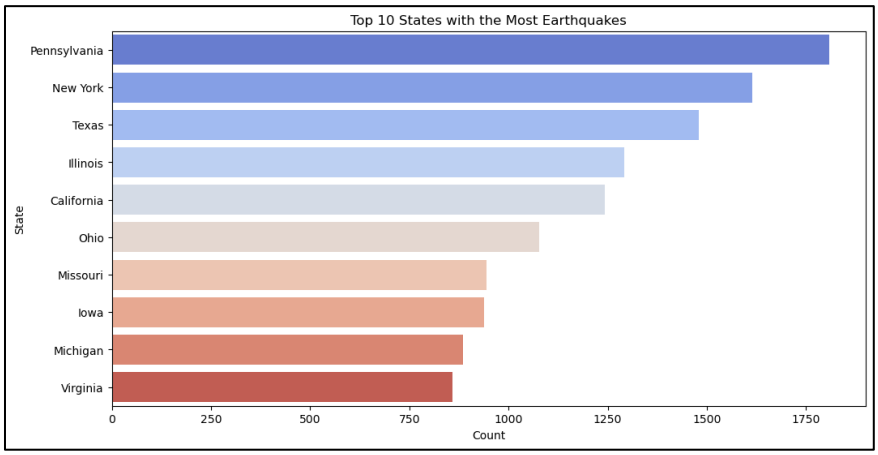

Figure 3: Exhibits Top 10 States with the Most Earthquakes

This diagram displays the top 10 U.S. states with the most recorded earthquakes. Among these, Pennsylvania is listed first, followed by New York, Texas, Illinois, and then California. California is supposed to have a very high rate of seismic activity due to its position along the Pacific Ring of Fire, yet here it landed fifth instead of first. This might mean that Pennsylvania and New York have a large number of smaller earthquakes, or that Pennsylvania and New York just don't count as many because of differences in data collection and measurement sensitivity or reporting standards. Ohio, Missouri, Iowa, Michigan, and Virginia fill out the top 10 and suggest seismic activity is not confined to traditionally earthquake-prone areas of the country. These findings underscore the widespread nature of seismic events across various states; hence, the need for adequate monitoring and preparedness against earthquakes should be holistic and not restricted to conventionally affected areas like California.

4. Predictive Modelling Approach

Model Selection

To devise and curate a reliable earthquake prediction algorithm for California, distinct machine learning models were considered, comprising Random Forest, XG-Boost, and Logistic Regression. Logistic Regression, Random Forest, and XG-Boost are probably the most popular algorithms in machine learning with their respective strengths and popular use cases. Logistic Regression is a linear model used for binary classifications. It predicts the probability of an input to be of a certain class by fitting the data onto a logistic function; hence, it does well in problems where features vary linearly or near-linearly with the outcome. The Random Forest is an ensemble learning technique wherein it builds many decision trees while training, based on their combination in making predictions. Several trees are used to reduce overfitting by averaging their output, hence more accurate results; this method then fits well for complex nonlinear relationships among data sets. Finally, XG-Boost is an acronym for the name Extreme Gradient Boosting, a new boosting algorithm that constructs an ensemble of weak learners sequentially one after another each subsequent model attempts to correct the errors of the previous one. It has used mostly gradient boosting and optimized computation for performance. Each of these algorithms has advantages depending on the underlying data and the difficulty of the task: logistic regression due to its simplicity and interpretability; Random Forest, because of its robustness and easy adaptation to problem specifics; and XG-Boost, in providing state-of-the-art performance in complex predictive tasks.

Training and Testing Framework

The dataset was then divided into training and testing sets to ensure the models generalize well on unseen data. Usually, it would have had an 80-20 split; the data is used 80% for training and kept 20% for testing. To make the model more reliable and avoid overfitting, cross-validation was applied to the training set, a technique known as k-fold cross-validation. With cross-validation, each model could be trained and validated multiple times by splitting the training set into multiple subsets, called folds. This procedure gave a robust assessment of the performance of models and reduced the risk of biased results. This would be important in identifying the most reliable model that could yield consistent predictive accuracy across different data segments.

Hyper-Parameter Tuning

Hyperparameter tuning was done for further optimization of model performance by applying different techniques such as grid search and random search. Grid search represents the exhaustive examination of a pre-defined hyperparameter investigation range in search of the best combination, while random search samples a random subset of the hyperparameters, thus much faster but often nearly as good. All the models went under hyperparameter tuning-scanning over the depth and the number of estimators in Random Forest, the learning rate and tree depth in XG-Boost, and hidden layer size in LSTM to make them as accurate and efficient as possible. These techniques helped optimize the configuration of each model to render the best predictive performance without having an overly simple model or an extremely complex model.

Evaluation Metrics

Finally, different metrics were considered for the evaluation of model performance. For each working model, accuracy, precision, and recall were considered prime metrics to measure their effectiveness in the prediction of earthquake occurrences. In a domain where the real risks for false positives and false negatives are different, precision and recall proved to be vital metrics. Comparison with a base model and other studies within seismic prediction were given to correctly position the accuracy and reliability of the present model's output. These results were compared with the best model intended to elevate better earthquake predictions and, most importantly, if the machine learning approach attained any value worthy of note over conventional predictive models. This trend towards a comprehensive approach to predictive modeling targets the development of a data-driven solution to support timely qualitative earthquake forecasting, enabling better disaster preparedness and risk mitigation in California.

5. Results

Pattern and Trend Analysis

Retrospectively, investigations into six decades of seismic data from California uncover some critical aspects of earthquake frequency, location, and magnitude. The first pertinent understanding derived is that earthquakes are usually fairly predictable; low-magnitude earthquakes occur every year with a somewhat consistent frequency. Regarding higher-magnitude earthquakes, periods of spiking can often be observed in the data following longer periods of low activity. This pattern is consistent with seismic theories on the evolution of stress along fault lines especially the San Andreas Fault, which undergoes an accumulation of tension that periodically results in major earthquakes. Spatial analysis also points to the existence of areas with high earthquake frequencies, such as the vicinity around the San Andreas, Hayward, and Calaveras faults, that highlight the position of these fault zones within the seismicity of California.

Temporal trends have interesting variations, both in terms of frequency and magnitude of earthquakes. For example, slight increases in earthquake occurrences have been noted for recent decades, which may be the result of better monitoring technology and increased sensitivity in seismic detection rather than an actual rise in seismic activity. Evidence of earthquake

epicenter migration is also present, with activity sometimes migrating along different segments of major fault lines. Magnitude trends reflect that, though low-intensity earthquakes (less than 4.0 magnitude) are very much more common and happen with periodic recurrences, the large ones do, too above magnitude 6.0. These are important to take into consideration when understanding earthquake risks and to develop models that may predict and prepare Californians and their infrastructures for future seismic events.

Predictive Model Performance

1. Logistic Regression

The following code snippet represents the Python implementation of the logistic regression classification algorithm used to predict the probability of binary outcomes. First, it instantiates a logistic regression model with at most 1000 iterations to converge. Then, this model is trained on training data, X_train, and y_trian. After training, the model makes predictions on the test set, X_test, and it stores the result in y_pred_lr. Then, the code calculates and prints the accuracy score, confusion matrix, and classification report to perform the model evaluation. The accuracy score calculates the proportion of correctly classified instances. A confusion matrix is a kind of visualization that summarizes the number of true positives, true negatives, false positives, and false negatives. The classification report provides precision, recall, F1-score, and support for each class in detail. In this example, logistic regression has been used for binary classification. The common performance metrics were shown to be used for the evaluation of the model.

```
Modeling with Logistic Regression
lr_model = LogisticRegression(max_iter=1000)
lr_model.fit(X_train, y_train)
y_pred_lr = lr_model.predict(X_test)
print("Logistic Regression Accuracy:", accuracy_score(y_test, y_pred_lr))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_lr))
print("Classification Report:\n", classification_report(y_test, y_pred_lr))
```

Table 2: Showcases the Logistic Regression Modelling

Output:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96      5507
           1       0.00      0.00      0.00       469

    accuracy                           0.92      5976
   macro avg       0.46      0.50      0.48      5976
weighted avg       0.85      0.92      0.88      5976
```

Table 3: Displays Logistic Regression Classification Reporting

This classification report depicts that the logistic model has a high accuracy of 0.92, suggesting that the model predicts the majority of the cases correctly. On taking a precise look at precision, recall, and F1-score for each class, the imbalance has been recorded. Class 0 has a perfect recall of 1.00, meaning all actual positive instances were correctly identified. In contrast, it has a lower precision of 0.92, suggesting some false positives. Correspondingly, class 1 has low precision and the model cannot properly identify instances of this class. These are supported by the macro average and weighted average metrics, considering both classes and further reduced performance compared to the overall accuracy. This overemphasizes the fact that the model is biased towards class 0. In general, while the model correctly attains high accuracy, this class imbalance may be an issue and further optimization may be required to improve the performance on class 1.

2. Random Forest

This Python code snippet below was used for classification under a random forest which is an ensemble learning method where multiple decision trees work together to produce results that may provide higher accuracy in prediction. The code first creates a random forest model with 100 decision trees at a fixed random state for reproducibility. The model should then train the samples of training data provided by X_train and y_train. The model is trained and predicts on test data stored in X_test. It stores in y_pred_rf. Further, it calculates the accuracy score, confusion matrix, and classification report to evaluate the performance of the model. It then predicts the accuracy score, showing the percentage of correct predictions, a confusion matrix showing the distribution of true positives, true negatives, false positives, and false negatives, and a classification report that shows precision,

recall, F1-score, and support for each class in detail. Essentially, this code shows the use of Random Forest for classification tasks and also its evaluation using some common performance metrics.

```
Modeling with Random Forest
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)
print("Random Forest Accuracy:", accuracy_score(y_test, y_pred_rf))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_rf))
print("Classification Report:\n", classification_report(y_test, y_pred_rf))
```

Table 4: Portrays Random Forest Modelling

Output:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      1.00      0.96      5507
           1       0.80      0.07      0.14       469

    accuracy                           0.93      5976
   macro avg       0.86      0.54      0.55      5976
weighted avg       0.92      0.93      0.90      5976
```

Table 5: Exhibits Random Forest Classification Report

From the classification report, the random forest model had an overall accuracy of 0.93, showing that most of the cases were predicted correctly. However, there is a class imbalance; Class 0 shows better precision and recall when compared to Class 1. Class 0 enjoys perfect recall (1.00), indicating that all actual positive instances were well-identified. Class 1 will have lower precision and recall, indicating the model's difficulty in identifying this class correctly. The macro and weighted averages are well below the overall accuracy, further emphasizing that this model was oriented towards class 0.

3. XG-Boost

This code snippet below illustrates the XG-Boost modeling, an ensemble learning algorithm on a classification task. An instance of an XG-Boost classifier is initialized with the parameters for disabling label encoding and metrics for log loss. The model is then fitted on the train data represented through X_train and y_train. After successful training, the model makes the prediction on testing data-which is X_test-and the result is stored in y_pred_xgb. Then, it computes and prints the accuracy score, confusion matrix, and classification report to analyze model performance. The accuracy score measures the proportion of correct predictions, the confusion matrix visualizes the distribution between the true positives, true negatives, false positives, and false negatives, and the classification report provides detailed metrics like precision, recall, the F1-score, and support for each class. The code above is a general example of using XG-Boost for classification and then evaluating its performance with commonly used metrics.

```
Modeling with XGBoost
xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='mlogloss', random_state=42)
xgb_model.fit(X_train, y_train)
y_pred_xgb = xgb_model.predict(X_test)
print("XGBoost Accuracy:", accuracy_score(y_test, y_pred_xgb))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_xgb))
print("Classification Report:\n", classification_report(y_test, y_pred_xgb))
```

Table 6 portrays the XG-Boost Modelling.

Output:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96      5507
           1       0.50      0.04      0.08       469

    accuracy                           0.92      5976
   macro avg       0.71      0.52      0.52      5976
weighted avg       0.89      0.92      0.89      5976
```

Table 7: Depicts the XG-Boost Classification Report

The classification report from the XG-Boost model has an overall accuracy of 0.92, indicating the model correctly predicts most cases; however, there is a class imbalance in class 0 instances that have higher precision and recall than the class 1 instances. The recall of Class 0 -and as shown is perfect at 1.00, implying all actual positive instances in the class were correctly identified. By contrast, Class 1 is less precise and has lower recall, showing that the model fails to identify this class correctly. Where the macro and weighted averages depict performance as lower compared to overall accuracy, it surely follows that the model tends toward class 0. The high values for accuracy, class imbalance, and lower performance concerning the class 1 forecast may indicate areas for additional improvements as using techniques to balance the class or other hyperparameters.

6. Discussion

Implications for Urban Planning and Disaster Preparedness

The findings from seismic activity analysis have deep implications for urban planning and disaster preparedness in California. Knowledge of the pattern, place, and magnitude of earthquakes over time will help urban planners and policymakers structure communities that can remain resilient during such calamities. For instance, higher earthquake frequencies would automatically call for increased stringency in building codes, especially along fault lines, to ensure that houses situated on such lines are designed to withstand major seismic activity. This operation may even limit or reallocate urban development out of high-risk zones into less risky zones. In terms of emergency response planning, the findings support the creation of comprehensive disaster response plans that include creating and maintaining evacuation routes, emergency shelters, and sustaining resources for rapid response. On this count, the policy recommendations may range from periodic retrofitting of sensitive infrastructure such as hospitals, schools, and bridges in line with modern seismic standards, to mandating periodic disaster preparedness drills of residents in high-risk areas by local governments. Possible policy recommendations to local governments and disaster response agencies would include the following:

Land-use zoning: Establishing zoning restrictions that reduce building in these hazard-prone areas.
Building Codes: Seismic retrofitting of old buildings due to strict building codes and utilization of efficient earthquake-resistant construction methods while constructing new ones.
Emergency Management Planning: Detailed plans of emergency response, shelter locations, and communication strategy.
Public awareness of earthquake hazards and safety measures can be disseminated through mass media.

Challenges and Limitations

This study points out several challenges and limitations; for instance, the dataset, about such a long period of six decades, is susceptible to sparsity and quality problems, especially the old records when the seismic recording technology was not so developed. This fact may create some problems in the temporal consistency of data; it might underestimate the frequency or magnitude of previous earthquakes. While the predictive models used in this research are very promising, forecasting earthquake occurrences is intrinsically a very challenging task because seismic events are highly complex and mostly chaotic. Models may further fail to capture the fine details of earthquake triggers, particularly those of larger quakes, and can give very high false positives or false negatives; hence, such models would be less effective for practical purposes.

Future Research Directions

To enhance the accuracy of earthquake prediction algorithms, future research should explore the following directions:

Additional data sources: The integration of all types of data, including GPS measurements, radar interferometry, or geochemical indicators, will provide extensive information about seismic activity.

Advanced machine learning techniques: Advanced machine learning algorithms, like deep learning, make full use of complicated patterns in data to enhance the accuracy of predictions.

Real-time seismic data: This integration involves developing a real-time system to monitor seismic activity continuously, updating forecasts continually as new data arrives.

7. Conclusion

The primary objective of this research project was to assess historical seismic data to identify trends and patterns in California's seismic activity. Besides this research project aimed at developing predictive models that can provide insight into the possibility of seismic events in the future. To assess seismic activity in California, data were gathered from a credible and reputable source, most notably, the United States Geological Survey (USGS) Earthquake Database, which provides detailed records of earthquakes spanning over six decades. This dataset included all recorded seismic events in California, capturing the key details about the place of occurrence in latitude and longitude, the magnitude of the seismic event, the depth at which it happened, and the time when the seismic activity took place. To devise and curate a reliable earthquake prediction algorithm for California, distinct machine learning models were considered, comprising Random Forest, XG-Boost, and Logistic Regression. The findings from seismic activity analysis have deep implications for urban planning and disaster preparedness in California. Knowledge of the pattern, place, and magnitude of earthquakes over time will help urban planners and policymakers structure communities that can remain resilient during such calamities. For instance, higher earthquake frequencies would automatically call for increased stringency in building codes, especially along fault lines, to ensure that houses situated on such lines are designed to withstand major seismic activity. This operation may even limit or reallocate urban development out of high-risk zones into less risky zones.

References

[1] Al Banna, M. H., Taher, K. A., Kaiser, M. S., Mahmud, M., Rahman, M. S., Hosen, A. S., & Cho, G. H. (2020). Application of artificial intelligence in predicting earthquakes: state-of-the-art and future challenges. *IEEE Access*, *8*, 192880-192923.

[2] Asim, K. M., Moustafa, S. S., Niaz, I. A., Elawadi, E. A., Iqbal, T., & Martínez-Álvarez, F. (2020). Seismicity analysis and machine learning models for short-term low magnitude seismic activity predictions in Cyprus. *Soil Dynamics and Earthquake Engineering*, *130*, 105932.

[3] Buiya, M. R., Laskar, A. N., Islam, M. R., Sawalmeh, S. K. S., Roy, M. S. R. C., Roy, R. E. R. S., & Sumsuzoha, M. (2024). Detecting IoT Cyberattacks: Advanced Machine Learning Models for Enhanced Security in Network Traffic. Journal of Computer Science and Technology Studies, 6(4), 142-152.

[4] Calderón, A., & Silva, V. (2021). Exposure forecasting for seismic risk estimation: Application to Costa Rica. *Earthquake Spectra*, *37*(3), 1806-1826.

[5] Gitis, V., Derendyaev, A., & Petrov, K. (2021). Analyzing the performance of GPS data for earthquake prediction. Remote sensing, 13(9), 1842.

[6] Hasan, M. R., Islam, M. Z., Sumon, M. F. I., Osiujjaman, M., Debnath, P., & Pant, L. (2024). Integrating Artificial Intelligence and Predictive Analytics in Supply Chain Management to Minimize Carbon Footprint and Enhance Business Growth in the USA. Journal of Business and Management Studies, 6(4), 195-212.

[7] Hu, X., Bürgmann, R., Xu, X., Fielding, E., & Liu, Z. (2021). Machine-learning characterization of tectonic, hydrological, and anthropogenic sources of active ground deformation in California. *Journal of Geophysical Research: Solid Earth*, *126*(11), e2021JB022373.

[8] Iaccarino, A. G., Picozzi, M., Bindi, D., & Spallarossa, D. (2020). Onsite earthquake early warning: predictive models for acceleration response spectra considering side effects. *Bulletin of the Seismological Society of America*, *110*(3), 1289-1304.

[9] Islam, M. R., Shawon, R. E. R., & Sumsuzoha, M. (2023). Personalized Marketing Strategies in the US Retail Industry: Leveraging Machine Learning for Better Customer Engagement. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 14(1), 750-774.

[10] Islam, M. R., Nasiruddin, M., Karmakar, M., Akter, R., Khan, M. T., Sayeed, A. A., & Amin, A. (2024). Leveraging Advanced Machine Learning Algorithms for Enhanced Cyberattack Detection on US Business Networks. Journal of Business and Management Studies, 6(5), 213-224.

[11] Islam, M. Z., Shil, S. K., & Buiya, M. R. (2023). AI-Driven Fraud Detection in the US Financial Sector: Enhancing Security and Trust. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 14(1), 775-797.

[12] Khan, M. A., Debnath, P., Al Sayeed, A., Sumon, M. F. I., Rahman, A., Khan, M. T., & Pant, L. (2024). Explainable AI and Machine Learning Model for California House Price Predictions: Intelligent Model for Homebuyers and Policymakers. Journal of Business and Management Studies, 6(5), 73-84.

[13] Kourehpaz, P., & Molina Hutt, C. (2022). Machine learning for enhanced regional seismic risk assessments. *Journal of Structural Engineering*, *148*(9), 04022126.

[14] Morell, K. D., Styron, R., Stirling, M., Griffin, J., Archuleta, R., & Onur, T. (2020). Seismic hazard analyses from geologic and geomorphic data: Current and future challenges. *Tectonics*, *39*(10), e2018TC005365.

[15] Martinelli, G. (2020). Previous, current, and future trends in research into earthquake precursors in geofluids. Geosciences, 10(5), 189.

[16] Muhammad, D., Ahmad, I., Khalil, M. I., Khalil, W., & Ahmad, M. O. (2023). A generalized deep learning approach to seismic activity prediction. Applied Sciences, 13(3), 1598.

[17] Nicolis, O., Plaza, F., & Salas, R. (2021). Prediction of intensity and location of seismic events using deep learning. *Spatial Statistics*, *42*, 100442.

[18] Perez-Oregon, J., Varotsos, P. K., Skordas, E. S., & Sarlis, N. V. (2021). Estimating the epicenter of a future strong earthquake in Southern California, Mexico, and Central America using natural time analysis and earthquake nowcasting. Entropy, 23(12), 1658.

[19] Pinilla-Ramos, C., Pitarka, A., McCallen M. EERI, D., & Nakata, R. (2024). Performance evaluation of the USGS velocity model for the San Francisco Bay Area. *Earthquake Spectra*, 87552930241270575.

[20] Pro-AI-Rokibul. (2024). *California-Earthquakes-Prediction-And-Analysis-Using-Advanced-Machine-Learning-Algorithms/README.md at main · proAIrokibul/California-Earthquakes-Prediction-And-Analysis-Using-Advanced-Machine-Learning-Algorithms*. GitHub. https://github.com/proAIrokibul/California-Earthquakes-Prediction-And-Analysis-Using-Advanced-Machine-Learning-Algorithms/blob/main/README.md

[21] Rahman, A., Karmakar, M., & Debnath, P. (2023). Predictive Analytics for Healthcare: Improving Patient Outcomes in the US through Machine Learning. Revista de Inteligencia Artificial en Medicina, 14(1), 595-624.

[22] Reyes Canales, M., & van der Baan, M. (2021). Forecasting of induced seismicity rates from hydraulic fracturing activities using physics-based models for probabilistic seismic hazard analysis: A case study. *Pure and Applied Geophysics*, *178*(2), 359-378.

[23] Rundle, J. B., Donnellan, A., Fox, G., Crutchfield, J. P., & Granat, R. (2021). Nowcasting earthquakes: Imaging the earthquake cycle in California with machine learning. *Earth and Space Science*, *8*(12), e2021EA001757.

[24] Shawon, R. E. R., Rahman, A., Islam, M. R., Debnath, P., Sumon, M. F. I., Khan, M. A., & Miah, M. N. I. (2024). AI-Driven Predictive Modeling of US Economic Trends: Insights and Innovations. Journal of Humanities and Social Sciences Studies, 6(10), 01-15.

[25] Shawon, R. E. R., Miah, M. N. I., & Islam, M. Z. (2023). Enhancing US Education Systems with AI: Personalized Learning and Academic Performance Prediction. International Journal of Advanced Engineering Technologies and Innovations, 1(01), 518-540.

[26] Sumon, M. F. I., Osiujjaman, M., Khan, M. A., Rahman, A., Uddin, M. K., Pant, L., & Debnath, P. (2024). Environmental and Socio-Economic Impact Assessment of Renewable Energy Using Machine Learning Models. Journal of Economics, Finance and Accounting Studies, 6(5), 112-122.

[27] Sumon, M. F. I., Khan, M. A., & Rahman, A. (2023). Machine Learning for Real-Time Disaster Response and Recovery in the US. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 14(1), 700-723.

[28] Wang, Y., Gardoni, P., Murphy, C., & Guerrier, S. (2021). Empirical predictive modeling approach to quantifying social vulnerability to natural hazards. *Annals of the American Association of Geographers*, *111*(5), 1559-1583.

[29] Zeeshan, M. A. F., Sumsuzoha, M., Chowdhury, F. R., Buiya, M. R., Mohaimin, M. R., Pant, L., & Shawon, R. E. R. (2024). Artificial Intelligence in Socioeconomic Research: Identifying Key Drivers of Unemployment Inequality in the US. Journal of Economics, Finance and Accounting Studies, 6(5), 54-65.