Pacific Northwest Earthquake Prediction and Analysis

Project 4 Group 4
Sofonias Abebe
Asres Dagnew
Jenn Andrews
Leo Rios

Objective

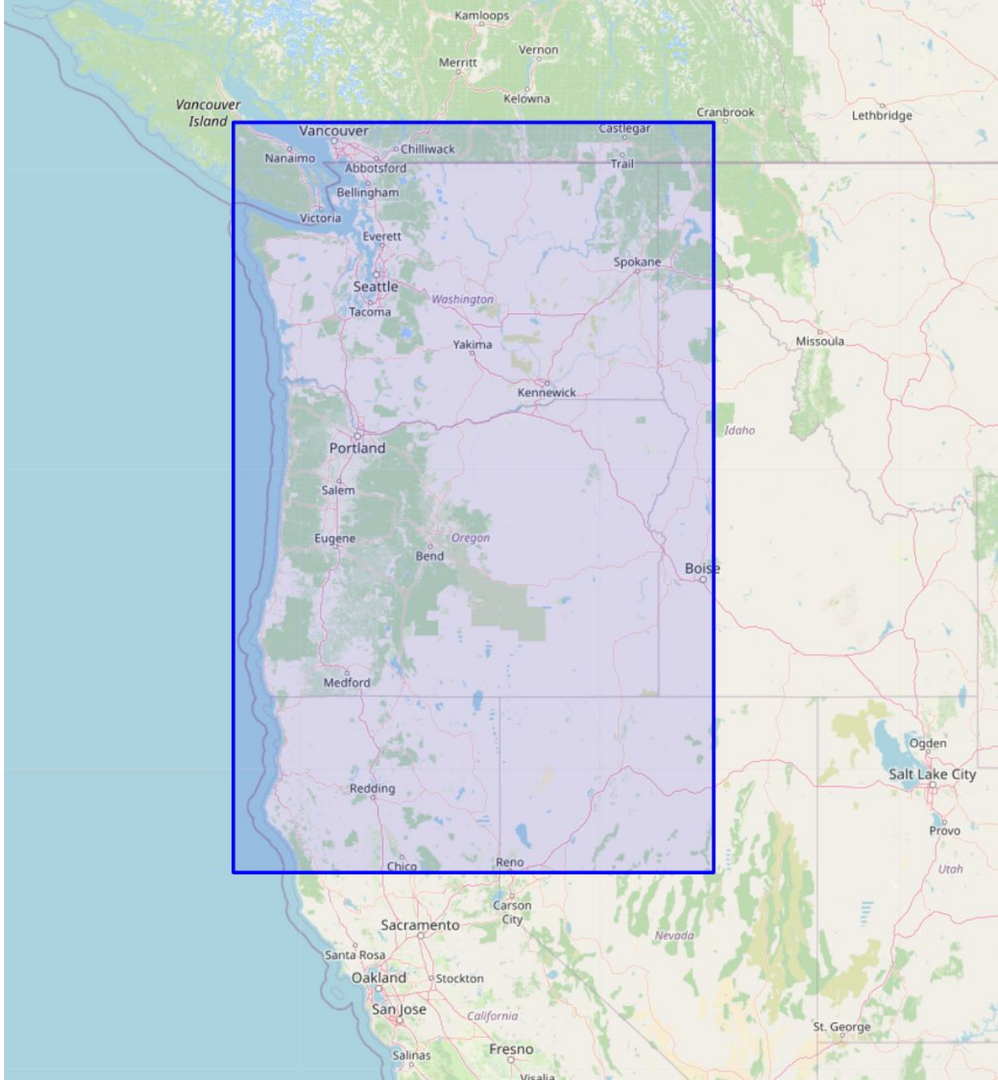
Analyze and predict earthquake patterns in the Pacific Northwest using historical data and machine learning.

This includes:

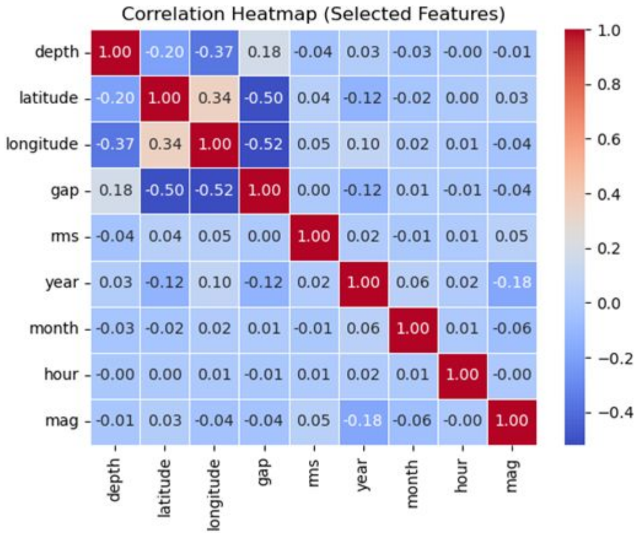
- ETL and database design
- Exploratory data analysis (EDA)
- ML modeling prediction
    - Occurrence
    - Magnitude
    - Occurrence + Magnitude prediction

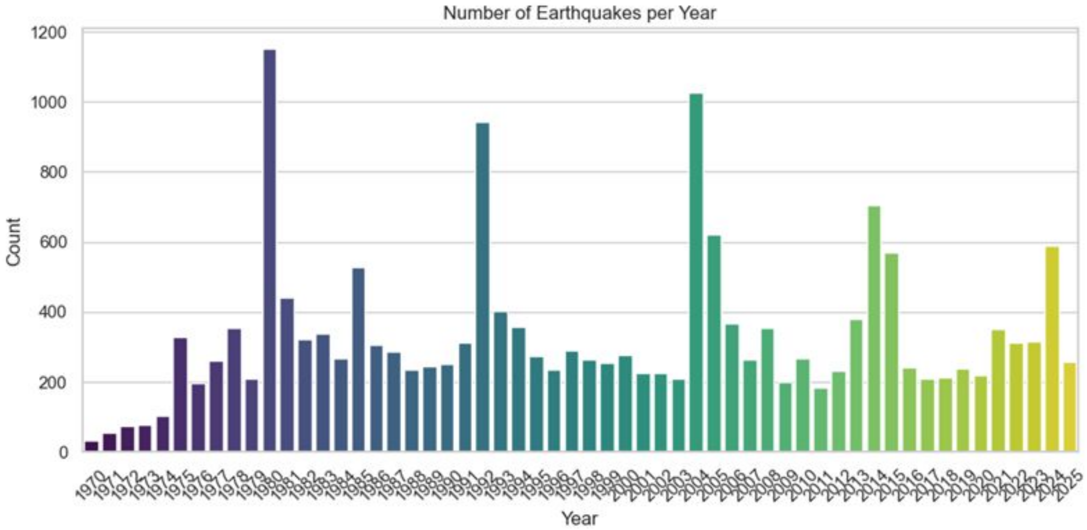
Data Sources and ETL

- Seismic activity dataset was obtained from the USGS Earthquake Hazards Program Catalog
- Filtered measurements for:
  - Magnitude >2.5
  - Jan 1970 – May 2025
  - Latitude: 39.5 to 49.5
  - Longitude: -125 to -116
- Exported to CSV and transformed with Pandas
- Created a PostgreSQL database with schema

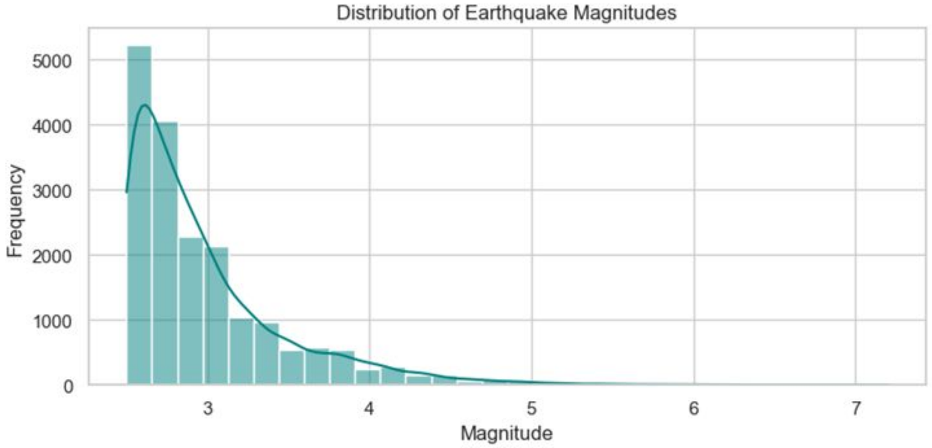


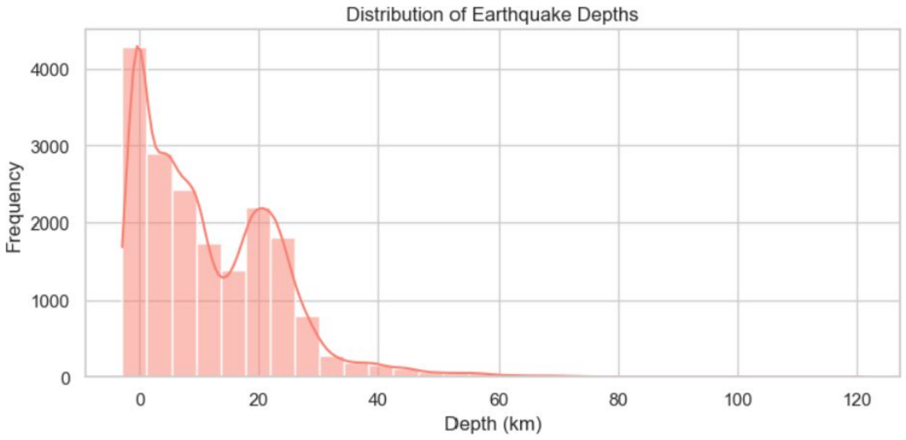
Data Exploration (EDA)





Data Exploration (EDA)





Initial Model

- Initial training using Linear Regression and Random Forest with Magnitude as the target feature

◆ Linear Regression

MAE: 0.35993099681809165

R^2: 0.027332519223118967

◆ Random Forest Regressor

MAE: 0.3253886634199134б

R^2: 0.19025735218760542

Initial Model

Feature Expansion and Grid Search

- Encode ‘magType’ as categorical
- Hyperparameter tuning

```
Best Params: {'max_depth': 20, 'min_samples_split': 5, 'n_estimators': 200}
MAE: 0.2399893804481306
R²: 0.5441280184488366
```

Initial Model

Switching to Classification

📈 Classification Report

Metric	Class 0 (Low)	Class 1 (High)
Precision	0.96	0.71
Recall	0.99	0.41
F1-score	0.98	0.52
Support	3470	226

Two Stage-Pipeline Modeling

Stage 1 Pipeline : Occurrence classification

Goal: Predict whether an earthquake will occur in a given time/location bin

Applies a Random Forest Classifier to predict whether an earthquake will occur in a specific region and time frame.

Earthquake Classification Model Setup

- ➢ Model: Random Forest Classifier
- ➢ Spatial Binning: 0.5° latitude × 0.5° longitude
- ➢ Temporal Binning: Monthly intervals
- ➢ Class Balancing: SMOTE (Synthetic Minority Oversampling Technique)

Features Used:

- Previous month’s quake count and max magnitude
- Spatiotemporal bins (`lat_bin`, `lon_bin`)
- Seasonal signals (`month_sin`, `month_cos`)
-

- ➢ Training Data: Pre-2015
- ➢ Testing Data: 2015 and onward

Two Stage-Pipeline Modeling

Magnitude Prediction

Goal: If a quake is predicted to occur, estimate its magnitude

Applies a Random Forest regressor to predict the maximum magnitude in a specific region and time frame.

Earthquake Magnitude Model Setup

- **Model**: Random Forest Regressor
- **Spatial Binning**: 0.5° latitude × 0.5° longitude
- **Temporal Binning**: Monthly intervals
- **Target Variable**: Maximum earthquake magnitude per bin
- **Features Used**:
  - max_mag_prev: Max magnitude from the previous time bin
  - count_prev: Earthquake count from the previous time bin
  - lat_bin, lon_bin: Spatial identifiers
  - month_sin, month_cos: Seasonal (cyclical) time features

- **Training Data**: Pre-2015

- **Testing Data**: 2015 and onward

Occurrence classification performance summary

Overall Metrics

Overall Accuracy: 96.0%

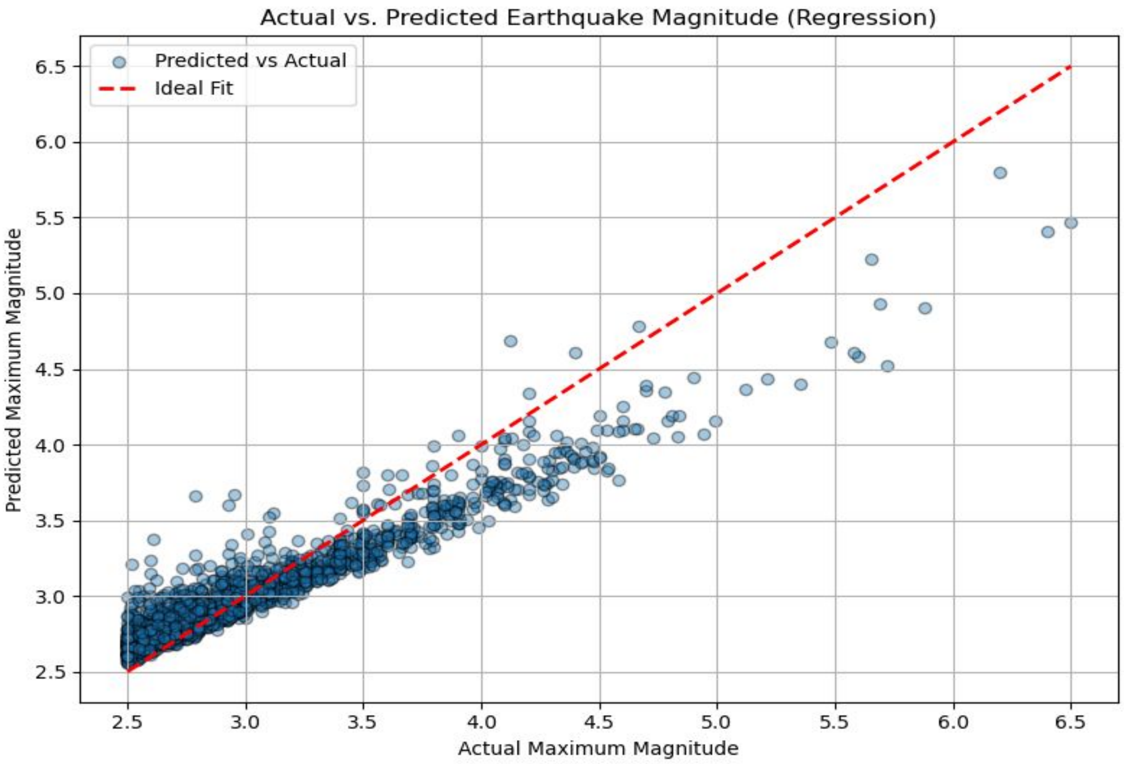
Regression Model Performance Summary

Earthquake Magnitude Regression Results:

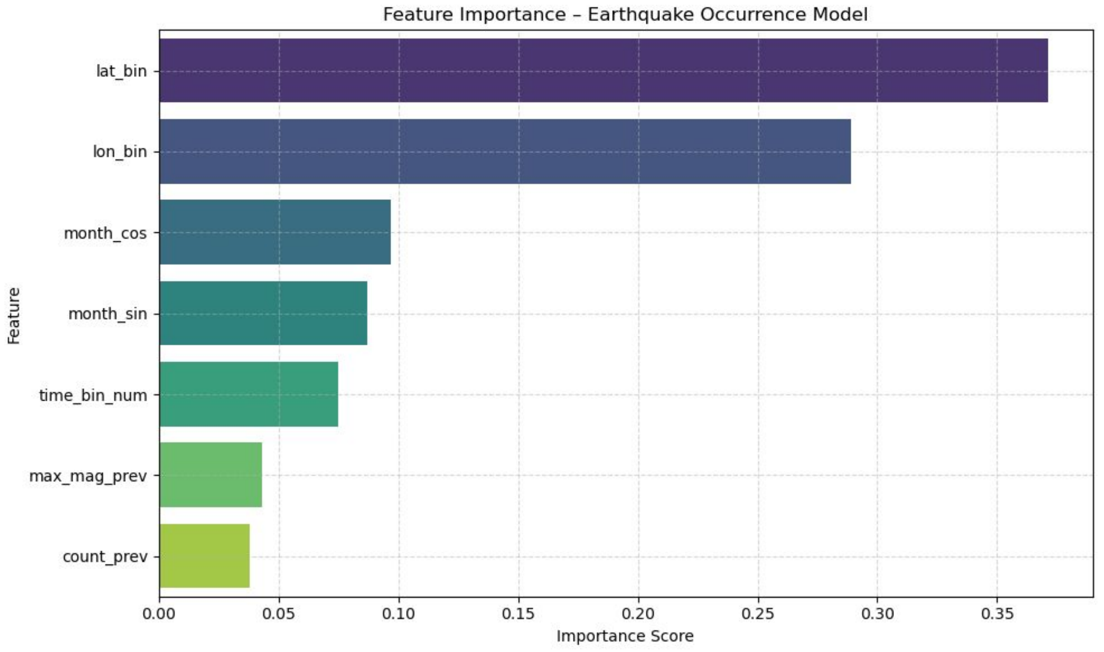
RMSE: 0.191

R^2 Score: 0.866

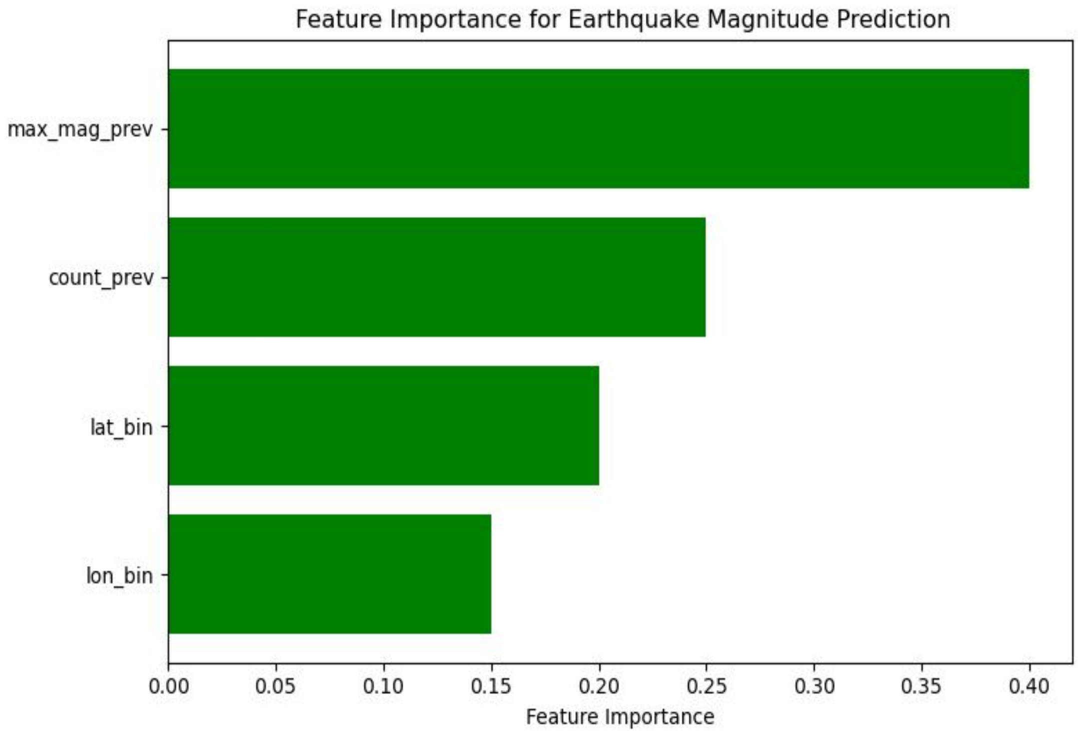
Regression Model Performance Summary



Key predictors of Earthquake Occurrence



Key Predictors of Earthquake Magnitude



Interactive Earthquake Occurrence Prediction Error map

This map visualizes the absolute error between the model's predicted earthquake probability and the actual outcome.

Error = | predicted probability – actual label |

Highlights how accurate the model was at each location.

🟢 Green = Accurate

🔴 Red = Inaccurate

Size: Larger dots = Greater error

Why It Matters

Pinpoints areas with high prediction errors

Helps identify regions needing model improvement

Interactive Quake Magnitude Prediction Error Map

This map shows how accurately the model predicted earthquake magnitudes across the Pacific Northwest.

- **Green dots** = Accurate predictions (low error)
- **Red dots** = Inaccurate predictions (high error)
- **Larger dots** = Greater prediction error (larger residuals)

Each point represents a spatial bin, and hovering reveals:

- Location (latitude & longitude)
- Actual vs. predicted magnitude
- Prediction error (residual)

Purpose: Highlights regions where the model performs well vs. areas needing improvement

Summary

- In the PNW, earthquake occurrence has increased slightly in recent years, possibly due to improved detection or tectonic shifts.
- Most earthquakes in the PNW are of low to moderate magnitude ($M<4$).
- Epicenters are concentrated along tectonic boundaries - especially the Cascadia Subduction Zone.
- Overall, our two-stage model achieved a 96% accuracy in predicting whether an earthquake would occur at a specific time and location, and reached an R^2 score of 85.2% for predicting the magnitude of those occurrences.
- Earthquake magnitude is best predicted by recent seismic activity such as maximum magnitude and number of occurrences in the previous time bin.

Challenges/Limitations

- Class imbalance made quake occurrence detection difficult despite measures taken to balance the dataset.
- Magnitude predictions underperformed for high-magnitude quakes.
- Geological features such as faults and stress maps were not included, limiting the model's performance.
- No real-time or streaming implementation was tested.

Further Research

- Integrate geophysical features (faults, ground types)
- Explore deep learning models for spatiotemporal patterns
- Improve magnitude modeling with quantile regression or weighted sampling