Predictive Modeling of Seismic Indicators using K-NET Data:

Machine Learning and Bayesian Approaches

Atishay Jain, Soumyadip Shyam,Darshan Nayak

June 6, 2025

# Introduction

The objective of this assignment is to develop robust models to predict key seismic indi- cators—Peak Ground Acceleration (PGA) and Peak Ground Velocity (PGV)—using the publicly available K-NET ground motion dataset from Japan. This project is critical for earthquake resilience and disaster management systems. The modeling framework inte- grates both traditional machine learning techniques and Bayesian probabilistic models to offer both predictive power and uncertainty quantification.

# Data Source

* **Dataset:** K-NET Dataset from the [Kyoshin Network (Japan).](https://www.kyoshin.bosai.go.jp/)
* **Total samples:** ∼11,820 events
* **Features:** Seismic waveform response, metadata (site/equipment/earthquake char- acteristics), derived features
* **Target variables:** log(PGA), log(PGV)

# Data Preprocessing

* **Leakage Columns:** Removed features derived from or directly representing the targets (e.g., maxvel, maxvelv, maxacc, maxaccv).
* **Missing Values:** Mostly negligible; filled with domain-specific constants or re- moved if uninformative.
* **Datetime Parsing:** Extracted hour, month, and derived night-time indicator from

jen origin time.

* **Categorical Features:** Dummy encoding for site and instrument type indicators.

# Cited Research Papers

**Reference Contribution Usage in Our Work**

Lin et al. (2023), EA- AI

S. Khoshnevis & D. Kamalian (2019), Soil Dynamics & Earth- quake Engineering

Kohrangi et al. (2017), Earthquake Spectra

Ghosh et al. (2022), Elsevier Sensors Wang et al. (2021),

Seismological Re-

search Letters

Chen et al. (2020), Journal of Earthquake Engineering

Zhang & Zhao (2018), Engineering Geology

Kumar et al. (2021), Natural Hazards

Song et al. (2023), Geophysical Journal International

Li et al. (2020), IEEE

Transactions on Geo- science and Remote Sensing

Used CNN+LSTM for soil response

Site-specific PGA predic- tion using GPR

ML-based fragility analy- sis

Uncertainty quantifica- tion via PyMC3

K-NET preprocessing best practices

Ensemble ML models for intensity prediction

PCA for dimensional- ity reduction in seismic studies

Applied LightGBM for damage prediction

Statistical modeling of PGV in complex terrains

Multi-source data fusion for seismic regression

Inspired feature extraction and response modeling Used for Moment Tensor

Features

Used for energy based fea- tures

Applied Bayesian Linear Regression and GPR

Used for waveform filtering and response metrics

Helped in feature creation

Informed PCA use before GPR

Supported LightGBM inte- gration in pipeline

Used for terrain-based fea- ture engineering

Motivated feature engineer- ing diversity

Table 1: Summary of Research Papers Referred and Their Usage

# Feature Engineering

Extensive feature engineering was performed based on waveform analytics, signal statis- tics, geospatial reasoning, and domain insights.

## Categories of Features

* **Waveform Features:** Log transforms, skewness, kurtosis, spectral slopes, peak response times.
* **Energy-based Features:** Arias intensity, cumulative RMS, early vs. late energy decay ratios.
* **Geospatial Features:** Site-to-event distance, angle offsets, strike/dip/magnitude ratios.
* **Soil Metadata:** AVS30, D1100, D1400 gradients, sensor depth vs elevation.
* **Moment Tensor Features:** Mxx, Mxy, Mzz, Moment Energy.

**Final feature count after engineering and deduplication:** 548 features.

# Feature Selection

* **Variance and Correlation Thresholding:** Removed near-constant features and highly correlated features (correlation *>* 90%).
* **LightGBM Importance Analysis:** Ranked all features by gain importance.
* **Final Selection:** Plotted top *K* features vs average MSE to determine optimal *K*. Two-step process: (1) Bins of 50 features, (2) Bins of 10 features from 100 to 180. **Optimal** *K* = 150.

# Data Imbalance Handling

* **Target Distributions:** Both log(PGA) and log(PGV) were continuous and skewed (log(PGV) more skewed).
* **Quantile Binning:** Applied decile binning for both targets.
* **Inverse Frequency Weights:** Computed and applied during training for all ML models (average of weights for PGV and PGA).

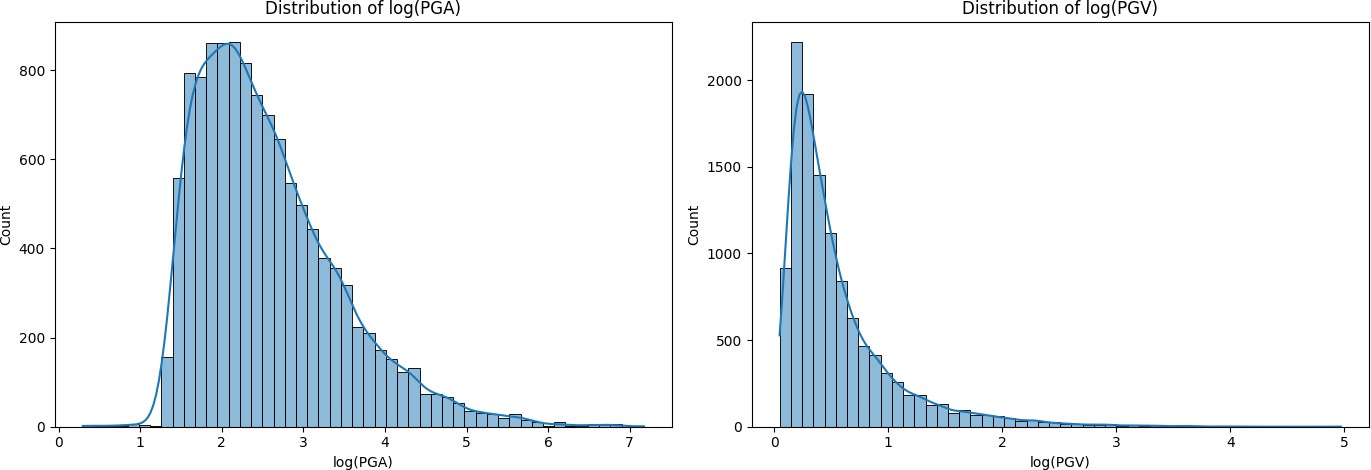


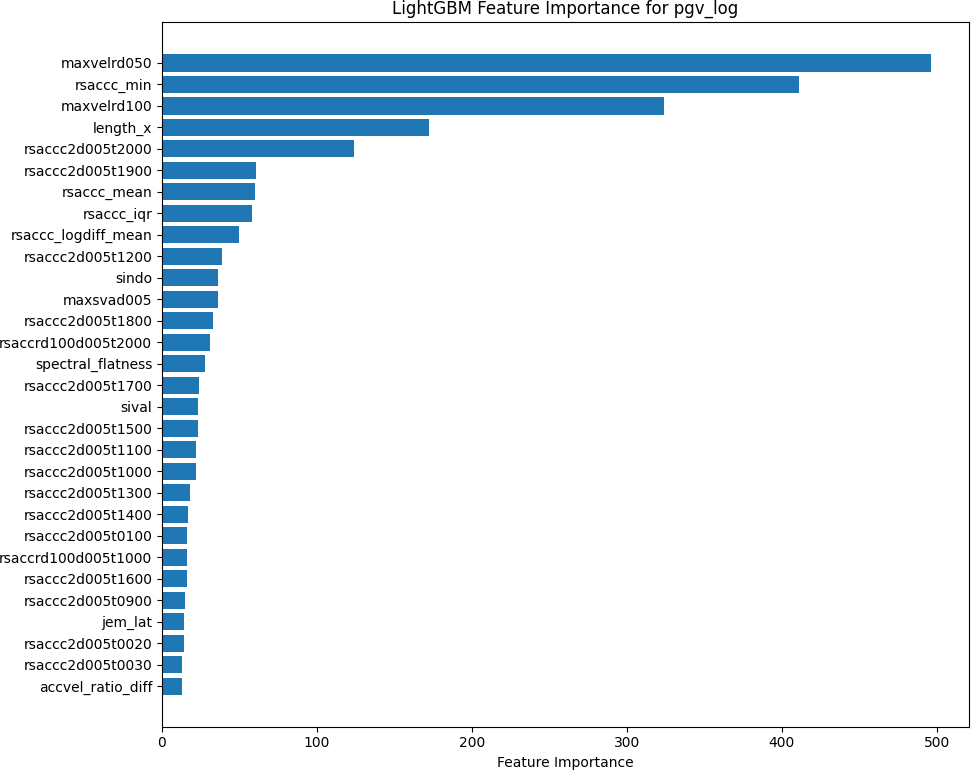
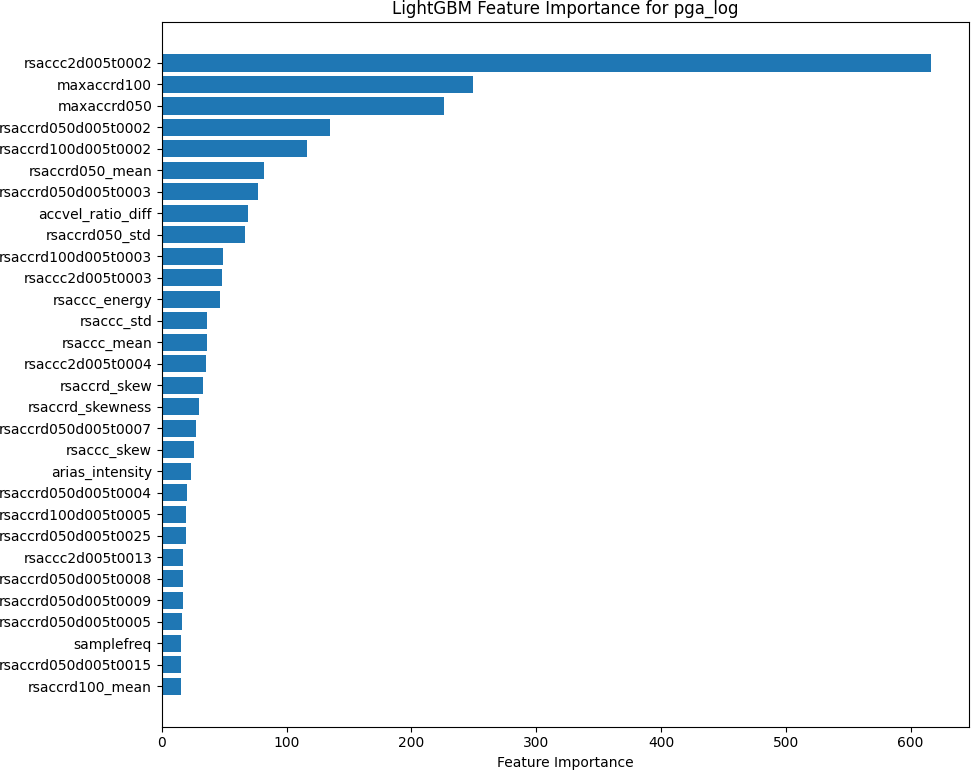
Figure 1: Distributions of log-transformed PGA (left) and PGV (right) in the dataset

# Model Training and Evaluation

## Machine Learning Models

Trained three traditional regressors using sample weights:

* Random Forest Regressor
* XGBoost Regressor
* LightGBM Regressor



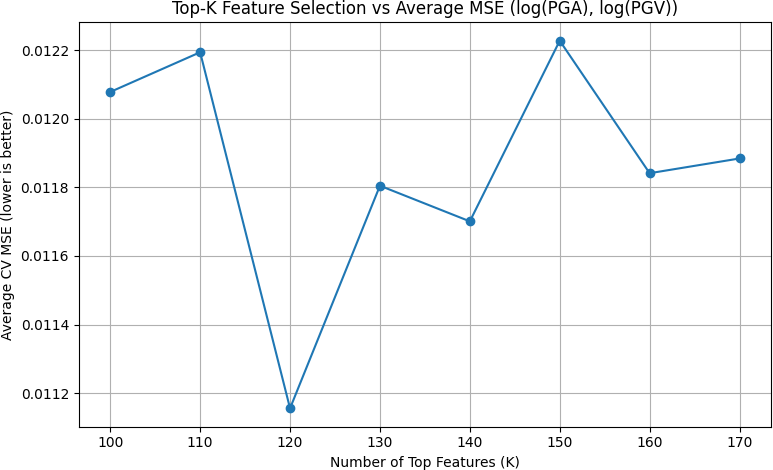
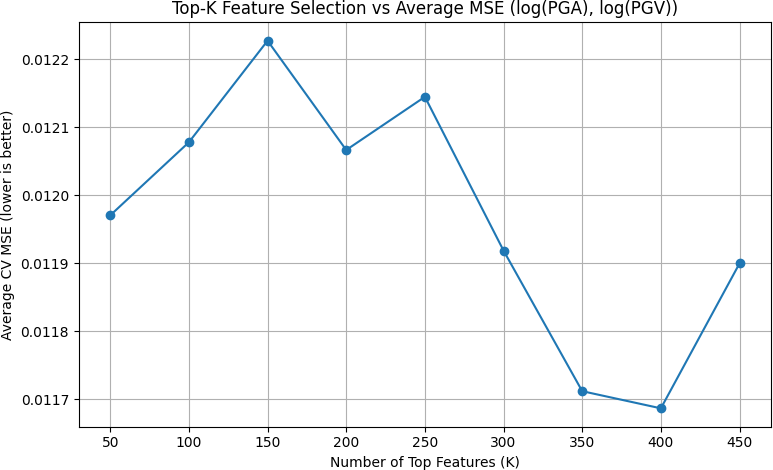
(a) PGA (b) PGV

Figure 2: Comparison of LightGBM Feature Importances

## Evaluation Metrics

* Mean Absolute Error (MAE)
* Root Mean Squared Error (RMSE)
* *R*2 Score (Coefficient of Determination)

Models were trained separately for log(PGA) and log(PGV) targets.



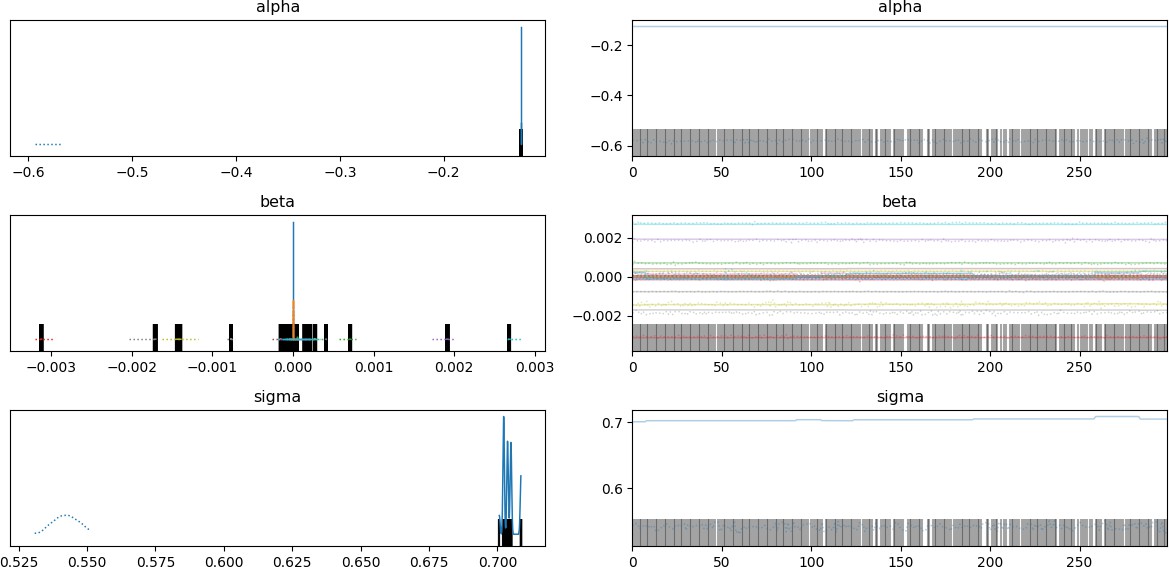
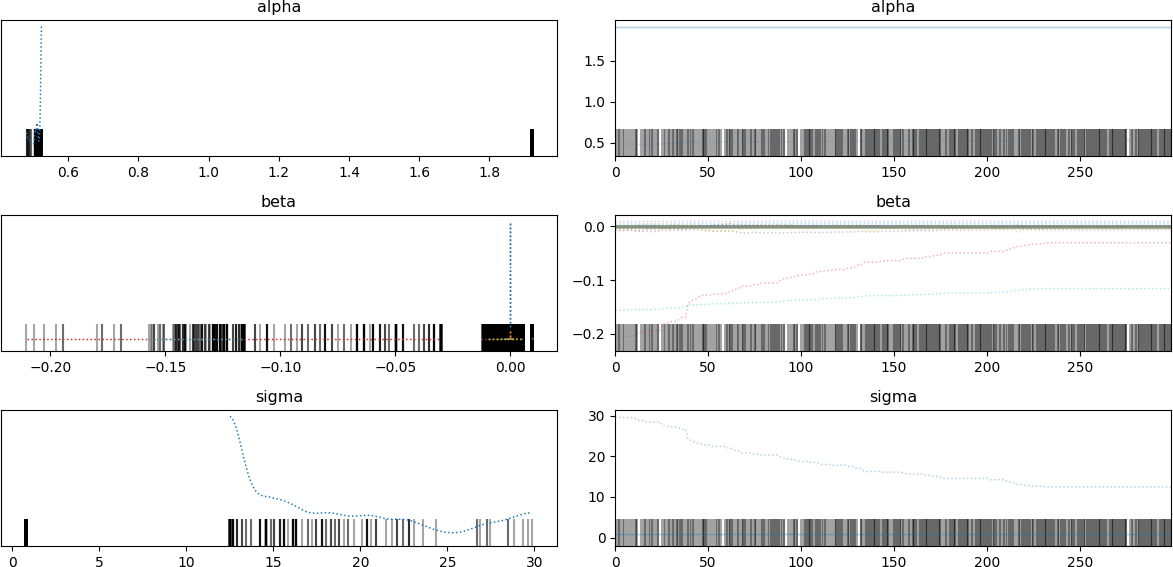
(a) PGA (b) PGV

Figure 3: Average MSE vs K

## Bayesian Models

To quantify uncertainty:

* **Bayesian Linear Regression (PyMC):** Used PCA (20 components) to balance computational efficiency and model quality. Defined priors for weights and noise, posterior sampling via NUTS, and generated credible intervals for predictions.
* **Gaussian Process Regression (GPR):** Attempted but found computationally intensive; limited to sampled data with few features, which degraded performance and led to discontinuation.



(a) PGA (b) PGV

Figure 4: Trace and posterior plots for Bayesian linear regression parameters

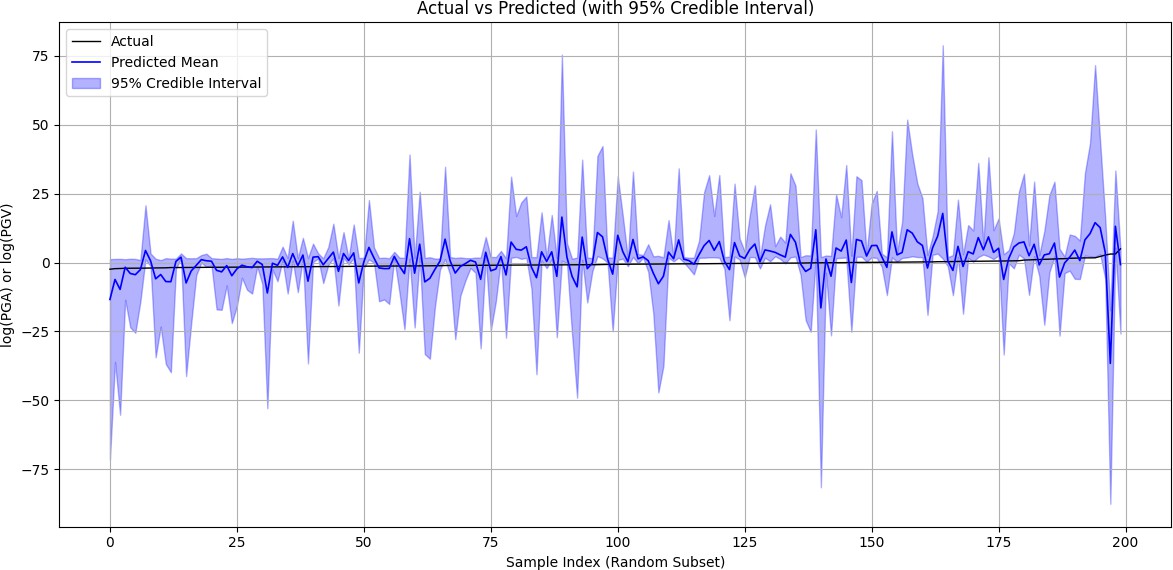


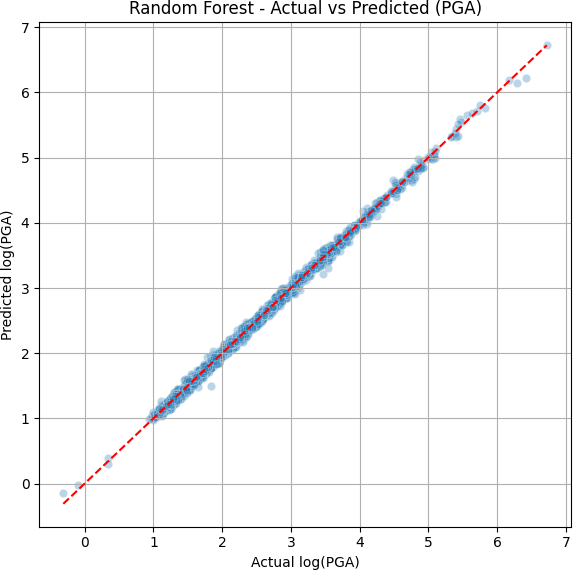
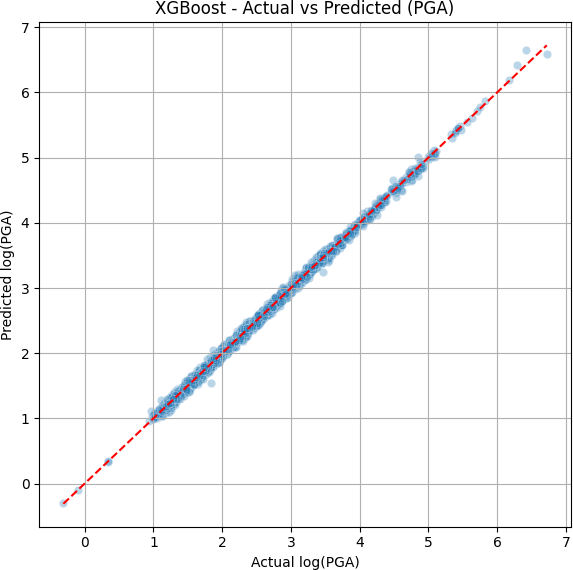
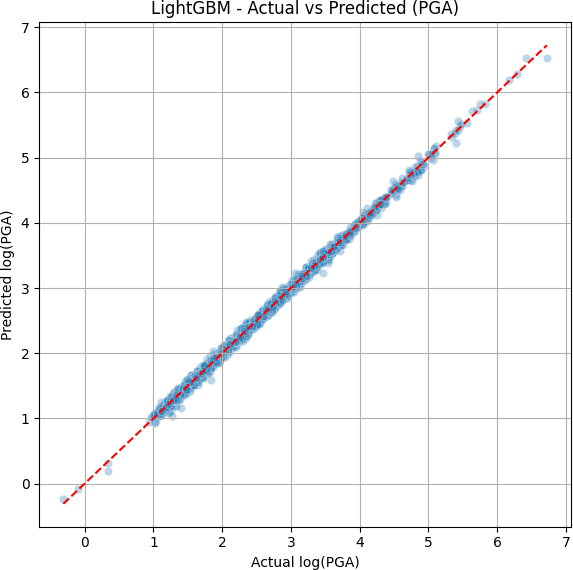
Figure 5: Predicted vs. Actual log(PGA) or log(PGV) with 95% Credible Intervals for Bayesian Model

# Evaluation Summary

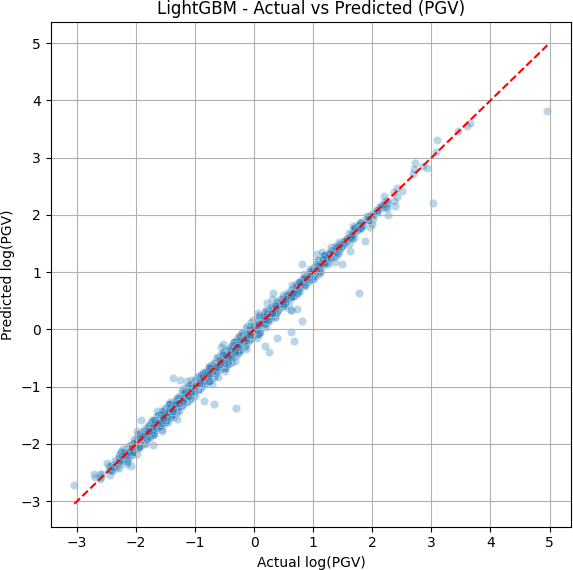
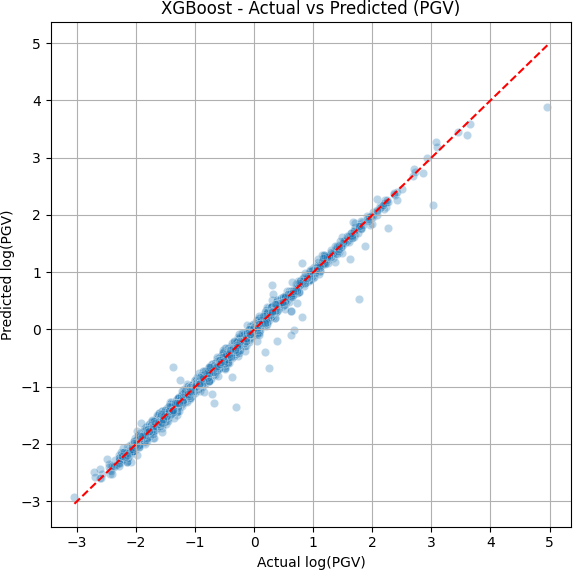
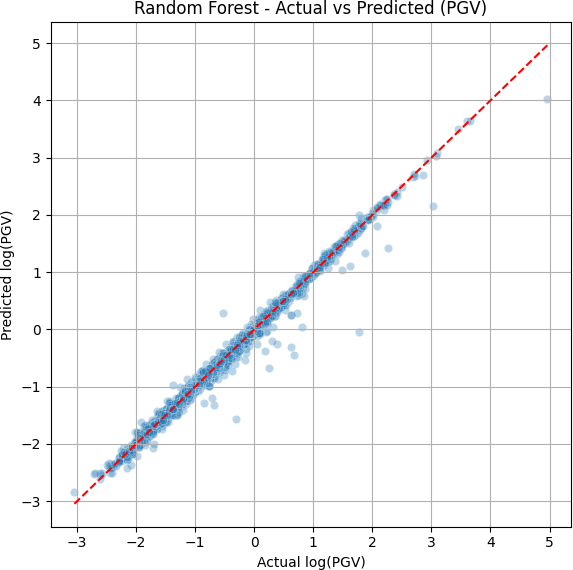
* Visualization of log-transformed target variable distributions.
* LightGBM feature importances.
* Actual vs. predicted plots with 95% credible intervals (Bayesian models).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Target** | **MAE** | **RMSE R2** | |
| RF | PGA | 0.0325 | 0.0443 | 0.9977 |
| RF | PGV | 0.0527 | 0.1020 | 0.9899 |
| XGB | PGA | 0.0338 | 0.0437 | 0.9978 |
| XGB | PGV | 0.0516 | 0.0889 | 0.9923 |
| LGBM | PGA | 0.0352 | 0.0456 | 0.9976 |
| LGBM | PGV | 0.0530 | 0.0879 | 0.9925 |
| BLR[1](#_bookmark0) | PGA | 4.5571 | 8.8484 | −88.7103 |
| BLR | PGV | 0.4532 | 0.7285 | 0.4855 |

Table 2: Performance Metrics for Different Models and Targets

(a) Random Forest PGA (b) XGBoost PGA (c) LightGBM PGA



(d) Random Forest PGV (e) XGBoost PGV (f) LightGBM PGV

Figure 6: Predicted vs. Actual Values for PGA and PGV Across Different Models

# Conclusion

This project demonstrates an integrated approach combining traditional machine learning and Bayesian methods for seismic indicator prediction using K-NET data. The workflow emphasizes robust preprocessing, feature engineering, and uncertainty quantification, pro- viding a strong foundation for earthquake resilience and disaster management systems.