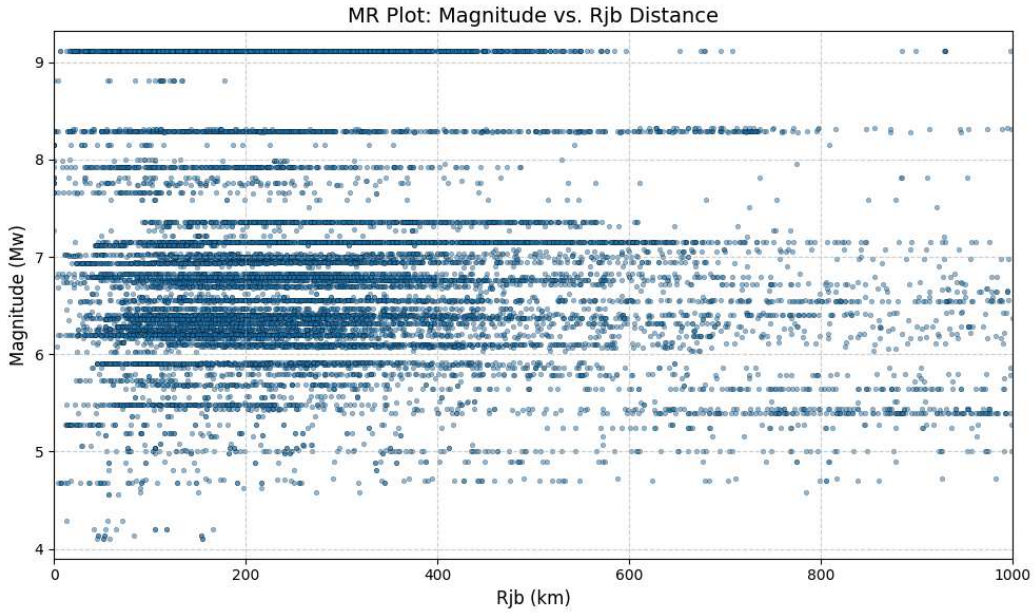
Prediction of Spectral Acceleration Using Gradient Boosting

1. Introduction

This study develops a Gradient Boosting model to predict 20 spectral acceleration (SA) values based on five input ground motion features: magnitude (mag), rupture distance (rjb), logrjb , logvs30 , and event type (inter-intra). The model includes a careful preprocessing pipeline, model training with early stopping, residual decomposition using mixed-effects modeling, Residual analysis,Ground motion physics,Importance,SHAP analysis for explainability.

2.Magnitude vs Rjb Scatter Plot:

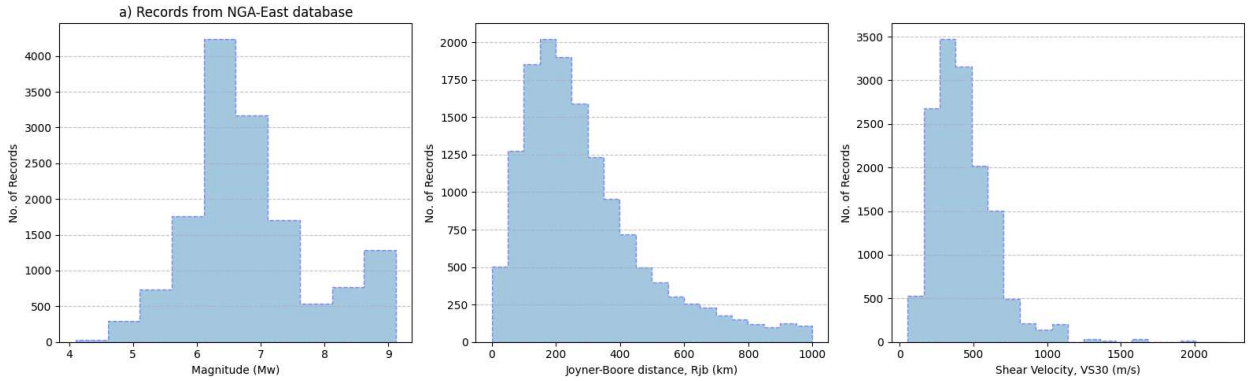
This scatter plot visualizes the distribution of events across different magnitude (mag) and Joyner-Boore distance (rjb) combinations in the dataset used for training and evaluation.



- The plot shows a dense cluster of data points for **moderate magnitudes (5.0–6.5)** and **short-to-moderate distances (0–100 km)**, which is typical of recorded ground motion datasets like NGA.
- Fewer data points appear at **larger distances (>200 km)** or for **larger magnitudes (>7.0)**, consistent with the relative rarity of such records.
- The coverage ensures that the model is well-trained across the critical near-field range but may have increased uncertainty for predictions at far distances or large magnitudes due to data sparsity.

3.Histograms of Input Features:

This figure presents histograms of three key input parameters—Moment Magnitude (Mw), Joyner-Boore distance (Rjb), and Shear-wave velocity at 30 m depth (Vs30)—from the NGA-East database used in this study.



- **Magnitude (Mw)** is concentrated around 6.0–6.5, reflecting a dataset dominated by moderate earthquakes.
- **Rjb** is right-skewed, with most recordings within 0–300 km, ensuring good coverage of near-field motions.
- **Vs30** peaks around 300–500 m/s, indicating a prevalence of stiff soil and soft rock sites in the data.

4.Summary Statistics of Input and Output:

Input Parameters:

Parameter	mag	rjb	logrjb	logvs30	intra_inter
min	4.1	0.01	-2	1.7243	0
max	9.12	999.0898	2.9996	3.3483	1
mean	6.8318	289.7475	2.352	2.5906	0.4232
std	1.0028	196.9747	0.3695	0.2032	0.4941
skewness	0.7859	1.2926	-3.3307	-0.087	0.3107
kurtosis	0.3906	1.535	33.8885	0.1169	-1.9035

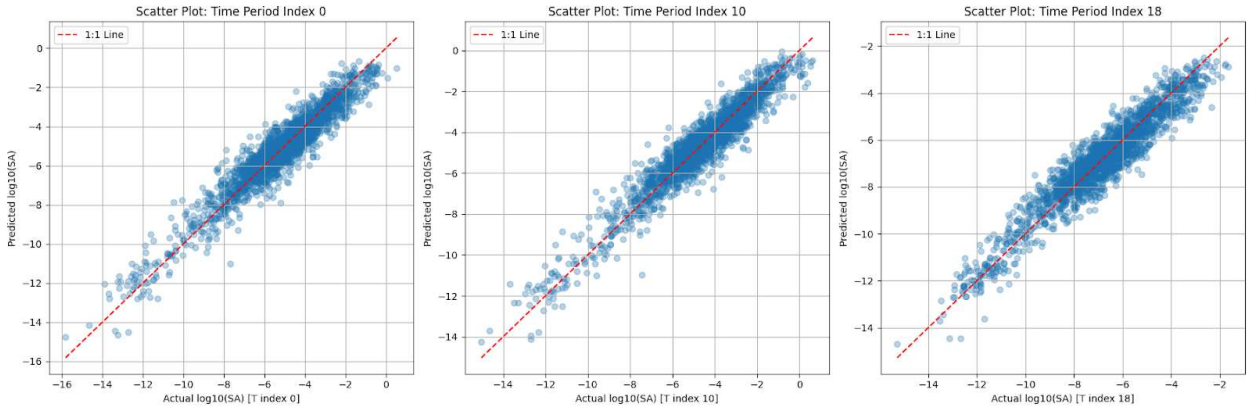
- **Magnitude (mag):** Ranges from 4.1 to 9.12, with a mean of 6.83, showing variability in seismic event intensity. Slight positive skew (0.79) and near-normal distribution.
- **Rupture Distance (rjb):** Varies widely from 0.01 to 999.09, with a mean of 289.75, showing high variability and positive skew (1.29).
- **Log of Rupture Distance (logrjb):** Range from -2.00 to 2.99, mean of 2.35, with a highly negative skew (-3.33) and heavy-tailed distribution (high kurtosis).
- **Log of Shear-Wave Velocity (logvs30):** Ranges from 1.72 to 3.35, with a mean of 2.59, close to normal distribution.
- **Intra-Inter Event Flag (intra_inter):** Ranges from 0.00 to 1.00, with a mean of 0.42, indicating mixed intra- and inter-event data, with light tails in distribution

Output Parameters:

Parameter	T0pt010S	T0pt020S	T0pt030S	T0pt050S	T0pt075S	T0pt100S	T0pt150S	T0pt200S	T0pt300S	T0pt400S	T0pt500S	T0pt750S	T1pt000S	T1pt500S	T2pt000S	T2pt500S	T3pt000S	T3pt500S	T4pt000S	T5pt000S
min	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
max	2.5801	2.7391	3.5567	4.9801	5.9791	3.6631	5.8752	6.2565	5.252	4.234	3.0608	2.27	1.2481	1.3501	1.2663	0.6708	0.3824	0.3857	0.2931	0.2265
mean	0.0304	0.0311	0.033	0.0396	0.0499	0.0608	0.0715	0.0738	0.0678	0.0591	0.0515	0.0382	0.03	0.0198	0.0143	0.0108	0.0084	0.0068	0.0055	0.0039
std	0.085	0.0884	0.099	0.128	0.1542	0.1829	0.2191	0.2259	0.2036	0.1681	0.1412	0.0969	0.074	0.0509	0.0373	0.0272	0.0213	0.0176	0.014	0.0098
skewness	8.2602	8.5575	10.0383	11.3257	10.0677	7.6269	8.5839	8.8185	8.9577	7.8704	6.9268	6.591	6.028	7.8387	8.7858	6.9784	6.2945	6.8688	6.221	5.9755
kurtosis	120.298	128.9357	190.275	242.2775	208.1898	82.9457	117.0851	124.5151	131.3643	99.5985	68.6172	68.6205	51.906	106.6552	156.0315	87.9723	60.6166	79.6645	60.9231	57.1025

Most parameters show high skewness (>7) and heavy kurtosis, suggesting significant outliers and concentrated distributions around low values. Parameters like **T0pt010S to T0pt100S** have lower mean values, while others (e.g., **T0pt150S to T0pt500S**) show increasing variability.

5.Plots of Actual vs Predicted log10(SA) Across Time Periods:



The scatter plots show actual vs. predicted log10(SA) for the Gradient Boosting model at time period indices 0, 10, and 18. Across all periods, predictions align well with the 1:1 line, indicating strong model performance.

- **Index 0 (short period):** Predictions are tightly clustered around the 1:1 line, showing high accuracy and low variance.
- **Index 10 (mid period):** Slightly more scatter is observed, suggesting increased prediction uncertainty, but alignment remains good.
- **Index 18 (long period):** Similar to index 10, with moderate dispersion and reliable predictions overall.

In summary, the model performs best at short periods and remains accurate, though slightly less precise, at longer periods.

6.Model Architecture:

The model architecture consists of a GradientBoostingRegressor wrapped in a MultiOutputRegressor to handle multiple target variables simultaneously. Key features include:

- GradientBoostingRegressor (GBR):
  - Builds 3000 trees with a learning rate of 0.01 and a max depth of 3 to avoid overfitting.
  - Uses 80% of the data for training each tree and limits the minimum number of samples per leaf to 10 for better generalization.
  - The model minimizes Mean Squared Error (MSE).
- MultiOutputRegressor: Handles multiple target variables by training separate models for each target while sharing the same base estimator.

This architecture is efficient for multi-output regression tasks, where multiple related continuous targets are predicted simultaneously, and it prevents overfitting through careful hyperparameter tuning.

7.Model Performance Metrics for Target Variables:

- **R^2**: Values between **0.8890 and 0.9166** suggest strong predictive power across all targets, with the model explaining a significant portion of variance in the target variables.
- **Inter-Std (τ)**: The inter-event standard deviation decreases from **0.4898 to 0.3534**, indicating less variability in the model's predictions for higher target values, meaning more consistency.
- **Intra-Std (φ)**: The intra-event standard deviation also decreases from **0.6348 to 0.5694**, showing that the model becomes more precise for individual events as the target increases.
- **Total Std**: The combined standard deviation decreases from **0.8018 to 0.6701**, indicating overall reduced uncertainty as the model predicts larger target values.

The model performs well with high R^2, and its predictive uncertainty decreases for higher targets, showing better precision and stability as it predicts larger values.

Target Variable	R^2	Inter-Std (τ)	Intra-Std (φ)	Total Std
T0pt010S	0.9098	0.4898	0.6348	0.8018
T0pt020S	0.9093	0.4921	0.6373	0.8052
T0pt030S	0.908	0.4999	0.644	0.8153
T0pt050S	0.9024	0.5221	0.6731	0.8518
T0pt075S	0.8939	0.5441	0.7192	0.9018
T0pt100S	0.889	0.5591	0.7488	0.9346
T0pt150S	0.8918	0.54	0.7442	0.9195

T0pt200S	0.8964	0.5206	0.7261	0.8934
T0pt300S	0.9052	0.4966	0.6788	0.841
T0pt400S	0.9098	0.4954	0.648	0.8157
T0pt500S	0.9107	0.4866	0.6327	0.7982
T0pt750S	0.9063	0.4666	0.625	0.78
T1pt000S	0.9004	0.4595	0.6307	0.7804
T1pt500S	0.8914	0.4358	0.6447	0.7782
T2pt000S	0.8903	0.4244	0.6426	0.7702
T2pt500S	0.8933	0.4149	0.6342	0.7579
T3pt000S	0.8974	0.4048	0.6249	0.7446
T3pt500S	0.9018	0.4001	0.6143	0.7331
T4pt000S	0.907	0.3804	0.5983	0.709
T5pt000S	0.9166	0.3534	0.5694	0.6701

8.Residual Analysis:

Inter-event Residual vs Magnitude (Top Row)

- Across all periods (0.1s, 1.0s, 3.0s), the inter-event residuals show no strong trend with magnitude (Mw), indicating that the model captures magnitude scaling well.
- The mean residuals are generally close to zero with moderate spread, showing unbiased event-specific performance.

Intra-event Residual vs Rjb (Middle Row)

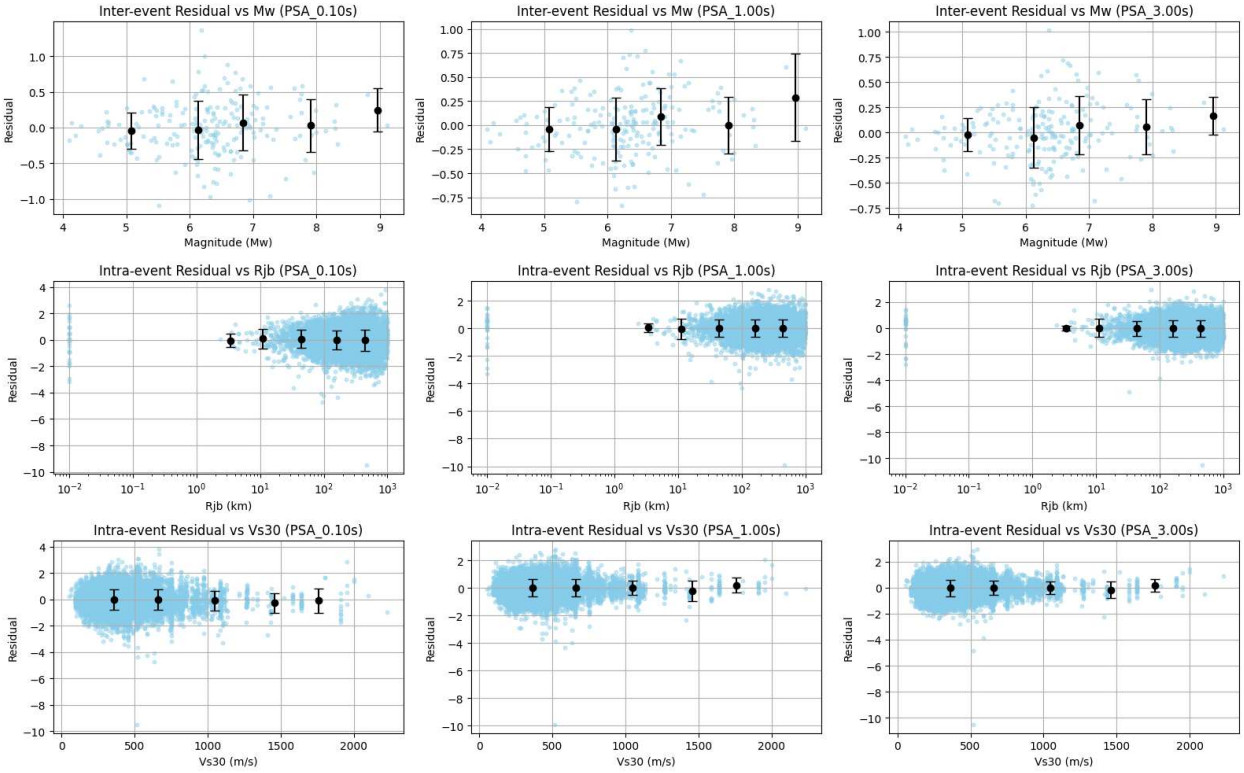
- Residuals slightly decrease with increasing distance (Rjb), especially beyond ~10 km, suggesting mild underprediction at farther distances.
- The variability is larger at shorter distances but reduces at greater distances, which is typical in ground motion models due to signal attenuation.

Intra-event Residual vs Vs30 (Bottom Row)

- Residuals show a negative trend with Vs30, particularly for lower Vs30 values (< 1000 m/s), indicating underprediction at soft sites.
- This trend weakens at higher Vs30 values, suggesting the model is more accurate for stiffer sites.

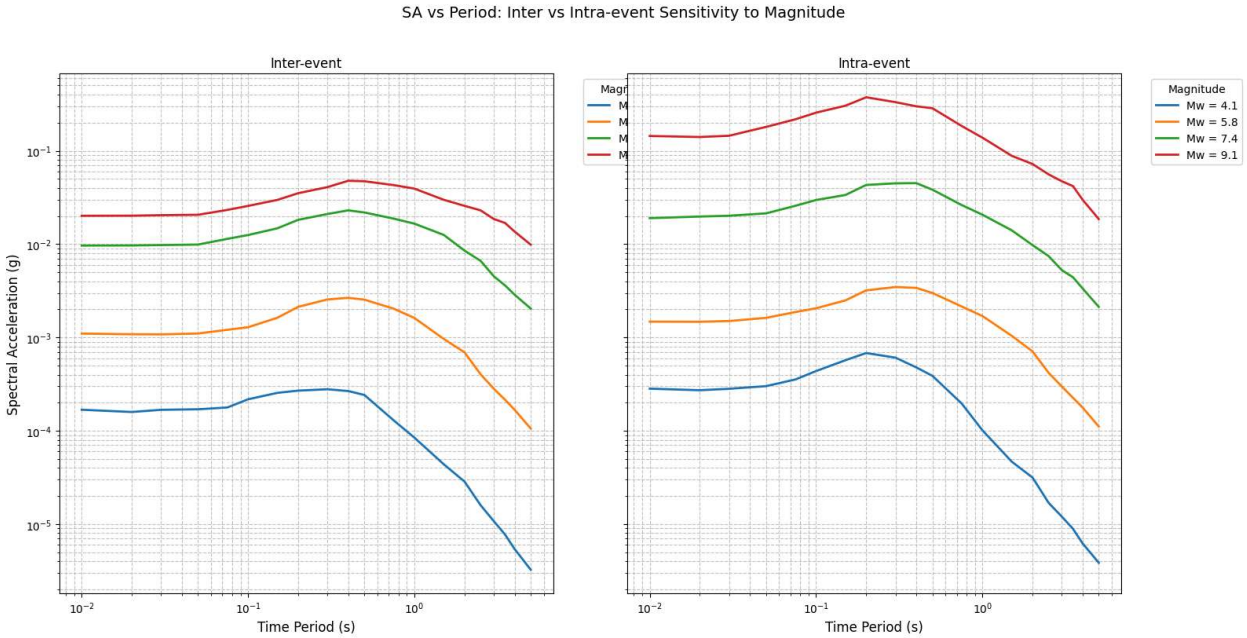
Summary

The Gradient Boosting model performs robustly with respect to magnitude but shows minor biases with distance and site conditions, especially underpredicting for soft soils and at greater distances.

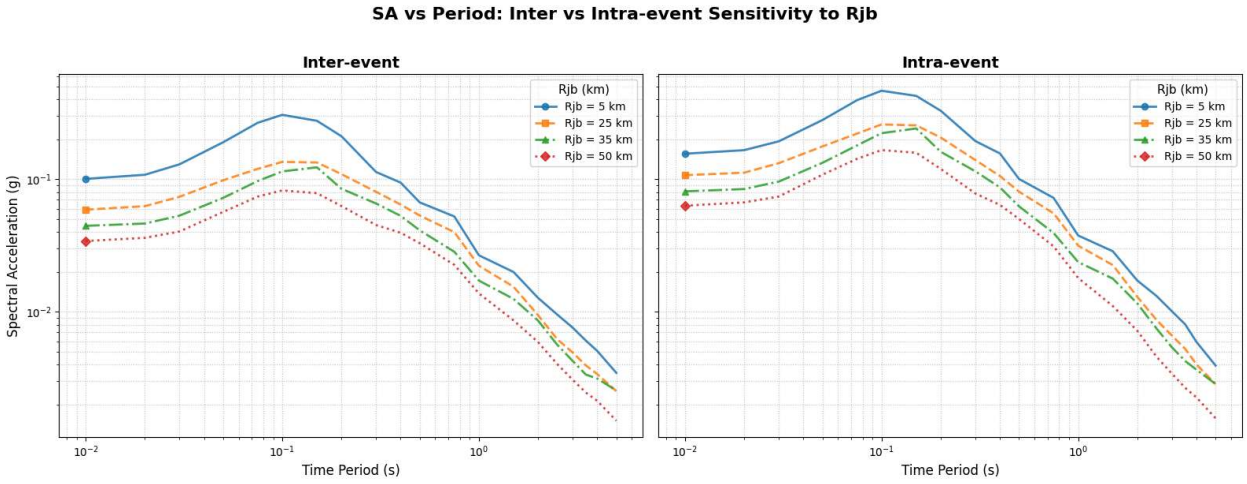


9.Magnitude Sensitivity Plot:

- **Magnitude Sensitivity:** Both inter- and intra-event components show increasing SA with magnitude, especially at longer periods (Mw 7.4 and Mw 9.1).
- **Period Dependence:** SA peaks between 0.1s and 0.5s for most magnitudes, with faster decay at higher frequencies for smaller magnitudes (e.g., Mw 4.1).
- **Inter vs Intra-event Comparison:** Intra-event SA is higher than inter-event SA at most periods, indicating dominant site/path effects. The intra-event plot shows clearer magnitude separation, suggesting better modeling of within-event scaling.



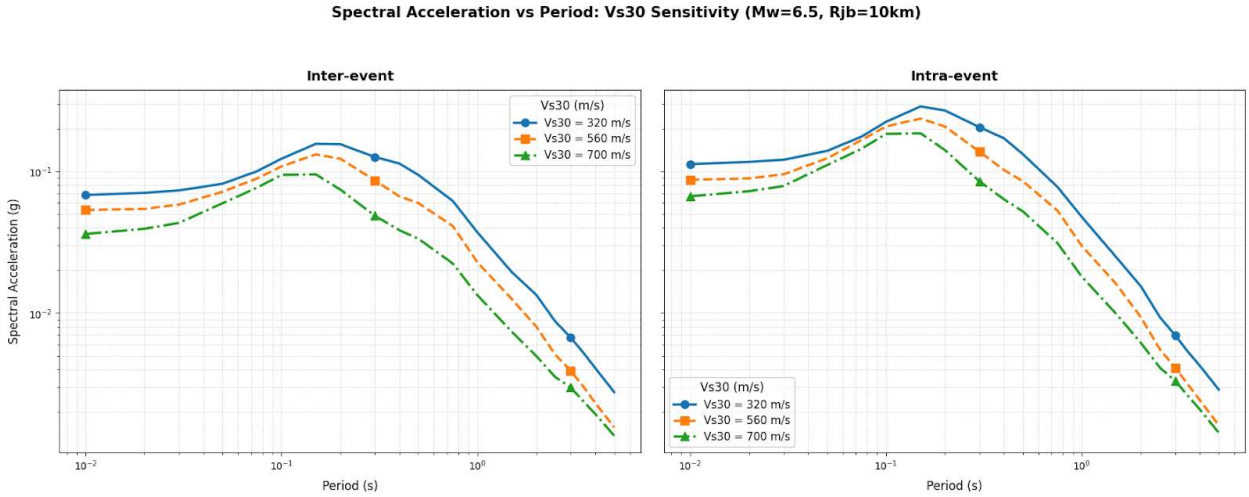
10.Rjb Sensitivity Plot



- **Distance Sensitivity (Rjb):** Both inter- and intra-event components show a clear dependency on the Rjb distance, with spectral acceleration (SA) decreasing as distance from the rupture (Rjb) increases.
- **Period Dependence:** At shorter periods (0.01s - 0.1s), SA is highest for Rjb = 5 km. At longer periods (greater than 0.1s), the curves for different Rjb values begin to converge, indicating reduced sensitivity to distance at higher periods.
- **Inter vs Intra-event Comparison:** The intra-event plot shows a more pronounced effect of distance on SA, with clearer differences between Rjb values, especially at shorter periods. In the inter-event plot, the distance effect is less distinct across periods, suggesting that intra-event variability has a stronger impact on the SA response.

Spectral acceleration decreases with increasing distance (Rjb), particularly at shorter periods. The intra-event variability exhibits a stronger distance dependence, while the inter-event component shows a less pronounced distance effect. The model captures the expected behavior of spectral acceleration with respect to both Rjb and period.

11.Vs30 Sensitivity Plot:



Vs30 Sensitivity (Mw=6.5, Rjb=10 km)

- SA decreases with increasing Vs30, indicating stronger site amplification for softer soils (lower Vs30).
- Sensitivity is most pronounced at mid-periods (0.1–1s), where differences between Vs30 values are largest.
- At short and long periods, the curves converge, showing reduced Vs30 influence.

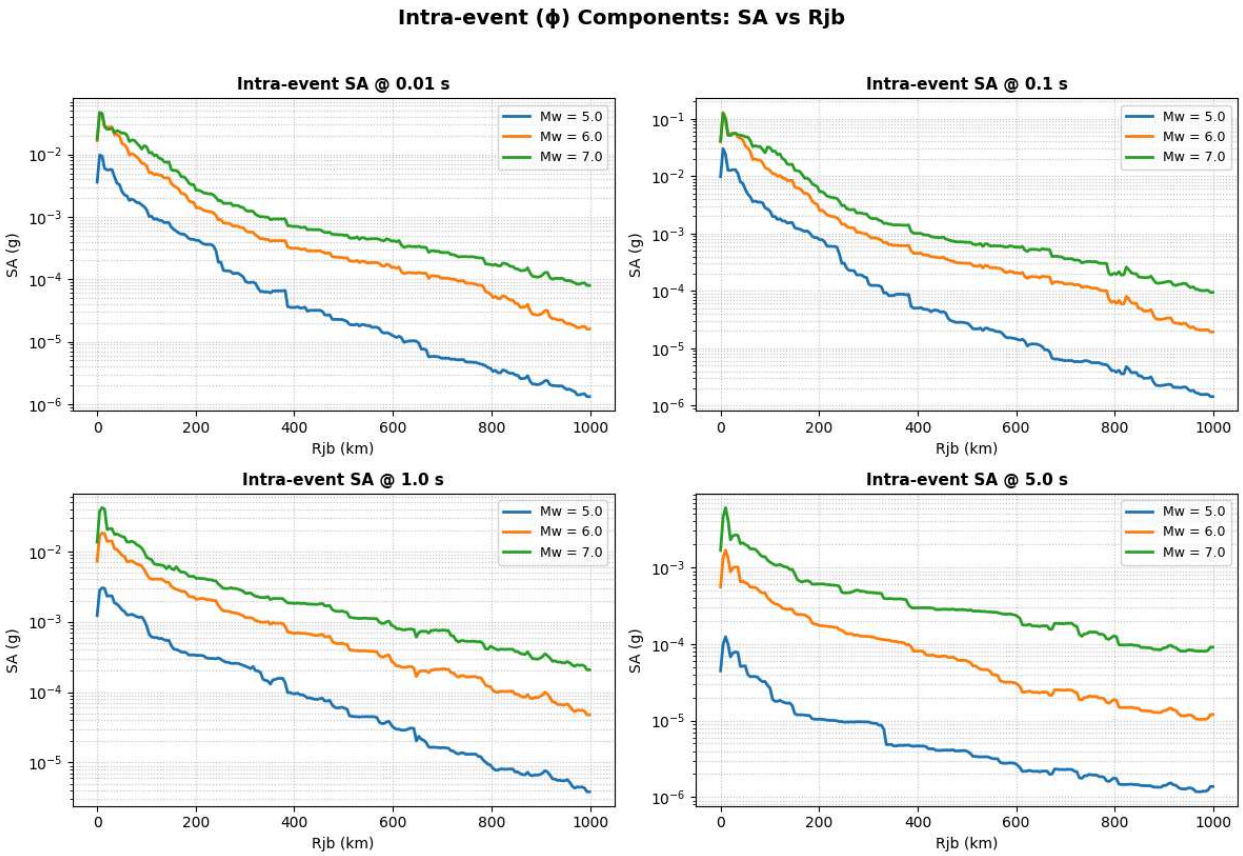
Inter vs Intra-event:

- Intra-event plot shows clearer separation between Vs30 curves, indicating stronger site-specific effects.
- Inter-event plot exhibits milder Vs30 sensitivity, suggesting less event-to-event variation related to site conditions.

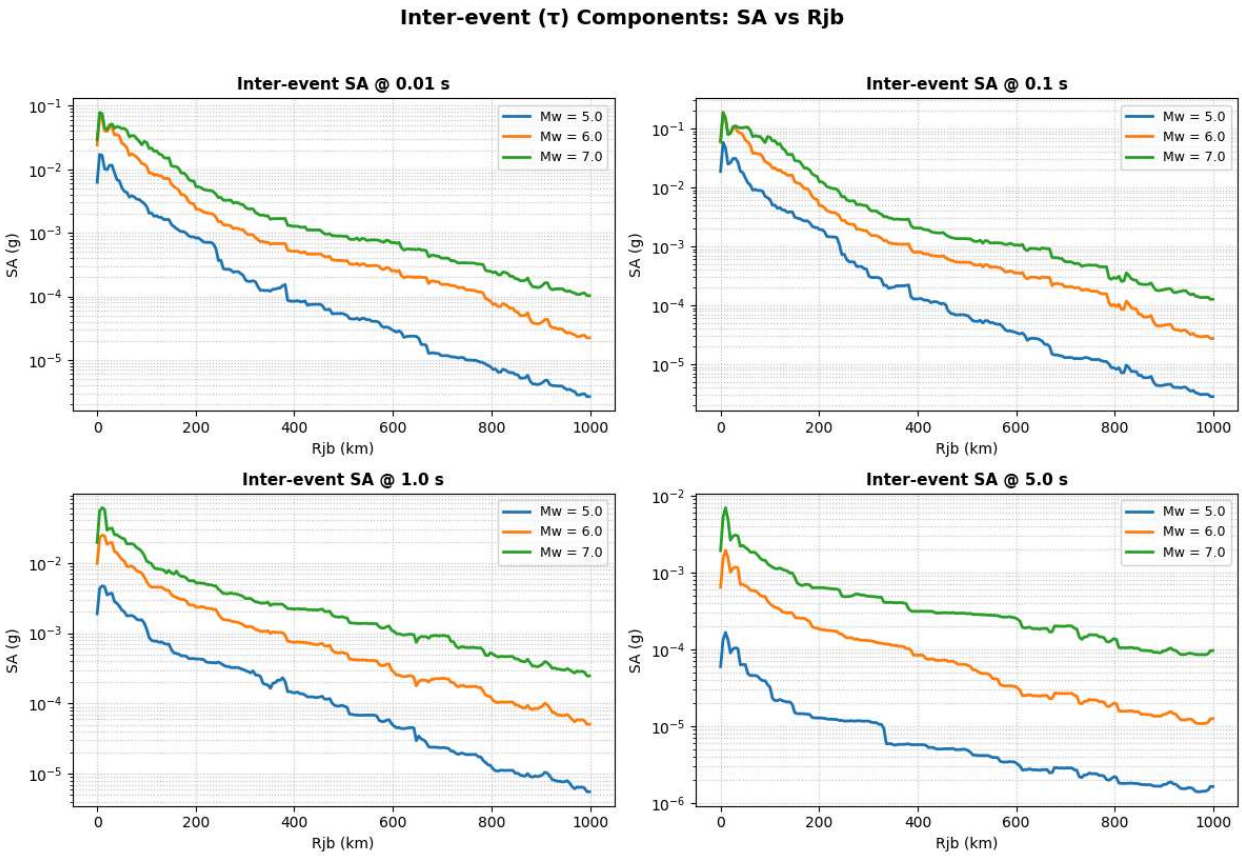
The model captures expected behavior—lower Vs30 amplifies SA more, especially within individual events.

12.SA @ T vs Rjb:

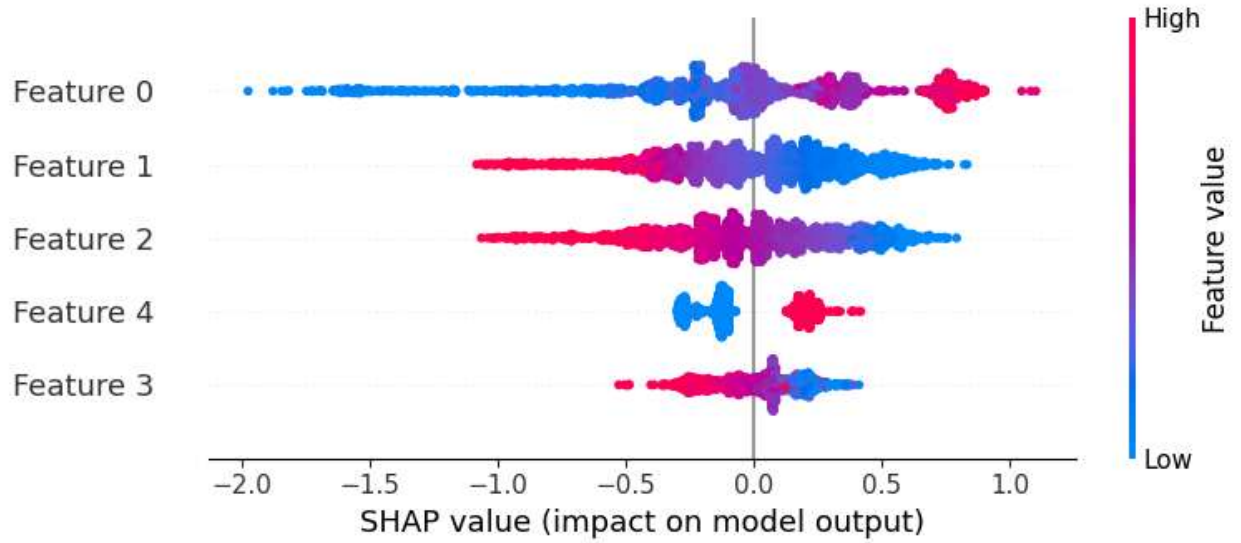
- **Intra-event (ϕ) Components: SA vs Rjb**
  - Spectral acceleration (SA) decreases with increasing rupture distance (Rjb) across all periods. For a given Rjb, SA increases with earthquake magnitude (Mw). Intra-event variability is higher at shorter periods and lower for distant sites.

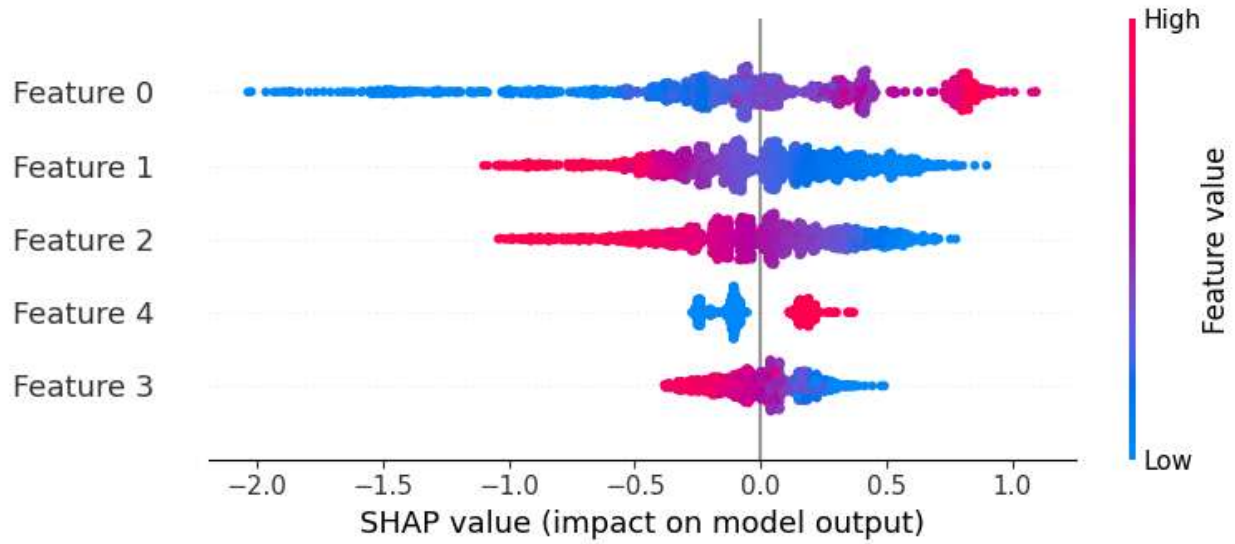


- **Inter-event (τ) Components: SA vs Rjb**
  - Similar to intra-event trends, SA decreases with Rjb and increases with Mw. However, these plots capture event-to-event variability. The separation between magnitudes is more consistent across distances, showing systematic inter-event differences.



13.SHAP Analysis Summary





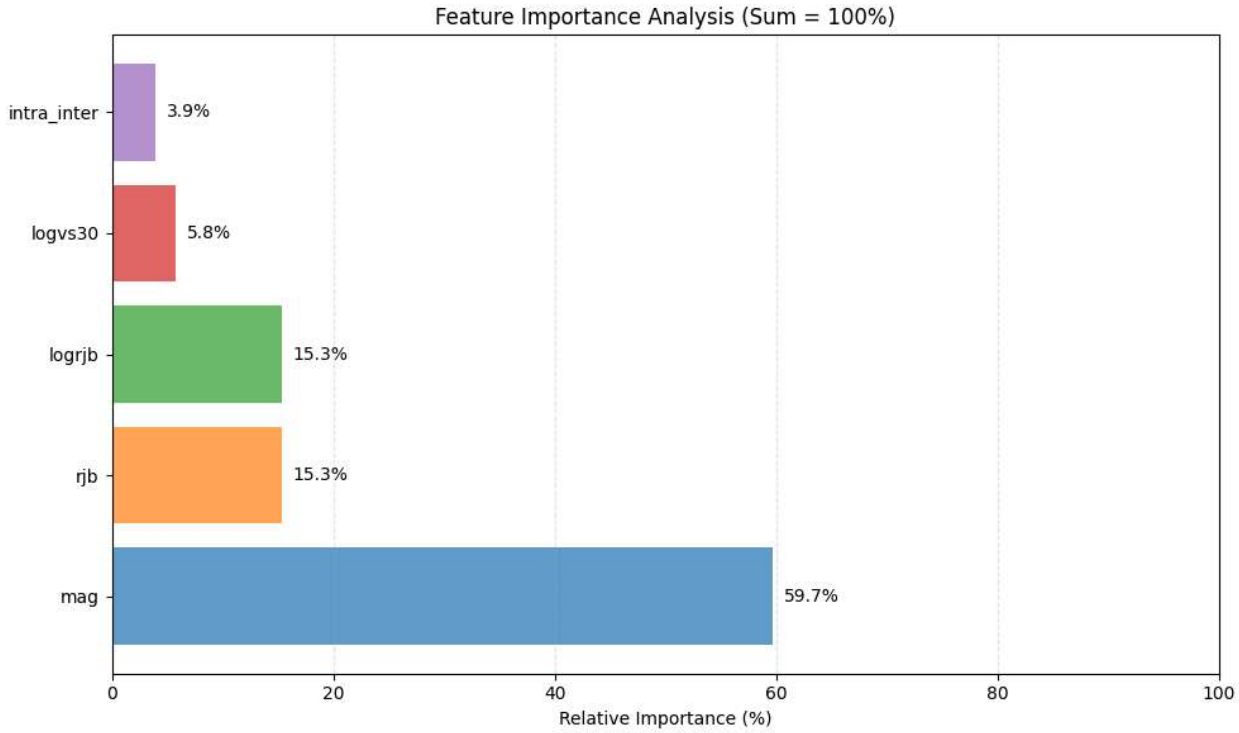
Each point represents one instance (data sample), showing:

- X-axis: SHAP value — how much that feature affected the model's output (positive or negative impact).
- Color: Feature value (blue = low, red = high).
- Y-axis: Features, ranked by mean importance.

Observations:

- Feature 0:
  - Has the highest influence on the model.
  - Higher values (red) tend to increase the prediction (right of center).
  - Lower values (blue) tend to decrease the prediction (left of center).
- Feature 1:
  - Has a moderate influence.
  - High values reduce the prediction (concentrated on the left side).
- Feature 2:
  - Similar behavior to Feature 1, with a mix of impact depending on the value.
- Feature 3 and 4:
  - Have the least impact.
  - Influence is mostly centered near zero, indicating low contribution to predictions.

14.Feature Importance Summary:



- **Magnitude (mag)**: Most significant feature, contributing **59.7%** to the model predictions.
- **Rupture distance (rjb)**: Contributes **15.3%.**
- **Log-transformed rupture distance (logrjb)**: Also contributes **15.3%**, highlighting redundancy or complementary effect.
- **Site condition (logvs30)**: Moderate influence with **5.8%** importance.
- **Event type (intra_inter)**: Least influential, contributing only **3.9%.**

Code: [Gradient Boosting Model]