

# Earthquake Prediction using XAI

---

Rishabh Jain

IIT Delhi

Research Intern

National University of Singapore

Guide: Vishal Srivastava

National University of Singapore

| 8th August 2024

# Outline

---

1. Tree-based Models
2. Deep Learning Models
3. Conclusion

---

# Tree-based models

# Experiment

- Extracted the following features from each data point i.e. array of 500 signal values: mean, standard deviation, min, max, median, 25th Percentile, 75th Percentile
- Results:

XGBoost Results:  
Train Accuracy: 0.9095723898415454  
Test Accuracy: 0.8309963099630996

Classification Report (Test Data):

	precision	recall	f1-score	support
0	0.86	0.94	0.90	3232
1	0.64	0.39	0.49	833
accuracy			0.83	4065
macro avg	0.75	0.67	0.69	4065
weighted avg	0.81	0.83	0.81	4065

LightGBM Results:  
Train Accuracy: 0.859387887996527  
Test Accuracy: 0.8337023370233703

Classification Report (Test Data):

	precision	recall	f1-score	support
0	0.85	0.96	0.90	3232
1	0.69	0.34	0.45	833
accuracy			0.83	4065
macro avg	0.77	0.65	0.68	4065
weighted avg	0.82	0.83	0.81	4065

Mean: 0.06737364083528519  
Std Dev: 0.4041627049446106  
Min: 0.11716023832559586  
Max: 0.10022184997797012  
Median: 0.10996411740779877  
25th Percentile: 0.09498073905706406  
75th Percentile: 0.1061367467045784

Mean: 228  
Std Dev: 594  
Min: 579  
Max: 520  
Median: 419  
25th Percentile: 330  
75th Percentile: 330

# Experiment

## Decision Tree Results:

Train Accuracy: 1.0

Test Accuracy: 0.7943419434194342

## Classification Report (Test Data):

	precision	recall	f1-score	support
0	0.87	0.87	0.87	3232
1	0.50	0.50	0.50	833
accuracy			0.79	4065
macro avg	0.68	0.68	0.68	4065
weighted avg	0.79	0.79	0.79	4065

## SVM Results:

Test Accuracy: 0.795079950799508

## Classification Report (Test Data):

	precision	recall	f1-score	support
0	0.80	1.00	0.89	3232
1	0.00	0.00	0.00	833
accuracy			0.80	4065
macro avg	0.40	0.50	0.44	4065
weighted avg	0.63	0.80	0.70	4065

## Gradient Boosting Results:

Test Accuracy: 0.813530135301353

## Classification Report (Test Data):

	precision	recall	f1-score	support
0	0.83	0.96	0.89	3232
1	0.61	0.25	0.36	833
accuracy			0.81	4065
macro avg	0.72	0.60	0.62	4065
weighted avg	0.79	0.81	0.78	4065

## CatBoost Results:

Test Accuracy: 0.8130381303813038

## Classification Report (Test Data):

	precision	recall	f1-score	support
0	0.83	0.96	0.89	3232
1	0.62	0.23	0.34	833
accuracy			0.81	4065
macro avg	0.72	0.60	0.62	4065
weighted avg	0.79	0.81	0.78	4065

# Random Forest

- Random Forest was the **best** performing tree-based classifier
- Yielded highest f1-score as well as a highest test accuracy
- Model architecture: 100 trees
- Results:

## Random Forest Results:

Train Accuracy: 0.9999131756023443  
Test Accuracy: 0.8378843788437884

## Classification Report (Test Data):

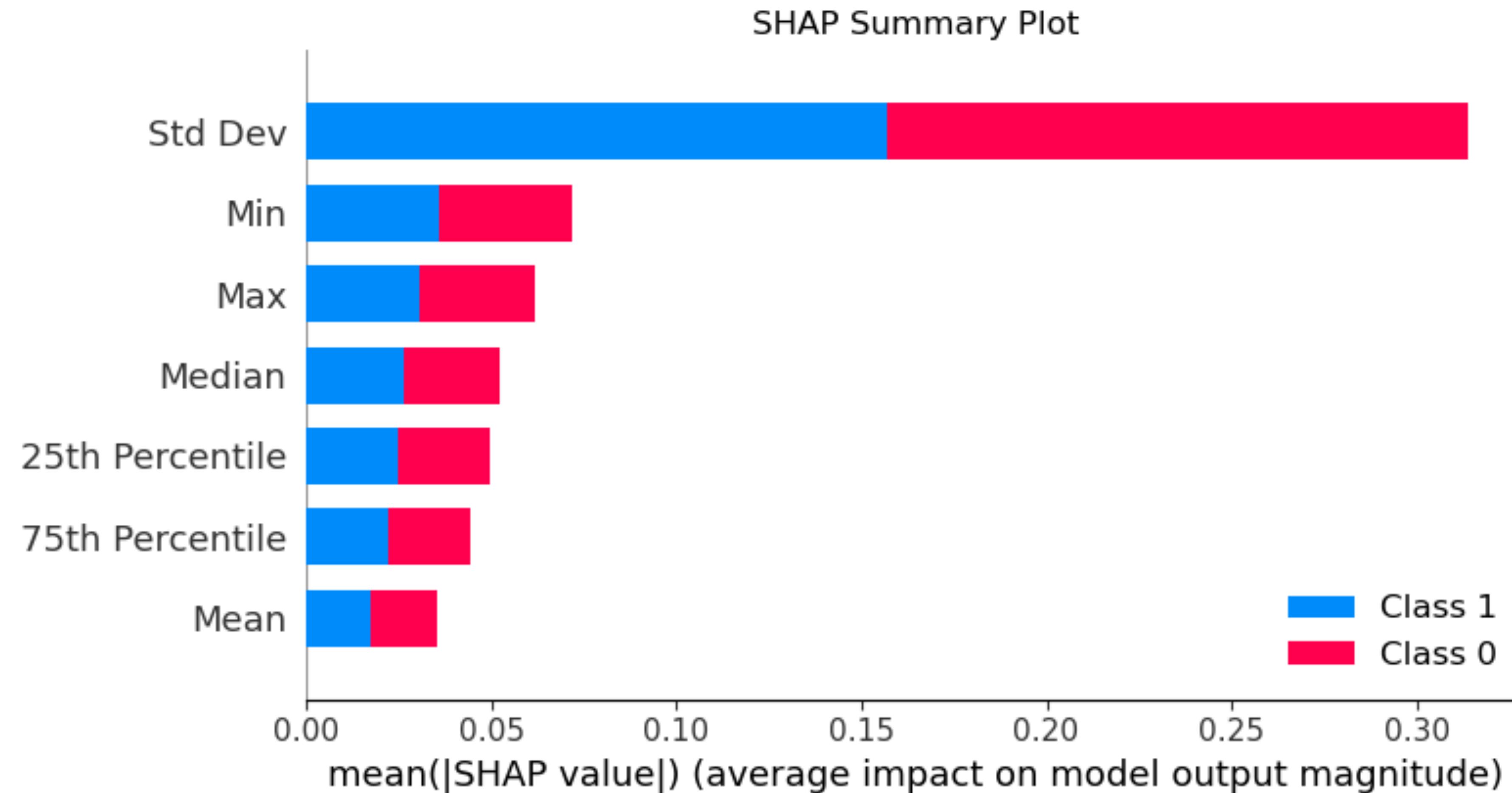
	precision	recall	f1-score	support
0	0.87	0.93	0.90	3232
1	0.64	0.48	0.55	833
accuracy			0.84	4065
macro avg	0.76	0.70	0.72	4065
weighted avg	0.83	0.84	0.83	4065

## Feature Importance scores:

Mean: 0.10695665950830259  
Std Dev: 0.2954301652002547  
Min: 0.1281152413886958  
Max: 0.12803158230704936  
Median: 0.11764776119316997  
25th Percentile: 0.1074469192637853  
75th Percentile: 0.1163716711387423

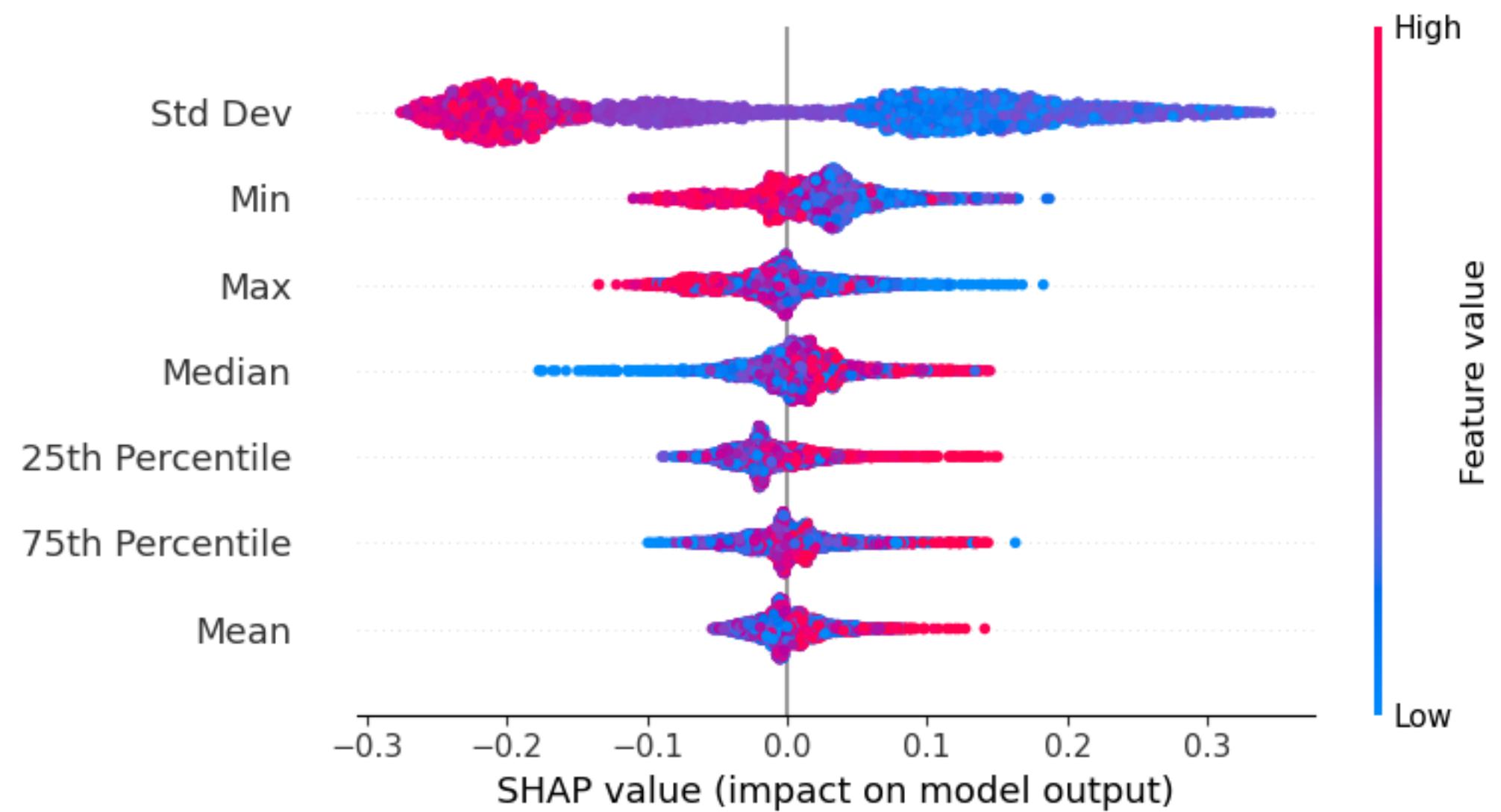
# Shap Plots

---

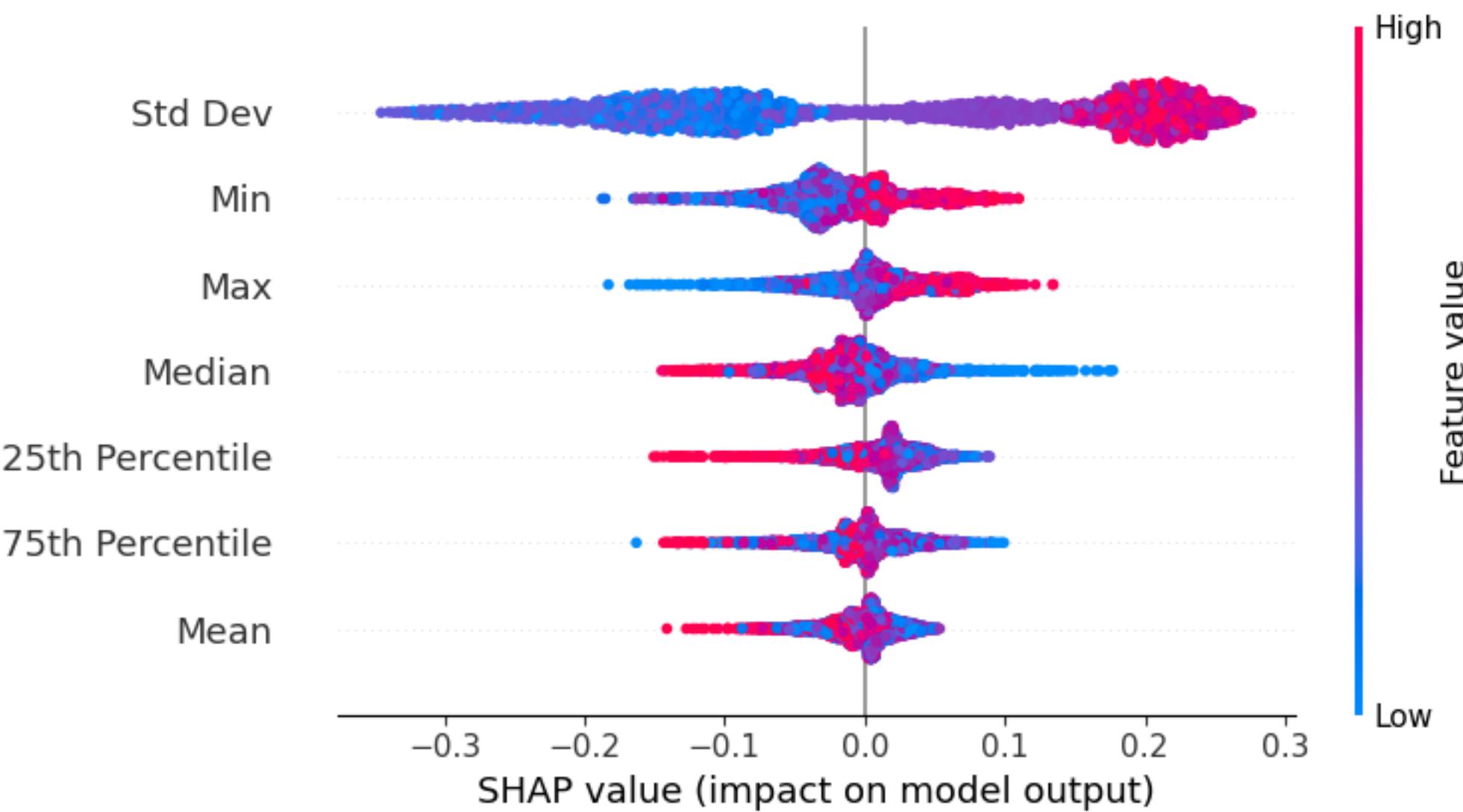


# Shap Plots

Shap Values for positive class (Class 1)



Shap values for negative class (Class 0)



- Standard deviation (std) has highest impact on predicting target values; where high std values (red) strongly push prediction towards class 0 while **low std values (blue) push prediction towards toward class 1** (earthquake occurs).
- Min and Max values have moderate impacts with lower values (blue) pushing prediction towards class 1.

# More Features

- Since prediction showed high correlation with standard deviation, added **kurtosis** and **skewness** metrics as new features in the random forest classifier model.
- This led to an increase in Precision (Class 0: +3% & Class 1: +8%) and Recall (Class 0: +1% & Class 1: +11%) as well as overall test accuracy (+3%).

**Random Forest Results:**

Train Accuracy: 1.0

Test Accuracy: 0.8691266912669127

**Classification Report (Test Data):**

	precision	recall	f1-score	support
0	0.90	0.94	0.92	3232
1	0.72	0.59	0.65	833
accuracy			0.87	4065
macro avg	0.81	0.77	0.78	4065
weighted avg	0.86	0.87	0.86	4065

**Feature Importance scores:**

Mean: 0.055970161319407366

Std Dev: 0.2925182213256333

Min: 0.058535206696417326

Max: 0.06529345492605461

Median: 0.05595424392442817

25th Percentile: 0.05276260462111531

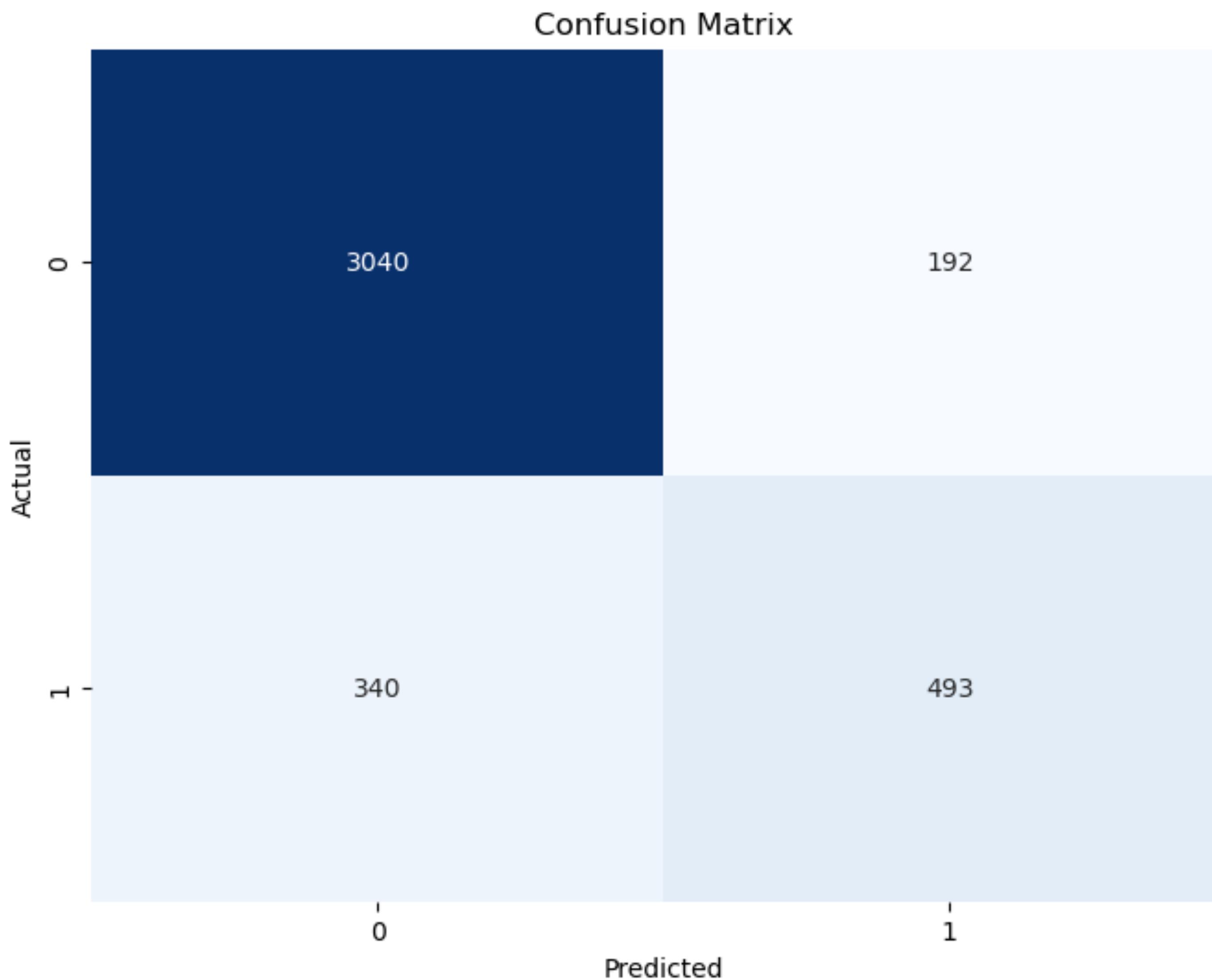
75th Percentile: 0.05907783415489771

Kurtosis: 0.16104207454625677

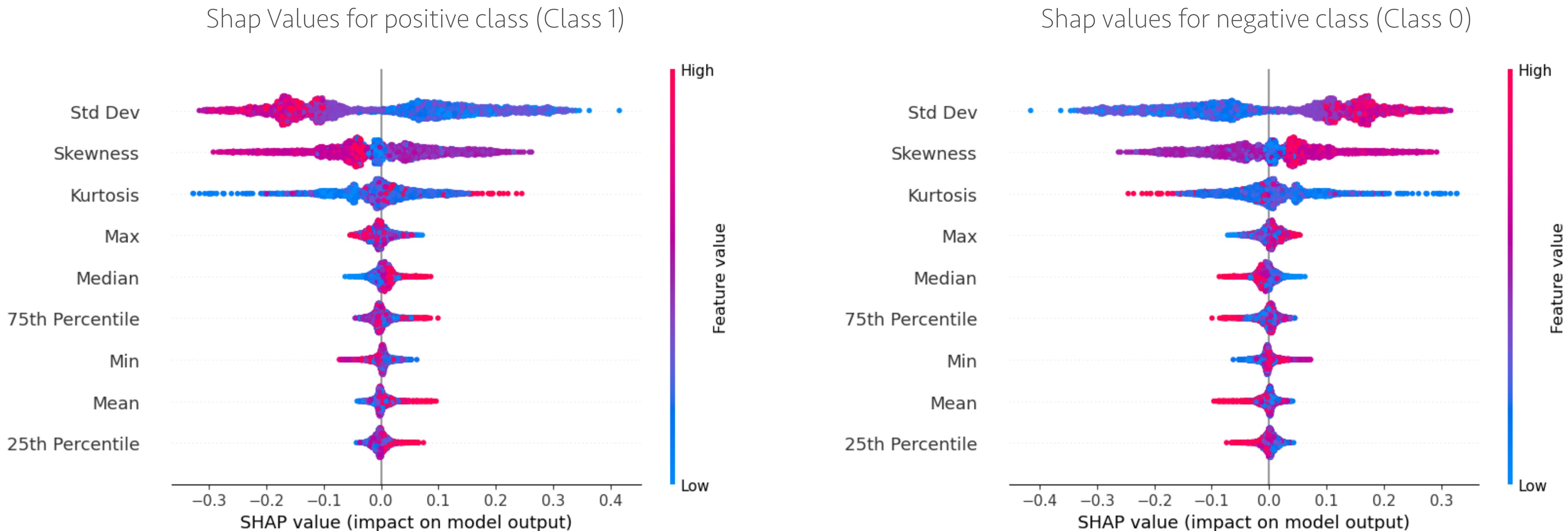
Skewness: 0.1988461984857895

# More Features

---



# More Features



- High skewness values are pointing towards class 0, though the correlation is not extremely strong.
- Low kurtosis values in some cases show high positive shap value for class 0. (No earthquake for low kurtosis)

---

# Deep Learning models

# 1D CNN and Bi-LSTM

- Continue to give poor results.
- Tried out various heights and depths for both networks, results continue to remain futile.

1D CNN

119/119 [=====] - 2s 13ms/step				
	precision	recall	f1-score	support
class 0	0.80	0.99	0.88	3010
class 1	0.37	0.02	0.04	784
accuracy			0.79	3794
macro avg	0.58	0.51	0.46	3794
weighted avg	0.71	0.79	0.71	3794

LSTM

119/119 [=====] - 24s 182ms/step				
	precision	recall	f1-score	support
class 0	0.79	1.00	0.88	3010
class 1	0.00	0.00	0.00	784
accuracy			0.79	3794
macro avg	0.40	0.50	0.44	3794
weighted avg	0.63	0.79	0.70	3794

# Transformers

- So far, the best performing model has been a transformer. Here are the results:

**Test Accuracy: 0.8984**

	<b>precision</b>	<b>recall</b>	<b>f1-score</b>
0	0.95	0.92	0.93
1	0.72	0.83	0.77
<b>accuracy</b>			0.9
<b>macro avg</b>	0.84	0.87	0.85
<b>weighted avg</b>	0.91	0.9	0.9

Comparison with Random Forest

- The **recall** for class 1 has drastically improved (+24%)
- The f1-score increased by 20%
- Overall accuracy also increased by 3%

---

# Interpretability

---

# DeepLift

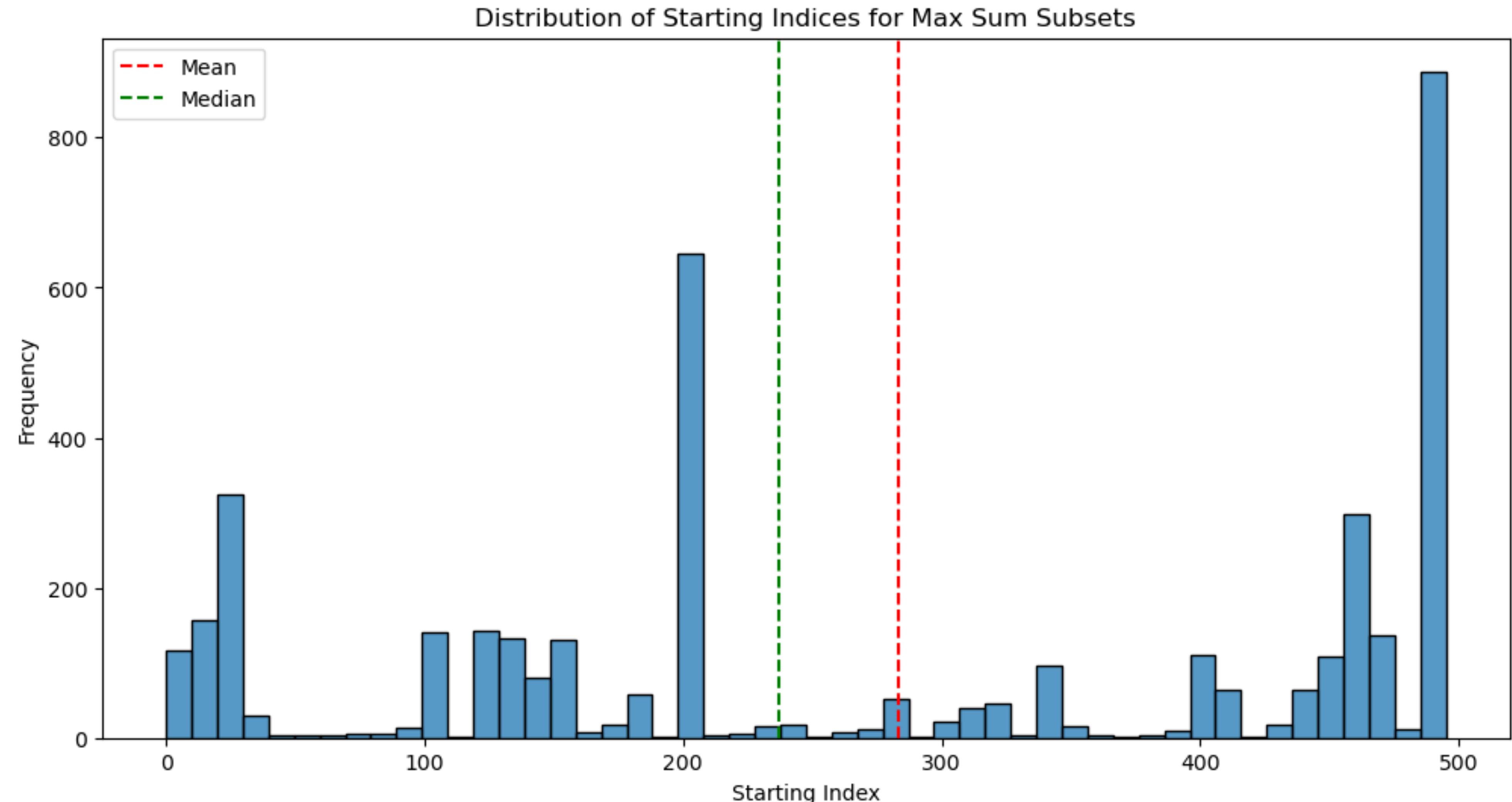
	0	1	2	3	4	5	6	7	8	9	...
Unnamed: 0	0.000000	1.000000	2.000000	3.000000	4.000000	5.000000	6.000000	7.000000	8.000000	9.000000	...
signal_12655_feature_0	0.018635	0.001580	0.058518	0.111104	0.070719	0.012744	0.033324	0.005458	0.108375	0.207180	...
signal_2075_feature_0	0.000065	-0.003093	0.002543	0.002082	-0.003373	-0.001869	-0.010125	-0.001255	-0.002398	-0.003798	...
signal_13200_feature_0	0.001094	0.014676	-0.004718	0.000529	0.002340	-0.000006	-0.000095	0.006663	0.005519	0.009260	...
signal_7233_feature_0	0.286033	-0.279386	-4.685286	0.200493	0.120609	0.009532	0.018782	0.053006	0.078470	0.203260	...
...	...	...	...	...	...	...	...	...	...	...	...
signal_3902_feature_0	-0.007039	0.027481	0.076063	0.001241	0.002870	-0.000188	-0.000467	0.000165	0.001686	0.010322	...
signal_23908_feature_0	-0.000037	-0.002503	0.002725	0.004413	-0.000570	-0.001585	-0.007744	0.000602	-0.000316	0.000517	...
signal_6776_feature_0	-0.000020	-0.000623	0.000441	0.010269	-0.003379	-0.001961	-0.015673	-0.000077	-0.001285	-0.001763	...
signal_18023_feature_0	0.020197	-0.001264	-0.033216	-0.029223	-0.022870	-0.013012	-0.011949	-0.022280	-0.019988	-0.020988	...
signal_25662_feature_0	0.002480	0.002809	0.019012	0.011678	0.003477	0.000548	0.003350	0.005651	0.006281	0.009170	...

4066 rows × 500 columns

- Above is the table of relevance scores derived using DeepLift method, each row represents a test example (total 4065) and each column represents a time step (total 500).
- To identify the key time steps contributing to earthquake prediction for each set of signals (each test datapoint), I chose a **subset of 5 consecutive time steps** (out of 500) which had the maximum average of relevance scores. I have plotted the results on the next slide.

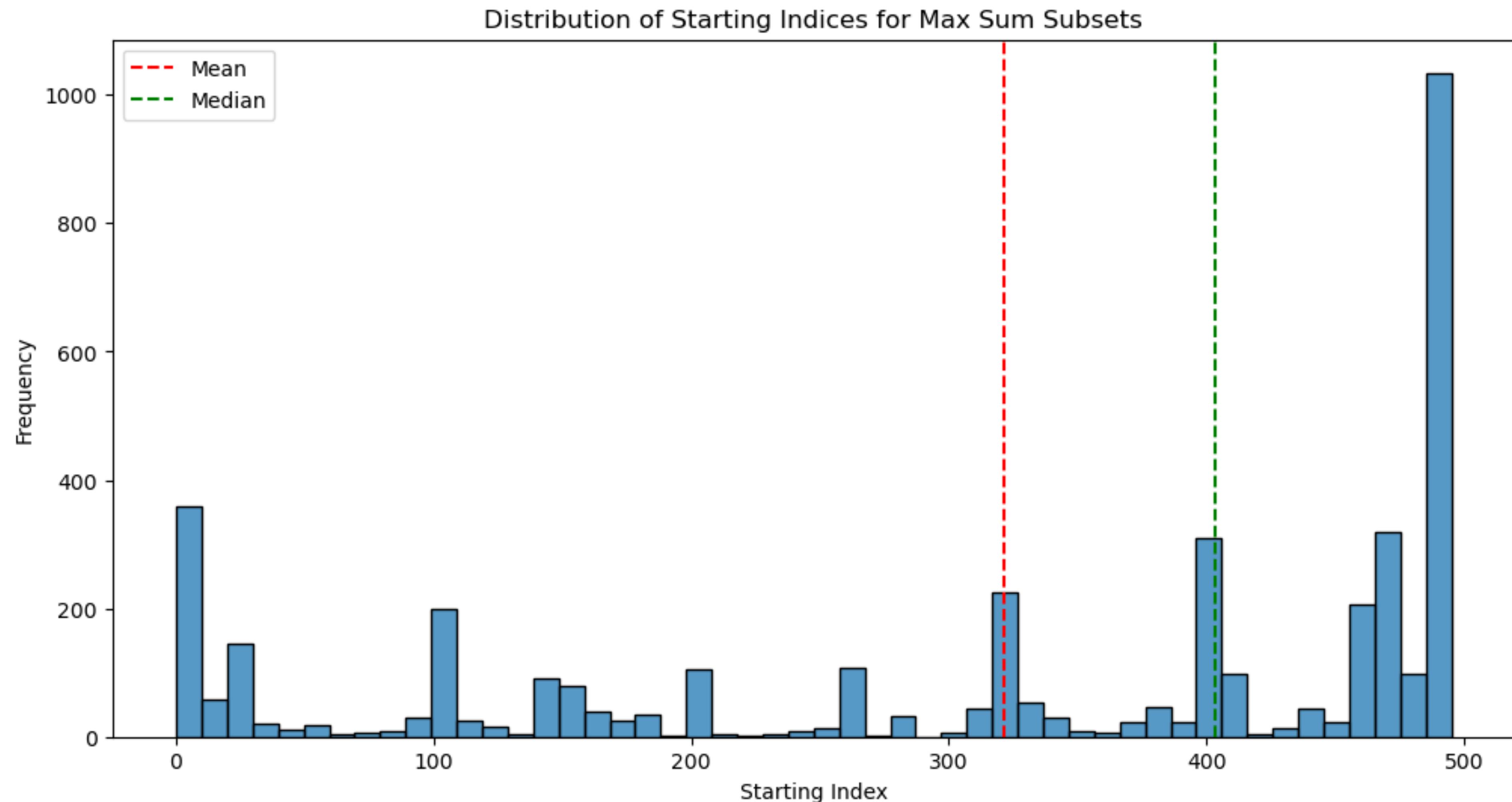
# DeepLift

- **Starting index:** of subsets (length 5) that have the maximum average of relevance score.
- **Frequency:** number of occurrences of each starting index value.
- It is observed that signal values near time step 490 have the greatest impact, followed by those around time step 200, and then time step 20.



# GradSHAP

- This is a similar plot for GradSHAP values.
- As observed for DeepLift, signals around time step 490 are most impactful.
- This is followed by signals around time step 10 and then around 400.



# Conclusion

---

- A reasonable model (Transformer) has been achieved
- The interpretability
- We should employ our model to lab generated data using transfer learning