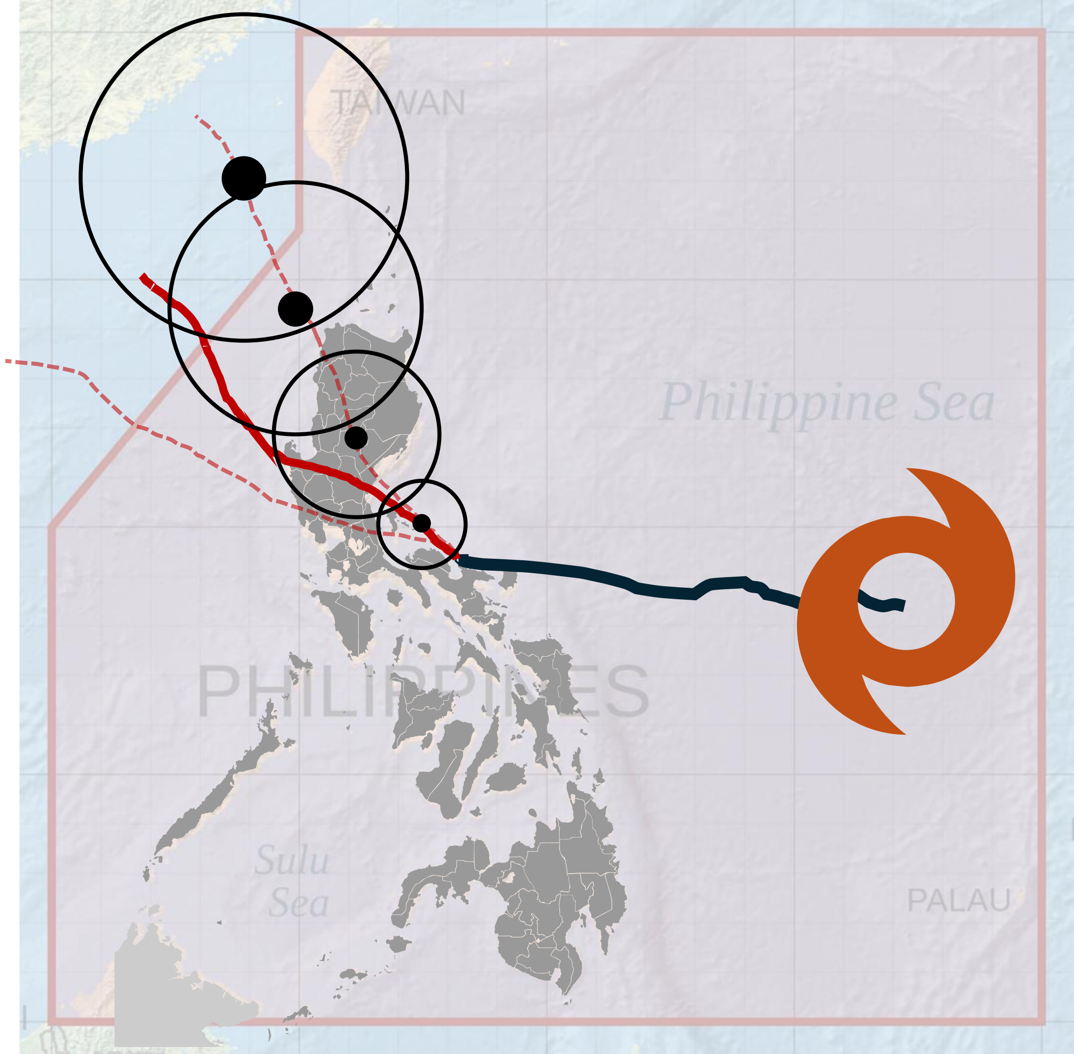
Typhoon Dataset - Time Series Forecasting

Presented by: Francis S. Dela Cruz
September 08, 2025

Table of Contents

Statement of the Problem

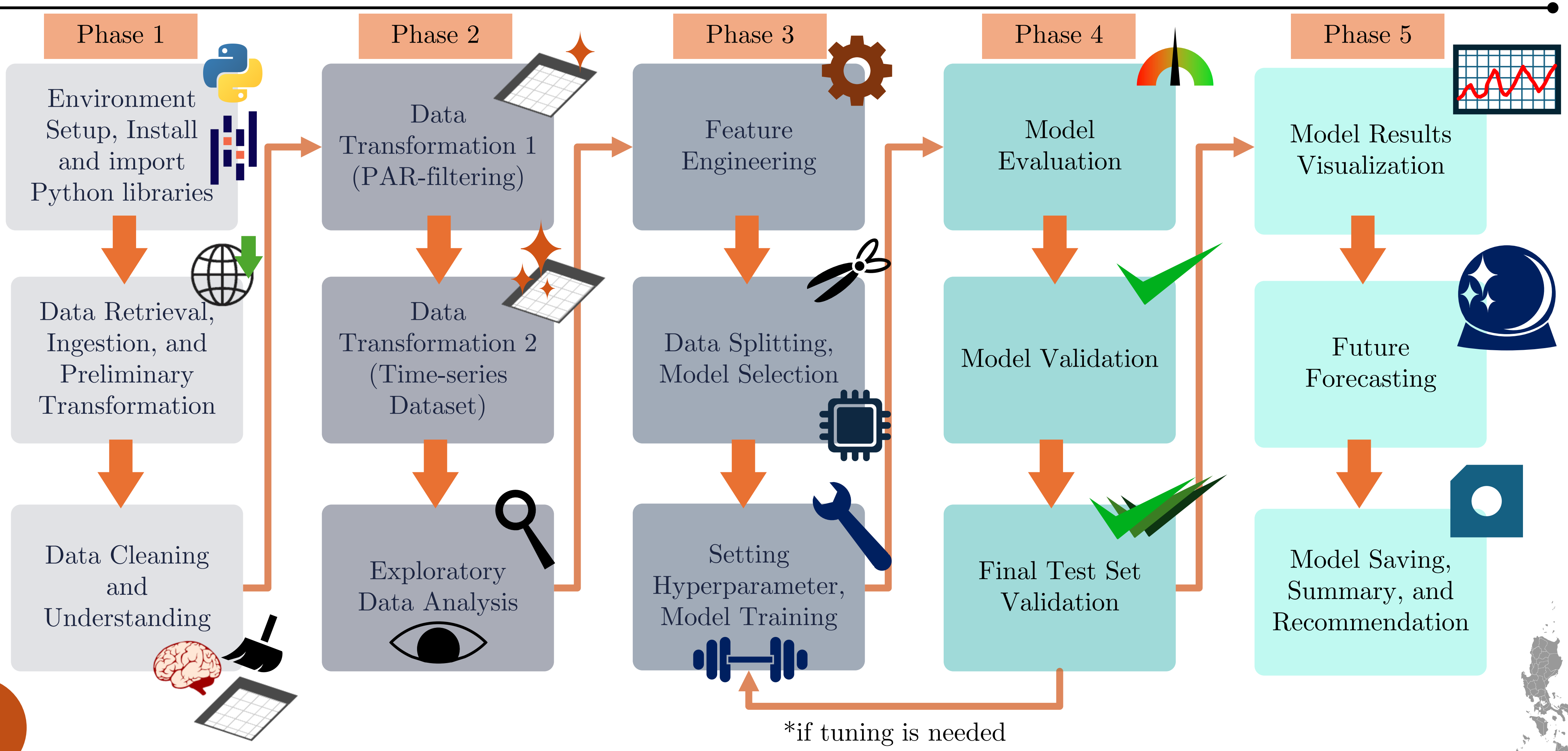


- The Philippines is known to be susceptible to typhoons and resulted to loss of lives and properties, and an estimated economic loss between **12.3B USD – 100B USD** from 2003-2022, and possible higher due to other factors.

- To mitigate and prepare for upcoming calamities, is it possible to predict or estimate the number of typhoons that may affect us within a 2-year window period?

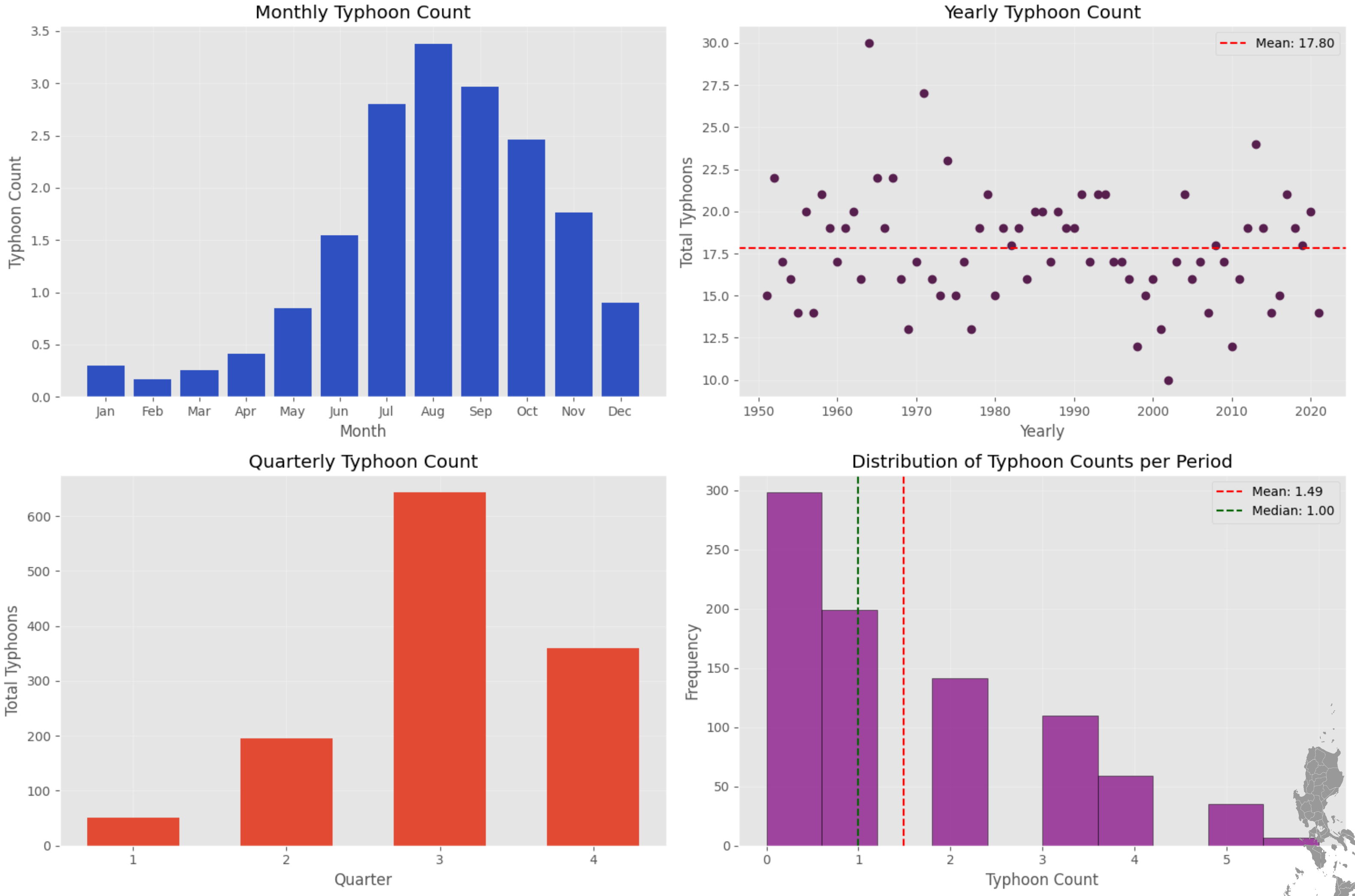
- Tasks: Conduct a Time-Series Forecasting using the RSMC Tokyo-Typhoon Center - Best Track Data https://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/trackarchives.html

Technical Methodology

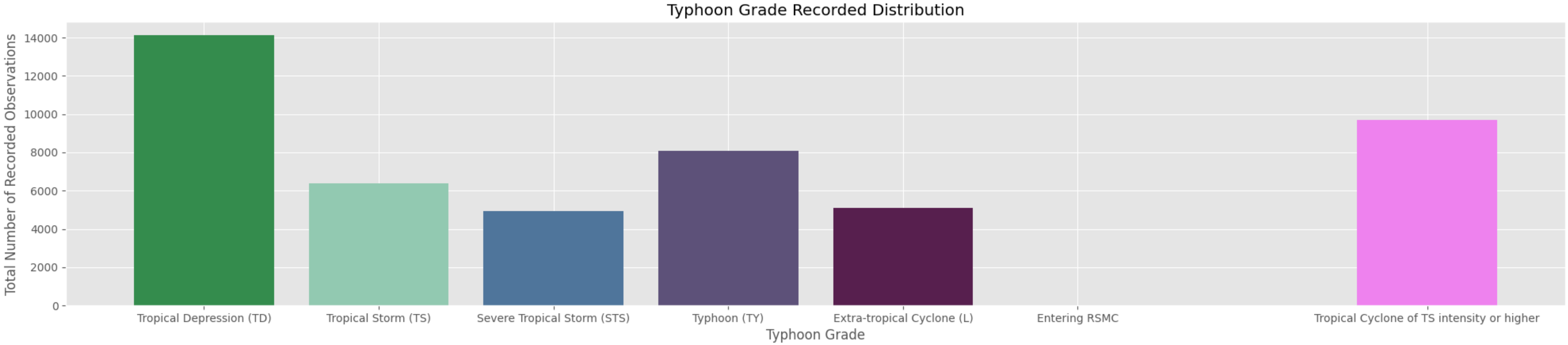


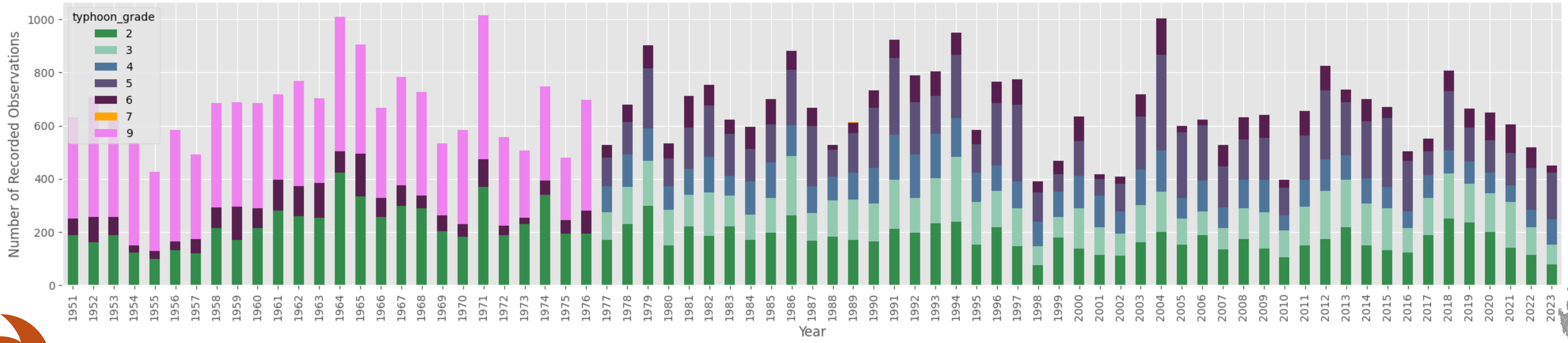
Exploratory Data Analysis

- July – September – peak typhoon season
- Yearly average of around 18 typhoons entering PAR
- Q3 – "Habagat" Season – experiences lots of typhoon
- Monthly average of 1.49 typhoons



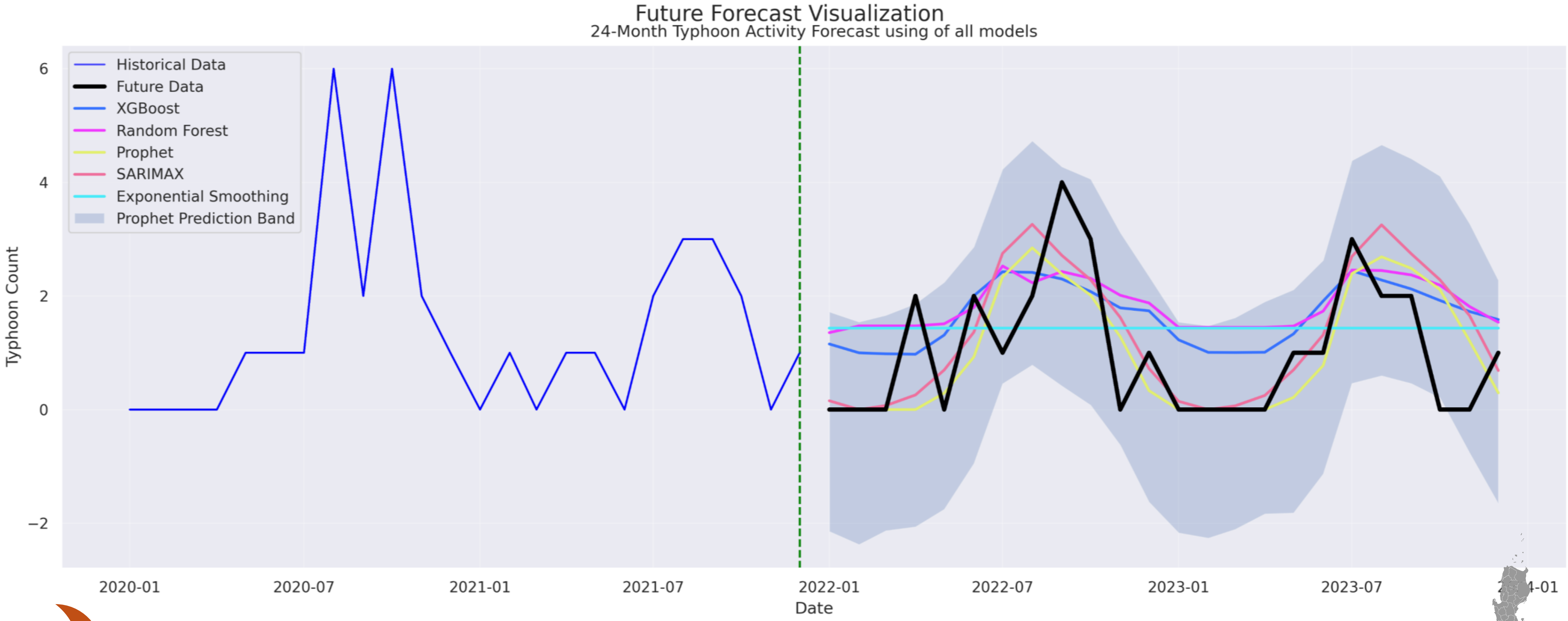
Exploratory Data Analysis (Typhoon Grade Distribution)



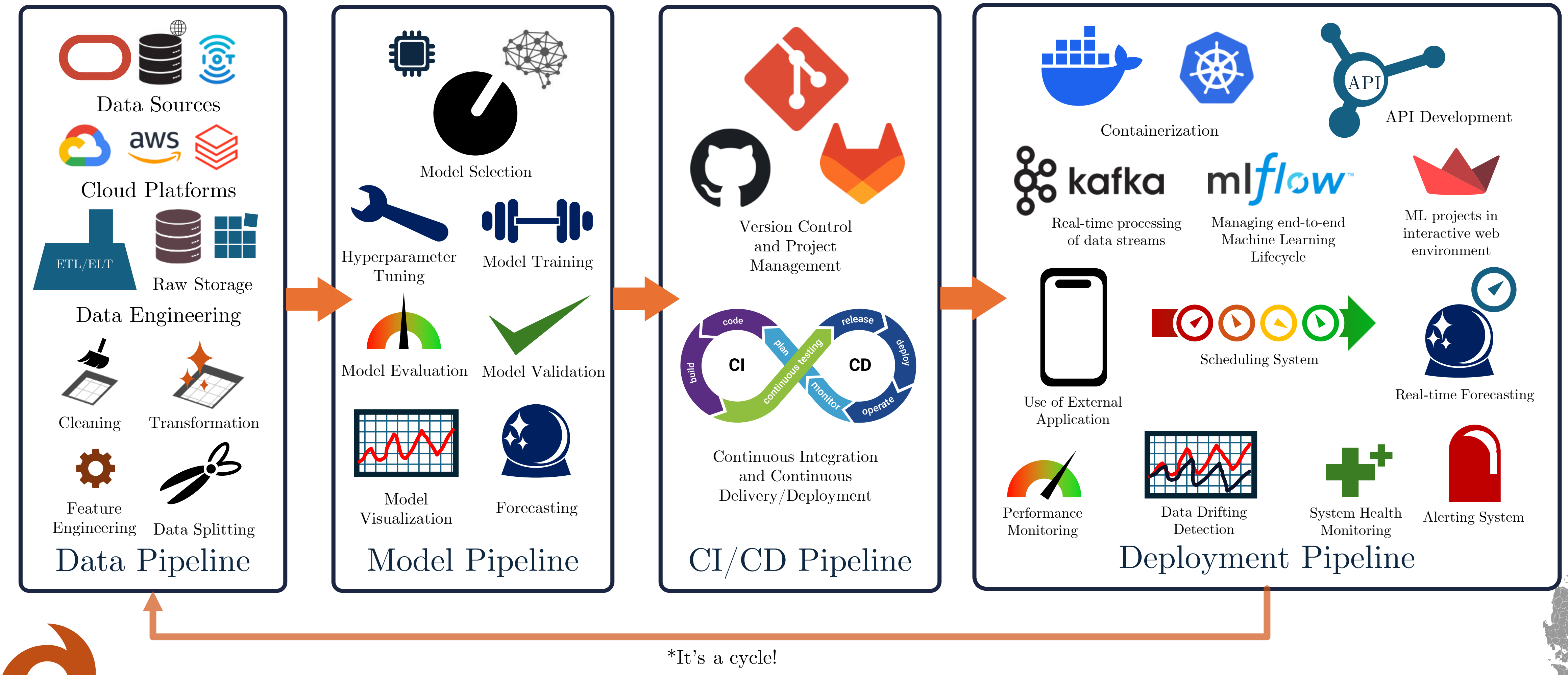


Time Series Forecasting

Results and Discussion (Forecast)

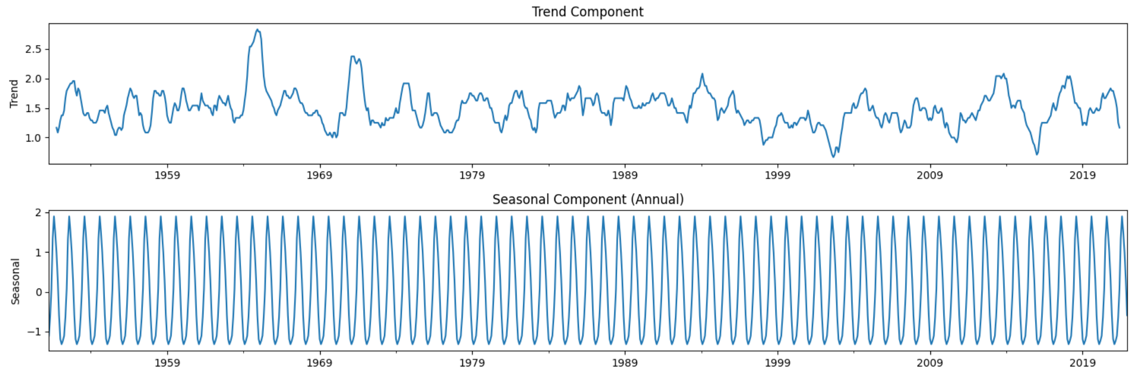


Deployment Software Architecture

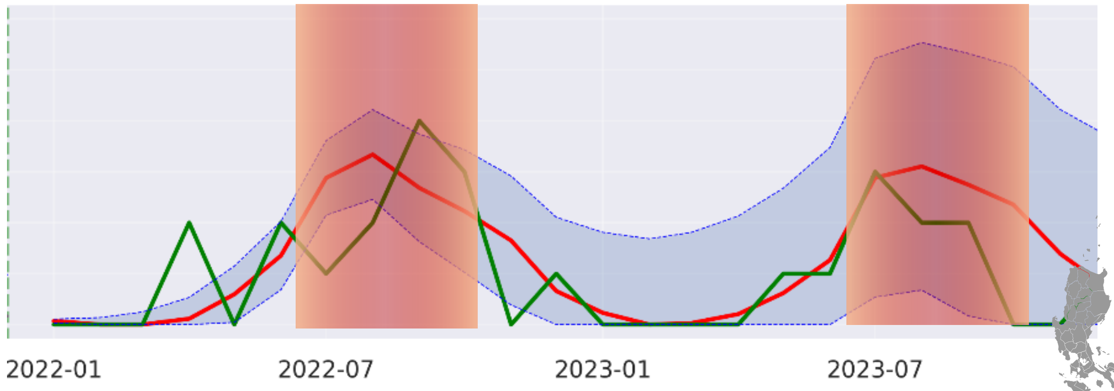


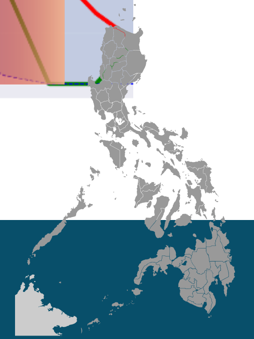
Policy Recommendations

- The historical data, upon decomposition, showed no increasing nor decreasing trend, and exhibits strong yearly seasonality.
- It is expected within the Habagat Season, known for humid and hot weather with heavy rainfall, we can expect more possible typhoons to enter within the third quarter between months of July and September.
- Various models predict two-peak counts within the next 2 years, allowing the government to have enough time to prepare and allocate resources for mitigation, monitoring, response, and infrastructure improvements that largely reduces impact to people and overall national growth.

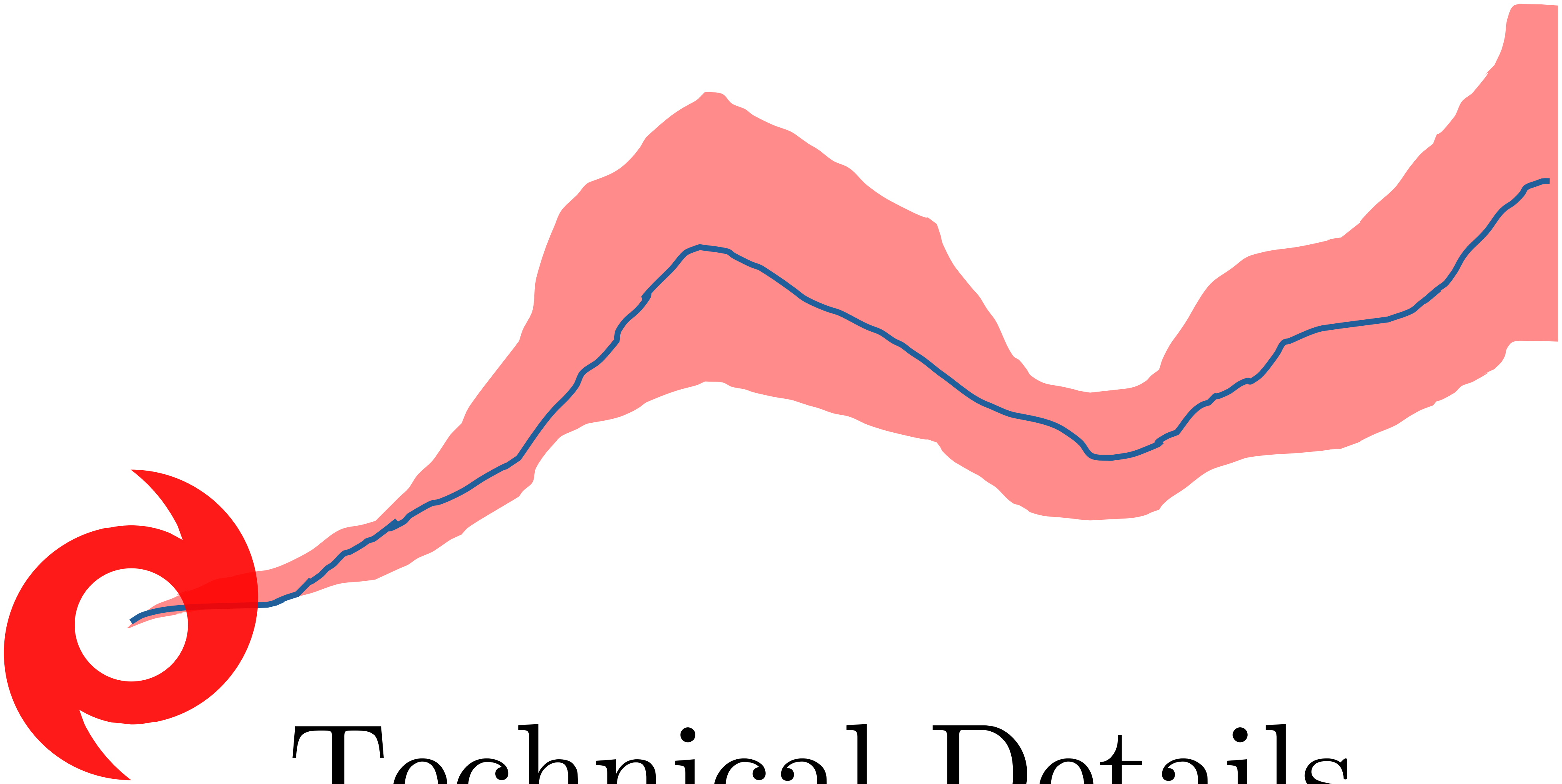


Jul – Aug - Sept



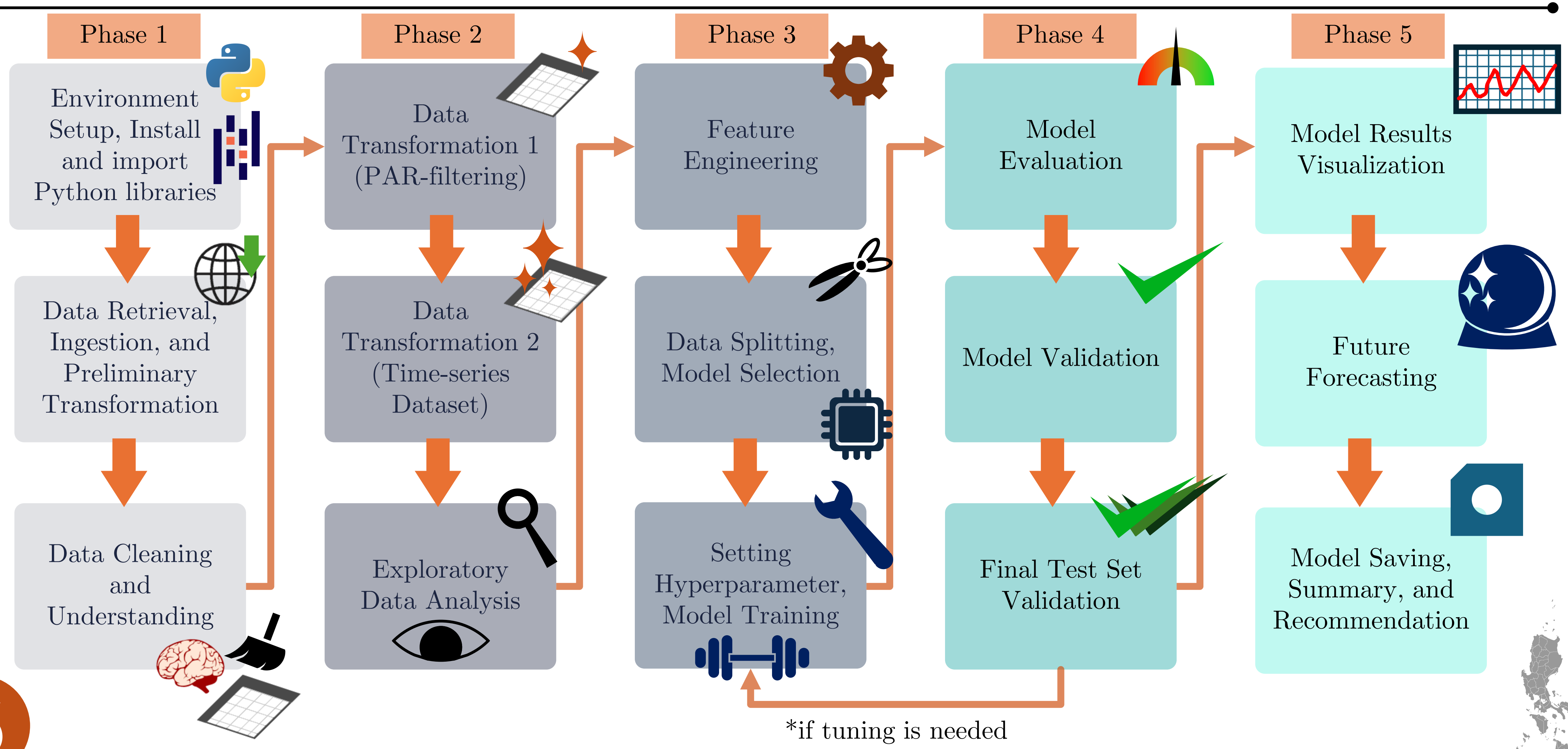


Thank You!



Technical Details

Technical Methodology



Technical Details (Install and Import Libraries)

Environment Setup, Install and import Python libraries

Data Retrieval

Data Retrieval, Ingestion, and Preliminary Transformation

気象庁
Japan Meteorological Agency

Data Processing

By KRYS

kaggle

CSV

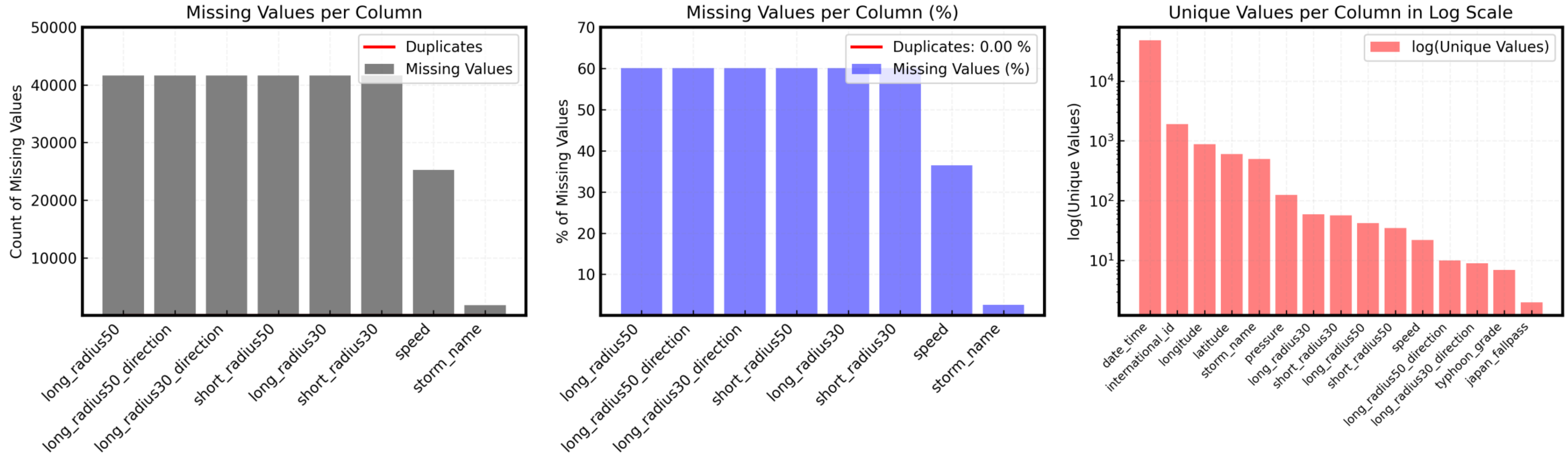
	international_id	date_time	storm_name	typhoon_grade	latitude	longitude	pressure	speed
67886	2122	2021-12-20 06:00:00	RAI	4	18.1	111.4	985	55
67887	2122	2021-12-20 12:00:00	RAI	3	19	111.9	1000	35
67888	2122	2021-12-20 18:00:00	RAI	2	19.9	112.8	1006	0
67889	2122	2021-12-21 00:00:00	RAI	2	20.8	114	1008	0
67890	2122	2021-12-21 06:00:00	RAI	2	21.3	115.3	1008	0

Shape of the kaggle typhoon ds: (67891, 15)

Technical Details (Cleaning)

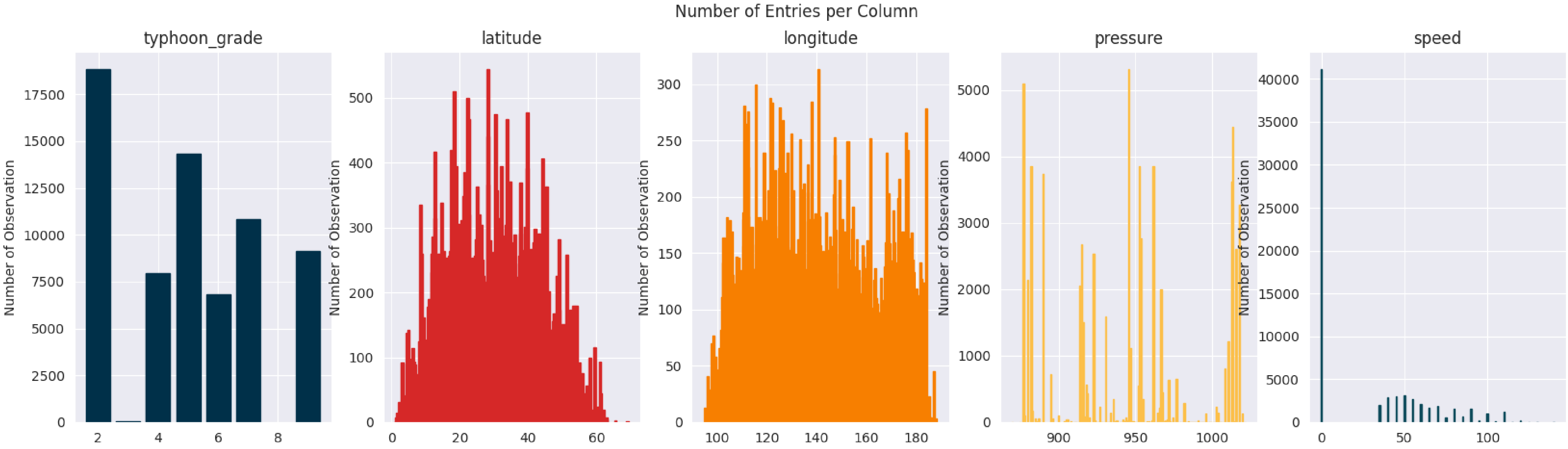
Data Cleaning and Understanding

Figure: Missing, Duplicates, and Unique Values



Technical Details (Cleaning)

Data Cleaning and Understanding

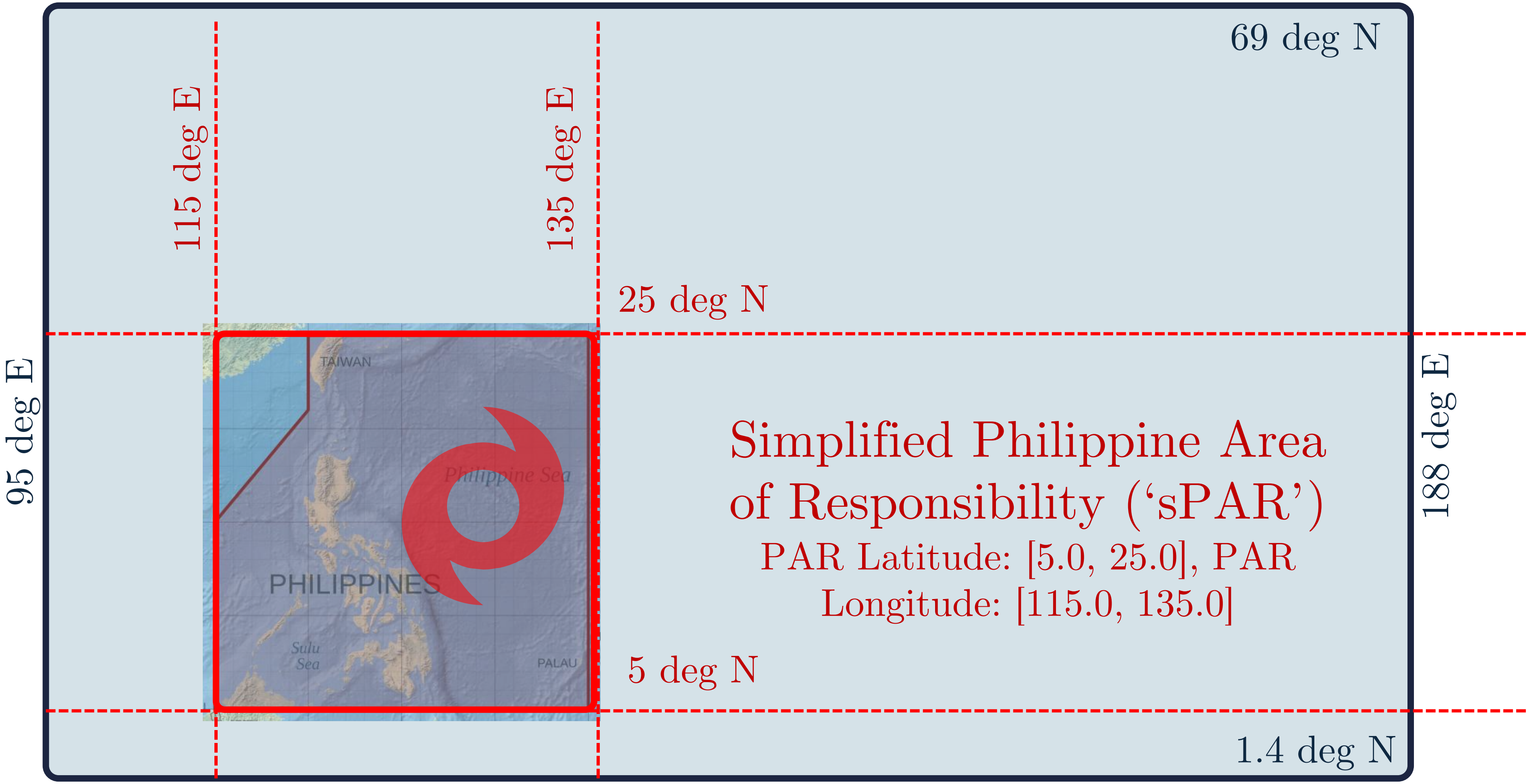


Technical Details (PAR Filtering)

Data Transformation (Spatial Filtering)

Philippine Area of Responsibility (PAR)

Point	Latitude	Longitude
a	5° N	115° E
b	15° N	115° E
c	21° N	120° E
d	25° N	120° E
e	25° N	135° E
f	5° N	135° E

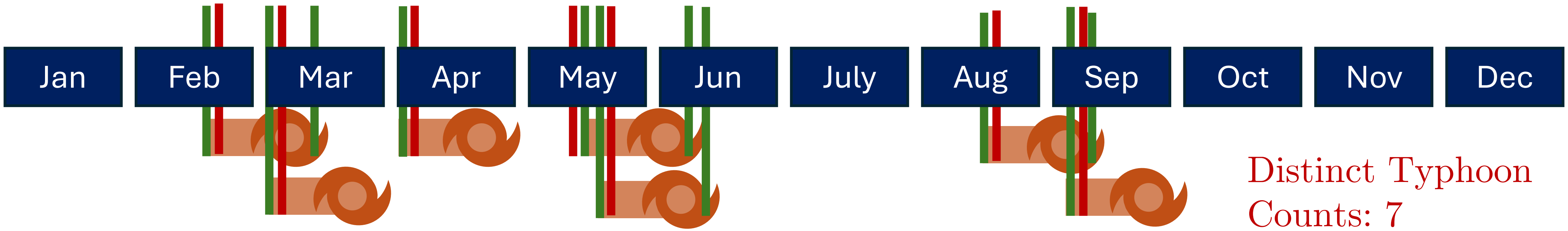


RSMC-Tokyo Center Area Coverage

RSMC Latitude: (1.4, 69.0), RSMC Longitude: (95.0, 188.0)

Distinct Typhoon Counts vs. Observation per Month

datatime	year	month	typhoon_obs	typhoon_ids
1951-04-01 00:00:00	1951	4	2	[np.int64(5103), np.int64(5104)]
1951-05-01 00:00:00	1951	5	1	[np.int64(5104)]
1951-06-01 00:00:00	1951	6	1	[np.int64(5106)]
1951-07-01 00:00:00	1951	7	3	[np.int64(5106), np.int64(5107), np.int64(5109)]
1951-08-01 00:00:00	1951	8	4	[np.int64(5109), np.int64(5110), np.int64(5111), np.int64(5112)]



Distinct Typhoon Counts: 7

Observation per Month: 11

Data Transformation (Observations to Count Data)

datatime	year	month	typhoon_obs	typhoon_counts
1951-04-01 00:00:00	1951	4	2	2
1951-05-01 00:00:00	1951	5	1	0
1951-06-01 00:00:00	1951	6	1	1
1951-07-01 00:00:00	1951	7	3	2
1951-08-01 00:00:00	1951	8	4	3

Technical Details (Decomposition)

Exploratory Data Analysis

Feature Engineering

Lags: [1,2,4,6]

Rolling Values: Mean, STD, Max, Sum: [6,12]

Temporal: Year Quarter Month

Cyclical Encoding: Year Half-Year Quarter Month

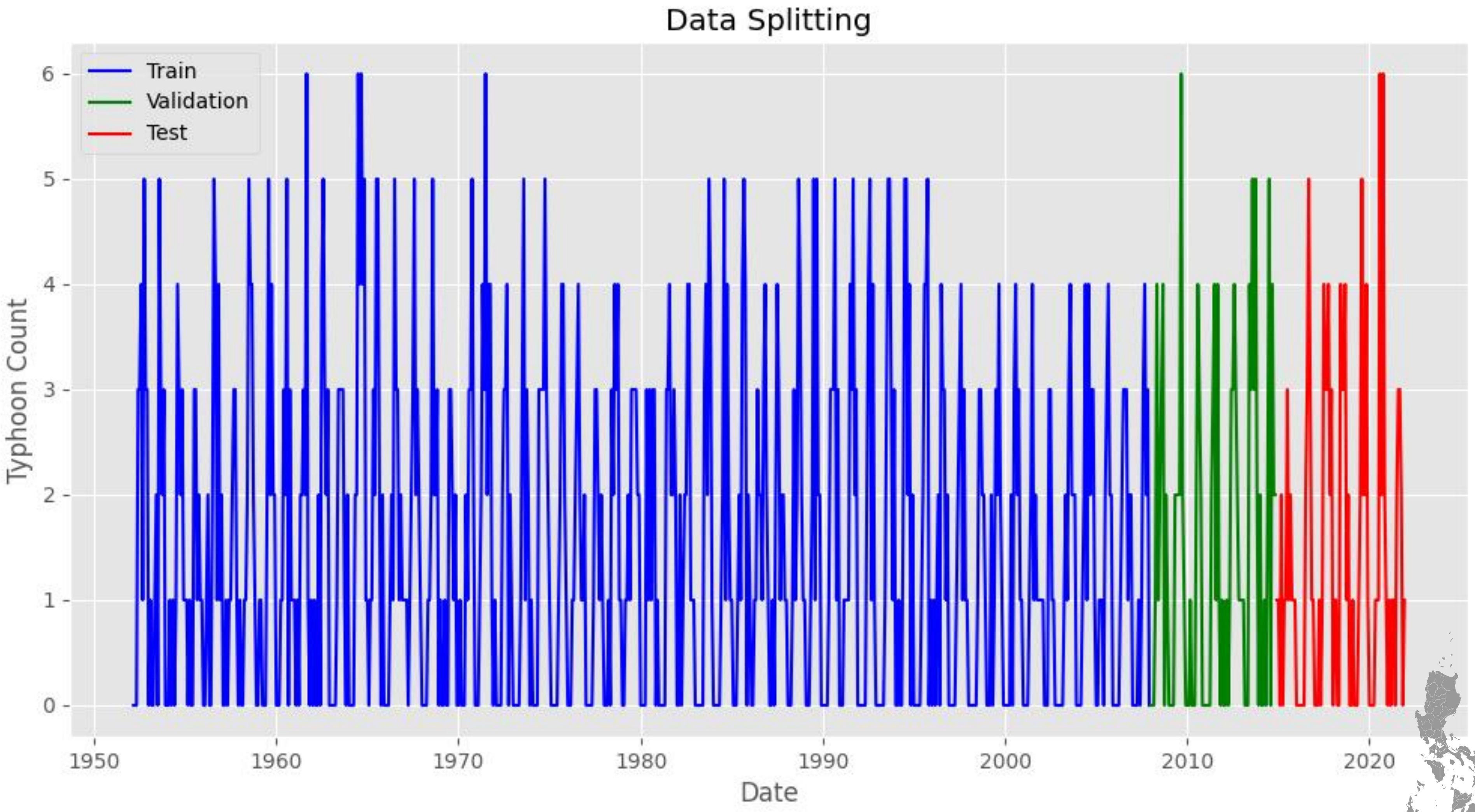
Binary Encoding: New typhoons [1,4,6] Total Typhoons [6,12]

Interaction Features: Month * Rolling Mean at 12-month

	typhoon_count	lag_1	lag_2	lag_4	lag_6	rolling_mean_6	rolling_std_6	rolling_max_6	rolling_sum_6	rolling_mean_12
1952-03-01 00:00:00	0	0	0	1	2	0.833333	0.983192	2	5	1.25
1952-04-01 00:00:00	0	0	0	2	2	0.5	0.83666	2	3	1.08333
1952-05-01 00:00:00	0	0	0	0	1	0.333333	0.816497	2	2	1.08333
1952-06-01 00:00:00	3	0	0	0	2	0.5	1.22474	3	3	1.25
1952-07-01 00:00:00	3	3	0	0	0	1	1.54919	3	6	1.33333

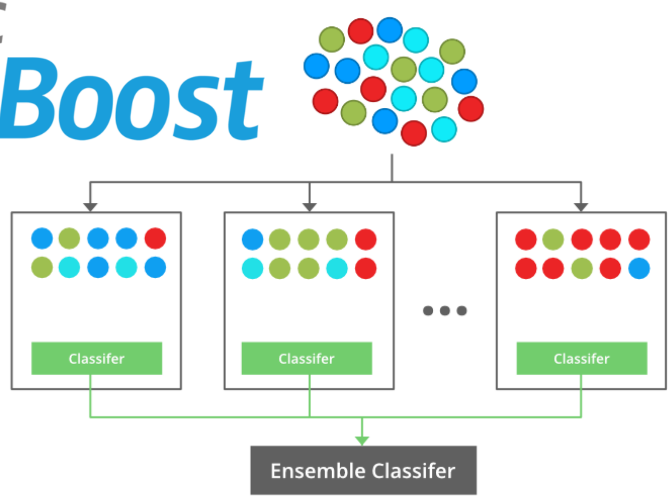
Technical Details (Data Splitting)

Train: 80%
Validation: 10%
Test: 10%

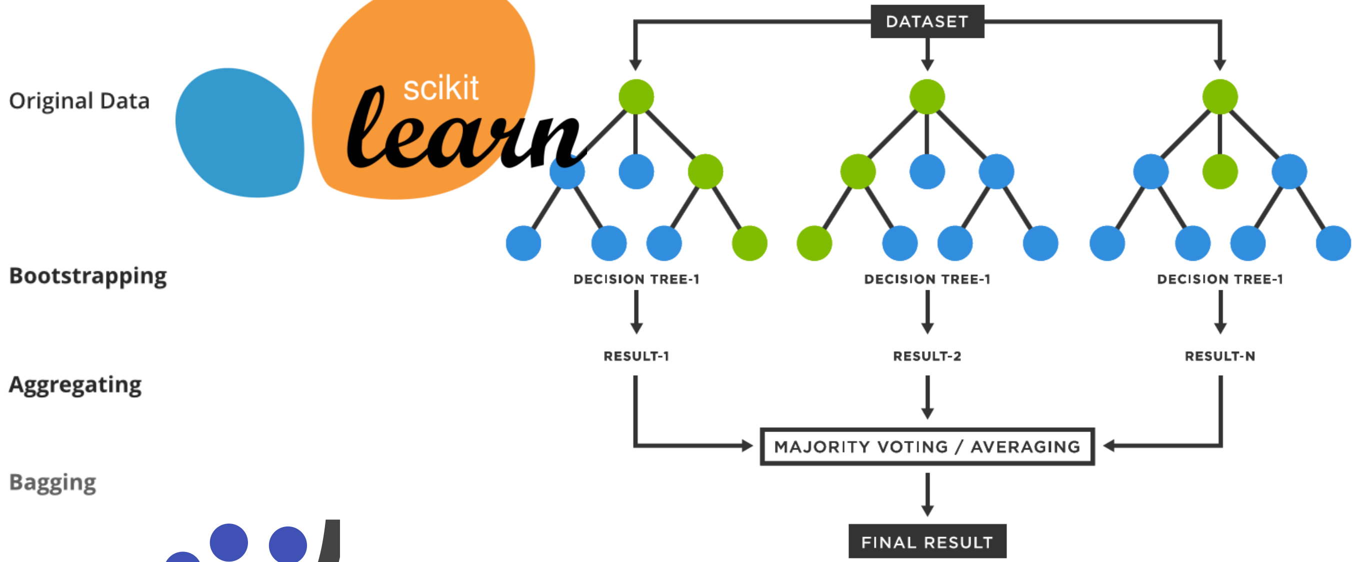


Technical Details (Models Used)

dmlc XGBoost



scikit learn

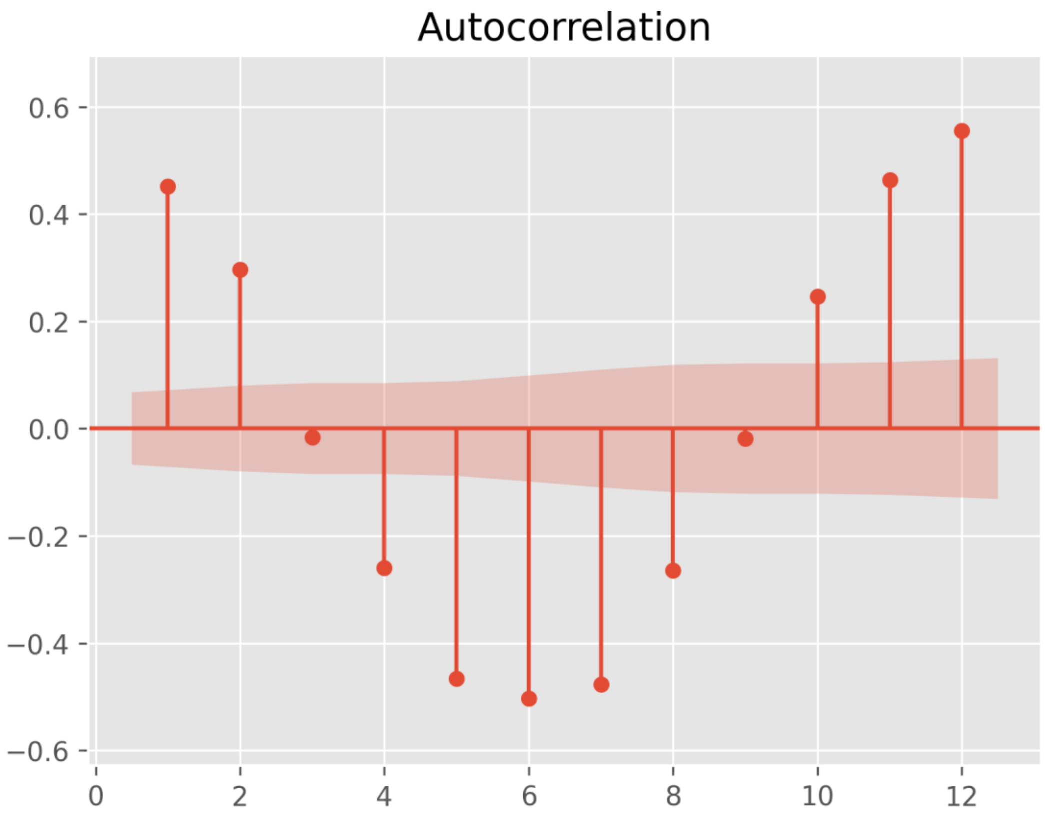


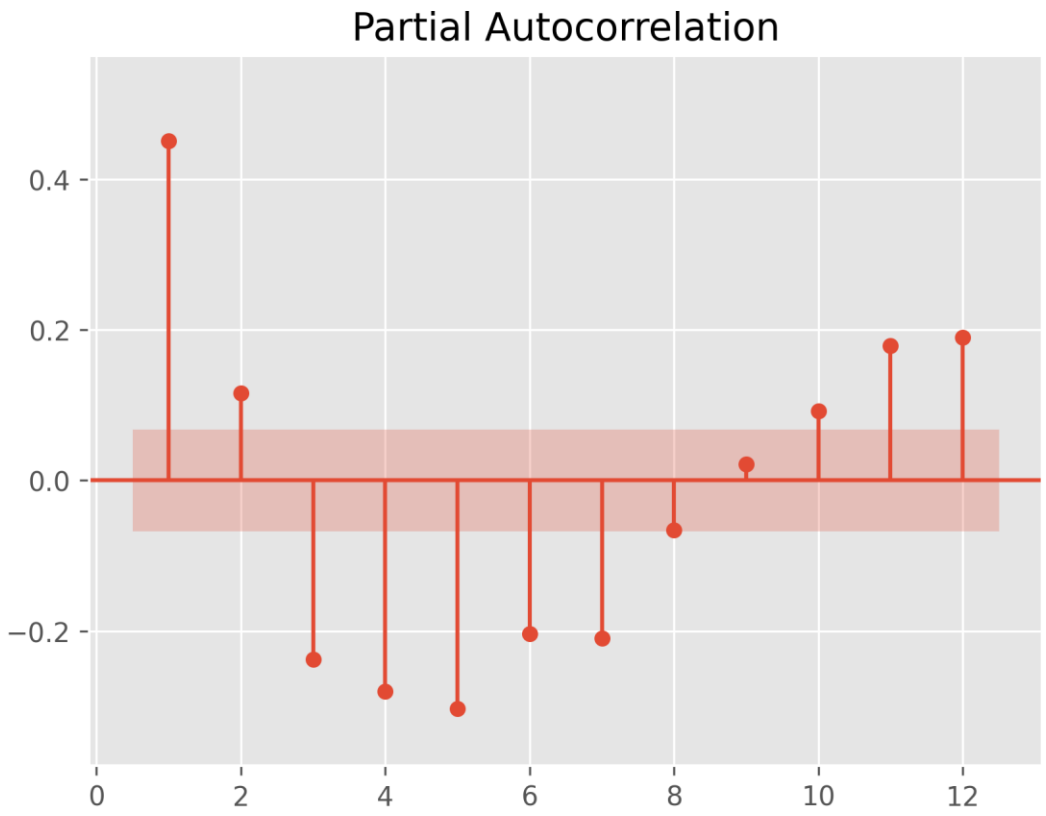
PROPHET

statsmodels
(SARIMAX, Exponential Smoothing)

Additional: Moving Averages, Seasonal Naïve Method

Technical Details (For SARIMAX: (p,d,q), (P,D,Q,m))



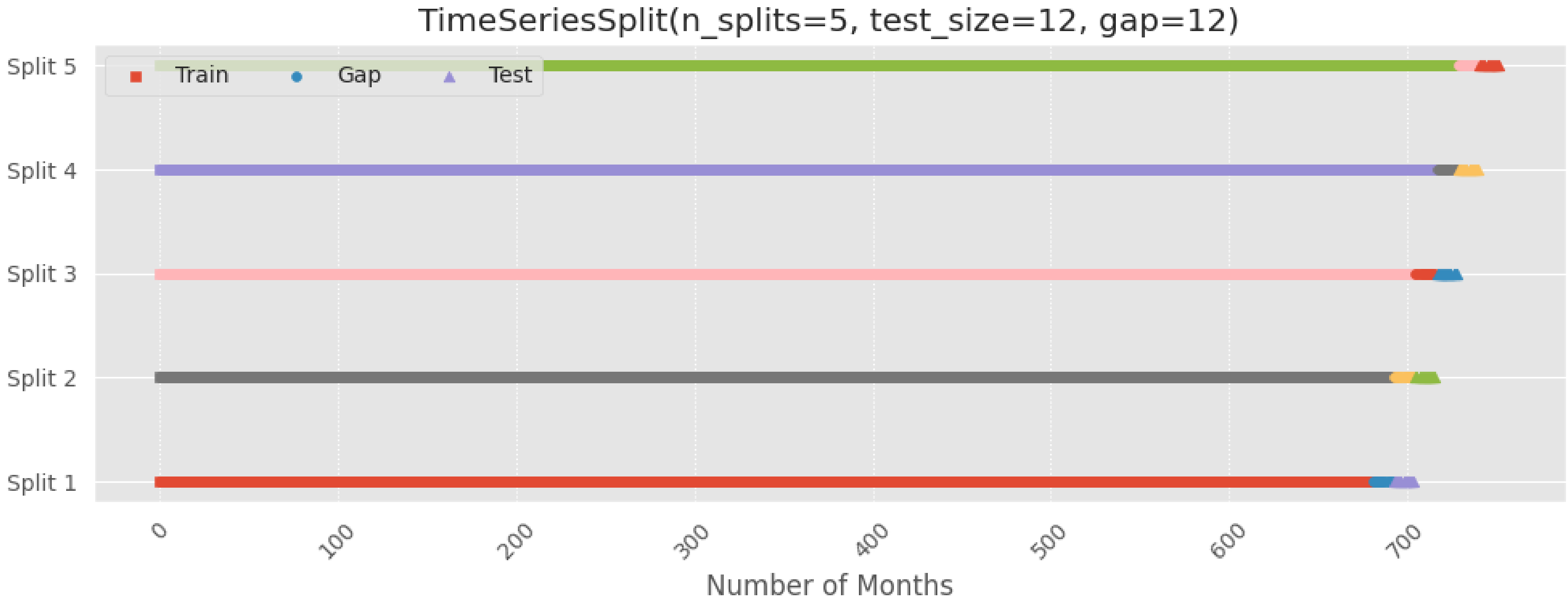


Technical Details (Model Evaluation on Validation Set)

Model	RMSE	MAE	MAPE (%)	R^2	Poisson Dev
XGBoost	0.569	0.337	27.218	0.869	10.880
RandomForest	0.625	0.354	32.054	0.841	12.966
SARIMAX	1.146	0.812	47.396	0.468	53.987
ExpSmoothing	1.148	0.818	47.070	0.466	49.875
Prophet	1.169	0.816	50.476	0.446	53.558
SeasonalNaive	1.414	0.988	63.628	0.189	36.410
MovingAvg_2 months	1.618	1.199	66.683	-0.061	61.599

Metric: Lower is better for RMSE/MAE/MAPE/Poisson Dev; **higher is better** for R^2

Technical Details (Walk-Forward Validation)



Cross-Validation Results:

	Model	CV_Mean_RMSE	CV_Std_RMSE
0	XGBoost	0.703829	0.300551
1	RandomForest	0.562857	0.168472
2	Prophet	1.070317	0.295922

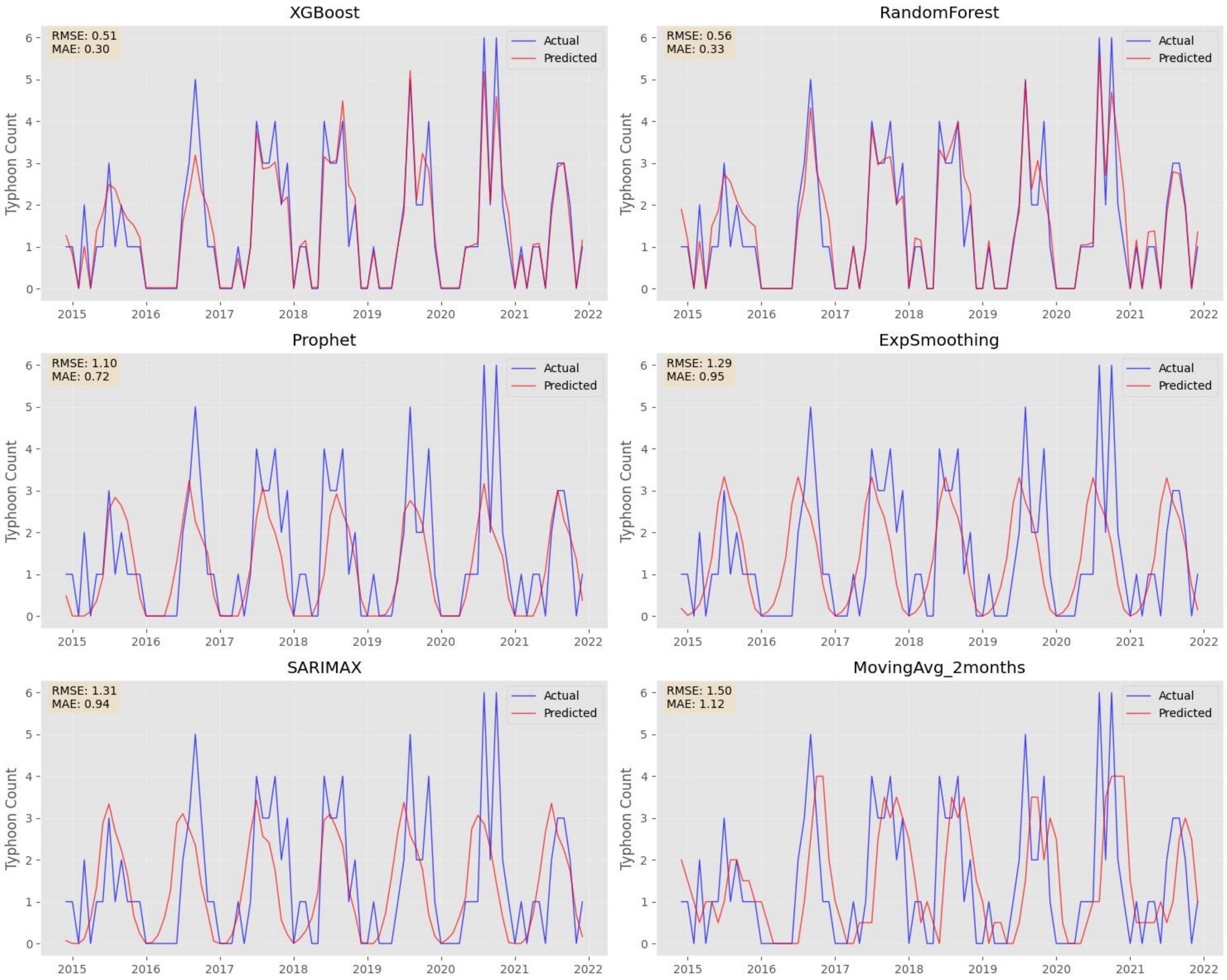
Technical Details (Test Set Evaluation)

Model	RMSE	MAE	MAPE (%)	R^2
XGBoost	**0.507**	**0.298**	**24.962**	**0.885**
RandomForest	0.558	0.334	32.871	0.861
Prophet	1.100	0.717	51.656	0.461
ExpSmoothing	1.293	0.953	63.949	0.256
SARIMAX	1.300	0.951	64.197	0.248
MovingAvg_2months	1.503	1.118	69.430	-0.006
SeasonalNaive	1.782	1.318	83.246	-0.414

Metric: Lower is better for RMSE/MAE/MAPE; **higher is better** for R^2

Model Validation

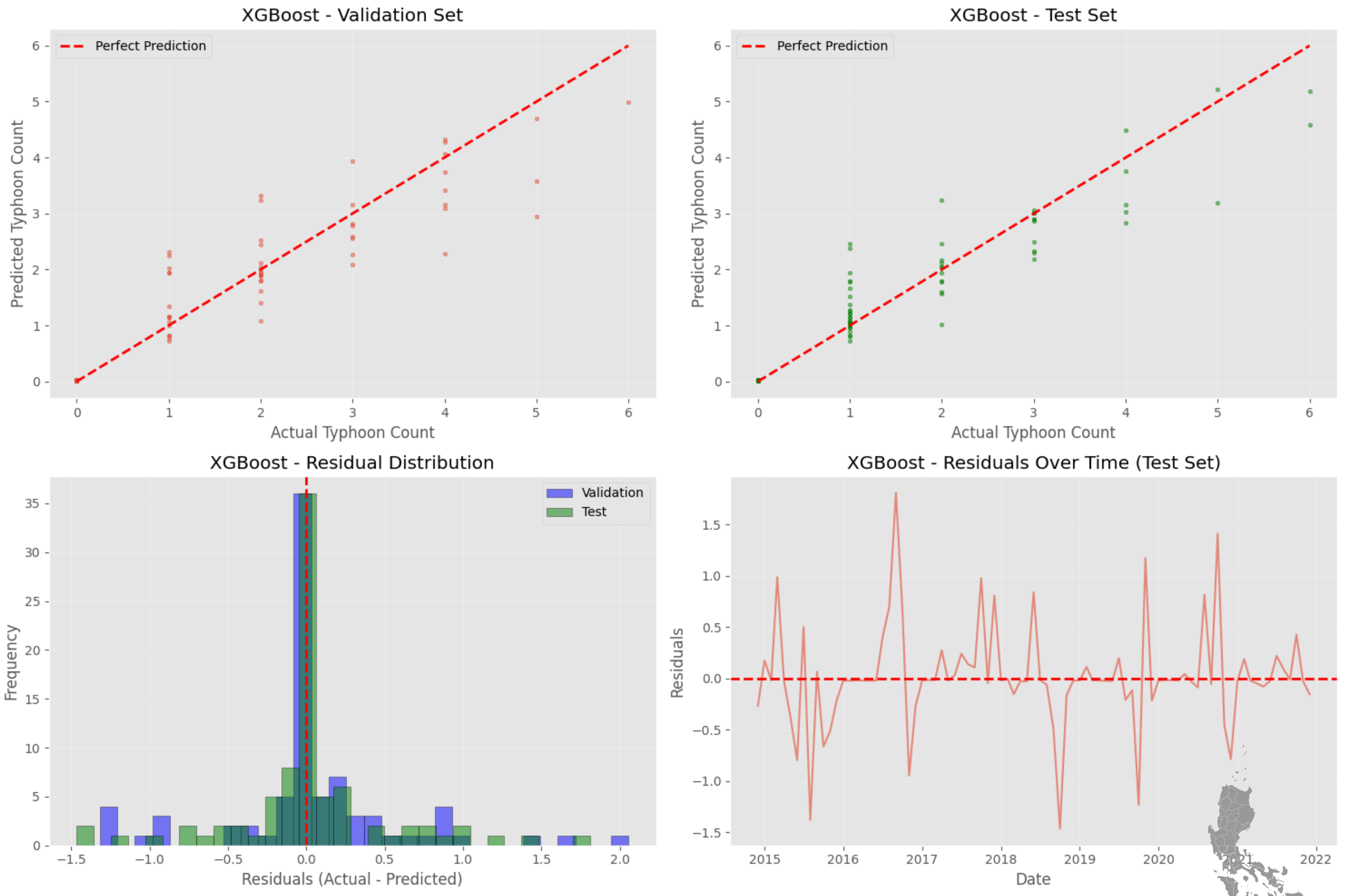
- XGBoost and Random Forest fits well (probably overfitting, need further tuning)
- Prophet, Exponential Smoothing, and SARIMAX captured the seasonality.
- 2-month Moving Average somehow captures the seasonality but not reliable.



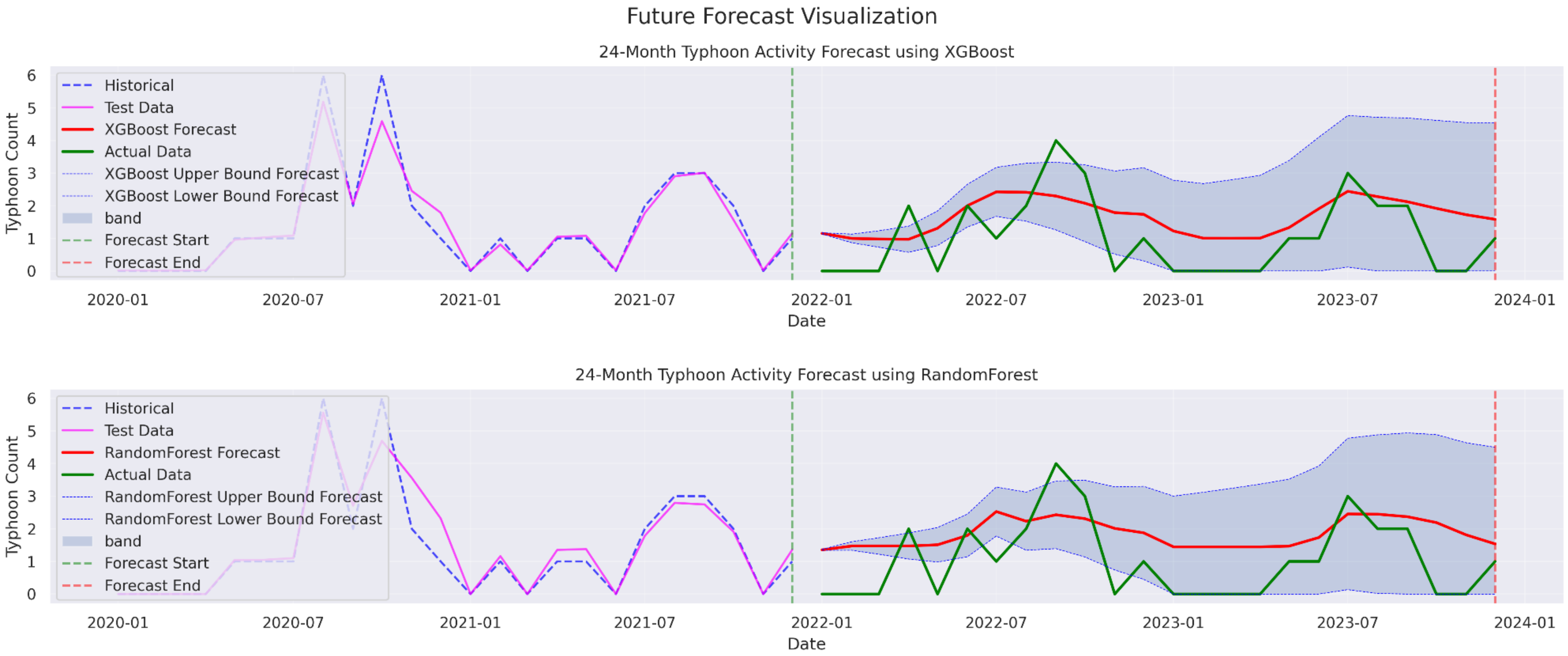
Model Analysis

XGBoost: Best Model

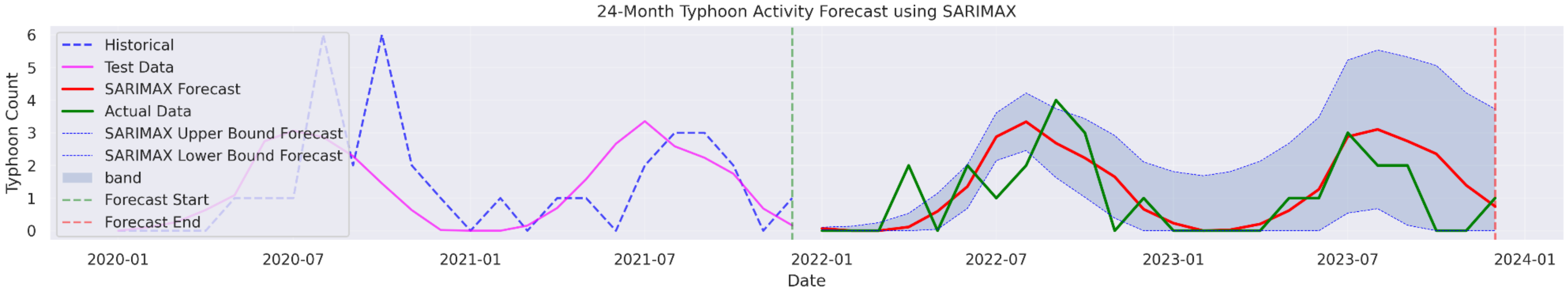
- Well distributed points along the 45-deg line, but with drifting at 4-6 counts
- Prophet, Exponential Smoothing, and SARIMAX captured the seasonality.
- 2-month Moving Average somehow captures the seasonality but not reliable.

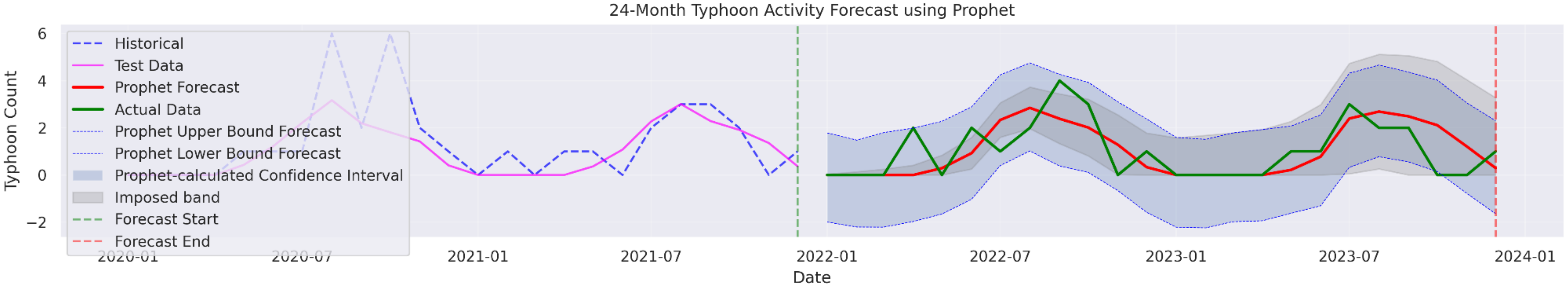


Forecasting Across Models

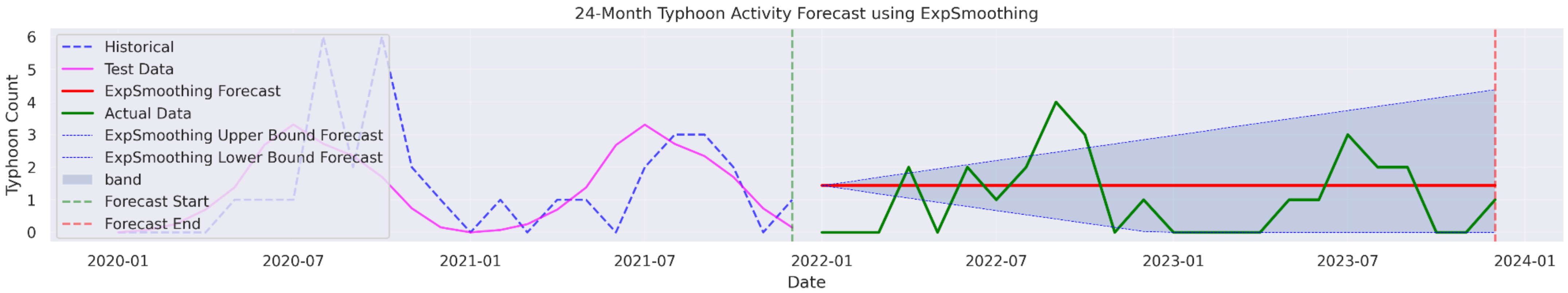


Forecasting Across Models





Forecasting Across Models



Takeaways:

- SARIMAX and Prophet shows promising as forecasting model
- While XGBoost and Random Forest Regression tops the model validation, it decently perform in the forecasting task, but not as tight compared to classical or hybrid ones.
- For the Exponential Smoothing, the mean of test prediction was used to show its capability even if it's a simple forecasting method