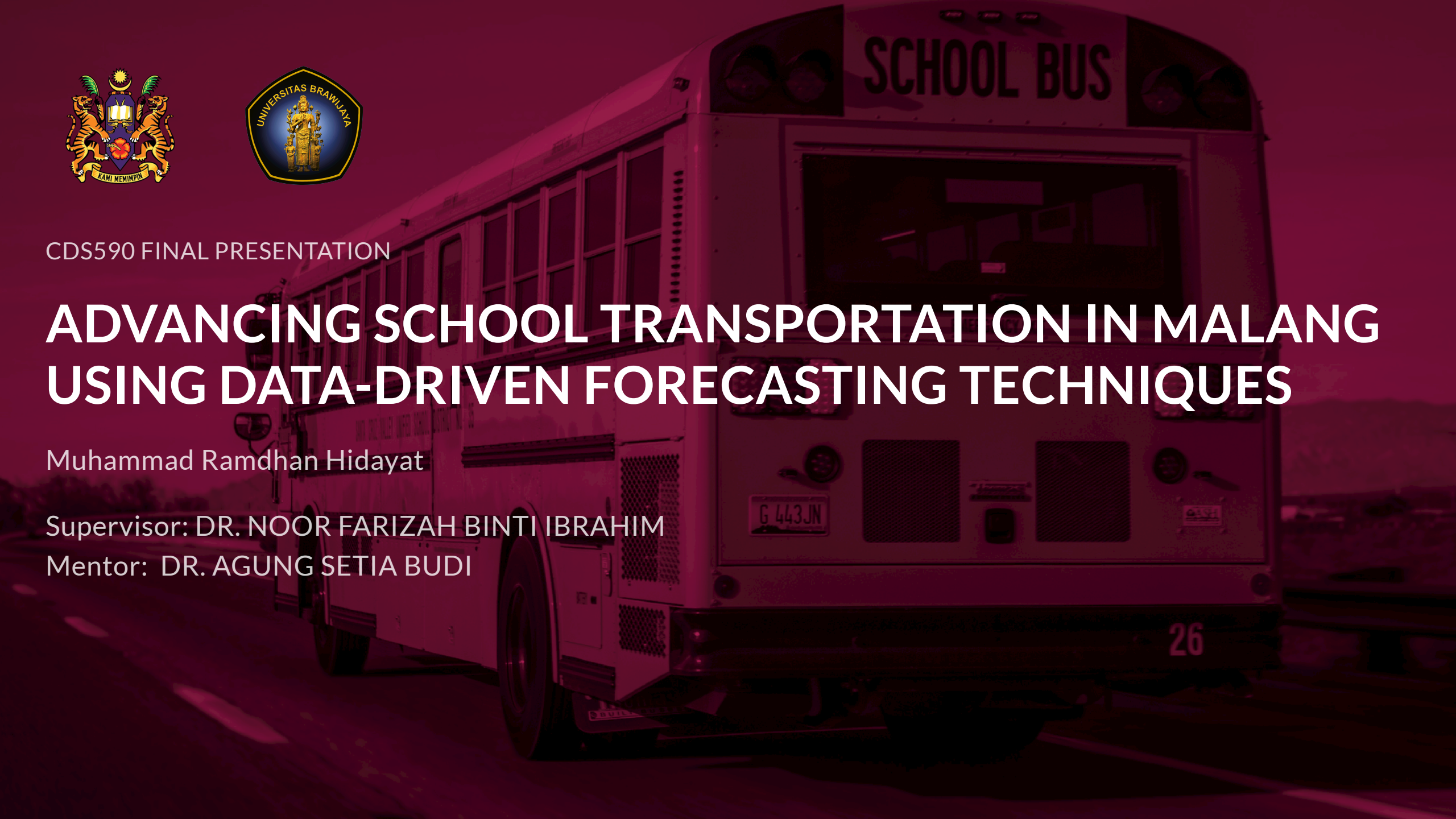


CDS590 FINAL PRESENTATION

ADVANCING SCHOOL TRANSPORTATION IN MALANG USING DATA-DRIVEN FORECASTING TECHNIQUES

Muhammad Ramdhan Hidayat

Supervisor: DR. NOOR FARIZAH BINTI IBRAHIM
Mentor: DR. AGUNG SETIA BUDI

— Section 1 —

1 | Background

2 | Objectives

— Section 2 —

3 | Related Works

— Section 3 —

4 | About the Data

5 | Data Preparation

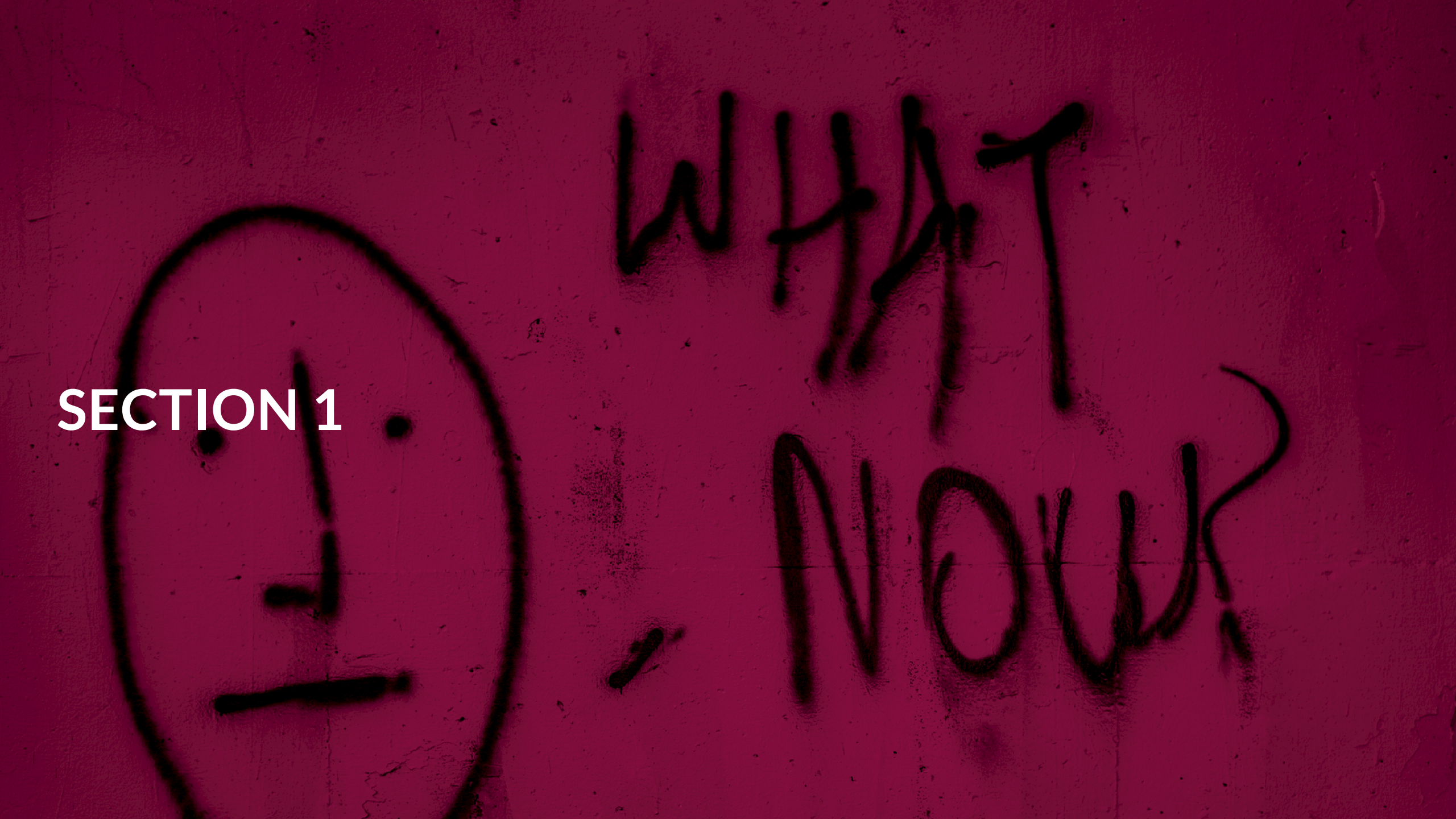
6 | Feature Engineering

7 | Experiment Design

— Section 4 —

8 | Results

9 | Conclusion



SECTION 1

BACKGROUND

Free School Bus in Malang

Malang, a city in Indonesia, has a public school bus system designed to transport students to and from school.

No real-time Bus Arrival Prediction

The absence of accurate bus arrival time predictions causes inconvenience for students and parents, leading to unnecessary waiting and scheduling difficulties

University of Brawijaya (UB) Stepping In

In January 2024, UB starts to collect GPS data

Develop a Prediction Model for Bus Arrival

UB kindly shared the GPS data for us to analyze and then to develop prediction model from it

OBJECTIVES

Research Questions

Q1: Which machine learning model provides the best performance?

Q2: What are the most informative features for predicting school bus arrival times?

Q3: To what extent does incorporating cyclical temporal features improve the model predictive performance?

Objectives

1. To develop a robust prediction model that can accurately predict the bus arrival time while addressing the challenges posed by the data quality (Q1)
2. To identify important predictive features that can be used for further study. (Q2 & Q3)

SECTION 2

RELATED WORKS

RELATED WORKS

1 Machine Learning Models

- **Linear Regression:** Predict travel times (Elhenawy et al., 2018)
- **GBDT:** Freeway travel time prediction (Cheng et al., 2018)
- **CatBoost:** Superior performance for driver arrival time estimation (Sergoyan, 2020)

2 Feature Selection Methods

- Recursive Feature Elimination (RFE): Effective for high-dimensional datasets (Kuhn and Johnson, 2013)
- LASSO LARSIC: applied on time-series prediction (Nourizad et al., 2021)
- Genetic Algorithm (GA) Feature Selection: applied on vast feature spaces (Pudjihartono and Li, 2013)

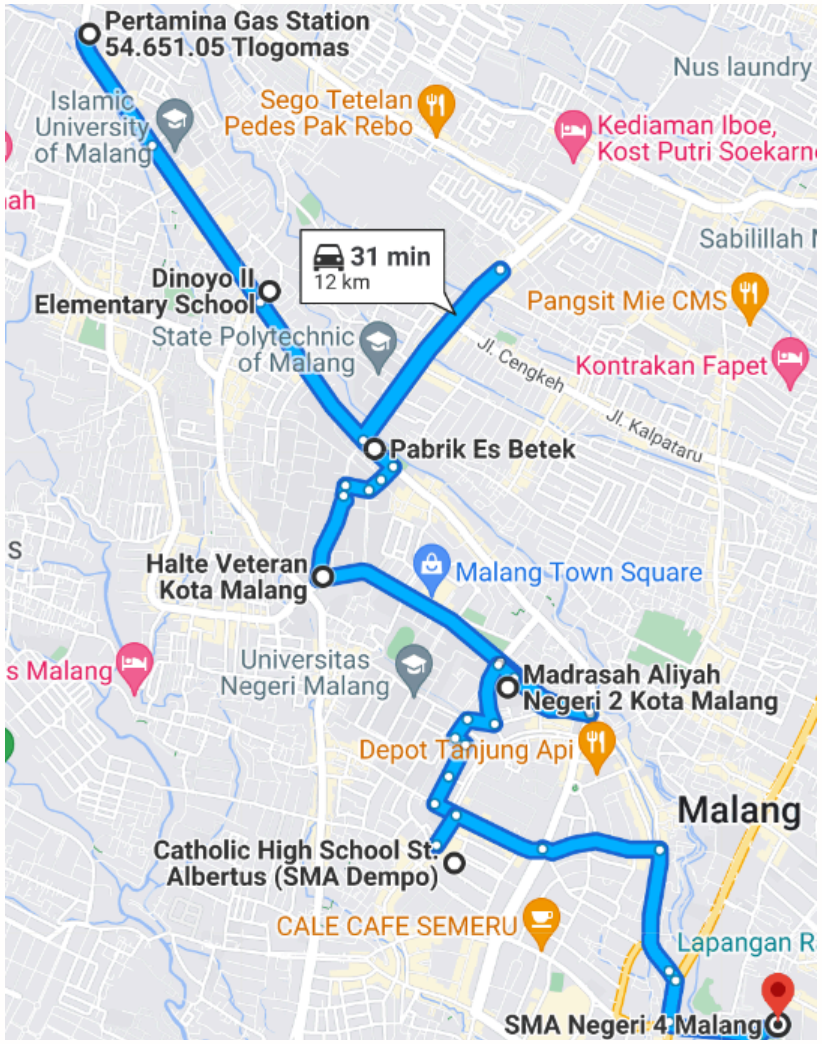
3 Trigonometric Coding

- Cai et al. (2021) achieved 95% accuracy with deep neural networks.

SECTION 2

METHODOLOGIES

DATA PROFILE



Context

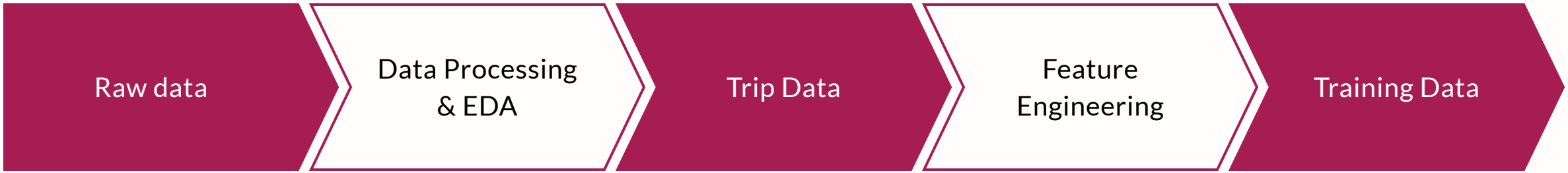
- Event-based GPS beacons data collected from a school bus operation over 4 months
- GPS records are transmitted each time the bus arrives at a designated stop.
- The school bus service **operates on weekdays**, covering 7 distinct stops daily.
- The bus operates in two shifts:
  - Morning Shift: Starts at 7:00 AM for student pick-up and school drop-off
  - Afternoon Shift: Begins at 3:00 PM for student pick-up from schools and drop-off

RAW DATA

bus_id	route_id	imei	latitude	longitude	speed	time	flag
33	10	d2:18:57:29:37:7	-7.93	112.6	10	2024-01-17 14:27:47	1
33	10	fb:fd:2a:a8:e2:b	-7.94	112.63	10	2024-01-17 14:49:56	0
33	10	fb:fd:2a:a8:e2:b	-7.94	112.63	10	2024-01-17 14:50:06	0
33	10	fb:fd:2a:a8:e2:b	-7.94	112.63	10	2024-01-17 14:50:21	1
33	11	fb:fd:2a:a8:e2:b	-7.94	112.63	10	2024-01-17 14:50:41	0
33	11	fb:fd:2a:a8:e2:b	-7.94	112.63	10	2024-01-17 14:50:51	0

- **Bus ID:** Unique identifier for each bus (e.g., 33).
- **Route ID:** Indicates the direction of the bus (10 for one direction, 11 for the opposite).
- **IMEI:** Unique identifier for the bus's GPS device.
- **Latitude and Longitude:** Geographic coordinates indicating the bus's location.
- **Speed:** Bus speed at the time of recording.
- **Time:** Timestamp of the GPS data record.
- **Flag:** Indicator for data relevance or redundancy (1 for relevant, 0 for redundant).

DATA PREPARATION



- Redundant Records
- Irrelevant Records
- Coordinate Discrepancies

- Duplicate Removal
- Weekend Data & Outlier Removal
- Clustering -> Central measure -> standardized lat lon

- 548 Rows
- 11 Columns

- Historical weather data
- Feature Extraction
  - Trigonometric coding
  - Speed -> Category
- Minimize multicollinearity

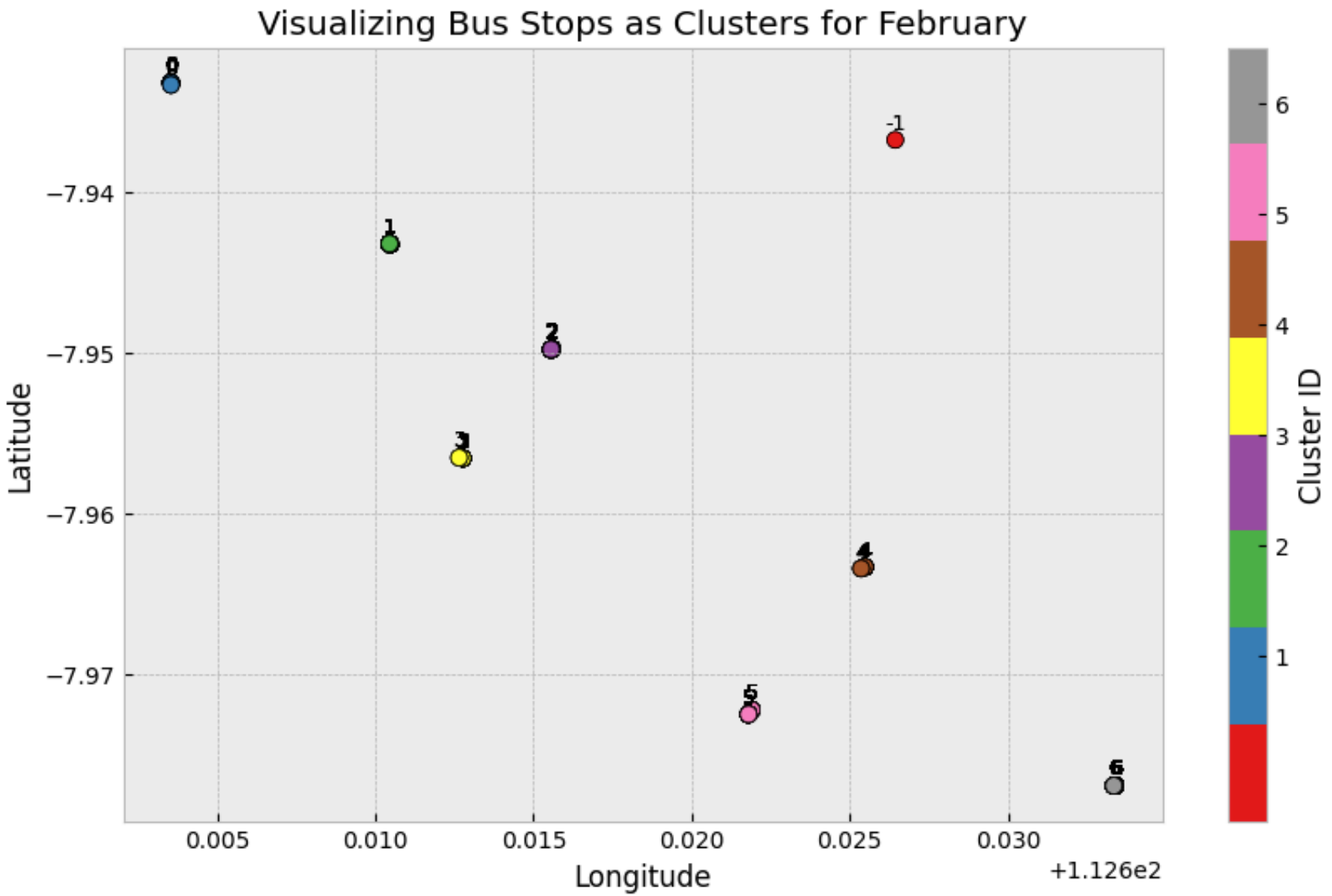
- 548 Rows
- 17 Features

STANDARDIZING COORDINATES

Effectiveness of DBSCAN

- On February data, DBSCAN identified certain bus stop records as noise. Initially we're not certain about this classification

- Data provider confirmed on May that these records were indeed invalid due to recent policy changes by the local government

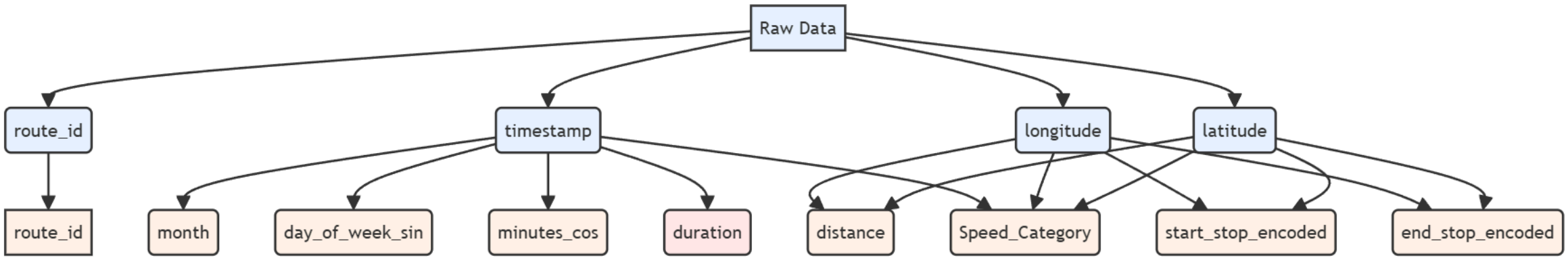


TRIP DATA

route_id	trip_id	start_time	end_time	start_stop	end_stop	duration	distance	day_of_week	month	avg_speed
10	350	2024-04-04 07:43:02	2024-04-04 07:44:40	MAN 2	SMA Dempo	1.633333	1.091282	3	4	40.087911
10	117	2024-02-22 07:09:28	2024-02-22 07:10:46	SMA 9	SD Dinoyo 2	1.300000	0.914033	3	2	42.186149
10	348	2024-04-04 07:34:34	2024-04-04 07:38:46	SD Dinoyo 2	SMA 8	4.200000	1.491343	3	4	21.304903
10	491	2024-05-08 07:17:31	2024-05-08 07:27:10	SMA 8	SMP 4	9.650000	3.200574	2	5	19.899941
11	73	2024-02-28 17:16:15	2024-02-28 17:22:17	SMA 9	SD Dinoyo 2	6.033333	0.914033	2	2	9.089833

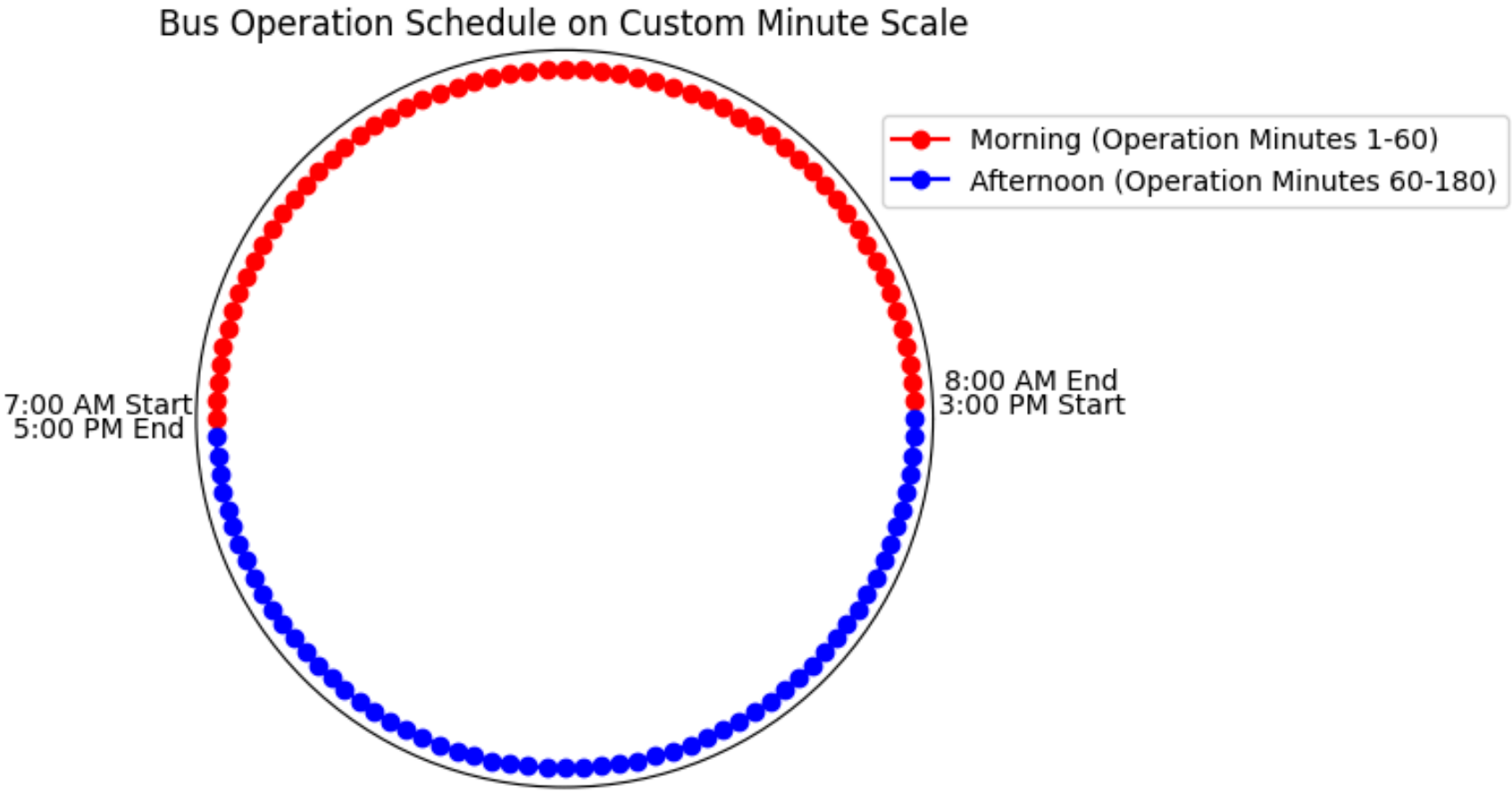
FEATURE ENGINEERED DATA

route_id	distance	month	day_of_week_sin	minutes_cos	temperature	precip_mm	humidity	visibility	pressure	cloud_cover	wind_speed	Speed_Category	start_stop_encoded	end_stop_encoded	duration
10	1.348795	2	0.433884	0.87462	18	0	90	10	1014	40	3	1	6	1	1.883333
10	0.914033	2	0.433884	0.857167	18	0	90	10	1014	40	3	0	1	3	9.616667
10	0.914033	2	0.433884	0.766044	18	0	90	10	1014	40	3	1	3	1	1.35
10	1.491343	2	0.433884	0.743145	18	0	90	10	1014	40	3	0	1	2	3.933333
10	1.589263	2	0.433884	0.694658	18	0	90	10	1014	40	3	0	2	0	6.916667
10	1.091282	2	0.433884	0.615661	18	0	90	10	1014	40	3	1	0	4	1.833333


Raw Data
route_id
timestamp
longitude
latitude
route_id
month
day_of_week_sin
minutes_cos
duration
distance
Speed_Category
start_stop_encoded
end_stop_encoded

FEATURE ENGINEERING

Cyclical Compressed Time Scale



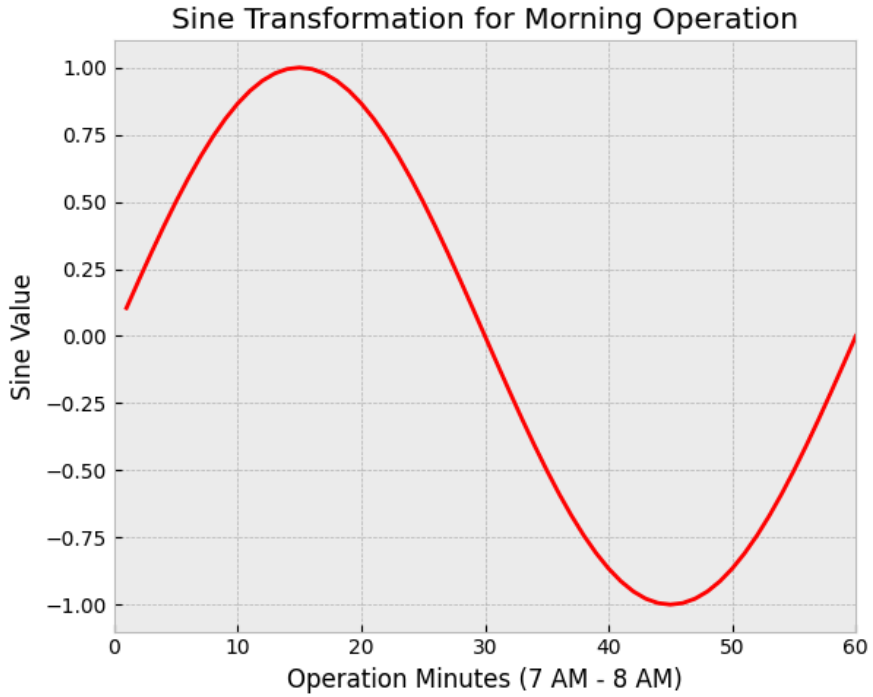
FEATURE ENGINEERING

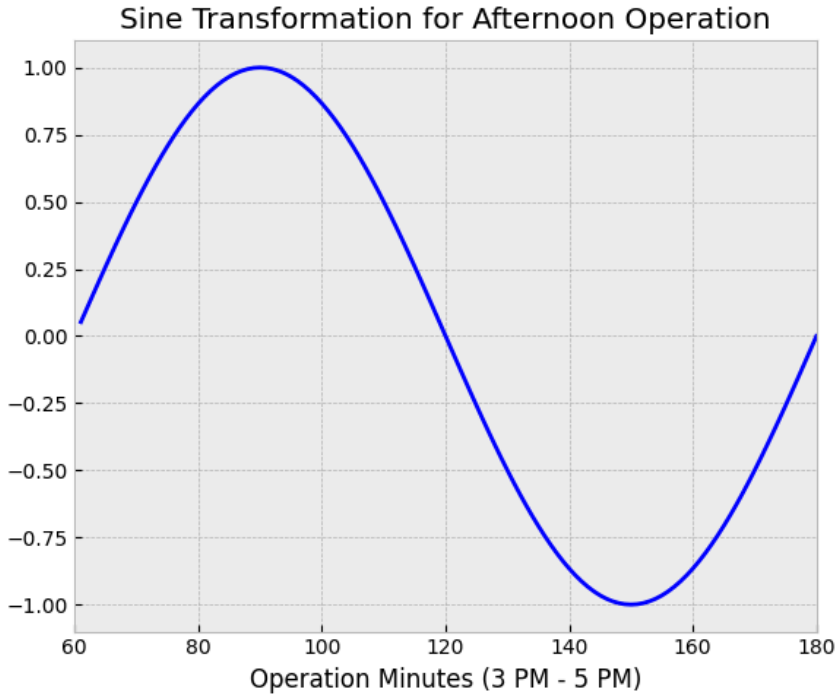
Sine & Cosine transformation

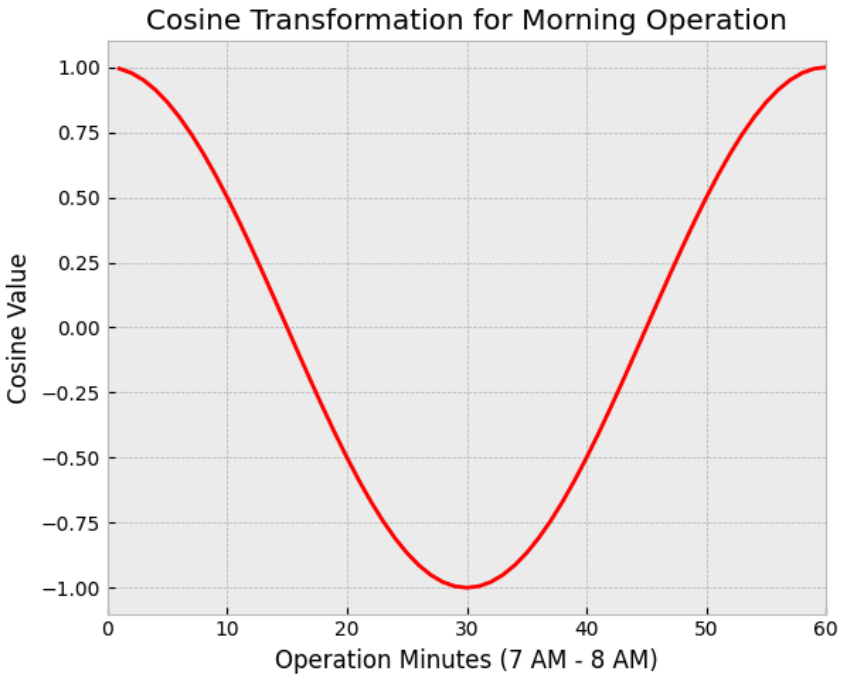
- Operation minutes are mapped to a 2π radian scale

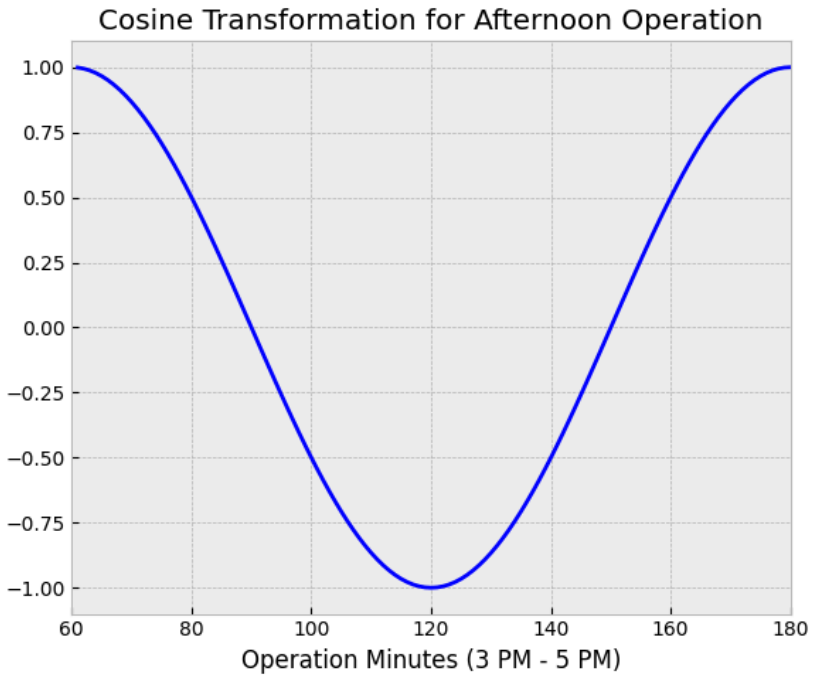
- Each minute is converted into radians based on the total cycle of 120 minutes corresponding to 2π radians

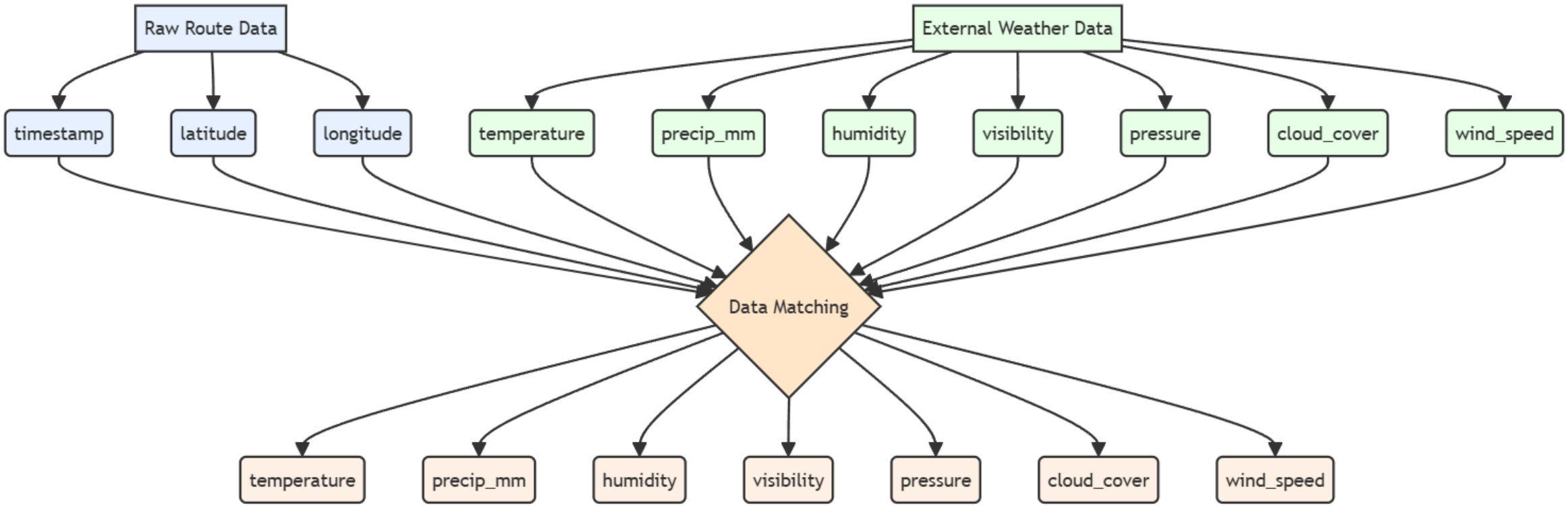
- This transformation is also applied on days information (day_of_week)



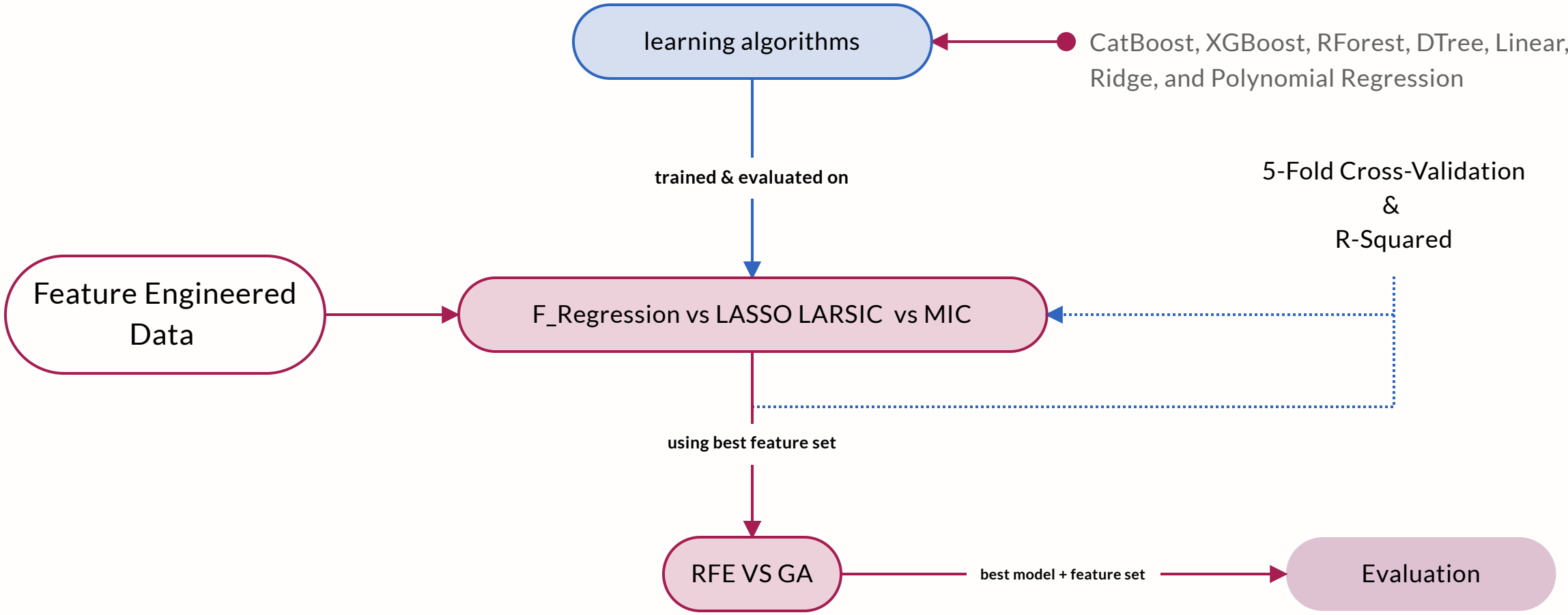






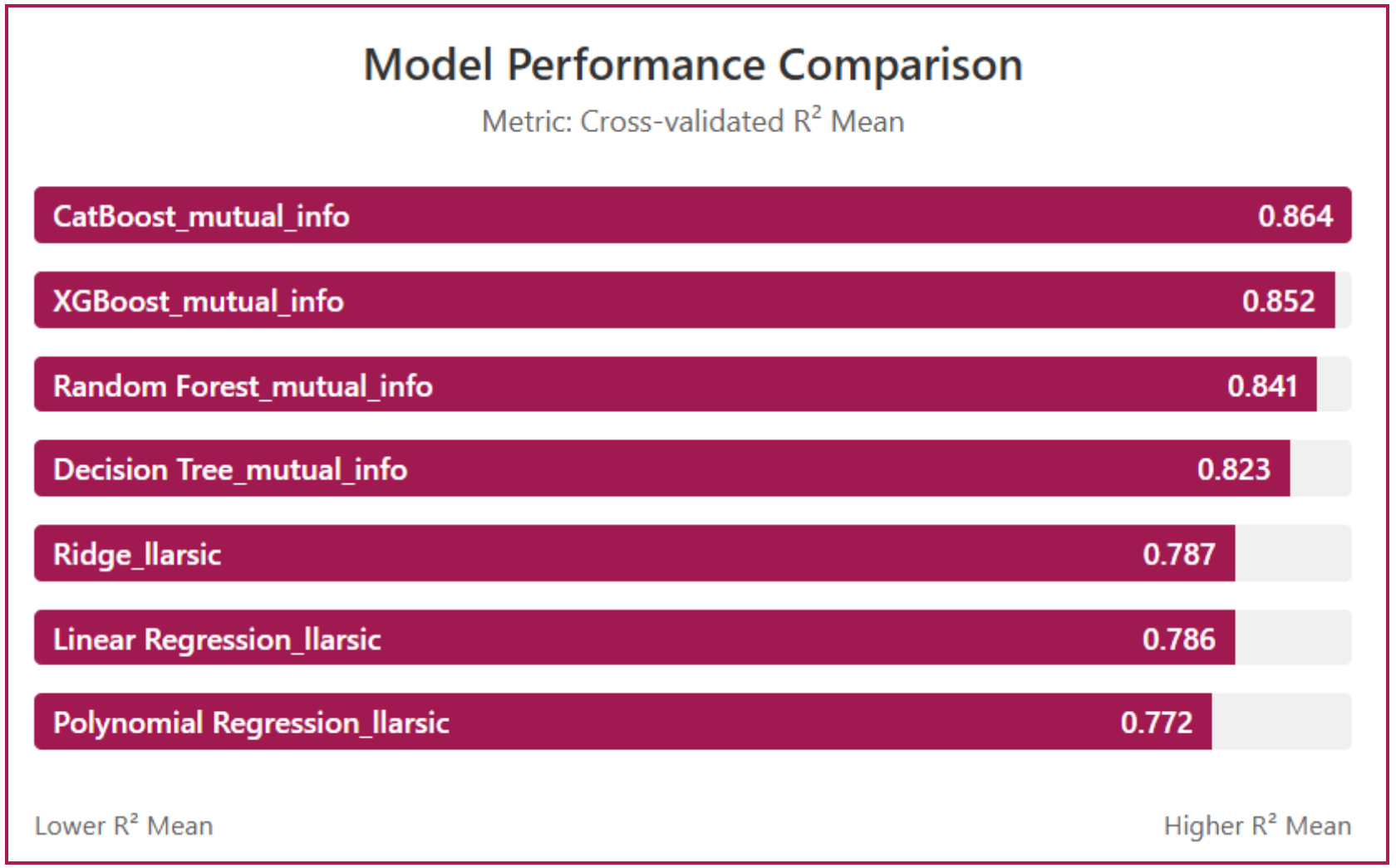

Raw Route Data
External Weather Data
timestamp
latitude
longitude
temperature
precip_mm
humidity
visibility
pressure
cloud_cover
wind_speed
Data Matching
temperature
precip_mm
humidity
visibility
pressure
cloud_cover
wind_speed

EXPERIMENTAL DESIGN



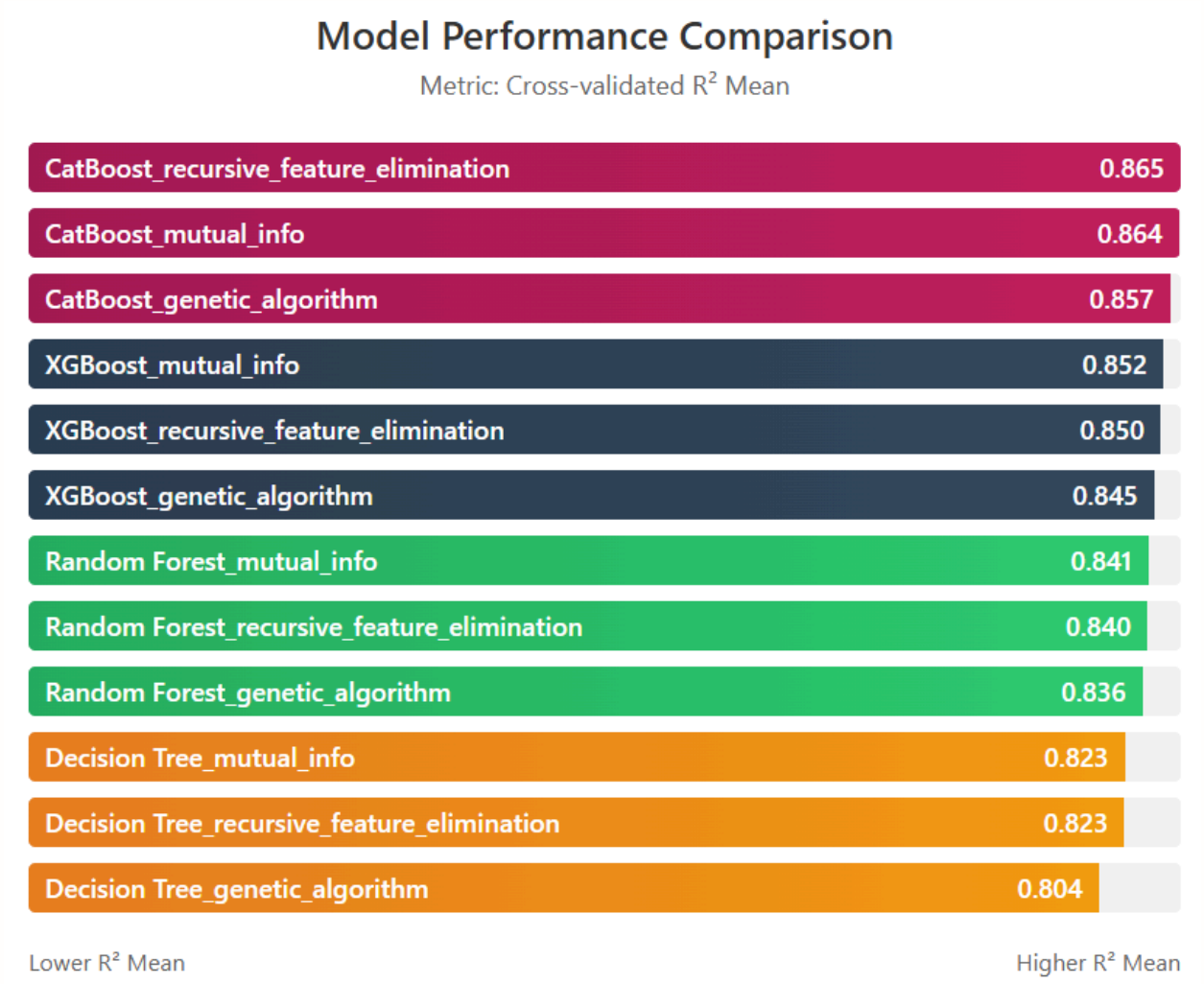
RESULTS

MODEL AND FS METHOD EVALUATION 1



- Boosting algorithms (CatBoost and XGBoost) perform the best, with CatBoost slightly outperforming XGBoost
- Tree-based models (Random Forest and Decision Tree) come next in performance.
- The simpler regression models (Ridge, Linear, and Polynomial) show lower performance compared to the more complex algorithms.

MODEL AND FS METHOD EVALUATION 2



- For most algorithms, RFE and MIC performs better than GA
- For most algorithms, MIC outperforms RFE
- However, since the aim is to get the best model, CatBoost with RFE feature sets is picked as the final solution

FEATURE IMPORTANCE ANALYSIS

feature	cb_feat_imp	rf_feat_imp
distance	47.56%	71.62%
Speed_Category	11.93%	14.04%
minutes_cos	9.90%	2.78%
start_stop_encoded	9.61%	1.66%
route_id	8.00%	0.81%
precip_mm	3.38%	1.21%
pressure	2.76%	1.57%
end_stop_encoded	2.63%	4.13%
humidity	2.61%	1.37%
day_of_week_sin	1.60%	0.81%

- distance is the most important predictor
- 'Speed_Category' is the second most important feature in both models
- CatBoost is better at capturing temporal patterns (see minutes_cos)

TRIGONOMETRIC-CODED FEATURES

1 | ## minutes_cos:

Strong Predictor:

- Ranked among top 3 most important features
- Suggests cosine transformation effectively captures relevant time-based variations

2 | ## day_of_week_sin

R^2 Increase: Marginal improvement observed (~0.1%) despite being the least important feature

CONCLUSION

CONCLUSION

1. Catboost Outperformed Other Models

   This aligns well with (Sergoyan, 2020)

2. Trigonometric-Coded Features works well with Catboost

   As suggested in permutation feature importance analysis

3. Implementing the optimal model will enhance convenience for both students and parents

Q & A

THANK YOU!

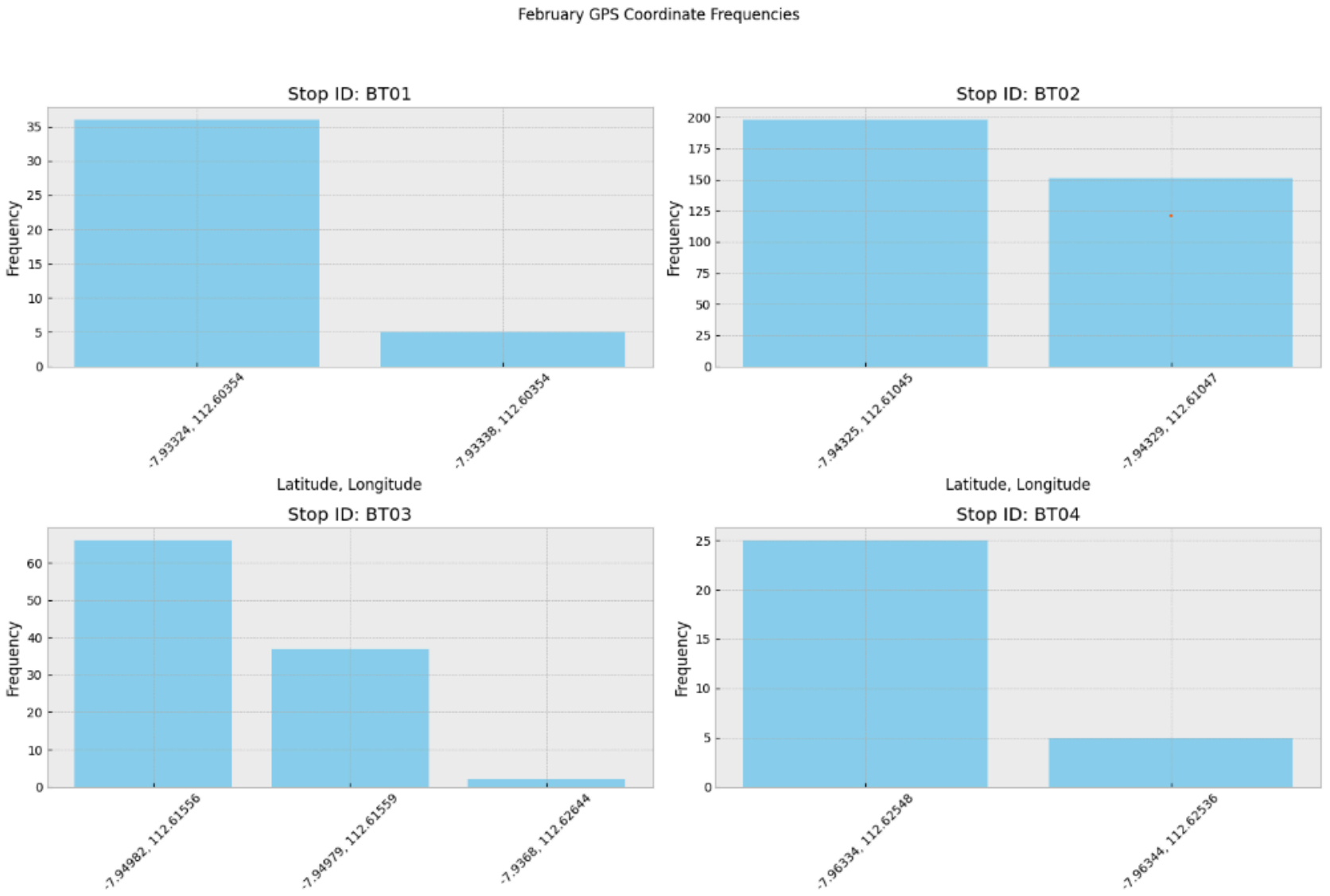
APPENDIX

STANDARDIZING COORDINATES

Mode as central measure

- The arithmetic mean might be influenced by outliers or anomalous data points.

While the median is robust against outliers, it is not suitable to data points that are fixed in nature



5-FOLD CROSS-VALIDATION