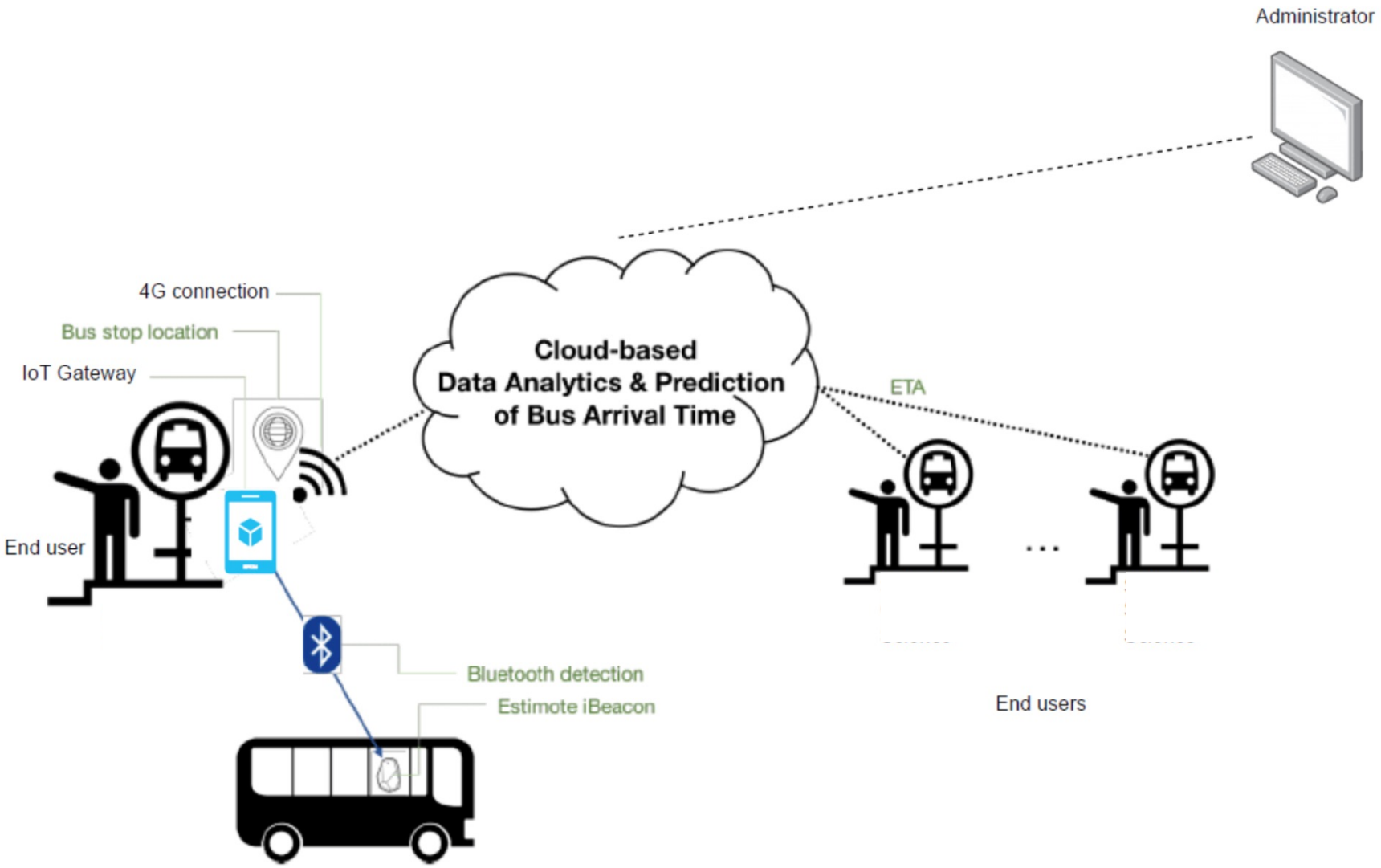
UNIVERSITI SAINS MALAYSIA

Predicting Estimated Time of Arrival Using Boosting Models

Say-Hong Kam (USM, Malaysia), **Yung-Wey Chong (USM, Malaysia)**,
Noor Farizah Ibrahim (USM, Malaysia), Sye-Loong Keoh (UoG, Singapore),
Somnuk Phon-Amnuaisuk (UTB, Brunei),
Sharul Kamal Abdul Rahim (UTM, Malaysia)

Introduction

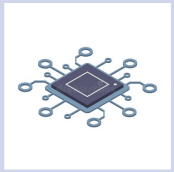


UNIVERSITI SAINS MALAYSIA

Problem Statement

The high installation and maintenance costs of On-Board Diagnosis (OBD) sensors.

Sensor data is known for its temporal and spatial correlation characteristics…. and extremely dirty

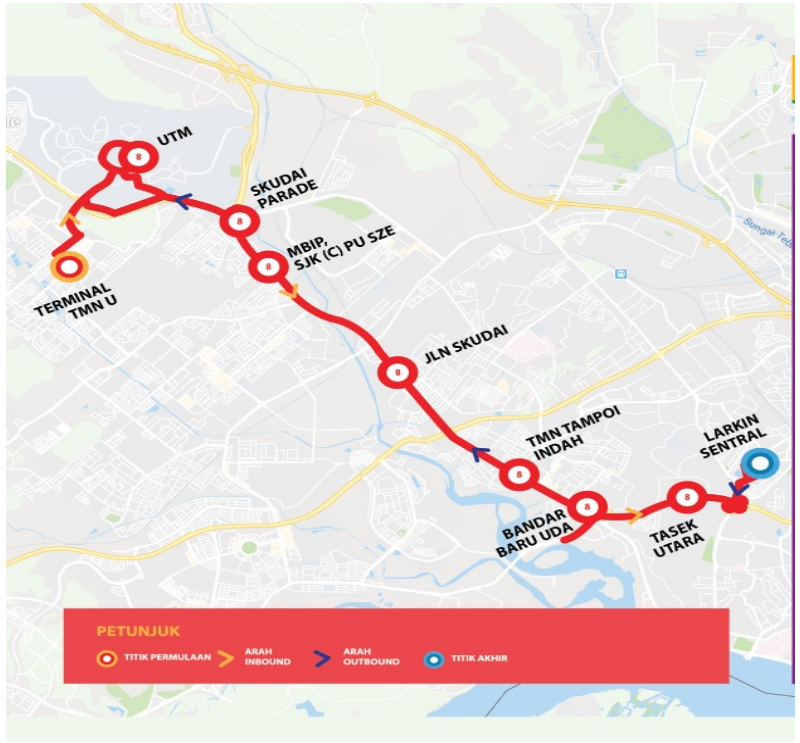


Current model requires high computational cost (deep learning, SVM, Time series models)

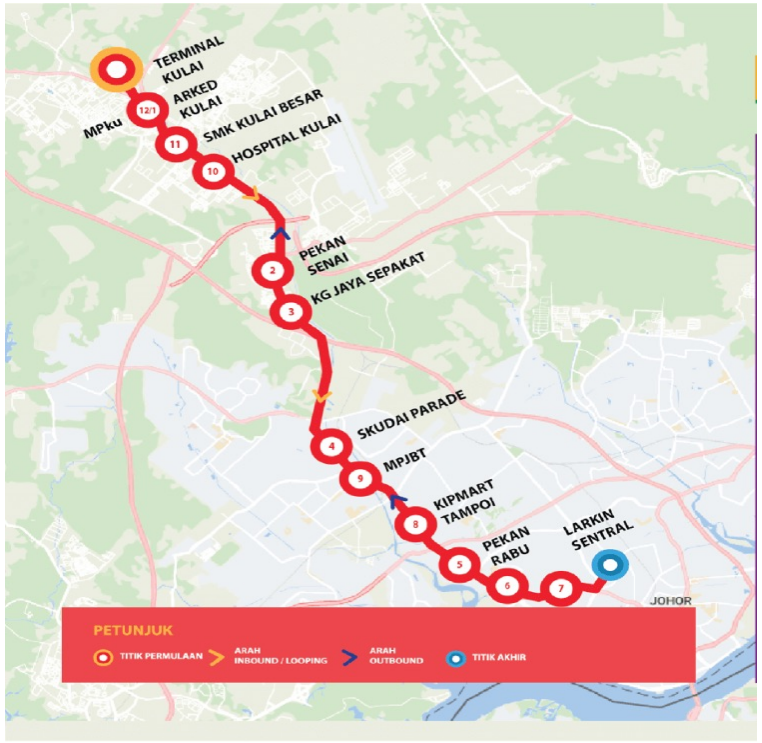
Contribution

- A data cleaning method to improve data quality by removing duplicate data from Bluetooth Low Energy (BLE) based system, breaking down bus routes into segments, and transforming data for training.
- Feature selection method to select the most relevant features through the use of Boruta.
- Boosting models to minimize computational power while preserving accuracy.

Methodology



P211 bus route from Larkin Sentral to Terminal Taman Universiti



P411 bus route from Larkin Sentral to Terminal Kulai

Data Cleaning → Feature Selection → Modelling

UNIVERSITI SAINS MALAYSIA

Data preparation

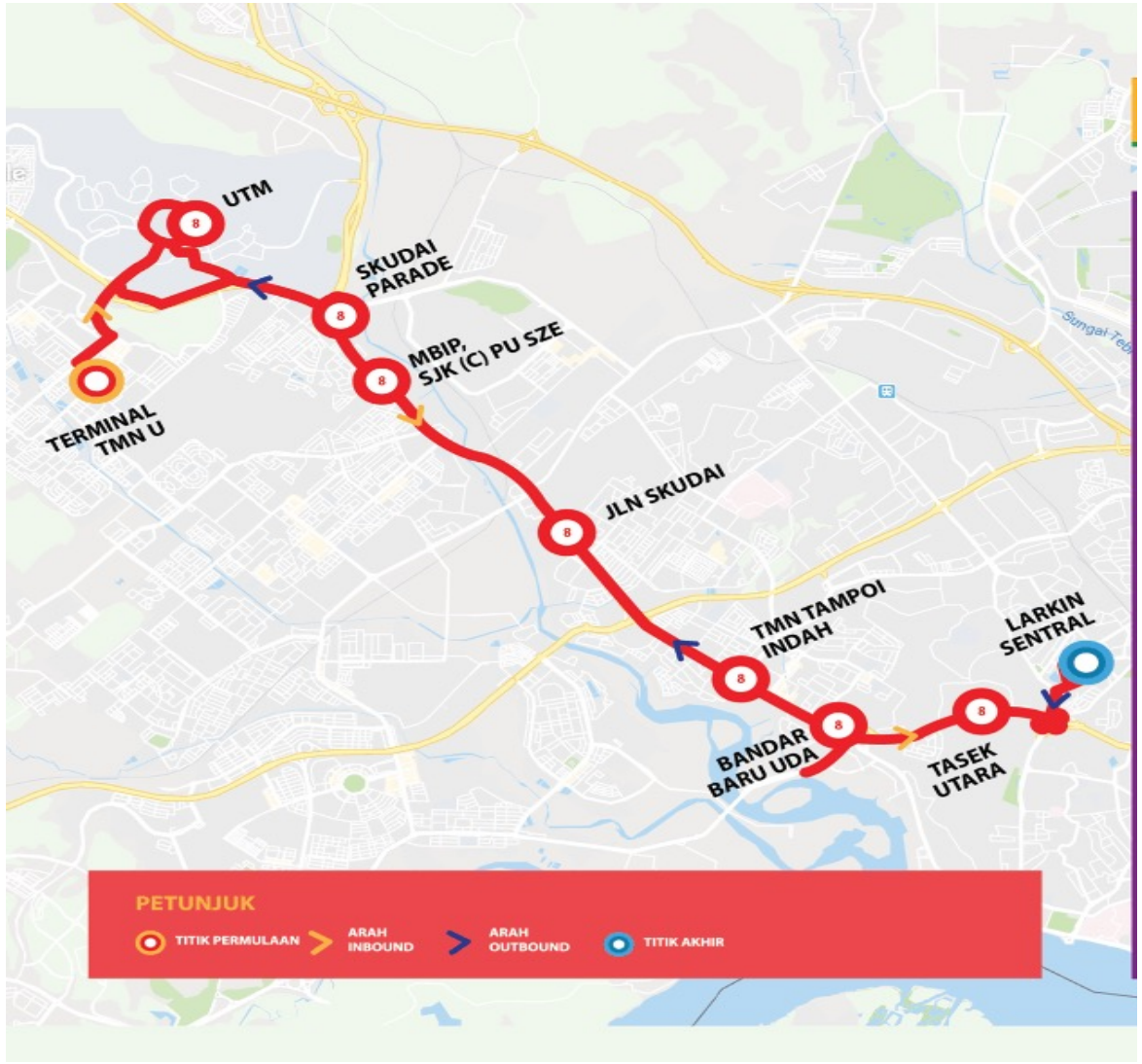
Map the location of the bus to the nearest bus stop

Remove duplicates and keep the latest record

Slicing route into sections and compute the required time to travel across

Encode cyclical data using sine / cosine method

Remove Outlier Using IQR method



UNIVERSITI SAINS MALAYSIA

Feature selection

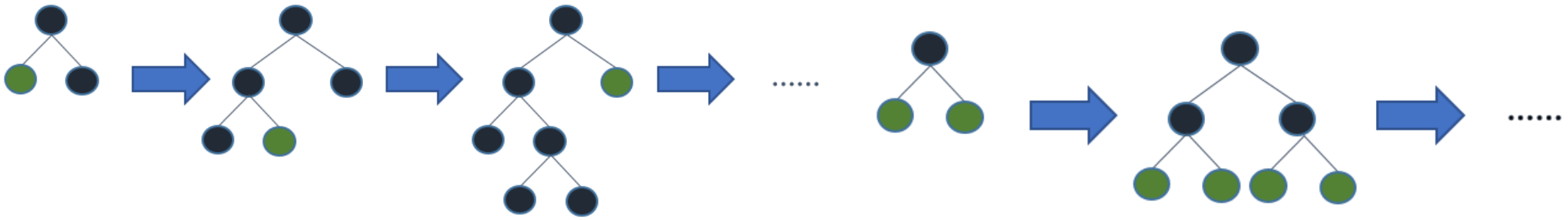
Features	Description
Bus id	The unique id of the bus
Route id	The unique id of the route
Latitude	Latitude of the bus
Longitude	Longitude of the bus
TimeStamp	Timestamp when the instance is recorded
Bus stop *	Bus stop no of the bus
Route order *	Current order of the route
Dest route order *	Destination of the bus
Name *	Nearest Bus stop name
Route *	Name of the route
dist_from_bus_stop *	Current distance of bus from the nearest bus stop
Sin half hour *	Hour encoded in sine
Cos half hour *	Hour encoded in cosine
Sin month *	Month encoded in sine
Cos month *	Month encoded in cosine
Sin_day *	Day encoded in sine
Cos_day *	Day encoded in cosine

Target : Time difference

UNIVERSITI SAINS MALAYSIA

Models

- LightGBM:
  - Leaf-wise tree growth.
  - It is designed to be memory-efficient and faster than traditional gradient boosting implementations.
  - Uses a technique called Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to improve training speed.
  - It works well with large datasets and is particularly effective in scenarios where data size and training speed are critical factors.
- XGBoost:
  - Level-wise tree growth.
  - It utilizes a gradient boosting framework that builds a series of weak learners (typically decision trees) sequentially.
  - Uses a variety of regularization techniques to prevent overfitting, such as shrinkage (learning rate) and column subsampling
- AdaBoost:
  - Ensemble learning method that combines multiple weak classifiers to build a strong classifier.
  - Assigns weights to instances and adjusts them at each iteration to focus on the difficult-to-classify instances.
  - It sequentially trains a series of weak learners (e.g., decision trees) and adjusts the weights of misclassified instances to emphasize their importance in subsequent iterations.



Leaf-wise tree growth

Level-wise tree growth

UNIVERSITI SAINS MALAYSIA

Results

Algorithms	Training				Testing				Training Time (s)
	R2	*MAE*	*MSE*	*RMSE*	*R2*	*MAE*	*MSE*	*RMSE*	
With cyclical encoding features									
LightGBM	0.8705	26.9656	1529.5820	39.1009	0.8468	30.0727	2079.6053	45.6027	3.75
AdaBoost	0.5409	53.6239	5424.0889	73.6484	0.4899	64.2462	6923.4502	83.2073	0.2031
XGBoost	0.8361	30.4841	1936.6205	44.0071	0.7481	37.8071	3419.7583	58.4787	9.1875
Without cyclical encoding features									
LightGBM	0.8219	31.9949	2104.1626	45.8712	0.8145	32.7940	2517.8129	50.1778	2.4219
AdaBoost	0.5409	53.6239	5424.0889	73.6484	0.4899	64.2462	6923.4502	83.2073	0.1719
XGBoost	0.7858	34.8125	2530.1913	50.3010	0.7184	41.4807	3822.7701	61.8286	6.5156

Discussion

- LightGBM achieves the highest accuracy with moderate training time compared with XGBoost and AdaBoost.
- Time differences exist between routes and hours.
- Cyclical features improves the R2 value about 3% and 3-5 seconds in terms of RMSE value.

Limitation and Future Work

- Limitations : Does not include parameters such as flood, traffic condition, accidents.

- Future works : Stacking to mitigate the noisy data issue

UNIVERSITI SAINS MALAYSIA

Thank you