Bus Travel Time Prediction using Extreme Gradient Boosting

Prashansa Agrawal[1], Jahnavi Yarlagadda[2], Antony Franklin[1], and Digvijay Pawar[2]
[1]Department of Computer Science & Engineering, IIT Hyderabad, India
[2]Department of Civil Engineering, IIT Hyderabad, India
{cs18mtech11037, ce17resch11013}@iith.ac.in, antony.franklin@cse.iith.ac.in, dspawar@ce.iith.ac.in

Objective

- This study aims to analyze various parameters that affect travel time in Indian public transportation scenario.
- We propose an optimized tree-based ensemble algorithm called Extreme Gradient Boosting (XGBoost) for the bus arrival time prediction.

Motivation and Data Collection

- Inspired by the need, various Model-Driven & Data-Driven approaches are proposed in literature.
- However, only limited studies talk about ensemble models in transportation, specifically predicting travel time for Indian traffic conditions.
- We propose an *optimized tree-based ensemble algorithm* called *Extreme Gradient boosting (XGboost)*, which was originally proposed by Chen and Guestrin (2016).
- A study stretch of 26 km is selected near Hyderabad city on the four-lane divided national highway (NH-65).
- The data is collected for the passenger buses by mounting a high-end GPS data logger.
- Total 92 trips data were collected comprising of 69 travel hours and 2116 kilometers.

Figure 1: Experimental Route (IITH-BHEL X Road)

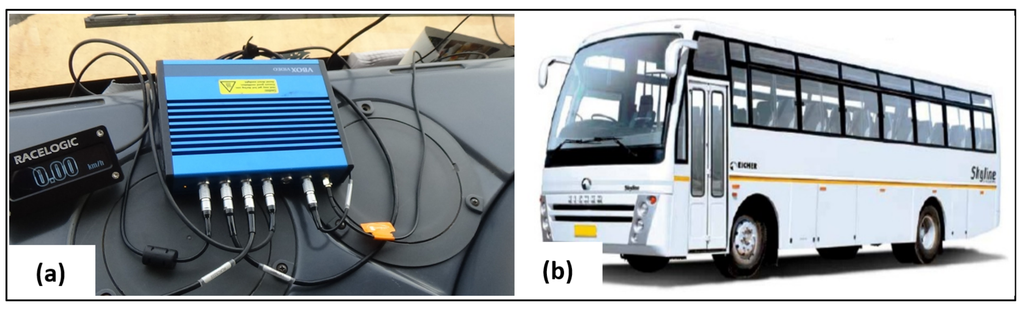


Figure 2: (a) GPS Data Logger. (b) Study vehicle used for data collection.

Proposed Methodology, Data Preparation and Analysis

- The latitude and longitudinal information is used to track the bus over the selected route and at various junctions on the route.
- To account the effect of different traffic conditions on bus travel time, the study stretch is divided into segments characterized by important junctions on the route namely IITH Main Road (I), Isnapur (II), Patancheru (III), and BHEL-X-Road (IV)
- The GPS coordinates of the bus are tied to the junctions on the road to locate the bus at various segments using Map matching algorithm.
- Haversine formula is used to determine distance between two coordinates.
- Map matching algorithm iterates through each coordinate in the GPS log (log_co) and identifies the nearest junction (seg_co) which satisfies the reference distance thresholds.
- For link travel time, the study stretch is segmented, and a unique ID is assigned to each segment.
- Along with the selected parameters, the segment ID, distance, and the time required to travel previous segment are used as parameters.
- The day timings are divided in three intervals as morning (peak hours), afternoon (off peak hours), and evening (peak hours).
- There is no significant trend observed on different weekdays, a parameter called weekday/weekend is used to differentiate between weekdays and weekends.
- The direction of travel is also considered as a parameter to account the travel time variations due to change in traffic conditions with respect to direction of travel.

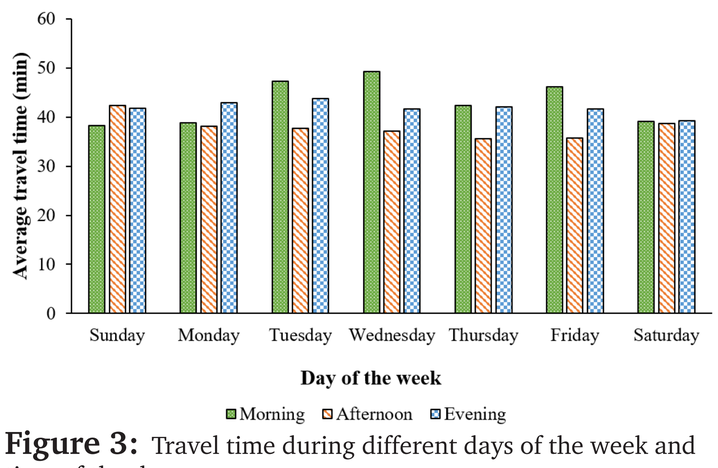


Figure 3: Travel time during different days of the week and time of the day

Conclusion and Future Work

- Extreme Gradient Boosting has not been evaluated in the present studies on GPS data for bus arrival time prediction.
- We explored XGBoost to model travel time on GPS data using various parameters on which the bus is running to implicitly learn the traffic patterns.
- Incorporating the mentioned parameters, XGBoost is found to predict significantly better than other benchmark models such as Gradient Boosting Machine, Random Forest and Support Vector Regressors.
- Exploring the impact of temporal correlation on prediction accuracy by considering time series data along with spatial correlation is the future scope of this work.

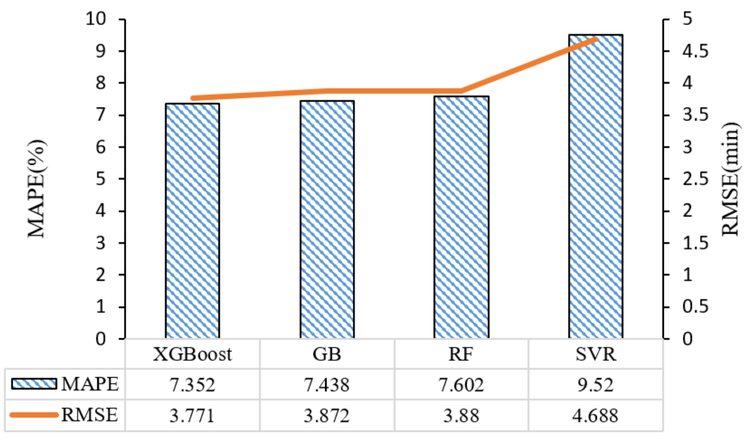
Experimental Evaluation

- In case of whole stretch, weekday/weekend is observed to be the most important attribute, while the direction and time of travel had similar importance.
- In case of link travel time prediction, feature importance of distance is evident as it is directly proportional to travel time.
- However, it is interesting to see the impact of travel time of previous segment on current segment.
- Other than these, time of travel also seemed to be an important feature.

Whole stretch	
Parameter	**Description**
Direction	0 and 1 for opposite directions
Time	Morning/Afternoon/Evening
Weekday/weekend	1 for weekend, 0 for weekday

Links	
Parameter	**Description**
Segment id	Each segment assigned with unique id
Distance	Segment length
Direction	0 and 1 for opposite directions
Time	Morning/Afternoon/Evening
Weekday/weekend	1 for weekend, 0 for weekday
Prev_segment_tt	Travel time required for previous segment

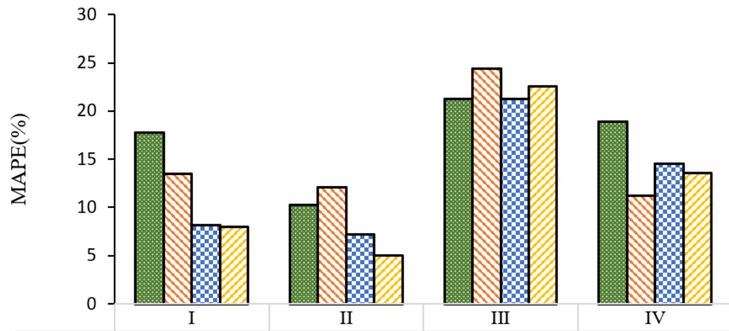
- XGBoost is observed to outperform GB, RF, and SVR with MAPE of 7.35%.
- While GB is observed to give second best performance at MAPE of 7.43%.
- Accuracy of RF is also observed to be very close to GB at MAPE of 7.6%.
- SVR is observed to give least performance when compared to mentioned models, with MAPE of 9.5%.



	XGBoost	GB	RF	SVR
MAPE	7.352	7.438	7.602	9.52
RMSE	3.771	3.872	3.88	4.688

Figure 4: Results for prediction on whole stretch (10-fold cross validation)

- In segments (links) I, II, and IV, we can observe a clear improvement in accuracy using XGBoost.
- On segment III, we can observe that SVR and GB outperform RF and XGBoost, while XGBoost gives better overall performance.



	I	II	III	IV
SVR	17.72	10.28	21.289	18.88
RF	13.524	12.088	24.364	11.261
GB	8.16	7.21	21.231	14.53
XGBoost	8.01	5.038	22.55	13.58

Figure 5: Results for link travel time prediction