MACHINE LEARNING BASED TRAVEL TIME PREDICTION OF URBAN BUS TRANSIT USING GPS DATA

Ratneswaran Shiveswarran

228017T

Master of Science (Major Component Research)

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

August 2023

MACHINE LEARNING BASED TRAVEL TIME PREDICTION OF URBAN BUS TRANSIT USING GPS DATA

Ratneswaran Shiveswarran

228017T

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Science (Major Component Research)

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

August 2023

DECLARATION

I declare that this is my own work and this Thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: *UOM Verified Signature* Date: 09/08/2023

The supervisors should certify the Thesis with the following declaration.

The above candidate has carried out research for the Master of Science (Major Component Research) Thesis under our supervision. We confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr Uthayasanker Thayasivam

Signature of the Supervisor: Date: 09/08/2023

Name of Supervisor: Dr Sivakumar Thillaiampalam

Signature of the Supervisor: *UOM Verified Signature* Date: 10/08/2023

ACKNOWLEDGEMENT

I am sincerely grateful for the advice, guidance and constant support given by my beloved supervisor Dr T. Uthayasanker. Without his help and encouragement, this research project would not have been completed. I would like to thank him for taking time out of his busy schedule to be available any time to support my research works and motivate me in my downs.

I would also like to thank my co-supervisor Dr T. Sivakumar for his valuable feedbacks and suggestions. With this, I would like to thank the faculty members in the DataSEARCH research centre, Prof. Amal Kumarage and Dr Shehan Perera for their generous support, feedback and continuous guidance and the other researchers in the centre for their assistance in my progress.

I am extending my gratefulness to the progress review committee members, Dr A. Shehan Perera and Dr Chitraranjan. Their valuable insights and guidance helped me immensely in the progress of this research.

This research was supported by a grant from the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Higher Education & the World Bank. I am sincerely thankful for their financial support during this project.

Last but not least, a tremendous thanks to my family and all my friends who are always back with me.

ABSTRACT

An accurate and reliable arrival time prediction of buses to the next bus stops is a valuable tool for both passengers and operators. Existing studies have some limitations in bus travel time prediction. They focus little on three aspects such as heterogeneous traffic flow conditions, dwell time prediction and interpretation of explanatory variables. Consequently, we break down the prediction problem into sub-models for running times and dwell time prediction and incorporate a feature engineering framework that generates features related to the running bus, the prediction day, and immediate and long historical time variations to capture heterogeneous traffic conditions. We propose a multi-model stacked generalisation ensemble model by leveraging the advantages of best-performing models in homogeneous conditions such as Extreme Gradient Boosting (XGBoost) and convolutional long short-term memory (ConvLSTM) models. It outperformed the state-of-the-art models by 11% in mean absolute error (MAE) on average. It can predict extreme conditions in bus journeys more accurately.

Nevertheless, the input data for the machine learning model should be the historical travel times of the route. We proposed two simple novel algorithms to extract bus trips and match bus stop sequences towards extracting dwell times and running times from the raw crude GPS data generated at a medium sampling frequency of 15 seconds. Those algorithms incorporate various challenges like non-uniformity, poor network coverage, discontinuities in streaming and skipping of bus stops. In addition, we attempted to interpret the feature importance of the generated features. We found insights like driver behaviour and the immediately preceding dwell time influence the stopping pattern and the prediction model, which pave the way for strategic management by authorities.

Keywords: Bus travel time prediction, Machine learning, Multi-model ensemble, Ensemble Learning, GPS data processing, Heterogeneous traffic

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	**Description**
AFC	Automatic Fare Collection
ANN	artificial neural network
APC	Automated Passenger Counts
ATIS	Advanced Travel time Information System
AVL	Automatic Vehicle Location
DNN	deep neural network
ES	exponential smoothing
ETL	Extract Transform Load
GPS	Global Positioning Systems
GTFS	General Transit Feed Specifications
ITS	Intelligent Transportation Systems
KF	Kalman filter
kNN	k-nearest neighbor
LSTM	long short-term memory
MAE	Mean absolute error
MAPE	Mean absolute percentage error
ML	Machine Learning
MSE	Mean squared error
OSM	Open Street Map
ReLU	Rectified Linear Unit
RF	Random forest
RMSE	Root mean square error
RMSProp	Root Mean Square Propagation
RNN	Recurrent neural network
SVM	support vector machines
XGBoost	Extreme Gradient Boosting