MODELING AND ENHANCING FUEL ECONOMY OF FLEET VEHICLES BASED ON DATA ANALYTICS

Sandareka Kumudu Kumari Wickramanayake

158040G

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

December 2018

MODELING AND ENHANCING FUEL ECONOMY OF FLEET VEHICLES BASED ON DATA ANALYTICS

Sandareka Kumudu Kumari Wickramanayake

158040G

Thesis submitted in partial fulfillment of the requirements for the degree Master of Science by Research (Part-time) in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

December 2018

DECLARATION

"I declare that this is my own work and this dissertation 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. Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books)."

Signature: Date:

The above candidate has carried out research for the Masters Dissertation under my supervision.

Name of the supervisor: Dr. H.M.N.D Bandara

Signature of the supervisor: Date:

Name of the supervisor: Mr. Nishal Samarasekara

Signature of the co-supervisor: Date:

Abstract

Fuel consumption of a vehicle depends on several internal factors such as distance, load, vehicle characteristics, and driver behavior, as well as external factors such as road conditions, traffic, and weather. Moreover, not all of these factors are easily obtainable for the fuel consumption analysis. Therefore, fuel-fraud is relatively easier to conceal; thus, considered a significant threat to the fleet industry by managers. This research model and evaluate the fuel consumption of fleet vehicles based on vehicular data and suggest suitable process improvement actions to improve the fuel economy. We first model and predict the fuel consumption to identify possible frauds. We considered a case where only a subset of the factors mentioned above is available as a multivariate time series from a long-distance public bus. An evaluation of several machine learning techniques revealed that Random Forest could predict fuel consumption with 95.9% accuracy. To verify the detected cases of possible fuel fraud, we propose to use different indicators such as speed profile, the frequency of harsh events, total idle time, and day of the week. Further, we propose a solution to promote fuel-efficient driving through real-time monitoring and driver feedback. A classification model, derived from historical data, identifies fuel inefficient driving behaviors in real-time. The model considers both the driver-dependent and environmental parameters such as traffic, road topography, and weather in determining driving efficiency. If an inefficient driving event is detected, a fuzzy logic inference system is used to determine what the driver should do to maintain fuel-efficient driving behavior. The decided action is conveyed to the driver via a smartphone in a nonintrusive manner. We demonstrate that the proposed classification model yields an accuracy of 85.2% while increasing the fuel efficiency up to 16.4%.

Acknowledgment

I would like to dedicate my sincere thanks to my supervisor Dr. H.M.N Dilum Bandara for his dedicated support for the success of this research. This research would not have been a success without your guidance from the initial stage to the final phase of the research. This research was supported by Nimbus Venture (Pvt) Ltd, Sri Lanka providing related data sets and their insightful expertise knowledge. Thank you for the support given by providing domain knowledge and feedback throughout this research. I would like to thank the entire academic and non-academic staff of the Department of Computer Science and Engineering for their kindness extended to me in every aspect. Last but not least, I thank my parents, my husband and all my friends who supported me for the success of this piece of work. Your support was always precious.

Contents

List of Figures

List of Tables

List of Abbreviations

%IncMSE	Percentage of increment of MSE
3G	The third generation of wireless mobile telecommunications technology
ANN	Artificial Neural Network
API	Application Programming Interface
ARIMA	Auto-regressive Integrated Moving Average
COD	Code of Determination
FC	Fuel Consumption
GAM	Generalized Additive Models
GB	Gradient Boosting
GLM	Generalized Linear Models
GPS	Global Positioning System
HHDDT	Heavy Heavy-Duty Diesel Truck
HMI	Human Machine Interface
M2M	Machine-to-Machine
MAE	Mean Absolute Error
MARS	Multivariate Adaptive Regression Splines
MSE	Mean Squared Error
NSE	Nash-Sutcliffe efficiency
PD	Proportional-Derivative
REST	Representational State Transfer
RF	Random Forest
RMSE	Root Mean Squared Error
RPM	Revolutions per minute
WWO	World Weather Online

1 INTRODUCTION

1.1 Overview

With rising fuel prices and fast diminishing global fossil fuel reserves, it is imperative to improve the fuel efficiency of vehicles while looking for reliable, alternative energy sources. Although electric and hybrid vehicles are becoming more popular, the world is still dependent mostly on fossil fuel sources [1]. From this fossil fuel consumption, a significant portion is attributed to transportation. For example, in Sri Lanka, 82.3% of the total fossil fuel consumption is attributed to the transportation sector [2]. Hence, it is imperative to improve the fuel efficiency of vehicles to reduce fuel usage and thereby save money. Besides, saving fuel means protecting the environment.

Numerous factors contribute to the fuel consumption of a vehicle. Demir et al. [3] classified the factors influencing fuel consumption into five categories such as vehicle, environment, traffic, driver, and operations. While the efficient engine and vehicle designs (usually improved aerodynamics) can gain substantial fuel savings, further fuel savings are achievable by optimized driving and scheduling. Such optimization in the context of fleet management includes adapting driving patterns that could save fuel, reducing wastage and fraudulent activities, routing vehicles around traffic and optimized fleet scheduling.

Many related work have emphasized the impact of driver behaviors on fuel consumption of vehicle [4, 5, 6, 7, 8, 9]. For example, Gondar et al. [4] showed that efficient driver behaviors could provide up to 20% fuel savings. While drivers can be educated on general guidelines, further savings can only be achieved by individual feedback about driving patterns. Furthermore, if drivers can be provided suggestions in real-time to adjust their driver behaviors into fuel-efficient driver behaviors, a significant improvement in fuel economy could be achieved. However, recommendations provided should be practical; thus, driver behaviors should be evaluated for fuel efficiency considering as many as possible internal and external factors such as driver behavior, route details, road traffic conditions, weather and load information.

Another way to reduce/ eliminate fuel waste is to detect and prevent fuel frauds. Fuel being the single major cost of fleet industry, fleet managers can save millions per year by avoiding fuel frauds. Even when vehicles are refilled at authorized or on-site fuel stations, drivers manage to siphon diesel from the tank along the route and sell. Notably, long distance fleets such as buses, distribution trucks, and heavy vehicles (e.g., ready-mix concrete trucks) traveling under different road and traffic conditions

are more vulnerable to fuel fraud. Fleet managers complain that just by installing traditional fuel level monitors would not solve the problem as now driver deceits those sensors by pumping soap water into fuel tanks. All these raise the need for an intensive analysis for fuel fraud detection.

This research attempted to find solutions for the complications mentioned above via vehicular data analytics. Such analysis typically requires high-resolution data from GPS vehicle tracking system units, various sensors and other external sources such as weather data, and traffic data sources that are gathered across multiple days and vehicles. Given the volume, diversity and uncertainty of data, sophisticated data mining techniques are required to model fuel consumption and present identified recommendations to individual drivers.

Data analytics solutions can bring immense benefits to the fleet industry. First, the most important and prominent advantage is the cost reduction. Avoiding aggressive driving, also known as eco-driving, is the best way to reduce fleet management cost. Eco-driving reduces fuel expenditure as eco-driving is fuel efficient, reduces vehicle maintenance cost as eco-driving protect vehicle health and reduces insurance premium as usage-based insurance determines premium based on driver behaviors and records. We can encourage divers to eliminate aggressive driving by analyzing their diver behaviors and providing advises or suggestions. Large volume of data generated by various sensors in fleets, as well as other external sources such as weather and traffic data generators make it possible to do a comprehensive driver performance analysis. Second, fleet managers can optimize route planning for their fleets using insights produced by data analytics. The third benefit is, an intensive data analysis can reveal various fraudulent activities happening in the fleet industry and save millions of dollars per year. Further, better public perception is the key to success of a fleet company, especially taxi companies. Reduced aggressive driving and optimized route planning allow fleet companies to provide impressive customer service and protect their public image which is an indirect benefit of the vehicular data analysis.

1.2 Problem Statement

The primary problem to be addressed by this research is how to enhance the fuel economy of fleet vehicles via vehicular data analysis. In that direction we focus on two main tasks (a) identifying fuel frauds and (b) encouraging fuel-efficient driving behaviors via real-time individual feedback.

To detect fuel frauds, we predict instantaneous fuel consumption, compare the pre-

dicted value against recorded fuel usage and consider cases with significantly higher recorded fuel usages as possible fuel frauds. Formally, given a set of influencing factors $X = x_1, x_2, ..., x_n$ for a given point of time t we need to predict the fuel consumption of the vehicle y at t. Thus, task is to find a function f that captures the relationship such that:

$$y = f(x_1, x_2, ..., x_n) \tag{1.1}$$

Also, given the recorded instantaneous fuel consumption y^* we need to find a function $g(y, y^*)$ that determines if a fuel fraud might have happened.

In the second task, given a set of influencing factors $X = x_1, x_2, ..., x_n$ within a time period t and fuel consumption of the vehicle for the same period y, we need to classify whether this driving event is fuel-efficient or not. Thus, the task is to develop a classifier c such that:

$$c(x_1, x_2, ..., x_n, t, y) \rightarrow efficient/inefficient \tag{1.2}$$

If the driving event is classified as *inefficient* we further need to suggest a possible driving action to bring the vehicle back to the *fuel-efficient* state.

1.3 Research Objectives

The objectives of this research can be stated as follows:

1. Detect possible fuel frauds — The fleet industry is vulnerable to fuel frauds such as illegal fuel pull-outs. Drivers not only pull-out fuel but also refill the tank with soapy water to deceit fuel sensors. Therefore, simple monitoring of fuel level would not reveal fuel frauds. Fleet managers find these illegal fuel siphons as severe threats because not only they cause extra cost but also affect the vehicle health adversely. All these raise a need for a sophisticated mechanism to detect fuel frauds.

2. Verify identified prospective fuel frauds — Fuel consumption of a vehicle depends on many internal and external influences. Weather condition, road traffic, number of passengers and road topography are some of the factors that are out of control of the driver which can vary fuel consumption significantly even for same trips. Therefore, just by looking at the excess fuel usage, one can't conclude that a fuel fraud has taken place. Instead, in practical situations, drivers are given a chance to explain the detected variability in fuel usage of a journey. The driver

may provide various excuses such as extensive traffic and severe weather. Then the data analytics solution should be able to confirm or reject these claims based on the data. For example, driving under extensive traffic can be determined based on vehicle speed profile and acceleration. Hence, it is vital to integrate various sources of data such as traffic and weather with data collected from the devices on the vehicle and devise necessary indicators to validate drivers' claims.

3. Direct drivers to maintain a better fuel economy for their vehicles — While several internal and external factors influence the fuel economy of vehicles, driving behavior is the most economically controllable influencing factor. Even though drivers can be taught about fuel-efficient driving behaviors via training in general, individual real-time feedback can encourage a driver more effectually to adhere to fuel-efficient driving behaviors. How can be data analytics used to provide real-time personal driver feedback?

1.4 Outline

In this thesis we carry out a vehicular data analysis to derive insights and use them for the betterment of the fleet industry particularly in terms of fuel saving.

In Chapter 2 we discuss existing work in the area of vehicular data analysis. We present our literature survey under main three topics. First, we examine factors which influence the fuel economy of a vehicle. Second, we explore current work to predict instantaneous fuel consumption of vehicles and assumptions and limitations of those work. Finally, we look into existing efforts to monitor driver behavior, explore the impact of driver behavior on fuel economy and the proposed methods to mitigate adverse effects.

Details about our dataset and exploration analysis of the dataset are presented in Chapter 3. We discuss in detail how internal and external factors influence fuel consumption of the vehicle.

In Chapter 4 we develop a machine-learning model for instantaneous fuel consumption prediction with the intention of identifying possible fuel frauds. In this chapter, we present details about three different models we examined for the task, namely Random Forest, Gradient Boosting and Multi-Layer Perceptron. We also discuss what are the factors we selected to develop these models, experimental results and their analysis

In Chapter 5 we propose a novel framework to analyze driver behavior in real-time,

identify fuel-inefficient driving patterns and provide continuous feedback to the driver so that they can continuously maintain a fuel-efficient driving behavior. We propose to use historical data to derive heuristics to classify driver behavior for fuel efficiency. This framework is especially useful in the fleet industry not only to save money in terms of saving fuel and maintaining better health of vehicles, but also to do better appraisals for their drivers. Fleet managers can use statistics from our framework to identify better drivers and appreciate them, which would encourage drivers to follow fuel efficient driving patterns.

Finally, in Chapter 6 we summarize our work, discuss the limitations of the proposed approaches, remaining challenges and the path forward.

2 LITERATURE REVIEW

Several factors influence the fuel consumption of fleet vehicles; hence, essential to understand before attempting to predict fuel consumption. Moreover, it is also important to consider data analysis techniques to monitor driver behavior and promote eco-driving behaviors. Therefore, the literature survey focuses on these areas and associated technologies. Section 2.1 discusses factors which influence fuel consumption of fleet vehicles. Section 2.2 presents existing work to predict fuel consumption of vehicles. A summary of these current work is given at the end of this section indicating findings and limitations. Further, this section includes a brief introduction to machine-learning techniques used in this research. Section 2.3 illustrates current work which use data analysis to monitor driver behavior and promote eco-driving behaviors. Finally, in Section 2.4 a summary of the literature review is given.

2.1 Factors influencing fuel economy of fleet vehicles

The fuel economy of a vehicle is defined as the relationship between the distance traveled and the amount of fuel consumed by that vehicle [10]. Fuel economy can be measured in terms of volume of fuel per unit distance or the distance traveled per unit volume of fuel consumed. Numerous factors influence fuel consumption of a fleet vehicle. Broadly these factors can be categorized as internal factors; elements controllable by the diver and external factors; elements out of control of the driver.

Ahn et al. [11] identified numerous variables that influence vehicle energy and emission rates. They have classified these variables into six main categories as:

i. Travel-related – Distance traveled and number of trips went in the considered period

ii. Weather-related – Temperature, humidity and wind effects

iii. Vehicle-related – Engine size, the condition of the engine, whether the vehicle is equipped with a catalytic converter, whether the vehicle's air conditioning is functioning and the soak time of the engine

iv. Roadway-related – Roadway grade, Surface roughness

v. Traffic-related – Vehicle-to-vehicle and vehicle-to-control interaction

vi. Driver-related – Differences in driver behavior and aggressiveness

Viswanathan in [12] has studied reasons for the variation in fuel consumption of heavy vehicles running on highways. She has categorized those reasons into four as:

i. Due to vehicles

ii. Due to road

iii. Due to the usage (Driver)

iv. Due to ambient conditions (Temperature, wind, etc.)

2.2 Fuel consumption prediction using vehicular data analytics

Ahn et al. [11] presented several hybrid regression models to estimate fuel consumption and emission rates of hot stabilized light-duty vehicles and light-duty trucks. For this analysis, they have used a laboratory generated data set. Even though the laboratory may not precisely capture the real-world scenario, one plus point about this data set is it contains harsh accelerations; as per the industry threshold accelerations higher than $3ms^{-2}$ are considered as harsh accelerations. To develop a prediction model, they have selected instantaneous speed and acceleration as explanatory variables arguing that those two can capture most of the impact on fuel usage and emission rates of a vehicle.

The authors of this research have tried out different mathematical models using linear, quadratic, cubic and quartic combinations of instantaneous speed and accelerations. With empirical evidence, they have finalized the model as a combination of linear, quadratic, and cubic combinations of speed and acceleration, which has given reasonable results. However, in some situations, this model has predicted minus values for the dependent variable. To eliminate getting minus values, they have considered a log-transformed regression model. In testing, they have seen a significant variation in the difference between predicted value and the actual value of positive accelerations vs negative accelerations. Ahn et al. have explained this variation saying it is because acceleration exerts energy, but negative acceleration or deceleration does not. To address this situation, they propose to use two different models for the two cases. The proposed models for Measure of Effectiveness(MOE) are as follows.

$$ln(MOE) = \sum_{i}^{3}\sum_{j}^{3}(L_{i,j}^{e} \times s^{i} \times a^{j}) for a \geq 0 \tag{2.1}$$

$$ln(MOE) = \sum_{i}^{3}\sum_{j}^{3}(M_{i,j}^{e} \times s^{i} \times a^{j}) for a < 0 \tag{2.2}$$

where L and M are constants, s denotes speed and a denotes acceleration.

The coefficient of determination (coefficient of determination indicates how well the regression model approximates the actual data points) of this final model is from 0.92 – 0.99. However, this model may not demonstrate the same results when it is applied to a real-world scenario as the model has been developed using a laboratory generated dataset. Impact of real-world external factors might not be reflected in this dataset. Further, this proposed model assumes vehicles to be hot stabilized which is not a realistic assumption in the fleet industry.

Viswanathan [12] has tried to develop a predictive model to identify and classify usage and driving parameters that affect fuel consumption. She has analyzed a dataset of trucks traveling on highways. She has used Random Forest and Gradient Boosting machine learning models for this analysis praising their predictive power and simplicity. Performance of each model has been evaluated using Nash- Sutcliffe efficiency.

Random Forest, the first algorithm author of [12] has used is an ensemble algorithm that uses a collection of regression/ decision trees to make the prediction. Nash-Sutcliffe efficiency of fuel consumption prediction by Random Forest based model is 0.808 which indicates that Random Forest can capture the relationship between predictive parameters and dependent parameter very well. The other predictive model used in this research, Gradient Boosting is also an ensemble learning algorithm for regression and classification problems. The Nash-Sutcliffe efficiency for Gradient Boosting is 0.69. Therefore, this research concludes that Random Forest is better than Gradient Boosting to predict fuel usage of heavy vehicles running on highways. Further, the researcher has concluded that the most influencing factors for fuel consumption are Speed, Distance with cruise control, Distance with trailer, Maximum speed, Coasting. She has arrived this conclusion considering variable importance given by Random Forest (see Table 2.1) and selection frequency given by Gradient Boosting (Table 2.2)

However, this research is restricted to vehicles running on highways. Further, the author of [12] has not considered external influential factors on fuel consumption such

Table 2.1: Importance measurement of input variables by Random Forest.

Variable	**Importance Measurement**
Speed	75.575
Distance with trailer	13.944
Distance with cruise control	13.092
Max speed	12.374
Coasting	9.717

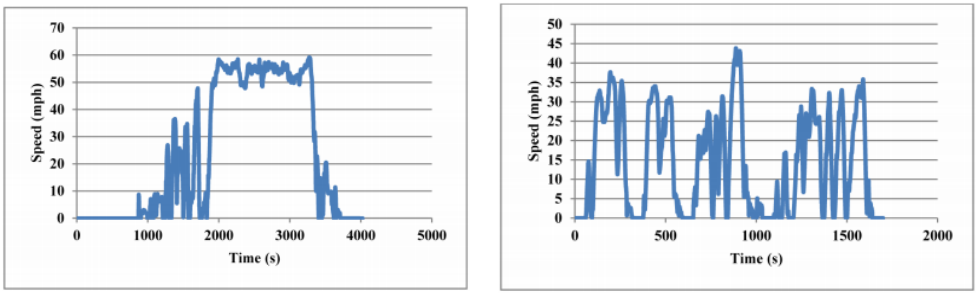
Table 2.2: GAM boosting- Selection frequencies.

Variable	**Selection Frequencies**
Speed	0.365
Coasting	0.185
Max speed	0.170
Distance with trailer	0.125
Distance with cruise control	0.65

as road conditions and ambient conditions. Furthermore, the prediction has been carried out in trip wise where the instantaneous impact of various factors was not captured.

In [13], the authors have evaluated four different predictive models in predicting fuel consumption of heavy/ medium duty vehicles based on driving cycle information,- vehicle speed, acceleration and road grade. The evaluated predictive models are a polynomial regression model, an artificial neural network, a polynomial neural network and multivariate adaptive regression splines (MARS). The primary assumption authors make in this research is that fuel consumption of a given vehicle only depends on known driving cycle properties. They justify their assumption using "road load equation" [13]. Using this equation, authors have shown that under constant vehicle weight fuel consumption is determined by driving cycle properties.

Based on the assumption that fuel consumption of a given vehicle only depends on known driving cycle properties, authors have used speed, acceleration and road grade to model fuel consumption. In this study, they have used four different driving cycles conditions as shown in Figure 2.1. Each of these driving cycles contains different driving characteristics such as idling, creep, transient and cruise mode [13]. Each of four models has been trained using one of four cycles as a training cycle and other three cycles as testing cycles. Among these four cycles, they could observe that data collected from Heavy Heavy-Duty Diesel Truck (HHDDT) driving cycle could contribute to developing a better predictive model. It is intuitive as HHDDT cycle contains a variety of driving characteristics. Among the four different predictive models, MARS has given the best predictive results with an average error percentage of -1.84% for the



(a) Heavy Heavy-Duty Diesel Truck (HHTDDT) speed trace.

(b) City Suburban Heavy Vehicle Cycle (CSHVC) speed trace.

(c) New York Composite Cycle (NYCC) speed trace.

(d) Hydraulic Hybrid Vehicle (HHV) speed trace.

Figure 2.1: Different driving cycles considered in related work [13].

laboratory generated data and -2.2% for real-world data.

However, the assumption in this analysis makes results less useful for real-world scenarios because the constant weight is not practical for most of the fleet vehicles. For example, the number of passengers on public transport buses or taxi service vehicles or the load of distribution trucks is not constant even throughout the same journey. Further, the influence of weather, an important critical factor towards fuel usage of a vehicle, is not considered in this study. A summary of fuel consumption prediction methods discussed above is given in Table 2.3.

2.2.1 Machine learning techniques for fuel consumption prediction

There are different machine-learning techniques to address complex classification and regression problems. Among these classification and regression models, “ensemble learning” methods have gained more significant interest. Ensemble learning can be defined as the process of generating multiple models such as classifiers and then aggregating their results to obtain better predictive performance [14]. Two well-known ensemble-learning methods are boosting and bagging [15]. In boosting, successive models give extra weight to training instances that were incorrectly predicted/ classified by previous models. While making the prediction/ classification, a weighted vote

Table 2.3: Summary of fuel consumption prediction work.

Ref	Dataset	Techniques	Conclusions	Limitations
[11]	Laboratory generated dataset	Hybrid regression models	Proposed model (a function of instantaneous speed and acceleration levels) estimates fuel usage with COD exceeding 90%	Only vehicles traveling on highways are considered Only for hot stabilized vehicles Dataset is not real-world
[12]	Usage and driving parameters	Random Forest, Gradient Boosting	RF is a good model to estimate fuel usage of heavy vehicles Speed has the highest impact Harsh acceleration & green band driving don't have a significant effect on fuel consumption	Instantaneous variation of FC was not reflected in data set as information is for per journey Only trucks running on highways are considered Impact of external factors are not considered
[13]	Laboratory generated dataset for training,& real time dataset for testing	Polynomial model, ANN, Polynomial NN, MARS	MARS is the best predictive model	Assume fuel consumption of a given vehicle depends only on driving cycle (Weather and traffic data are neglected)

is considered. Whereas in bagging, successive models do not depend on earlier models, instead, each model is independently constructed by a bootstrap sample of data. Prediction/ classification is developed based on a simple majority vote.

2.2.1.1 Random Forest

Random Forest (RF) proposed by Breiman [16], is an ensemble predictive model based on a collection of decision/regression trees. Instead of making the prediction based on just one tree, Random Forest uses a group of trees to take the decision. Being different from other bagging (bootstrap aggregation) techniques, Random Forest adds an additional layer of randomness to bagging. Like other bagging models, Random Forest also constructs each decision/ regression tree using a bootstrap of sample data with re-

placement. However, the tree building procedure is different. Instead of splitting trees using the best split among all variables, in a Random Forest each node is divided using the best among a subset of predictors randomly chosen at that node [15]. This strategy enables Random Forest to be robust against over-fitting and be outstanding among many other classifiers including discriminant analysis, support vector machines, and neural networks [14]. Further, facilitating the estimation of variable importance and outlier detection are other benefits of this algorithm. Random Forest is reasonably fast to obtain and can be easily parallelized [17]. A fine-tuned version of Random Forest can be obtained by backward-elimination of predictors based on the given variable importance.

Random Forest has been used in a myriad of domains to carry out predictions/ classifications and they are applicable for time series analysis as well. For example, Herrera et al. [17] used Random Forest to forecast hourly urban water demand in a city in southeastern Spain. Chen et al. [18] used Random Forest to predict droughts and demonstrated that Random Forest outperforms Auto-regressive Integrated Moving Average (ARIMA) in that context.

2.2.1.2 Gradient Boosting

Gradient Boosting (GB) is another ensemble predictive algorithm for regression and classification problems. Like other boosting algorithms, GB builds the model in stages and generalizes them by allowing optimization of an arbitrary differential loss function [19]. Different functions are used as the loss criteria; least square, least absolute deviation, and Huber-M loss function for regression and logistic likelihood for classification. Carrying out variable selection during the fitting process can be recognized as a key feature of GB [20]. Further GB algorithms provide prediction rules that have the same interpretation as common statistical models. This becomes a major benefit of GB over other machine learning algorithms such as Random Forest, which provides non-interpretable "black-box" predictions.

2.2.1.3 Artificial Neural Network

Artificial Neural Network (ANN) is a machine-learning technique inspired by biological neural networks and is mostly used to estimate or approximate complex functions that can depend on many inputs. Being analogs to neurons in the brain, ANN contains lots of processing units. ANNs can be used to derive linear and non-linear relationships

between a vast number of inputs and outputs. The architecture of the neural network, number of hidden layers and number of processing units (perceptrons) per each layer, loss function, activation functions and optimization function are critical considerations in designing a neural network for a problem [21]. ANNs are commonly used in a myriad of domains such as robotics, transportation, and finance.

Following advantages of ANN were considered while selecting it as one of the predictive models in this study. ANN requires less formal statistical training. Further ANN can implicitly detect complex nonlinear relationships between explanatory variables and response variables. Ability to identify all possible interactions between independent and dependent variables is also an advantage [22].

2.3 Analyzing driver behavior to improve the fuel economy of fleet vehicles

Walmun and Simonsen [6] examined the data generated by a fleet management system in Norway to find out determinants of fuel consumption by heavy-duty trucks. They concluded that even though under some conditions factors associated with infrastructure and vehicle have a more significant impact than driver behavior, in general driving pattern has a considerable influence on fuel usage of vehicles.

Furthermore, Fleetcarmar, a telematics platform provider, has published an analysis of real-world driving behavior [7] trying to quantify the effects of several facets of driver behavior on fuel consumption including speed, acceleration, breaking and idling. FleetCarma demonstrated that the impact of driving behaviors on fuel economy varies with the mass of the vehicle; higher the mass, higher the fuel consumption sensitivity to the driving behavior. Therefore, they bring up the point that fleet managers should address fuel economy problems of larger vehicles first to maximize savings. They have also shown that driver behavior can be enhanced through consistent feedback, and thereby fuel economy of the vehicle can be improved. Following are their suggestions for drivers to increase fuel economy.

- Reduce harsh acceleration and hard breaking – Try to keep harsh acceleration events less than 10% and hard breaking below than 15% of the acceleration during a journey
- Reduce idling time – Aim to eliminate idling event longer than 1 min.
- Reduce high speeds and use cruise control whenever possible

Following are the suggestions to motive drivers to adhere to fuel-efficient driving behaviors,

- Continuous feedback - Provide feedback on how drivers are currently doing and where they can improve on.
- Goal setting and management - Setting achievable goals for drivers is vital to improving driver behavior and reducing fuel spends. Managing goals makes it easy to reward drivers who meet the goal and plan additional driver behavior strategies for those who are having difficulty achieving the goal.

We have considered these recommendations in modeling our solution to tackle the problem of motivating drivers toward Eco-driving.

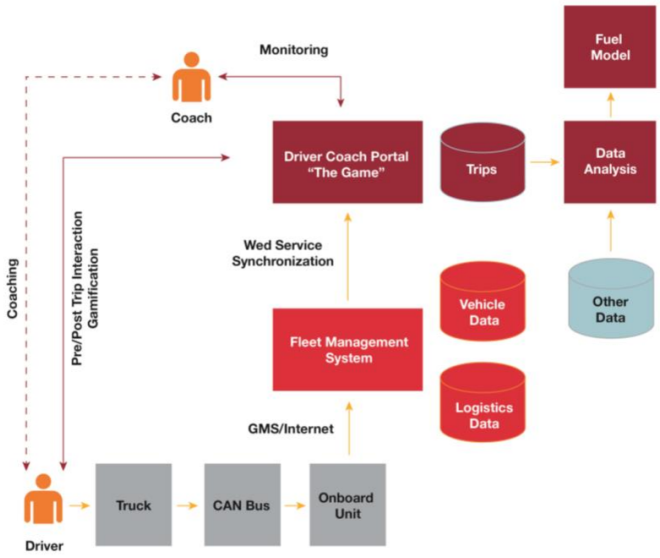


Figure 2.2: Eco-driving coaching service components.

Several current work have attempted to use individual driver feedback, based on data analytics, to improve the fuel economy of vehicles. A white paper published by CGI Group [5], a Canadian IT consulting company, talks about a driver-centric coaching program based on data analytics. Their client Scania, a fleet management company, has implemented a driver coach portal integrated with their fleet management system with the intention of encouraging drivers for fuel-efficient driving habits and other Eco-friendly driving patterns. Driver behaviors promoted by this portal are more roll-out, less hard braking, less idling, less hard accelerations, less high rpm, and more use of cruise control. Figure 2.2 shows the components of Eco-driving coaching portal. Trips are scored against benchmarks in the portal. Through a monitoring function, coaches

can see how their drivers are performing and provide instructions manually either by mobile phone or any other means. We attempt to automate this manual process.

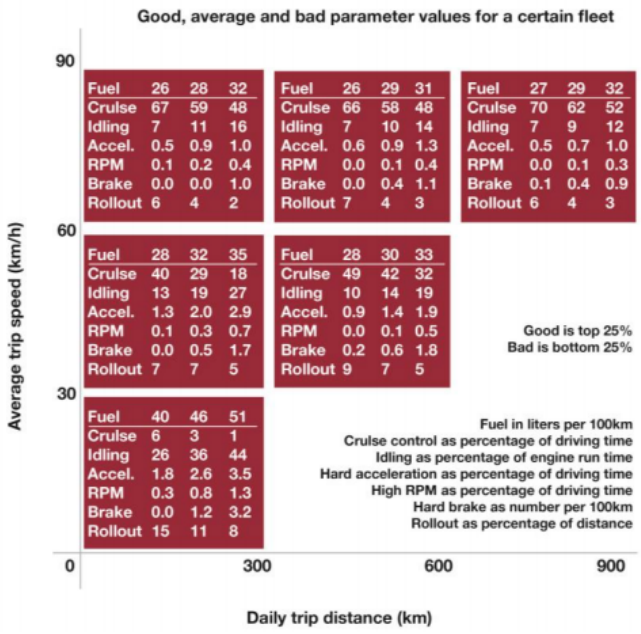


Figure 2.3: A cluster norm table for a particular fleet used for evaluating the trip performance.

CGI has tried to analyze the dataset generated through this portal to find out the relationship between driving behavior and fuel consumption. In that analysis, they have clustered data in a two-dimensional space of distance and average speed using some automatic and manual clustering techniques. Figure 2.3 shows the resultant cluster norm table. In each cluster, the first column indicates the best fuel consumption, next average and the last is the worst. An interesting fact to be noticed is that depending on the distance traveled and average speed, the figure of efficient fuel consumption changes. For instance, while 27 L/100 km is the efficient consumption for long-distance journeys, that figure for short-distance trips is 40L per 100km. The conclusion is, classifying driving behaviors just looking at fuel usage figures is not appropriate. Further, they have shown that depending on where the vehicle is traveling 10% - 30% of the variation in fuel consumption can be attributed to the driver behavior, e.g., driving behavior impact is higher for long-distance travels.

Although authors of this study considered weather information, only the average temperature and wind speed have been considered. However, other weather conditions like rain, fog, hail, and snow are also known to have a significant effect. In our study, we use the overall weather condition as the parameter representing the impact of weather.

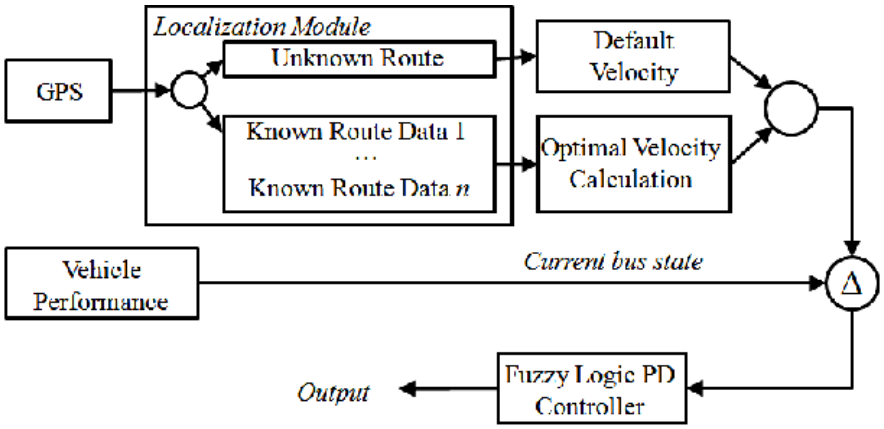


Figure 2.4: Architecture of the intelligent driver system introduced in [8].

Linda and Manic [8] proposed an Intelligent Driver System to improve fuel economy by learning historically fuel-efficient driving behaviors. The proposed system gradually builds a model of historically most fuel-efficient driving behavior for a fixed set of routes. This model studies both the vehicle performance and GPS data while modeling the most fuel-efficient driving behavior for a specific route. While driving, the velocity of the vehicle is compared with the calculated optimal speed for that particular location. In an inefficiency is identified, a fuzzy Proportional-Derivative (PD) controller is used to determining the best control action to take the vehicle to the optimal velocity (see Figure 2.4). This decision is either conveyed to the driver via a specialized Human Machine Interface(HMI) or used directly as predictive cruise control. If it is an unknown route, the velocity is compared with road default speed. However, being bound to known routes is a limitation of this solution. Not only that, it does not consider the impact due to weather or traffic conditions, which tends to change with different times of the day, week, and year even for a given route. Therefore, sometimes the control action determined by this system would not be the optimal action. Further, in some situations it would be unsafe to adhere to those instructions, e.g., suggesting to speeding up while the vehicle is under heavy rain.

Gilman et al. [9] have developed a reference architecture for context-aware driver assistance systems to provide personalized assistance for fuel-efficient driving. They have also developed a prototype as a proof of concept and that prototype collects, integrates, and analyses vehicle parameter, as well as diverse contextual data such as weather, and traffic data. They emphasize the importance of context-awareness of driver assistance systems to provide relevant feedback, especially for real-time assistance. They have summarized contextual information as driver context, environmental context and vehicle context. The summary is given in Figure 2.5. After analyzing these

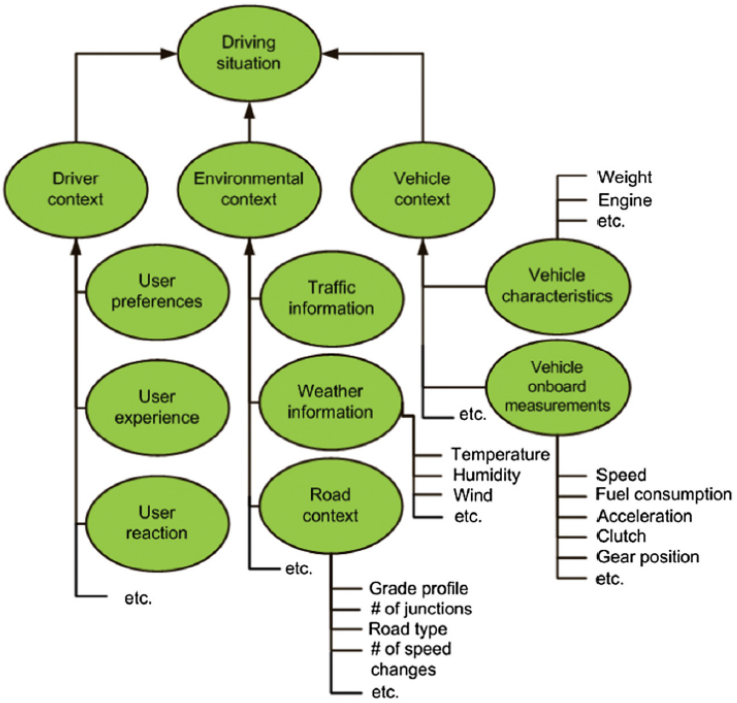


Figure 2.5: Context characterizing the driving situation illustrated by [9].

data, the proposed system provides off-line personalized feedback to improve future driving behaviors. While the analysis of multiple factors gives an enhanced model, it would be more effective if the feedback can be given in real time.

As the related work either select a limited set of relevant attributes that impact the fuel consumption or feedback is given in offline, there is still a need to build a model that is both comprehensive and can be used in real time.

2.4 Summary

The review of existing work in instantaneous fuel consumption prediction revealed that most of them had developed models for synthetic datasets and datasets collected under some controlled conditions, e.g., datasets collected from vehicles running on highways or from hot stabilized vehicles. Further, most of them have overlooked the impact of external factors even though external factors play a vital role in determining the fuel economy of a vehicle. However, such models might not be applicable in real-world situations. Therefore, still, there is a requirement for an instantaneous fuel consumption predictive model developed using a real-world dataset which covers typical influential factors.

Furthermore, since numerous factors influence fuel consumption and the relation-

ship between most of such factors and the fuel usage is non-linear, simple regression models such as logistic regression or regression trees are not suitable to model fuel consumption of a vehicle. Hence, exploring the existing work for different machine learning model, we identify model ensembles and artificial neural networks as the state-of-art models to handle such scenarios.

Survey on driver behavior monitoring and feedback systems raised the necessity for a framework which monitors driver behavior in real-time and provides continuous individual feedback in real-time. Even though existing work in this area have made some progress, they either select a limited set of relevant attributes that impact the fuel consumption, feedback is given in offline or feedback is given manually. Thus, there is still a need to build a framework that provides comprehensive and useful feedback and can be used in real time.

3 DATASET

In this chapter, we describe our dataset and present exploration analysis of it. In Section 3.1 we provide details on what are the parameters available in our dataset and how we collected data. In Section 3.2 we elaborate on how internal and external factors influence fuel consumption of the vehicle through the descriptive analysis.

3.1 Background

The dataset used for this research corresponds to a long distance, public bus in Sri Lanka. The bus starts from Depot around 4:00 pm and goes to Colombo (i.e., the commercial capital of Sri Lanka). Then bus leaves Colombo at 7:00 pm and travels along A2, A4, and AB10 roads and reaches the destination around 7:00 am on the following day. Figure 3.1 shows the route of the bus. Altogether, the bus travels around 365 km in one direction. The return journey is along the same route and typically between 4:00 pm to 7:00 am on the following morning. About one-third of the trip is through a mountainous region. This route captures almost all the external conditions a vehicle could encounter in real-world driving. For example, the bus goes through urban, rural, and mountainous areas, as well as driving times include peak, off-peak, and night driving. Therefore, we believe that this is a rich dataset to analyze influencing factors of fuel consumption.

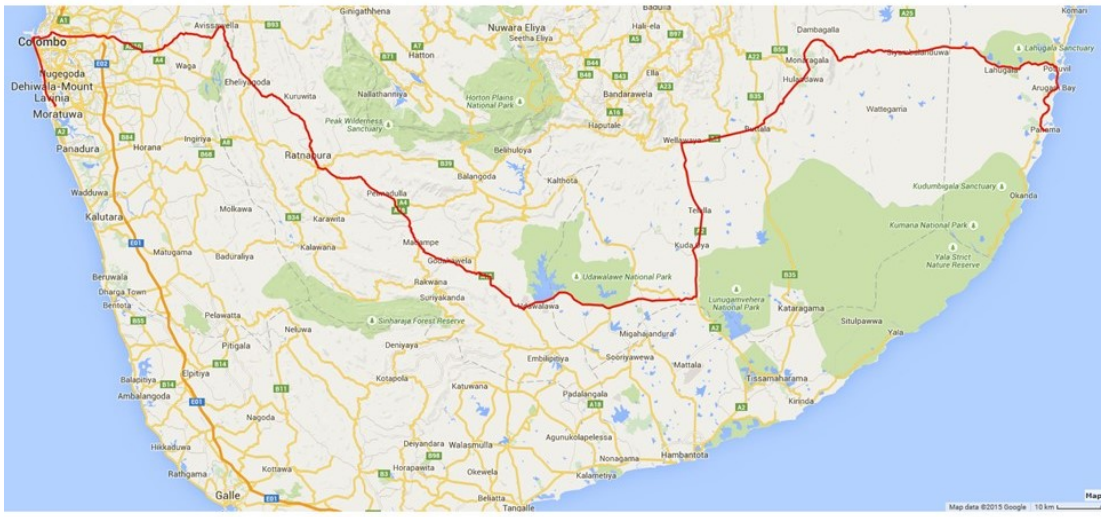


Figure 3.1: Route of the bus - from Katubedda to Panama.

The bus is fitted with a GPS-based tracking device and a capacitive, high-precision fuel sensor. Data collected via these devices is pushed to a cloud server in near real-

time over a 3G connection. The dataset consists of outward and inward journeys between May 13 and August 31, 2015. The dataset contains the following parameters:

- Timestamp (date and time)
- Longitude (Min: 5.918611^0N, Max: 9.835556^0N)
- Latitude (Min: 79.516667^0E, Max: 81.879167^0E)
- Bearing (00 to 3600)
- Elevation (Min: 0m, Max: 2,524m)
- Distance traveled (km) – between two readings
- Speed (kmh^{-1})
- Acceleration (kmh^{-2})
- Ignition status (1 – Ignition On or 0 – Ignition Off)
- Current battery voltage (Min: 0v, Max: 29v)
- Fuel level (Min: 0L, Max: 218L)
- Fuel consumption (L)

From the above parameters, timestamp, longitude, latitude, bearing and speed have been directly drawn from the GPS device mounted on the bus. Distance has been derived as the distance between nearest GPS locations while acceleration is derived from the speed. Elevation values have been obtained from Google API. The fuel sensor mounted on the bus sends voltage values which are respective to fuel level in the tank. Those voltage levels are converted to fuel levels using the following equation.

$$fuel\_level = \frac{|Voltage - Min\_voltage| \times tank\_size}{|Max\_voltage - Min\_voltage|} \tag{3.1}$$

The fuel consumption is calculated as the difference between the current fuel level and the previous fuel level. While this dataset set consists of many primary influencing factors of fuel usage of a vehicle, some crucial factors such as RPM value, the number of passengers for the journey are missing.

To capture the influence of external factors on fuel consumption, some parameters were derived from existing parameters. Road traffic is indicated by the time of day. Time of day was derived by rounding off the GPS timestamp to the nearest hour (e.g., 12:00, 13:00, 14:00 and so on). Change of elevation is used to capture the changes in road topography. Excessive idling is another parameter which causes an inefficient fuel economy. We considered the vehicle to be excessively idling if the speed is zero while the ignition status is on for more than one minute. We select one minute time interval to eliminate stopping for traffic lights or at bus stops. Another important but indirect environmental factor that affects the fuel economy is weather condition. When the weather is bad, drivers are forced to slow-down leading the lower fuel economy. The weather condition was obtained using a REST API provided by the World Weather Online (WWO) developer portal [23]. WWO API provides a detailed weather report for a given location, date, and time. However, for our analysis, we only extracted the weather descriptor (i.e., weather condition) for a given location, date, and time based on the bus's route and schedule. Weather descriptors provided by WWO include sunny, clear, partly cloudy, cloudy, overcast, patchy rain nearby, light drizzle, light rain shower, moderate rain, moderate or heavy rain, mist, and fog.

3.2 Descriptive Analysis

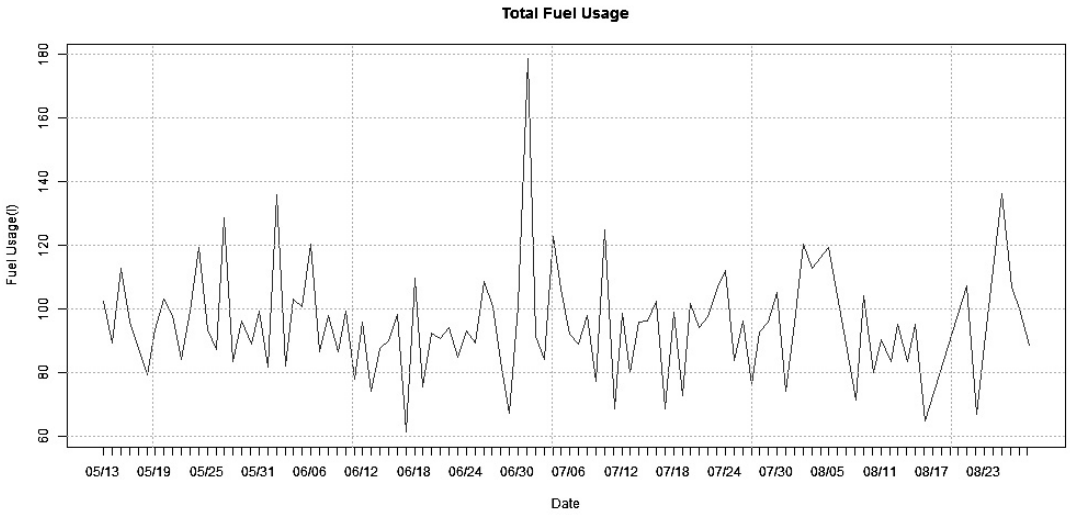


Figure 3.2: Total fuel consumption of each journey.

The first and most important step in developing a better data mining model is to understand the data through an exploratory data analysis. Looking at the total fuel consumption of the bus, we could see that even though the bus is going on the same

route for the almost same distance, there is a significant variation in fuel usage from one day to another, see Figure 3.2. Some of the exceptional cases were caused by changed routes, GPS device failures and bus breakdowns. Those were considered as outliers and removed from the analysis. However, even when outliers were eliminated, we could see a significant variation in fuel usage which was not intuitive.

Figure 3.3 shows the box plot of fuel consumption for outward and inward (i.e., return) trips. There is a significant difference in fuel consumptions for the outward vs. inward journey, average fuel usage for the inward journey is around 85L whereas for the outward journey it is 95L. Based on the analysis the following reasons were identified as the potential causes for this substantial difference between the two drives.

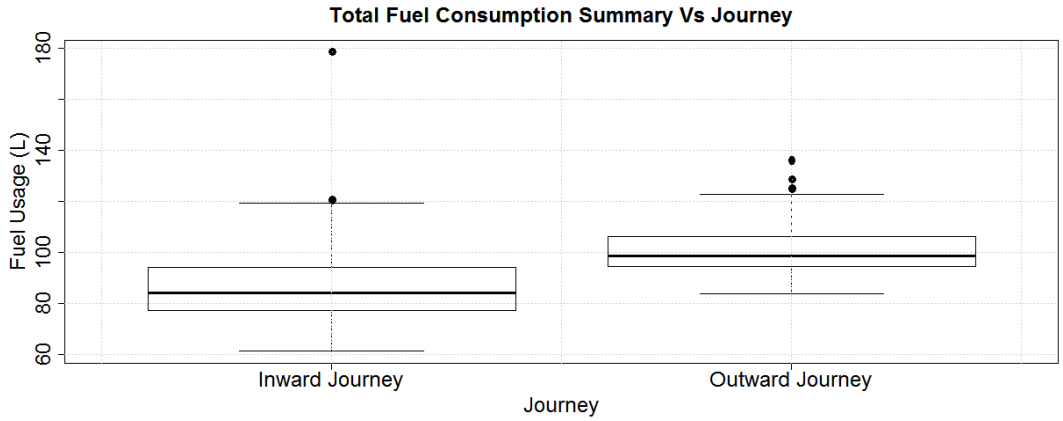


Figure 3.3: Fuel consumption for inward and outward journeys.

Different traffic conditions: In the outward journey, the bus starts the journey from Colombo in the evening around 4 p.m. which is a peak traffic hour. This bus begins its journey from Katubedda, a suburb of Colombo and goes to the central bus stand located in the center of Colombo. Close to Colombo, the bus might experience extremely high evening traffic. Then the bus starts its journey from the Colombo bus stand to Panama around 7.00 p.m. again through a road suffers from heavy traffic. In contrast, the bus may experience comparatively less traffic at Panama in the evening since it is a rural area. Alternatively, in the morning bus may experience more traffic while arriving in Colombo compared to Panama. However, since the bus reaches Colombo even before 7.00 a.m. in its return journey, the roads should be clearer than the evening. Traffic impedes buses getting their desired speed which contributes to most of the fuel consumption. Figure 3.4 and Figure 3.5 show the impact of traffic on fuel usage of each journey. Each graph illustrates the variation of mean speed and fuel usage for

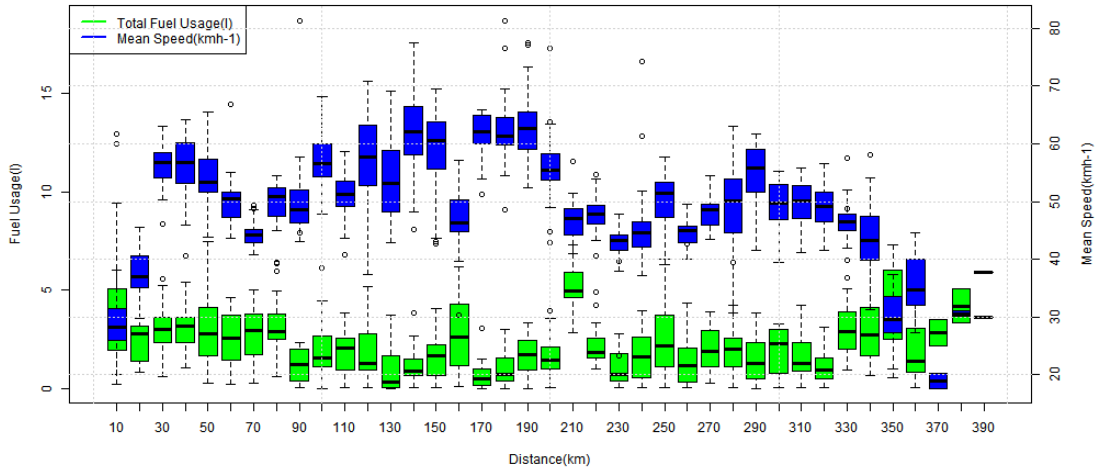


Figure 3.4: Comparing total fuel consumption and mean speed at each 10km for all trips from Panama to Colombo.

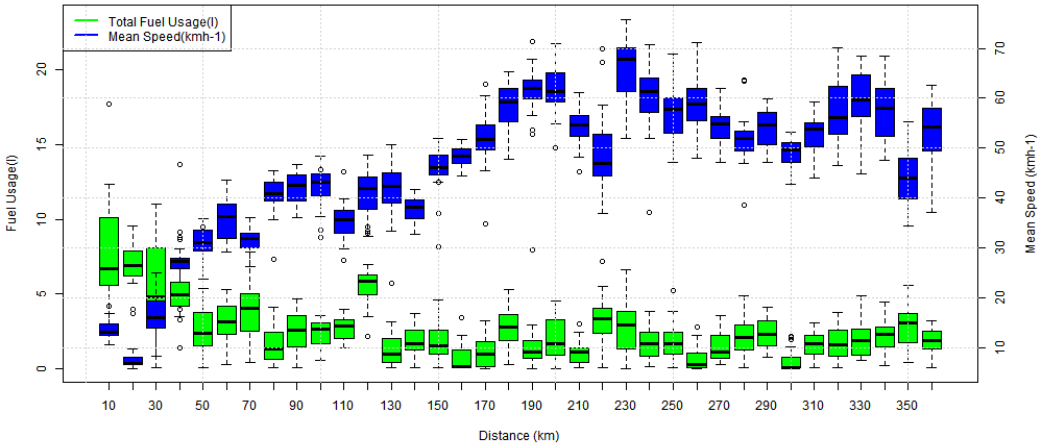


Figure 3.5: Comparing total fuel consumption and mean speed at each 10km for all trips from Colombo to Panama.

every 10km. Graphs indicate that fuel consumption in the urban area is higher than the rural area in the evening as well as in the morning. These observations are attributed to different traffic conditions in these two areas and time intervals.

The difference in experienced elevation in two journeys: According to the elevation graphs in Figure 3.6, the bus experiences steeper roads for the journey from Colombo to Panama when compared to the drive from Panama to Colombo. Steeper roads cause higher fuel consumption due to high acceleration and engine RPM.

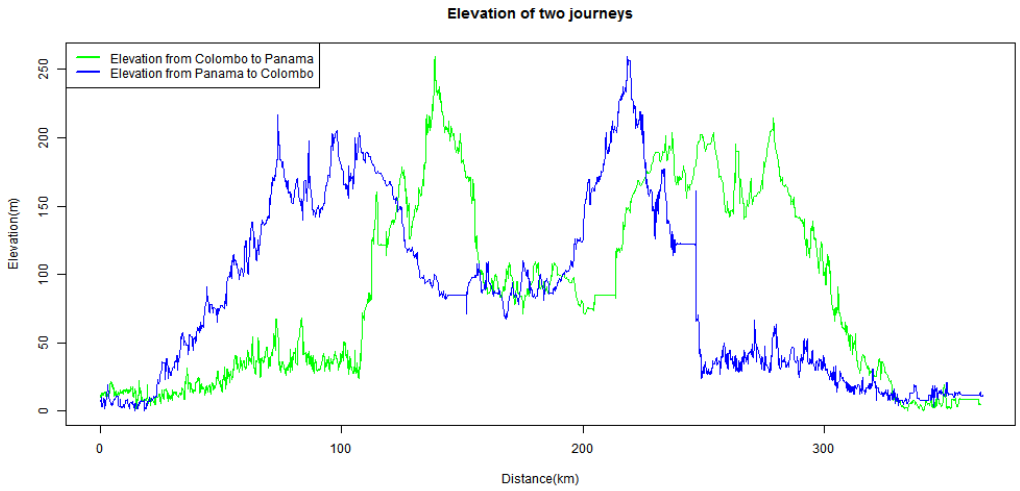


Figure 3.6: Change of elevation along the route in two directions.

Due to the significant difference between the two directions, the dataset was divided into two parts as inward and outward and the predictive models were developed separately.

Next, we investigate the relationship of vehicular dependent and external factors with fuel consumption of the bus. Among the available predictors, distance directly influences fuel consumption of the bus (see Figure 3.7).

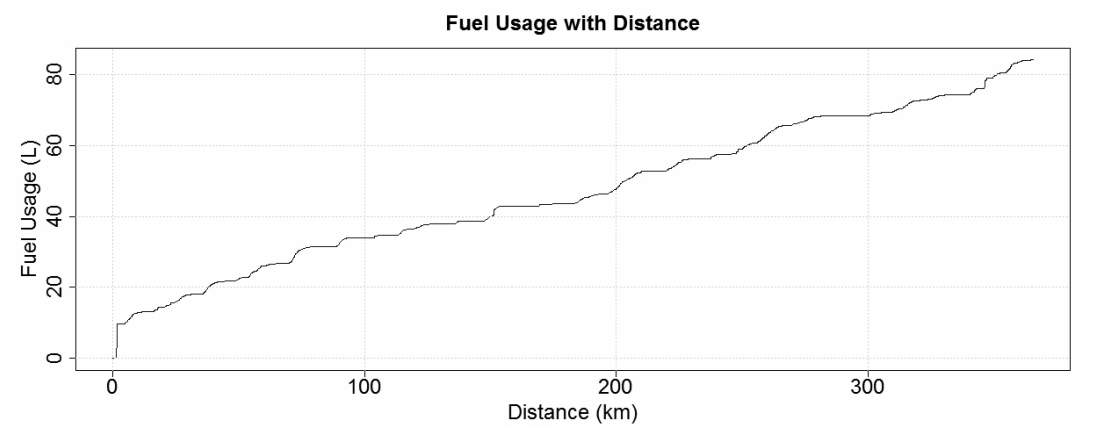


Figure 3.7: Impact of distance on fuel consumption.

Another crucial factor that affects the fuel economy is the speed of the vehicle [4, 24, 25]. Mean fuel consumption at each speed is depicted in Figure 3.8. According to Figure 3.8, the bus consumes the least amount of fuel when it is traveling around 50 kmh^{-1}. Therefore, $50kmh^{-1}$ can be considered as the most effective speed for this bus. When the speed is lesser or hight than the most effective speed, the bus consumes more

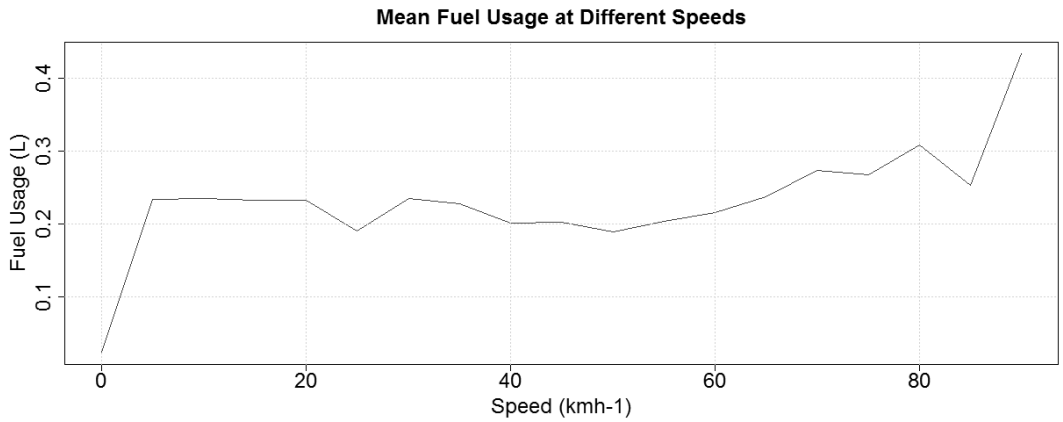


Figure 3.8: Mean fuel consumption at different speeds calculated across one minute samples.

fuel. Therefore, as illustrated in Figure 3.8, the relationship between fuel consumption and speed is not linear. Such findings are important in modeling the fuel consumption of a vehicle.

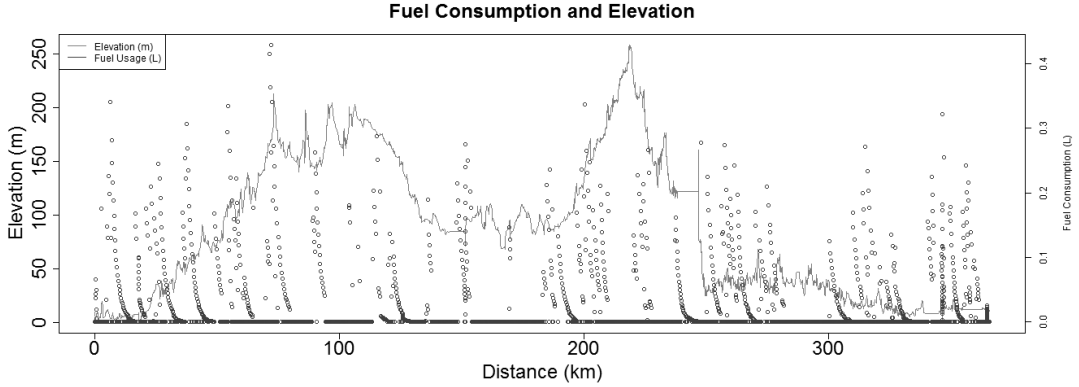


Figure 3.9: Fuel consumption variation with elevation.

Due to the traffic and elevation changes, location posses a substantial impact on fuel consumption. As mentioned earlier, a vehicle climbing up on a steeper road consumes more fuel due to high acceleration and engine RPM whereas when the vehicle is descending it's fuel consumption is relatively low. Figure 3.9 shows fuel usage and elevation throughout the journey from Panama to Colombo. Note high fuel usage at the beginning of the journey. Even though vehicle doesn't experience heavy traffic in this area, it is ascending to an area with higher altitude. Between 100km and 200km one can observe low fuel consumption due to the descent of the vehicle in this area. At the end of journey, despite vehicle is traveling in a relatively flat area, the fuel usage is high. This can be attributed to relatively heavy traffic experienced in Colombo area, the capital of the country. Further, according to Figure 3.9 the relationship between

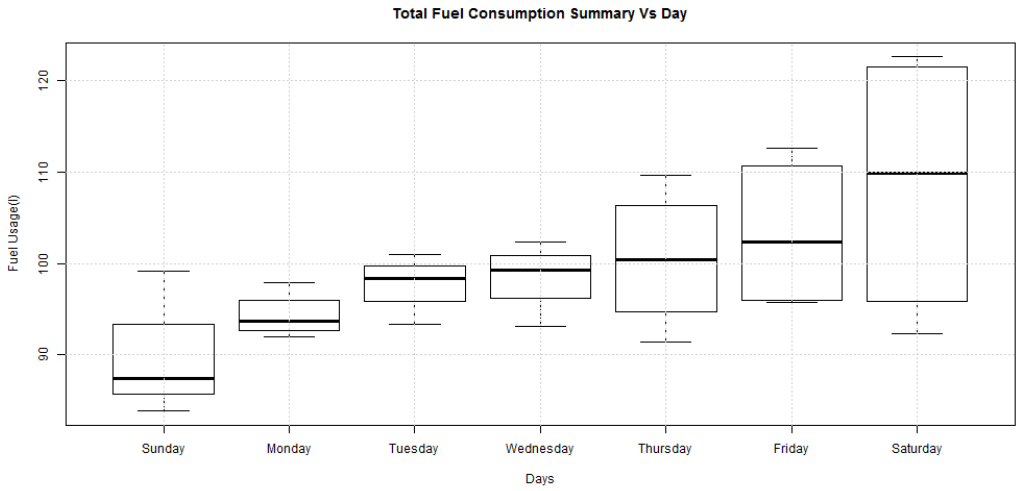


Figure 3.10: Variation of fuel consumption w.r.t. day of the week for journeys from Colombo to Panama.

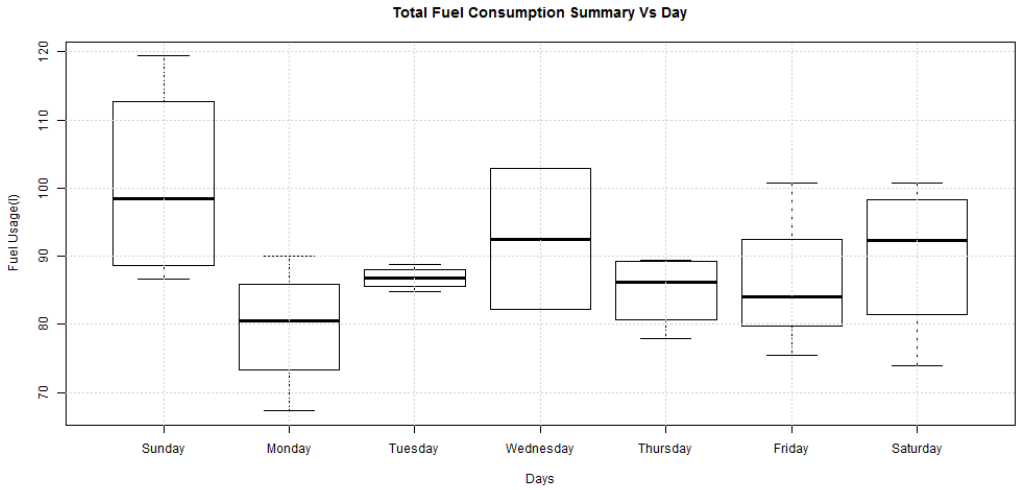


Figure 3.11: Variation of fuel consumption w.r.t. day of the week for journeys from Panama to Colombo.

fuel consumption and the elevation is also not linear.

Figure 3.10 and 3.11 show the impact of day of the week on the fuel consumption. It can be observed that depending on the day of traveling whether it is Sunday, Monday, etc., fuel consumption varies. Because this bus is going between Colombo, the capital of the country and Panama, a rural area, we expected to see increasing mean fuel consumption from Sunday to Friday in the outward journey because the most plausible scenario is people travel on Sunday to capital for work and leave the capital Friday for the weekend. We thought fuel consumption for Saturday would be higher than Sunday but less than Friday. Whereas for inward journey mean fuel consumption was expected

to decrease from Sunday to Friday. Colombo – Panama graph matches our expectation even though fuel consumption for Saturday is unexpectedly higher than that of Friday. A possible reason would be that employees from Panama area work Saturday as well and leave on Saturday. However, Panama-Colombo graph was much different than the expected graph. As expected Sunday has the highest figure, but fuel consumption for Wednesday also stands out. Sri Lanka calendar for this period provides the rationale behind this surprising observation. There were three holidays falling into Monday – Wednesday within these four months. Thus, the regular passengers of the bus might have enjoyed a long weekend and reported to work on Wednesday. The takeaway from this analysis is we would not be able to derive a linear relationship between the day of the week and fuel consumption of the bus.

The correlation matrix of variables in the data set is given in Figure 3.12. However, the correlation matrix was not useful to derive any important relationships other than obvious ones.

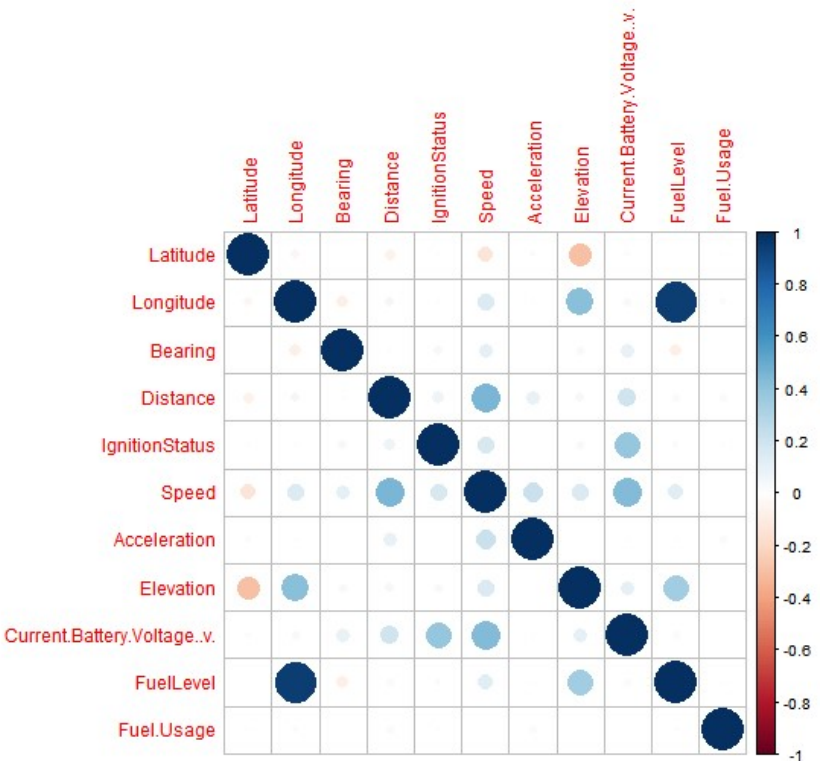


Figure 3.12: Correlation Matrix for parameters of the data set from Panama to Colombo.

This exploratory data analysis provides useful insights to select a suitable predictive model for this dataset. As some predictors have a non-linear relationship with the fuel consumption, a linear forecasting model such as linear regression model would not be appropriate.

4 FUEL CONSUMPTION PREDICTION

To identify the most suitable technique to predict fuel consumption, we carried out a comparative study evaluating a set of alternative models to predict the fuel consumption of the bus being considered. We first discuss data preprocessing and feature selection in Section 4.1. We assessed the appropriateness of random forest, gradient boosting, and artificial neural network algorithms for predicting fuel consumption in Section 4.2.1, Section 4.2.2 and Section 4.2.3, respectively.

4.1 Data preprocessing and feature engineering

For the purpose of prediction, the target variable is fuel consumption of the bus within a given time interval and the predictor variables are distance, speed, longitude, latitude, elevation, and day of the week. More formally, our goal is to approximate the unknown multiple regression function,

$$FuelConsumption = f(distance, speed, latitude, longitude, elevation, acceleration, day) \quad (4.1)$$

As mentioned in Section 3.2, the descriptive data analysis revealed that external factors for the outward journey (the journey from Colombo to Panama) are significantly different from those for the inward journey (the journey from Panama to Colombo). Hence, we divided the dataset into two and developed separate models for two kinds of journey. Outliers in the dataset due to known factors such as device breakdown and change route were excluded from model building. Data points correspond to situations where the bus engine was switch-off were also removed from the dataset.

Predictor variables to forecast the fuel consumption were selected based on the exploratory analysis described in Section 3.2 and the context knowledge. For instance, time and longitude have a high correlation in this dataset, thus time was removed. Day of the week seems to have more impact on fuel consumption than the date. Therefore, we derived day from the date. Further, Random Forest provides importance score for each variable which is useful in developing better predictive models. In the first iteration of model creation, we used all the variables to build a Random Forest model and considered variable importance indicated by it in selecting more influential parameters. Finally, we selected distance, speed, longitude, latitude, elevation, and day of the week as parameters of the instantaneous fuel consumption prediction model. The same set

of parameters were used to build all three different models. Furthermore, essential data preprocessing techniques such as handling missing values and scaling were applied to develop a more accurate model.

4.2 Implementation details

4.2.1 Random Forest

To evaluate the RF algorithm random forest package in R [15] was used. This package is based on Breiman's random forest algorithm for classification and regression. This predictive model can be fine-tuned using two parameters *ntree* – the number of trees within the ensembles and *mtry* – the number of variables randomly sampled for a split. The default value of *mtry* for regression is p/3, where p is the number of predictors. To find the optimal value of *mtry*, a parameter sweep was conducted. Based on the Out-Of-Bag (OOB) error estimate, *mtry* = 2 was selected as the best value. For *ntree* parameter, we considered values of 250, 500, and 750. In the first construction of the RF model, all the variables were fed into the model. Then variable importance was plotted as shown in Figure 4.1. In this image, the graph in the left shows the percentage of increment in MSE if a given variable is assigned values by random permutation. Higher the %IncMSE higher the impact of that variable on the dependent variable. The graph in the right shows the percentage of increment in node purity with respect to each variable. Node purity is measured by Gini Index which is the difference between Residual Sum of Squares before and after the split on a given variable. Again higher the %IncNodePurity higher the impact of that variable on the dependent variable. This variable importance graph was used to identify important variables or most relevant predictors. Besides, based on the context knowledge several parameters like fuel level and current battery voltage were removed from further analysis. The final model was developed considering these fine-tunes such that the accuracy of prediction enhanced.

4.2.2 Gradient Boosting

To evaluate the GB algorithm, we used *mboost* package in R, which implements methods to fit generalized linear models (GLMs), generalized additive models (GAMs), and generalizations using component-wise gradient boosting techniques. The mboost package can thus be used for regression, classification, time-to-event analysis, and a variety of other statistical modeling problems based on high- dimensional data. In this study, we used GAM to predict fuel consumption as the relationship between given predictors and fuel usage is non-linear. GAM can be accessed from *gamboost* function

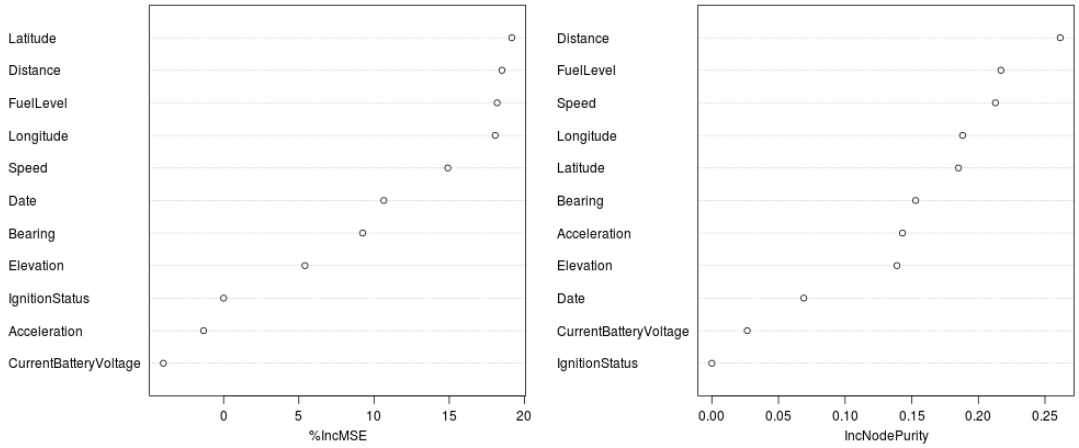


Figure 4.1: Variable importance given by Random Forest algorithm.

in mboost package. gamboost is flexible and provides efficient and helpful tools for fitting generalized additive boosting models. Different base leaners such as bols and bbs can be used to specify linear or non-linear relationship between independent and dependent variables. Linear or categorical effects can be specified by bols. Smooth effects can be defined in GB by the bbs base learner [26]. Therefore, based on the exploratory analysis in section II, the model given for gamboost function was derived as follows:

$$Fuelconsumption = bols(Long) + bbs(Lat) + bbs(Speed) + bols(Acc) + bols(Elev.Change) + bols(Dis) + bols(Day) \quad (4.2)$$

4.2.3 Neural Network

To evaluate how well an ANN can realize the relationship between predictors and response of this dataset, neuralnet R package was used. It has been built to train multi-layer perceptrons for regression analyses [21]. Theoretically, nuralnet can handle an arbitrary number of predictors and responses, as well as hidden layers and hidden neurons [21]. In this analysis, we used a multi layer perceptron with two hidden layers; three neurons at the first hidden layers and two neurons at the second based on empirical evidence; other structures had higher error values than this structure. Having few predictive parameters might have caused deeper and wider Neural Networks to over-fit and performed poorly in this problem.

4.3 Results and discussion

4.3.1 Prediction using Random Forest

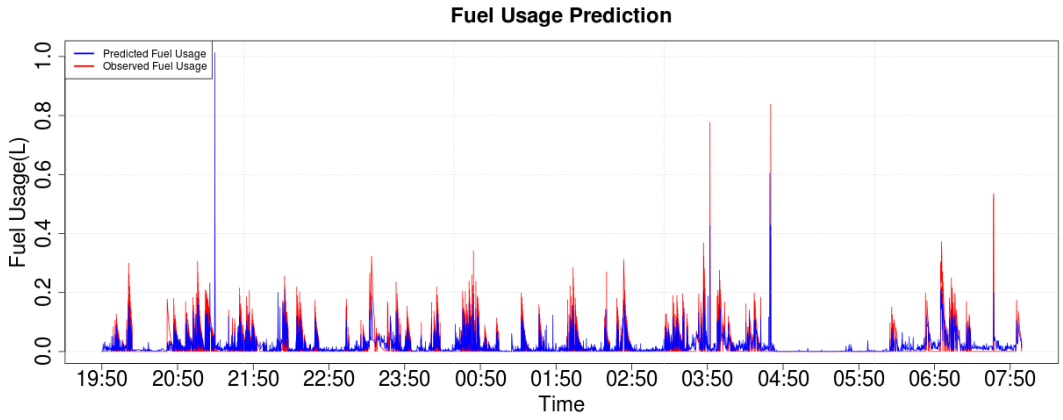


Figure 4.2: Predicted and observed instantaneous fuel consumption using Random Forest.

Figure 4.2 shows the predicted and observed instantaneous fuel consumption. By analyzing that graph, we could see that the algorithm is over predicting in some places especially when the observed value is zero or almost. When those instances were examined, we could see that the those are the cases when the bus is running at its optimal or near optimal speed. When the bus is running at its optimal speed, their fuel consumption is almost zero, but the Random Forest model is not capturing it well. Once we carried out post-processing step to replace this over predicted values with global average fuel consumption under the same situation, we were able to further reduce the prediction error.

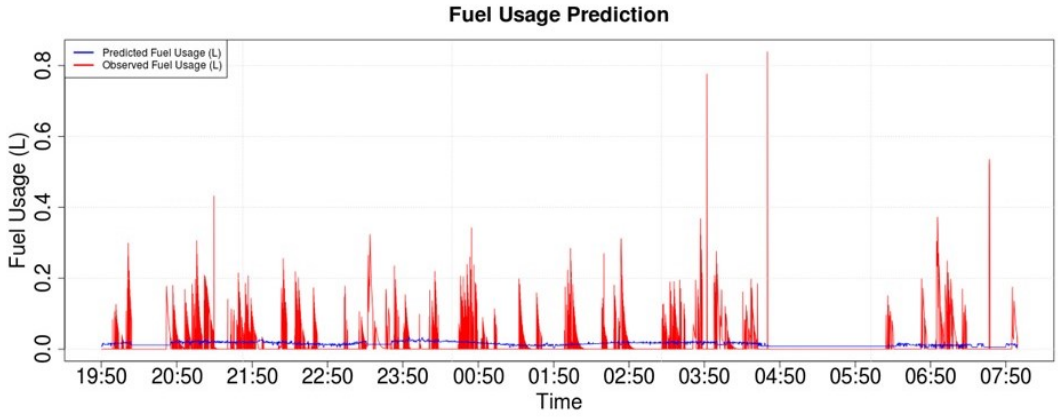


Figure 4.3: Predicted and observed instantaneous fuel consumption using Gradient Boosting.

4.3.2 Prediction using Gradient Boosting

Figure 4.3 shows the prediction results of Gradient Boosting technique. The graph shows the instantaneous observed fuel consumption vs. predicted fuel consumption. It is clear that Gradient Boosting based model has not been able to capture the relationship between fuel consumption and other influential factors. Instead, it just gives the average fuel consumption as the predicted value.

4.3.3 Prediction using Neural Network

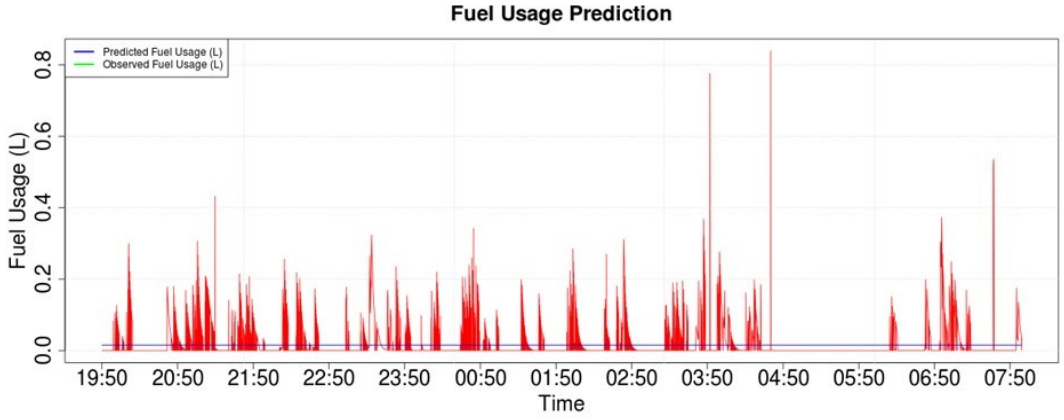


Figure 4.4: Predicted and observed instantaneous fuel consumption using Gradient Boosting.

Figure 4.4 shows the prediction results of Neural Network technique. As per the graph, Neural Network based model also has not been able to capture the relationship between fuel consumption and other influential factors.

4.3.4 Evaluation of prediction accuracy

We analyzed efficiency and error statistics of each predictive model to assess their accuracy. Efficiency measure can be carried out using different methods. We used Nash-Sutcliffe efficiency (NSE) coefficient to measure the predictive power of each model, which is defined as follows [27]:

$$Nash - Sutcliffeefficiency = 1 - \frac{\sum_{i=1}^{n}(EST_i - OBS_i)^2}{\sum_{i=1}^{n}(OBS_i - \bar{OBS})^2} \tag{4.3}$$

where EST_i and OBS_i denote the i-th estimated and observed fuel consumption values and n is the total number of samples. Three well-known error statistics were

also calculated to assess the accuracy of each predictive model. Those three statistics are Bias, Mean Absolute Error (MAE), and Root Mean-Squared Error (RMSE). Each is defined as follows:

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(EST_i - OBS_i) \tag{4.4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|EST_i - OBS_i| \tag{4.5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(EST_i - OBS_i)^2} \tag{4.6}$$

Table 4.1: Nash- Sutcliffe Efficiency.

Model	**Nash-Sutcliffe Efficiency**
Random Forest	0.26189
Gradient Boosting	-0.00240
Neural Network	-0.01304

Table 4.2: Error statistics of three techniques.

Error Statistic	**Random Forest**	**Gradient Boosting**	**Neural Network**
Bias	0.004768	0.0004498	0.002744
MAE	0.022955	0.0258532	0.027562
RMSE	0.040459	0.0471540	0.047404

NSE evaluates the predictive power of a model. While the efficiency of one being the perfect value, zero means the model is no better than just using the mean value of the observed data [27]. As shown in Table 4.1, RF model has the largest NSE coefficient. Further, it is the only predictive model with positive measure. GB and ANN have smaller negative coefficients closer to zero indicating that those models merely predict the mean value of the observed data. Standard error statistics in Table 4.2 also provide similar insights. The fuel consumption prediction graphs in Figure 4.2, 4.3, and 4.4 further verify insights of numerical analysis. While the RF captures the trend more accurately, GB and ANN are only predicting the fuel consumption in a conservative manner. Hence, in conclusion, the RF model has captured the relationship between predictor variables and fuel consumption better than other two models.

4.4 Identify prospective fuel frauds

Predicted fuel consumption values from a model like above can be used to identify prospective fraudulent activities by drivers of fleet vehicles. Suppose for a given day, F denotes instantaneous fuel consumptions recorded by the tracking device, where $F = \{f_1, ..., f_n\}$. Also, suppose F^* denotes instantaneous fuel consumption predicted by the model described above, $F^* = \{f_1^*, ..., f_n^*\}$. We define excess fuel usage ratio as follow:

$$\gamma = \frac{1}{n} \sum_{i=1}^{n} \frac{|f_i - f_i^*|}{f_i^* + \epsilon} \tag{4.7}$$

where ϵ is a small number used for numerical stability.

We also define a constant η based on the predictive model accuracy and domain expert consultation. We use the test dataset used for model evaluation to define eta in a similar manner to γ as indicated in the following equation.

$$\eta = \frac{1}{N} \sum_{i=1}^{N} \frac{|t_i - t_i^*|}{t_i^* + \epsilon} + 0.5 \tag{4.8}$$

, where N is the total number of instances in the test dataset, t_i and t_i^* indicate recorded and predicted instantaneous fuel consumption values respectively and 0.5 is a buffer added to compensate fuel sensor error as per domain expertise recommendation.

We suspect a fuel fraud may have occurred in a given day if $\gamma > \eta$.

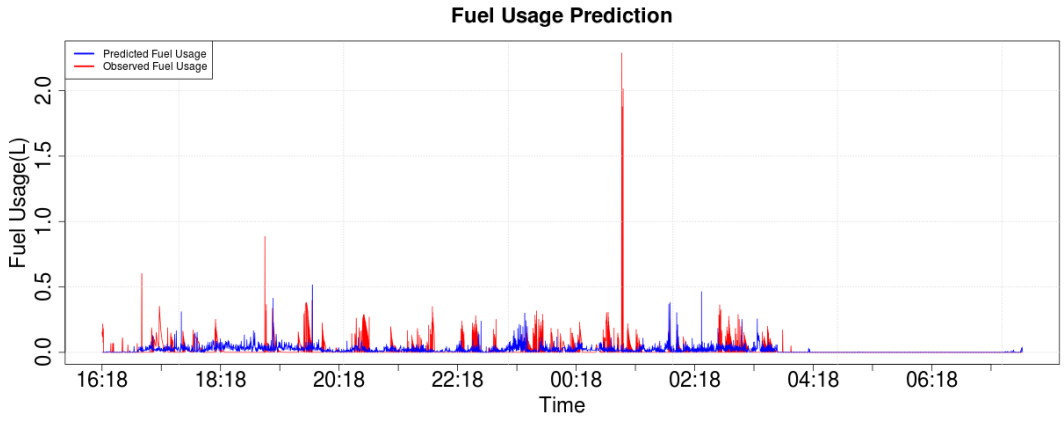


Figure 4.5: Predicted and observed instantaneous fuel consumption of 29/08.

For example, consider recorded and predicted fuel consumptions for 29/08/2015

shown in Figure 4.5. We can see that recorded fuel usage much deviates from the predicted values and there is a spike around 1:10 a.m. To quantitatively analyze this deviation, we calculate γ and η as in Equation 4.7 and 4.8, respectively. The value of γ for this journey is 1.8764. The value of η is 1.3365. Because $\gamma > \eta$, to understand what were the causes for this unusual fuel consumption of the vehicle, we have to enquire the bus driver and carry out further analysis as discussed in the next section.

4.5 Verify identified prospective fuel frauds

As we discussed in Section 3.2 fuel consumption of a vehicle depends on various factors. Some of these factors are controllable by the driver such as speed, accelerations and idling, but some of them are out of control of the driver such as road traffic, elevation changes and adverse weather conditions. Since the proposed predictive model does not consider all the influential factors, just because the predicted fuel consumption is lower than the actual fuel consumption, one cannot conclude that a fuel fraud has taken place. Instead, in practice, the driver of the particular journey is inquired to explain excess fuel usage. While the driver can justify abnormal fuel consumption for various reasons such as road traffic, extra load, severe weather condition, etc. fleet managers should be able to verify these claims. For that purpose, we propose the following key performance indicators to investigate suspicious fuel usages.

- Idling time
- Frequency of harsh events such as heavy acceleration and harsh breaking
- Speed profile
- Date and day of the week

Idling time Having the engine of a vehicle switch on without moving is known as *idling*. Idling has a dramatic effect on MPG figures and thus is a popular metric. However, considering all such situations for excess fuel investigations or driver performance analysis is not practical, because there are inevitable scenarios where the driver does not have other option but keeps the vehicle idling, e.g., waiting for traffic signals, waiting until passengers are getting on and off. Hence, here we redefine idling as having the engine switch on without moving for *one minute*. Drivers are accountable for excessive idling time throughout the journey.

Frequency of harsh events Harsh events such as harsh accelerations or harsh breaking result in high fuel spend. According to industry standards, accelerations higher than $3ms^{-2}$ are defined as harsh accelerations and decelerations greater than $3ms^{-2}$ are defined as harsh breaking. We propose to use the frequency of such events throughout the journey as a key performance indicator.

Speed profile Speed profile of a trip provides a lot of useful information about the trip. As key indicators, we propose to use the mean speed of the journey and skewness of the speed histogram. We consider an average speed between $40kmh^{-1}$ and $60kmh^{-1}$ (which we called optimum range) as a reasonable average speed for a bus (The values would be different for different vehicles. We choose this range since we are analyzing a dataset of a bus in this research). Mean speed less than $30kmh^{-1}$ might be an indicator of heavy road traffic. Skewness of the speed histogram indicates speed variations throughout a drive. If the skewness is positive or minor minus value that means the bus was driven in a lower speed for a significant portion of the journey. If the skewness is a large negative value bus has been driven in higher speed ranges. If the mean speed falls into optimum range and speed profile has larger negative value the driving behavior is good.

Date and day of the week Date and day of the week affect fuel consumption of a vehicle in different ways. Depending on the day of the week, e.g., whether it is a Friday or Saturday, road traffic varies significantly and cause varied fuel spends for vehicles. In general, we can expect the same traffic patterns for the same days in the work. However, this is not guaranteed, because depending on the date road traffic intensity might be changed. For instance, if a particular Friday is a holiday, the roads might be clearer in that specific Friday compared to other Fridays. Further, for fleets like public buses, generally, the load will be different for different days in the week. As an example, if the bus is traveling from a city area to a rural area, in a Friday the load will be higher than other days of the week. Then again if a particular Friday is a holiday, the same bus would have a higher load on Thursday of that week instead of Friday. Thus, it is important to check what is the date of the suspected fuel consumption and what day of the week it is.

4.6 Conclusions

We developed a fuel consumption prediction model which is useful in fuel frauds detection where the actual consumption of the vehicle can be compared against the pre-

dicted value. We evaluated the predictive ability of three predictive models, (RF, GB and ANN) in predicting the fuel consumption of a long-distance public bus. Among those three models, RF model could predict the fuel consumption more accurately while capturing the trends in data. The most important factors of the fuel usage prediction model were distance, location, elevation, speed, and day of the week. Finally, we introduced a set of key performance indicators to verify detected suspicious fuel usages.

One of the limitations of this work is ignoring the impact of some external conditions such as traffic, weather, and the load of the bus. Integrating such additional influencing factors would enhance the predictive ability even further and predicted value would be more reliable.

5 REAL-TIME MONITORING AND DRIVER FEEDBACK TO PROMOTE FUEL-EFFICIENT DRIVING

In this chapter, we propose a framework to analyze driver behavior in real-time, identify fuel-inefficient driving patterns and provide continuous feedback to the driver so that they can continuously maintain fuel-efficient driving behavior. We propose to use historical data to derive heuristics to classify driver behavior for fuel efficiency. This framework is especially useful in the fleet industry not only to save money in terms of saving fuel and maintaining the better health of vehicles but also to do better appraisals for their drivers. Fleet managers can use our statistics from our framework to identify the better driver and appreciate them which would encourage the driver to adhere to fuel-efficient driving patterns.

5.1 Overview of the proposed solution

Demir et al. [3] have stressed in their research that while most of the fuel consumption models pay attention to the impact due to vehicle, environment, and traffic, very little consideration is given for effects of driver behavior and operation. Gonder et al. [4] demonstrated that efficient driver behavior could provide up to 20% fuel savings. While drivers can be educated on general guidelines to save fuel, further savings can be achieved through individual feedback. This feedback is typically based on historical data, which is used for driver training and appraisals. While such training could improve the efficiency over time, it may saturate at a sub-optimal level. However, we believe that much more useful feedback can be given through real-time driver monitoring and feedback, where we could assist the drivers to change their driving behaviors while on the road, and maintain an efficient and safe driving behavior. The results could be immediate and more significant, as the driver is carefully pushed to reach a more optimum efficiency level. However, it is essential to do this in a non-intrusive manner while being aware of the environmental conditions, being conscious of impact on other traffic, and encouraging safe driving. One possibility is gamification [5], where the driver is rewarded for efficient and attentive driving than racing with other drivers.

Related work mostly considers only the driver dependent parameters while analyzing the impact of driver behavior on fuel usage. However, external factors such as weather, road traffic, road topography, and road conditions also influence the driving behavior. Therefore, to provide more practical and useful feedback one should consider both the driver-dependent and driver-independent factors on fuel usage. However, such

driver behavior and fuel consumption analysis typically requires high-resolution data from GPS-based tracking devices, various sensors, and other external sources. The data need to be gathered across multiple days and vehicles. Given the volume, diversity, and uncertainty of data, sophisticated data mining and data analytics techniques are required to identify fuel inefficient driver behaviors and to provide useful recommendations for individual drivers in real-time.

We propose a novel solution based on vehicular data analytics to encourage drivers to adhere to fuel-efficient driving behaviors. A classification model, developed based on historical data, evaluates driving behaviors for fuel efficacy considering vehicular, GPS, weather, and traffic data. If a driving behavior is detected to be inefficient, a fuzzy logic inference system decides what action the driver should perform to bring the vehicle back to a fuel-efficient state. The suggested action is conveyed to the driver via a mobile app as voice commands. A mobile app is selected to provide driver feedback due to the pervasiveness and with the intention of enhancing usability and flexibility, as well as to reduce costs. Voice commands are used, as they are non-intrusive and would not compensate for safety.

The crucial step in this solution is identifying driving events (i.e., periods of driving), which are fuel inefficient. While a classification model could be used, the primary challenge is how to label fuel efficiency of a given event considering all driver-dependent and driver-independent parameters. We address this problem by automatically clustering data points in a high dimensional space, and then analyzing those clusters for the fuel efficiency. Then we apply a label to the clusters and thereby individual data points are labeled. This labeling is more accurate than labeling just by looking at the fuel usage, as fuel usage depends on external conditions where fuel inefficient diving behaviors might be inevitable under some circumstances. For example, the driver might be forced to drive slowly due to heavy rain or excessive idling in a traffic jam.

We demonstrate the proposed technique using a dataset of a long-distance bus in Sri Lanka. This dataset provides an ideal test bench, as it includes all types of external conditions such as diving in urban and rural areas, driving within peak and off-peak hours, night driving, and driving through the mountainous region. As we do not use any parameter which is specific to this bus (e.g., load) or its route (e.g., latitude or longitude), our approach can be generalized to other cases and vehicles. We selected hierarchical clustering to cluster the data points based on their attributes, and then a random forest was used to classify the clusters as fuel efficient or inefficient. The developed classification model has an accuracy of 85.2%. To simulate the benefits

of the proposed mechanism, we compared the fuel economy of a journey with the historically best fuel economy for the same location, time, altitude, and weather, which resulted in 16.4% savings.

5.2 Proposed system architecture

To enhance the fuel economy of fleet vehicles, we propose a system that monitors drivers and provides individual feedback in real time. This system takes driver dependent and driver-independent influences into consideration in deciding whether a driving behavior is fuel efficient or not. Figure 5.1 illustrates a high-level overview of the proposed system. We assume vehicles are equipped with a GPS-based tracking system and a high-precision fuel sensor. Conventional floater-based fuel sensors are over sensitive to potholes and bumpers on the road, and upward and downward slopes, as well as rapid acceleration and deceleration. Therefore, high-precision fuel sensors are typically used to obtain more accurate fuel levels. Vehicle-related data such as speed, acceleration, the current location of the vehicle, and fuel level are first pushed to the cloud-based back-end in near real time. Data can be transferred through commodity technologies such as 3G/4G or Machine-to-Machine (M2M) communication.

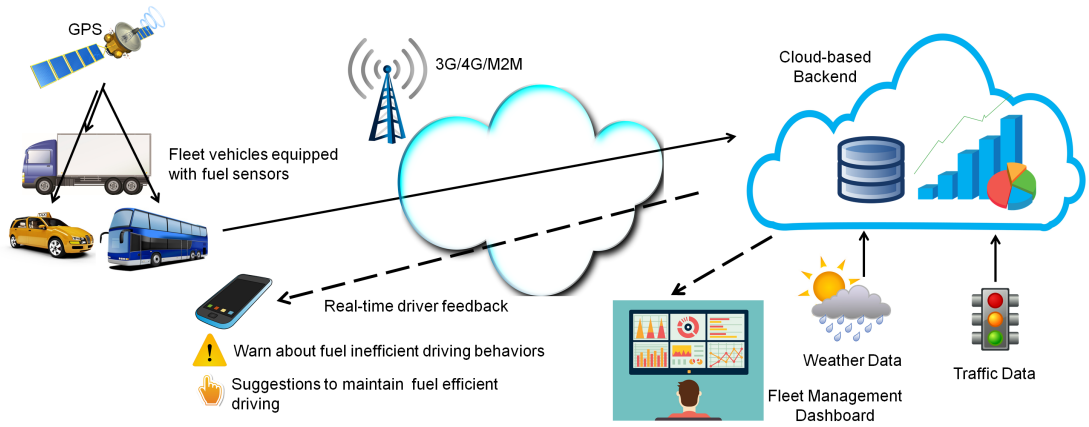


Figure 5.1: Overview of the proposed system for real-time monitoring and driver feedback to promote fuel-efficient driving.

An analytical engine running in the cloud-based backend combines GPS and fuel data, as well as other data such as weather and traffic conditions to estimate the current driving behavior. Today, relevant weather and traffic data could be pulled from third-party data sources using a REST API. If the driving behavior is determined to be ineffective, the analytical engine determines a suitable corrective action to take the vehicle back to the fuel-efficient state. The action is communicated to the driver as a voice alert using a mobile app.

Moreover, vehicle owners and fleet managers could be given a dashboard, which summarizes driving behavior and its impact to fuel consumption (see Figure 5.1). The dashboard may also indicate other metrics such as the percentage of time a driver adheres to fuel-efficient driving habits and causes for the inefficiency. Such a dashboard is useful in driver coaching and to appraise the performance of drivers based on driving behavior and fuel efficiency.

Figure 5.2 shows the analytical engine residing in the cloud-based backend. Once GPS, fuel consumption, and driving behavior related data arrive at the system, data are preprocessed. Preprocessing is required to clean the data and to derive new parameters such as isIdling (i.e., whether the vehicle is idling) and hour (i.e., time of day). Then vehicular data and weather data are fed into a classification model, which decides the fuel efficiency of the driving event. Driving behavior throughout one minute of time under given external conditions is defined as a driving event. If classifier classifies a driving behavior as fuel inefficient, then a fuzzy logic inference system decides the control action given the related data. Predicted control action is then transferred back to the driver as feedback. This process continues as new data points keep arriving from the vehicle.

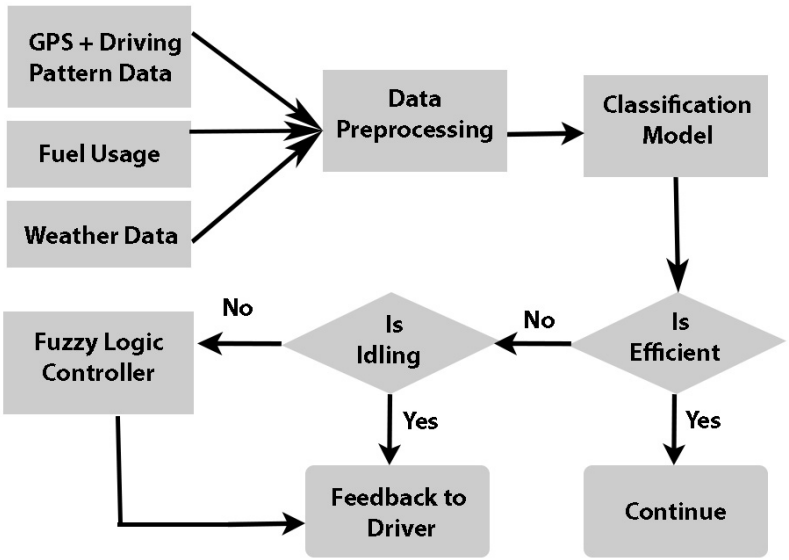


Figure 5.2: Data flow for one driving event.

The GPS-based tracking device sends data to the cloud when the bus takes a turn or every 17 seconds whichever occurs first. Tracking device tracks other events of interest such as rapid acceleration, deceleration, and ignition on/ off state too. However, they

are timestamped with the last time reading from the GPS (due to hardware constraints time is read from the GPS kit once every 5 seconds). This leads to uneven sampling. Therefore, to make the sampling rate consistent, the dataset was aggregated to one minute. Followings are the variables selected for further analysis.

- Speed (kmh^{-1})
- Acceleration (kmh^{-2})
- IsIdling

  Ignition status =1 and Speed = 0 for one 1 min
- Elevation change (m) - Road topography
- Time of the day - Traffic condition

  For the nearest hour
- Fuel mileage (kmL^{-1})
- Weather condition - Sunny, clear, partly cloudy, cloudy, overcast, patchy rain nearby, light drizzle, light rain shower, moderate rain, moderate or heavy rain, mist, and fog

These parameters were selected such that the driver impact (first three factors), road topography, road traffic and weather conditions are captured. Driver behavior impact is captured with speed, acceleration and idling. Here we considered a vehicle to be in idling status if it is not moving for one minute even though the engine is on. As mentioned above 1 minute was selected due to the information presented in [6] and also 1 min can differentiate idling due to stopping at traffic lights and idling due to driver carelessness. Time of the day captures traffic condition. Time is a better representation of traffic condition as this bus travels overnight. With distance and fuel usage we calculated fuel mileage for each driving event. The weather condition of the location such as rainy, sunny, etc. given time indicates the impact of weather.

5.3 Clustering

We to classify driver behaviors as fuel-efficient or inefficient while considering both the driver dependent and environmental parameters. However, it is important to recognize that sometimes the drivers will not be able to follow fuel-efficient driving behaviors due to external factors such as traffic and weather conditions. For instance,

consider the following two tables that indicate driving events in two different locations of the route. (As mentioned above in this analysis we considered 1mins driving events. Therefore, mileage in the table is an indication of the vehicle performance in that 1 minute instance.) Table5.1 contains information about driving events around Wellawaththa, a suburb of Colombo. The bus is traveling through Wellawaththa area around 5:00 p.m. which is a peak traffic time. Table5.2 contains information about driving events around Udawalawa, an area out to Colombo. The bus is going across this area in the night around 10:00 p.m., thus, the journey is not affected by traffic in this part of the drive. Further, Udawalawa is a flat area with negligible elevation changes. Hence, the bus is traveling at its optimal speed and fuel consumption for those driving events are very low which results in higher mileages.

Table 5.1: Driving events near Wellawaththa.

Distance (km)	**Speed (kmh^{-1})**	**Fuel usage (L)**	**Mileage (kmL^{-1})**
0.1803	7.2784	0.34078	0.5289
0.1214	6.0153	0.09747	1.2453
0.3734	15.719	0.35774	1.0438

Table 5.2: Driving events near Udawalawa.

Distance (km)	**Speed (kmh^{-1})**	**Fuel usage (L)**	**Mileage (kmL^{-1})**
1.6696	70.3359	0.0027	621.5171
1.3907	60.3271	0.0022	621.5171
1.3463	59.9492	0.0022	601.6583

When classifying driving events for efficiency, we must consider both driver-dependent and driver-independent parameters such as weather, traffic and elevation changes. When above two scenarios are considered driving events near Wellawaththa have lower mileage, not necessarily because the driver was not driving properly but because of the road traffic. In such a situation claiming it to be fuel inefficient and providing suggestions such as speed up the vehicle is not practical. Therefore, we cannot solely rely only on the fuel consumption to label driving behaviors for fuel efficiency, instead, we must consider as many as possible external factors in labeling our dataset.

Manually tagging individual data points as efficient or inefficient driving, considering all the influences is a tedious task. Therefore, we propose to use an unsupervised clustering technique to cluster the data points into different clusters in a high dimensional space. Then we take the assistance of a domain expert(s) to analyze those clusters and label them as either fuel-efficient or fuel inefficient considering not only the fuel efficiency, speed, and acceleration, but also weather, road condition, and traffic.

Once the clusters are labeled, respective data points can also be labeled based on their cluster membership.

To demonstrate the proposed mechanism, let us combine two significantly different set of data points along the bus route; data from a two hour (17:00 - 19:00) drive close to Colombo (an urban area) and a two hour (22:30 – 00:30) drive close to Udawalawa (a rural area). Figure 5.3 shows a scatter plot of the selected data points, where the fuel usage is plotted against longitude. Black dots indicate the data points close to Colombo, while red dots show the data points close to Udawalawa.

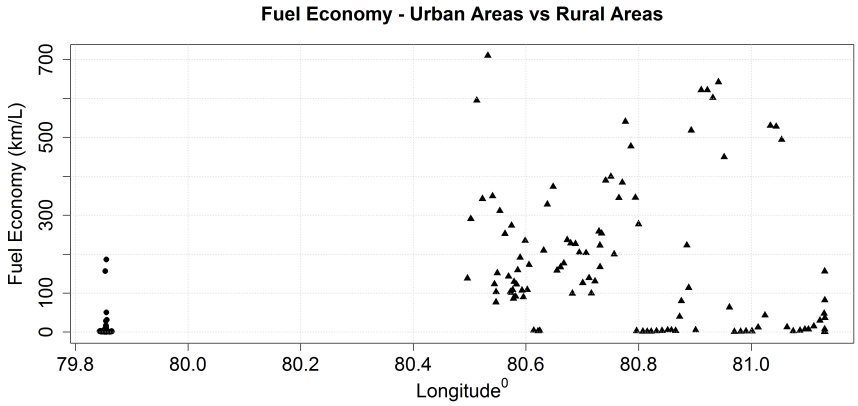


Figure 5.3: Fuel usage in an urban and a rural area. Dots – urban area and triangles – rural area.

If our unsupervised clustering algorithm can cluster the dataset accurately, it should identify at least two clusters, even without providing geographical information. However, a careful analysis of the distribution of data points reveals the following characteristics:

- The number of clusters is unpredictable for a given journey. Since fuel consumption depends on external factors, the number of clusters would vary widely.
- Clusters are not spherical.
- Clusters have uneven sizes. (i.e., different number of cluster members).
- Clusters have different densities, e.g., in Figure 5.3 the small cluster around 79.80 has a higher density than other clusters.
- Clusters do not adhere to a normal distribution.

We choose hierarchical agglomerative clustering, as it is more desirable for clustering data points with the above characteristics [28]. The key advantages of hierarchical agglomerative clustering techniques are the flexibility of exploring on different levels of granularity, easiness of handling any forms of similarity or distance, and applicability to any attribute type [28]. Moreover, if upper-level clusters on the dendrogram are not providing enough information to label the clusters, we can drill down to lower levels in the hierarchy, and then get smaller clusters and analyze their properties. This benefit of hierarchical clustering enables more accurate labeling of fuel efficient and inefficient events.

To implement agglomerative hierarchical clustering "hclust" function in R was used. We used Euclidean distance to measure the distance between data points. Empirical results showed that Euclidean distance provides better clusters compared to other distance measures such as maximum, Manhattan, Canberra, binary, and Minkowski. Similarly, ward.D2, single, complete, average, mcquitty, median and centroid algorithms [29] were used for agglomeration. Among the algorithms considered, ward.2D resulted in better clusters. In clustering fuel data, better clustering refers to the ability to cluster events with the same external conditions into the same cluster.

Dendrogram in Figure 5.4 shows cluster hierarchy for the dataset considered in Figure 5.3 (Each branch in the dendrogram represents a different cluster). We considered speed, acceleration, isIdling, elevation change, hour, and weather condition for this clustering. Cutting the tree at level four results in seven clusters. Figure 5.5 plots the same data points, labeled (colored) according to the clusters they belong to.

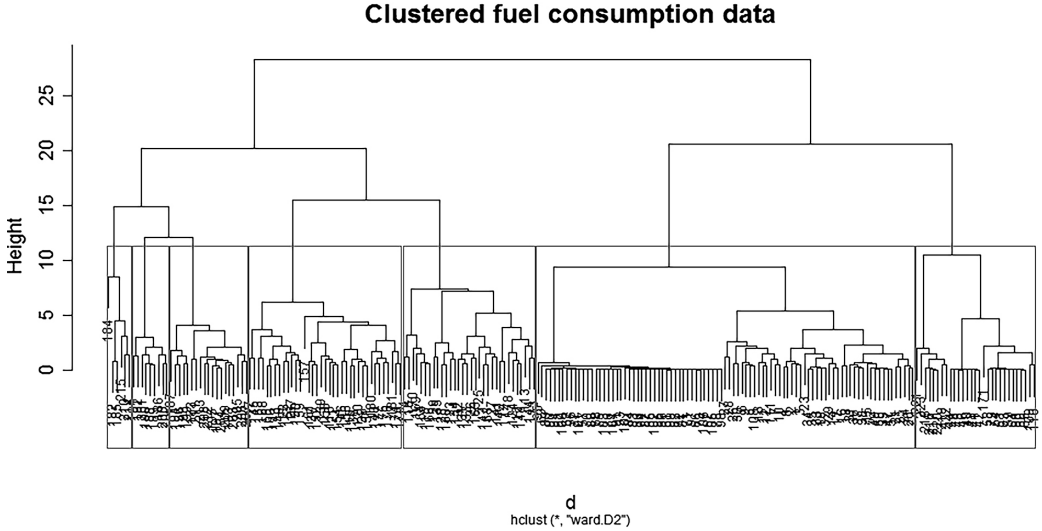


Figure 5.4: Dendrogram of clusters produced by hierarchical clustering.

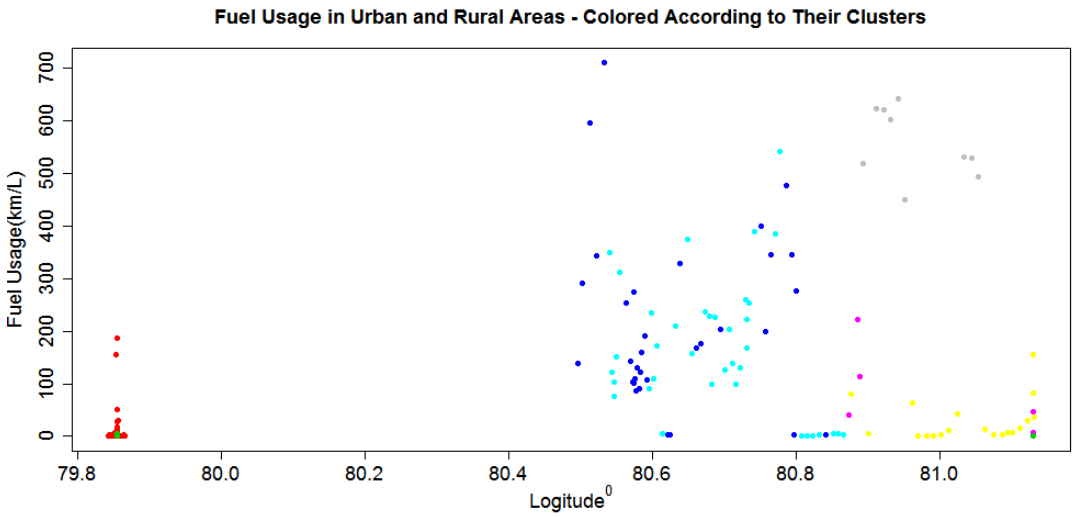


Figure 5.5: Seven clusters found in sample fuel consumption data points. Each color indicates a different cluster.

Table 5.3 presents a summary of identified clusters. One can directly label cluster two as inefficient, not only because of the lower mean fuel usage (5.28 kmL^{-1}), but also due to excessive idling at midnight. This is perhaps due to the driver having a break, while the engine is running. The table provides evidence to prove the argument that a driver cannot be accounted for inefficient fuel usage just based on the mean fuel usage. For instance, consider the first cluster. Both the fuel consumption (11.44 kmL^{-1}) and mean speed (6.86 kmh^{-1}) are low. However, the time of the day is 17:00. Therefore, one can conclude that road traffic might cause this lower speed. Thus, asking the driver to speed up the vehicle under this condition is not practical. Domain experts analyzed the resultant clusters and labeled for their fuel efficiency considering all the driver dependent and external factors. Consequently, cluster six (Table 5.3) was marked as inefficient because the mean fuel usage is not acceptable, even though all external conditions are desirable.

In labeling the whole dataset, each journey was clustered separately because different trips might have driven by different drivers. While it is known that the different drivers drive the bus on different days, the dataset did not contain information on who drove on a given day. Clustering each journey separately would eliminate specific driver behavioral impact on fuel consumption.

Table 5.3: Summary of each cluster derived using hierarchical clustering.

Cluster No	Mean Speed (kmh^{-1})	Mean Acc (kmh^{-2})	Mean Elevation Change (m)	IsIdling (Mode)	Time of the day (Mode)	Weather Condition (Mode)	Mean Mileage (kmL^{-1})	Fuel Efficiency
1	6.86	-14.56	-0.02	0	17	Clear	11.44	Efficient
2	0	0	0	1	0	Cloudy	5.28	Inefficient
3	45.89	-5.53	6.36	0	23	Mist	214.86	Efficient
4	45.99	-109.83	-7.379	0	23	Mist	167.25	Efficient
5	28.35	-6,818.00	5.025	0	0	Mist	71.88	Efficient
6	61.12	273	-0.96	0	0	Cloudy	29.57	Inefficient
7	62.77	252.54	-0.05	0	0	Cloudy	556.3	Efficient

5.3.1 Classification of Fuel Usage

Once the historical data points are labeled, the next step is to develop the classification model. The classification model should classify a data points as either fuel efficient or inefficient. The dataset we are analyzing is not linearly separable and is known to have outliers. Therefore, Random Forest (Random Forest) was selected as the classification technique, as it can handle non-linear features, high dimensional data, and many training samples. We used Random Forest algorithm available in R random forest package to build the classification model. *mtry*, the number of variables randomly sampled for a split, was set to three as it gave the least Out-Of-Bag (OOB) error estimate of 14.13%. *ntree*, the number of trees within the ensembles, was set to 500 based on the empirical evidence. Historical data that were labeled via clustering in the previous step was used to training the Random Forest-based classification model.

5.3.2 Determining the Control Action

If the classification model detects a driving behavior as fuel inefficient, the next task is to determine what the diver should do to bring the vehicle back to a fuel-efficient state. As seen in Figure 5.2 the decision-making process has two steps. First, the proposed system checks whether the inefficiency is due to idling. If it is due to the vehicle being idle, the system sends the feedback suggesting to stop the engine. If idling is not the reason for detected inefficiency, then the system uses a fuzzy logic inference system to determine the control action.

Many applications in the transportation domain use Fuzzy Logic inference systems successfully [30, 31, 32]. The reason for the popularity of Fuzzy Logic Controllers (FLC) is their ability to model real-world ambiguous reasoning. Nevertheless, FLC

emulates the need of an expert in the form of linguistic rules [33]. We choose driver-dependent influences on fuel usage, speed, and acceleration as inputs to the fuzzy logic system.

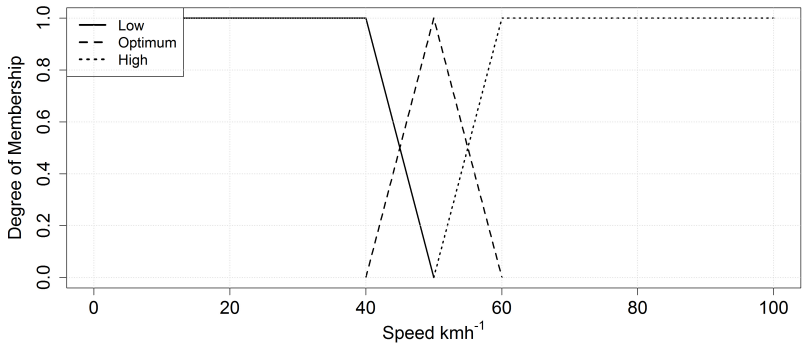


Figure 5.6: The membership function of speed.

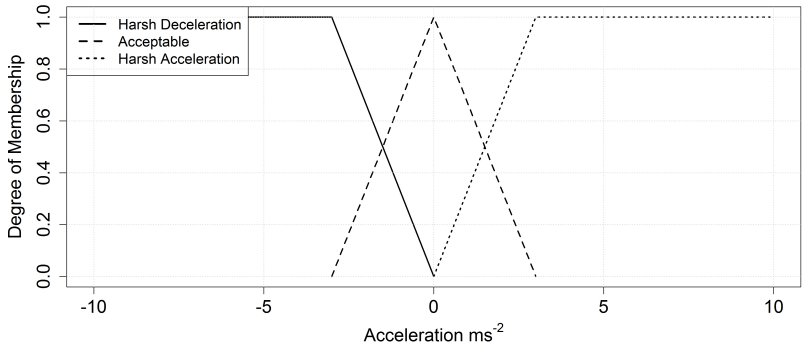


Figure 5.7: The membership function of acceleration.

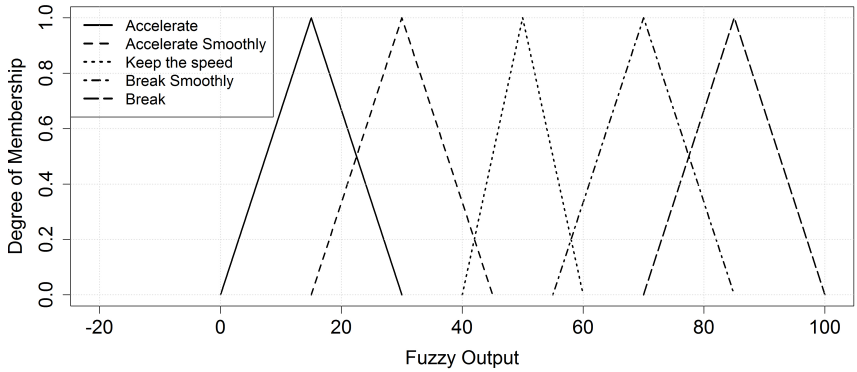


Figure 5.8: Fuzzy output membership function.

Figure 5.6 and 5.7 show the corresponding membership functions respectively. The fuzzy values of the speed are selected as Low (L), Optimum (O), and High (H) because through our exploratory analysis we could observe that fuel economy vs. speed curve follows a bell curve. The corresponding fuzzy values of acceleration are selected as Harsh Deceleration (HD), Acceptable (A), and Harsh Acceleration (HA) as the expert advice is to use gentle acceleration and breaking. We decided on these fuzzy values based on industrial norms.

Figure 5.8 shows the membership function of the fuzzy output, action to be taken by the driver. Finally, the fuzzy rules of the inference system (i.e., control actions) are derived as in Table 5.4 based on the fuzzy inputs and membership functions.

Table 5.4: Fuzzy rules.

Speed	**Acceleration**	**Control Action**
L	HD	Accelerate
L	HA	Accelerate Smoothly
O	HD	Keep the Speed
O	HA	Keep the Speed
H	HD	Break Smoothly
H	A	Break
H	HA	Break

5.4 Results

The performance of the classification model was assessed using 10-fold cross-validation. To analyze the test results standard statistics were calculated. The results are given in Table 5.5.

Table 5.5: Statistics of results of the classification model.

Statistical Measure	**Value**
Accuracy	85.16%
Kappa statistics	0.7011
Mean absolute error	0.2082
Root mean squared error	0.3191
Relative absolute error	41.82%
Root relative squared error	63.95%
Precision	0.852
Recall	0.852

Accuracy indicates the percentage of correctly classified test cases whereas Kappa statistic shows the agreement of prediction with the true calls. While mean absolute

error measures the average magnitude of errors of the prediction without considering the direction, root mean squared error gives the square root of the average squared error. Table 5.5 depicts that the RF-based classifier has higher accuracy, precision, and recall while having a lower error. Higher precision and recall mean that most of the driving events identified as inefficient driving events are true inefficient driving events and most of the inefficient driving events are correctly identified respectively. Therefore, high precision and recall grantee that the feedback sent to the driver is not intrusive and it is meaningful.

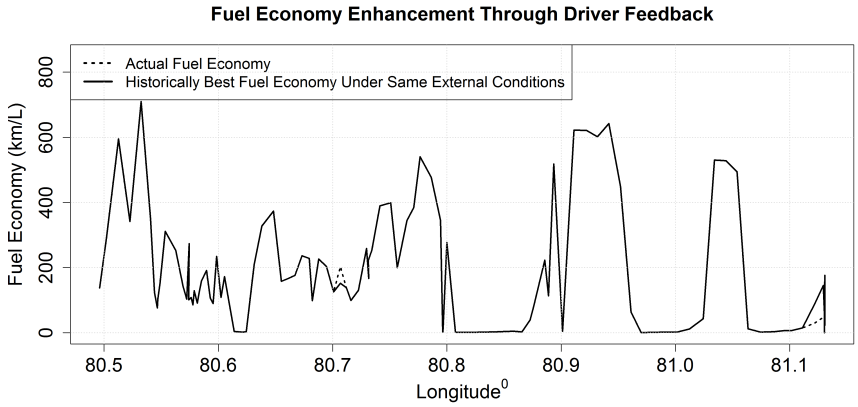


Figure 5.9: Actual fuel usage vs. adjusted fuel usage based on driver feedback for a selected journey

To test our solution's effectiveness in saving fuel, we carried out a simulation. We assessed to what extent the fuel economy of the bus can be increased if we would follow the historically best action in those detected inefficient driving events. Figure 5.9 shows both the actual fuel usage of the bus for a journey and the fuel usage when inefficient events are replaced by the historically best fuel economy under the same external conditions. We observed that the newly estimated fuel usage, based on historically best fuel economy with driver feedback is in average 16.36% higher than the actual fuel usage. This indicates an upper bound on expected gain in fuel efficiency, as on any given day the driver may not be able to drive at the best efficiency at each driving event. Nevertheless, aggregated saving over multiple days, routes, and vehicles could still be significant from the feet owners point of view, as the proposed solution is independent of the vehicle and route.

In the figure, there is one place where actual fuel economy is better than the historically best case. The possible reason can be that the classification model might have misclassified an efficient driving event as an inefficient driving event.

5.5 Conclusions

We proposed a novel framework based on vehicular data analysis to promote fuel-efficient driving behaviors among drivers via real-time driver monitoring and feedback. We demonstrated that by considering both driver dependent parameters such as speed and acceleration, as well as external parameters such as weather, road traffic, and road topography, more accurate and useful feedback can be given to the driver. The framework consists of a classification model and a fuzzy logic controller. The classification model classifies different driving behaviors as fuel efficient and inefficient. When a particular driving behavior is detected to be inefficient, the fuzzy logic inference system determines the corrective action to bring the vehicle back to a fuel-efficient state. Results demonstrate that the proposed solution can achieve significant fuel saving.

A possible avenue to improve is to integrate other driver independent parameters such as the load of the vehicle, road type, and real-time traffic data. Being able to do the classification and fuzzy inference only on the smartphone is also of interest. This would eliminate the need to send data to the cloud-based backend in (near) real-time saving both the bandwidth and power. While weather and traffic data need to be downloaded to the smartphone, it can be done less frequently.

6 CONCLUSION

6.1 Summary

In this research, we tried to model and evaluate fuel consumption of fleet vehicles based on vehicular data and data analytics methodologies and suggest suitable process improvement/ re-engineering actions to improve the fuel economy. This objective was narrowed down to three sub-objectives as follows:

- Detect possible fuel frauds
- Verify detected prospective fuel frauds
- Encourage drivers to maintain a better fuel economy for their vehicles

To achieve these objectives, we analyzed a dataset of a long-distance public bus running in Sri Lanka. To identify possible fuel frauds, we proposed to predict instantaneous fuel consumption and compare recorded fuel usage values against the predicted values. The predictive model is proposed to develop using various influential factors of fuel consumption. For this purpose, we evaluated the predictive ability of three machine-learning models. While the selected dataset has several essential parameters that directly influence fuel consumption, several other relevant parameters such as load, engine RPM, and traffic are not available. Even in the absence of such vital parameters, we demonstrated that Random Forest model could predict the fuel consumption more accurately (MAE of the final model is 0.0229) while capturing the trends in data. More precisely, given a set of parameters such as distance, location, elevation, speed, and day of the week, the Random Forest model informs us sensible fuel consumption for the journey. A thresholding mechanism is used to identify possible fuel frauds by comparing the recorded and predicted fuel usage values.

Nevertheless, as explained earlier fuel consumption of a vehicle is highly subjected to external factors such as weather, traffic, etc. Hence, we cannot accuse a driver of fuel frauds without a systematic verification. Thus, we had to identify better key performance indicators to assess fuel consumption on the day of interest to verify its validity. These indicators had to quantify the influence of each parameter - road traffic, driver behavior, elevation changes, and weather conditions - on fuel consumption. We proposed to use speed profile and frequency of harsh events as indicators of road traffic. Further, total idle time and day of the week also can be used as indicators.

Furthermore, we proposed a novel vehicular data analytics framework to promote fuel-efficient driving behaviors among drivers via real-time monitoring and driver feedback which is the third objective of the research. We demonstrated that by considering both driver dependent parameters such as speed and acceleration, as well as external parameters such as weather, road traffic, and road topography, we could provide useful and more accurate feedback to the driver. To achieve this, a classification model was developed to classify different driving behaviors as fuel efficient or inefficient. When a driving behavior is detected to be inefficient, a fuzzy logic inference system was proposed to determine the corrective action to bring the vehicle back to a fuel-efficient state. In our experiment using the dataset of a public bus, the accuracy of the classification model was 85.16% and estimated improvement of fuel economy by following provided feedback is in average 16.36%.

6.2 Research Limitations

In this section, we discuss the limitations of this research.

First, in this work, we have assumed that fuel consumption of a vehicle depends only on instantaneous factors. However, this assumption may not hold true always, because sometimes what happened in earlier time steps might still have an impact on the current fuel consumption. For example, acceleration might not affect fuel usage immediately. Another instance is, even though raining has stopped now, its impact might still be significant, e.g., slippery or flooded roads. Hence, developing a more reliable predictive model needs to release this assumption.

Second, the dataset used in this analysis possess some limitations. It does not contain some of the key influential factors such as the load of the bus, RPM value. Further, this dataset is suspected to contain fuel frauds. Even though we did not observe any exceptionally high instantaneous fuel usages as evidence, we identified a couple of days with unusual fuel consumptions due to obvious factors such as tracking device malfunctioning and change of route. Those were removed from the analysis. Nevertheless, the refined dataset might still contain fraudulent fuel consumptions which are hard to identify. Having such instances in the training dataset might have adversely affected the analysis because such fuel usages do not indicate the correct relationship between fuel consumption and explainer parameters.

The third limitation is, in the currently proposed framework for real-time driver monitoring and feedback, data analysis happens on the server side. This cause the system to suffer from network delay and would make it fail to provide timely feedback.

Thus the feedback might not be useful.

Finally, real-world implementation and evaluation of the proposed frameworks impose some limitations. Fraudulent activities identification and real-time feedback directly affect the career of fleet drivers; hence, the accuracy of fuel consumption prediction model and fuel efficiency classification model should be very high. Currently, the MAE of fuel usage prediction model is 0.0229 and the accuracy of fuel efficiency classification model is 85.16%. However, further improved models are preferable for real-world deployments and evaluations. False positives, false alarms about suspicious fuel spends or wrong feedback about driving patterns would cause for unsatisfied drivers. Ultimately, this would lead for wrong analysis. Even with perfect predictive models, challenges due to human factors should not be underestimated. How can we motivate fleet drivers to voluntarily agree to be monitored? How can we encourage drivers to follow recommendations given by our system? To answer these questions we might have to work closer with the fleet management industry and its domain experts. Domain expertise is required for development of the proposed frameworks as well, to verify fuel fraud and to label clusters for fuel efficiency. Nevertheless, the requirement of domain expertise raises another limitation. Acquiring domain expertise is costly and also it can add human biases to the process, e.g., verifying fuel frauds.

6.3 Future Work

As mentioned in Chapter 3, the dataset we analyzed did not contain some of the important influential factors of fuel consumption such as traffic, road type, RPM, and the load of the bus. Further, in this research fuel consumption was calculated based on fuel level measurements. Instead, one could consider fuel consumption estimation using OBD2 port as a better way of capturing real-time fuel consumption. Integrating above-mentioned additional factors and using OBD2 port would further improve the predictive model and would allow identifying fuel frauds in a more reliable manner. If variables such as traffic, road type, RPM, and the load of the bus are available, one can improve the classification model to classify driving events and provide more useful feedback to driver enhancing the usability of the system.

Modeling fuel consumption as a time series model considering the temporal relationship of variables as well would provide a more accurate predictive model. Recently, with the development of deep learning, Recurrent Neural Networks are used to model temporal relationships. Exploring applicability of deep learning to fuel consumption prediction would be a promising research direction.

Moreover, being able to carry out the classification and fuzzy inference on the smartphone itself is another avenue to explore. This would eliminate the need to send data to the cloud in (near) real-time saving both the bandwidth and power. While weather and traffic data need to be downloaded to the smartphone, it can be done less frequently. Further investigation of suitable mechanisms to derive driver feedback also would be a future direction.

References

[1] "Global energy trends – bp statistical review 2015." http://euanmearns.com/global-energy-trends-bp-statistical-review-2015/. Accessed: 2016-05-13.

[2] "Sri lanka sustainable energy authority, "sri lanka energy balance"." https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html. Accessed: 2016-08-11.

[3] E. Demir, T. Bektaş, and G. Laporte, "A review of recent research on green road freight transportation," *European Journal of Operational Research*, vol. 237, no. 3, pp. 775–793, 2014.

[4] J. Gonder, M. Earleywine, and W. Sparks, "Analyzing Vehicle Fuel Saving Opportunities through Intelligent Driver Feedback," in *SAE International Journal of Passenger Cars-Electronic and Electrical Systems*, no. April, pp. 24–26, 2012.

[5] CGI, "Modeling the Relation Between Driving Behavior and Fuel Consumption," tech. rep., 2014.

[6] H. J. Walnum and M. Simonsen, "Does driving behavior matter? an analysis of fuel consumption data from heavy-duty trucks," *Transportation research part D: transport and environment*, vol. 36, pp. 107–120, 2015.

[7] "Just how much does driver behavior actually affect fuel efficiency?." https://www.fleetcarma.com/driver-behavior-fuel-cost/. Accessed: 2017-03-07.

[8] O. Linda and M. Manic, "Improving vehicle fleet fuel economy via learning fuel-efficient driving behaviors," *International Conference on Human System Interaction, HSI*, pp. 137–143, 2012.

[9] E. Gilman, A. Keskinarkaus, S. Tamminen, S. Pirttikangas, J. R??ning, and J. Riekki, "Personalised assistance for fuel-efficient driving," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 681–705, 2014.

[10] "Fuel economy in automobiles." https://en.wikipedia.org/wiki/Fuel_economy_in_automobiles. Accessed: 2017-03-06.

[11] K. Ahn, H. Rakha, A. Trani, and M. Van Aerde, "Estimating Vehicle Fuel Consumption and Emissions based on Instantaneous Speed and Acceleration Levels," *Journal of Transportation Engineering*, vol. 128, no. 2, pp. 182–190, 2002.

[12] A. Viswanathan, "Data driven analysis of usage and driving parameters that affect fuel consumption of heavy vehicles," 2013.

[13] L. Wang, A. Duran, J. Gonder, and K. Kelly, "Modeling Heavy / Medium-Duty Fuel Consumption Based on Drive Cycle Properties," Tech. Rep. 2812, 2015.

[14] L. Rokach, "Ensemble-based classifiers," in *Artificial Intelligence Review*, pp. 1–39, 2010.

[15] a. Liaw and M. Wiener, "Classification and Regression by randomForest," *R news*, vol. 2, no. December, pp. 18–22, 2002.

[16] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] M. Herrera, L. Torgo, J. Izquierdo, and R. Pérez-García, "Predictive models for forecasting hourly urban water demand," *Journal of Hydrology*, vol. 387, no. 1-2, pp. 141–150, 2010.

[18] J. Chen, M. Li, and W. Wang, "Statistical uncertainty estimation using random forests and its application to drought forecast," *Mathematical Problems in Engineering*, vol. 2012, pp. 1–13, 2012.

[19] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[20] P. Bühlmann and B. Yu, "Boosting with the l 2 loss: regression and classification," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 324–339, 2003.

[21] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[22] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *Journal of Clinical Epidemiology*, vol. 49, no. 11, pp. 1225–1231, 1996.

[23] "Historical or past weather api."." `https://developer.worldweatheronline.com/api/docs/historical-weather-api.aspx`. Accessed: 2017-03-06.

[24] I. M. Berry, *The effects of driving style and vehicle performance on the real-world fuel consumption of US light-duty vehicles*. PhD thesis, Massachusetts Institute of Technology, 2010.

[25] J. S. Stichter, "Investigation of Vehicle and driver aggressivity and realtion to fuel

economy testing," tech. rep., 2012.

[26] B. Hofner, A. Mayr, N. Robinzonov, and M. Schmid, "Model-based boosting in R," tech. rep., 2012.

[27] J. E. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models part i—a discussion of principles," *Journal of hydrology*, vol. 10, no. 3, pp. 282–290, 1970.

[28] P. Berkhin, "Survey Of Clustering Data Mining Techniques," *Accrue Software, San Jose, CA*, pp. 1–56, 2002.

[29] "R documentation, "r: Hierarchical clustering." `https://en.wikipedia.org/wiki/Fuel_economy_in_automobiles`. Accessed: 2017-03-06.

[30] O. Linda and M. Manic, "Improving vehicle fleet fuel economy via learning fuel-efficient driving behaviors," in *International Conference on Human System Interaction, HSI*, pp. 137–143, 2012.

[31] A. Aljaafreh, N. Alshabatat, and M. S. Najim Al-Din, "Driving style recognition using fuzzy logic," *2012 IEEE International Conference on Vehicular Electronics and Safety, ICVES 2012*, pp. 460–463, 2012.

[32] D. Dorr, D. Grabengiesser, and F. Gauterin, "Online driving style recognition using fuzzy logic," *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, pp. 1021–1026, 2014.

[33] H. R. Berenji and P. Khedkar, "Learning and tuning fuzzy logic controllers through reinforcements," *IEEE Transactions on neural networks*, vol. 3, no. 5, pp. 724–740, 1992.