Fuel Consumption Prediction of Fleet Vehicles Using Machine Leaning: A Comparative Study

Sandareka Wickramanayake and H.M.N. Dilum Bandara
Department of Computer Science & Engineering
University of Moratuwa
Moratuwa, Sri Lanka
sandarekaw@cse.mrt.ac.lk, dilumb@cse.mrt.ac.lk

***Abstract*— Ability to model and predict the fuel consumption is vital in enhancing fuel economy of vehicles and preventing fraudulent activities in fleet management. 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. However, not all these factors may be measured or available for the fuel consumption analysis. We consider a case where only a subset of the aforementioned factors is available as a multi-variate time series from a long distance, public bus. Hence, the challenge is to model and/or predict the fuel consumption only with the available data, while still indirectly capturing as much as influences from other internal and external factors. Machine Learning (ML) is suitable in such analysis, as the model can be developed by learning the patterns in data. In this paper, we compare the predictive ability of three ML techniques in predicting the fuel consumption of the bus, given all available parameters as a time series. Based on the analysis, it can be concluded that the random forest technique produces a more accurate prediction compared to both the gradient boosting and neural networks.**

Keywords—artificial neural networks; fuel economy; gradient boosting; predictive model; random forest

I. INTRODUCTION

Ability to understand the factors that influence fuel consumption and then being able to predict it, is vital in enhancing fuel economy of vehicles and preventing fraudulent activities in fleet management. For example, let us consider a long distance bus that runs between a major city and a rural area. Along the way, the bus may encounter traffic and different terrain conditions such as going over a mountainous area. Moreover, the traffic intensity, weather conditions, and the load of the bus may vary depending on the day of the week. Furthermore, different drivers may drive the bus on different days. Consequently, the fuel consumption of the bus could drastically vary across days. Large variation on fuel consumption opens up opportunities for fuel fraud. For example, on a day that the road conditions were good and not many passengers took the bus, several liters of fuel could be pumped out of the tank without being noticed. Such activities are prevalent in economies where the fuel price is relatively high.

Whereas the bus owners may want to understand the factors that contribute to the fuel consumption, so that they can introduce suitable process reengineering steps to reduce the fuel consumption. Moreover, by being able to predict the fuel consumption the owners may also detect potential fuel fraud. With the advancement of Global Positioning System (GPS) based tracking devices and precision fuel sensors, vehicle owners are now able to capture high resolution, multi-variate time series datasets related to a vehicle location, speed, engine conditions, and fuel consumption. Such data can be also used to derive other Key Performance Indicators (KPIs) such as idling time, day of week, and driver acceleration and breaking profiles. However, still data related to several other influencing factors such as load, traffic, weather, and driver are not collected through the same tracking device, and are usually not measureable for each of the vehicle. While some of this information is available through other third-party systems/services, we consider a case where such information is not available (which is typical in rural areas and in developing countries). Hence, the challenge is how to model and predict the fuel consumption at the presence of only a subset of the key factors that contribute to the fuel consumption.

With evolvement of big data, powerful analysis tools are required to transform those data into meaningful insights. To reach conclusions from data, there are two cultures in statistical modeling; the data modeling culture and the algorithmic modeling culture [1]. The *data modeling culture* assumes a stochastic data model for the inside of the black box, into which independent variables are going in and response variables are coming out. The *algorithmic modeling culture* considers the inside of this black box to be complex and unknown. They start with the outcome and find a function $y = f(x)$ – an algorithm that operates on x (predictor variables) to predict the responses y [1]. When the relationship between the predictors and responses becomes increasingly complex and includes high dimensional data, interactions, and linear and non-linear relationships Machine Learning (ML) outperforms classical statistical models. Therefore, ML is suitable for fuel consumption prediction, as the model can be developed by learning the patterns in available data, which in-turn could be used for prediction.

In most of the previous fuel consumption researches using ML techniques, researchers have used datasets from

978-1-4799-1740-2/15/$31.00 ©2015 IEEE

vehicles travelling in highways. Viswanathan [2] studied the parameters that hold more importance in predicting fuel consumption of trucks travelling in highways based on ML techniques. However, it is relatively easier to model the fuel consumption in highways, as impact from external factors such as traffic and road conditions are stable throughout most parts of the journey. Moreover, a couple of explanatory parameters she had – coasting, distance with trailer attached, and distance with cruise control - are not available in the dataset analyzed in this research. In another study, Lindberg [3] attempted modeling fuel consumption of heavy vehicles by combining GPS data with road, vehicle, and weather data. Using this rich dataset author built a regression tree, random forest, boosted tree, and Support Vector Regression (SVR) models. However, author's findings had lower accuracy due to low-resolution data.

We evaluate three ML prediction models in predicting the fuel consumption of a long distance, public bus, given all available parameters as a time series. The dataset includes timestamp, speed, distance, location, bearing, ignition status, fuel level, elevation, and battery voltage. Fuel level is measured using a capacitive, high-precision fuel sensor. Formally, we attempt to predict an outcome variable y of a multivariate dataset including several explanatory variables x_1, x_2, x_3, …, x_n. Overall fuel consumption is derived by predicting the instantaneous fuel consumptions y, given x_1 to x_n, which includes distance, latitude, longitude, elevation, speed, etc. Random forest based model, gradient boosting model, and black box artificial neural network model are considered in the evaluation.

Rest of the paper is organized as follows. Section II presents the exploratory data analysis of the selected dataset. Selected ML techniques are described in Section III. Section IV presents the comparison of predictive ability of the three selected ML techniques based on the dataset. Predictive accuracy is evaluated in Section V while concluding remarks are given in Section VI.

II. Exploratory Data Analysis

A. Dataset

The dataset corresponds to a particular long distance, public bus in Sri Lanka. Bus starts from Depot around 4:00pm and then goes to Colombo (i.e., commercial capital). 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. Altogether, bus travels ~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 journey is through a mountainous region. The bus is fitted with a GPS-based tracking device and a capacitive, high-precision fuel sensor. Collected data is pushed to a cloud sever 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°N, Max: 9.835556° N)
- Latitude (Min: 79.516667° E, Max: 81.879167° E)
- Bearing (0^0 to 360^0)
- Elevation (Min: 0m, Max: 2,524m)
- Distance traveled (km) – between two samples
- 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)

B. Descriptive Analysis

The first step in model development is to understand the data, which can be achieved by an exploratory data analysis. Fig. 1 shows the fuel consumption for both outward and inward (i.e., return) journeys. It can be seen that there is a significant difference in fuel consumptions for outward vs. inward journey. Based on the analysis it was identified that this is due to the differences in traffic experienced by the bus (it experiences more traffic during the outward journey), time of day, and road conditions (outward journey through the mountainous region is steeper, where more acceleration is required). Due to this major difference, the dataset was divided into two parts as inward and outward and the predictive models were developed separately. In both the cases a couple of outliers, which were caused by a bus breakdown, GPS device failure, and bus taking additional routes were removed.

Among the available predictors, distance directly influences fuel consumption of the bus (see Fig. 2). Another crucial factor that affects the fuel economy is speed of the vehicle [4], [5], [6]. Mean fuel consumption at each speed is depicted in Fig. 3. Due to the traffic and elevation changes location create a substantial impact on the fuel consumption. As seen in Fig. 4 the relationship between fuel consumption and the elevation is not linear, which is important to be noted while creating predictive models. This exploratory data analysis is important in selecting a predictive model for this particular dataset. As some predictors have a non-linear relationship

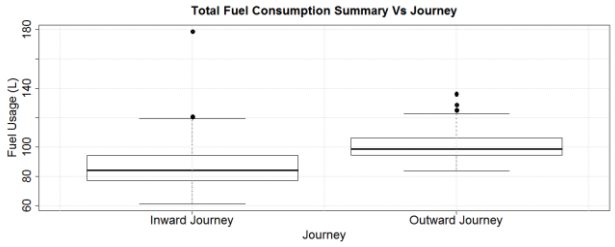

Fig. 1. Fuel consumption for inward and outward journeys.

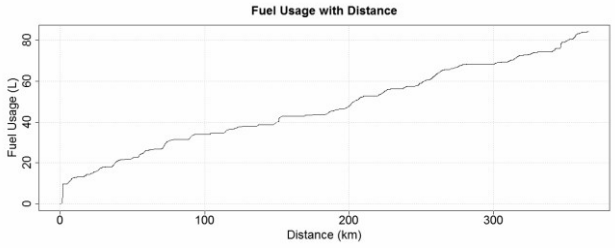

Fig. 2. Fuel consumption with distance.

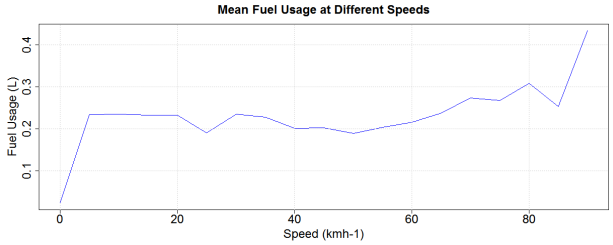

Fig. 3. Mean fuel consumption at each speed.

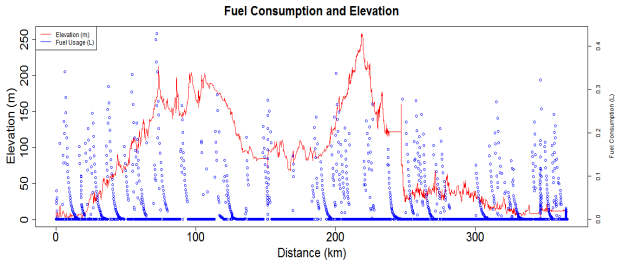

Fig. 4. Fuel consumption vairation with elevation.

with the fuel consumption, a linear forecasting model such as linear regression model would not be appropriate.

III. Machine Learning Techniques for Prediction

There are different ML techniques to address complex classification and regression problems. Among these classification and regression models, "ensemble learning" methods have gained greater 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 [7]. Two well-known ensemble-learning methods are boosting and bagging [8]. In *boosting*, successive models give extra weight to training instances that incorrectly predicted/classified by previous models. While making the prediction/classification a weighted vote is considered. Whereas in *bagging*, successive models are not dependent on earlier models, rather each model is independently constructed by a bootstrap sample of data. Prediction/classification is developed based on a simple majority vote. In this comparative study, we compare performance of two ensemble models Random Forest and Gradient Boosting and another ML technique Artificial Neural Network.

A. Random Forest

Random Forest (RF) proposed by Breiman, is an ensemble predictive model based on a collection of decision/regression trees [6]. Instead of making the prediction based on one tree, it depends on a collection of trees to take the decision. Being different from other bagging techniques, RF adds an additional layer of randomness to bagging. Similar to other bagging models RF also constructs each decision/regression tree using a bootstrap of sample data. However, the tree building procedure is different. Instead of splitting trees using the best split among all variables, in a RF each node is split using the best among a subset of predictors randomly chosen at that node [8]. This strategy enables RF to be robust against over fitting and be outstanding among many other classifiers including discriminant analysis, support vector machines, and neural networks [7]. Moreover, facilitating the estimation of variable importance and outlier detection are other benefits of this algorithm. Furthermore, RF is reasonably fast to obtain and can be easily parallelized [9]. A fine-tuned version of RF can be obtained by backward-elimination of predictors based on the given variable importance.

RF has been used in a myriad of domains to carryout predictions/classifications, and they are applicable for time series analysis as well. For example, Herrera el al. [9] used RF to forecast hourly urban water demand in a city in southeastern Spain. Chen et al. [10] used RF to forecast droughts, and demonstrated that particular prediction RF outperforms Autoregressive Integrated Moving Average (ARIMA).

B. Gradient Boosting

Gradient Boosting (GB) is another ensemble predictive boosting algorithm for regression and classification problems. It achieves optimal prediction by minimizing a loss function [11]. Similar to other boosting algorithms, GB builds the model in stages and generalizes them by allowing optimization of an arbitrary differentiable loss function. 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 [12]. Carrying out variable selection during the fitting process can be recognized as a key feature of GB [13]. Further GB algorithms provide prediction rules that have same interpretation as common statistical models. This becomes a major benefit of GB over other ML algorithms such as RF, which provides non-interpretable "black-box" predictions.

C. 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 a large number of inputs. ANNs can be used for nonlinear regression to realize complex relationships among variables. ANNs are commonly used in a myriad of domain such as medicine, transportation, and finance. Some examples include use of ANN to predict medical outcomes [14], model stock performance [15], and analyze Diesel engine performance and exhaust emission [16].

Following advantages of ANN were considered while selecting it as one of the predictive model in this study. ANN requires less formal statistical training. Further ANN is able to implicitly detect complex nonlinear relationships between explanatory variables and response variables. Ability to detect all possible interactions between independent and dependent variables is also an advantage [14].

IV. Fuel Consumption Prediction

In this section we describe the prediction task carried out to compare a set of alternative models for predicting the fuel consumption of the selected dataset. We evaluated the

appropriateness of RF, GB, and ANN algorithms for predicting fuel consumption.

In this prediction the target variable is fuel consumption of the bus within a given time interval. The predictor variables to forecast the fuel consumption were selected based on the exploratory analysis described in the section II and the context knowledge. For this particular case, we used distance, speed, longitude, latitude, elevation, and day of week. Further, with the intention of enhancing the accuracy of candidate models used, some predictor variables were dropped / added based on empirical experience; adding/ dropping of some variables increased accuracy. More formally, our goal is to approximate the unknown multiple regression function,

$$Fuel_{consum} = f(dist, speed, lat, long, elevation, day) \quad (1)$$

A. Prediction Using Random Forest

To evaluate the RF algorithm *random forest* package in R [9] 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 finding 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 seen on Fig. 5. Based on this important variables or most relevant predictors were identified. In addition, 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 all of these fine-tunes such that the accuracy of prediction enhanced. Fig. 6 shows the predicted and observed instantaneous fuel consumption and Fig. 7 shows the cumulative values of observed and predicted fuel usages.

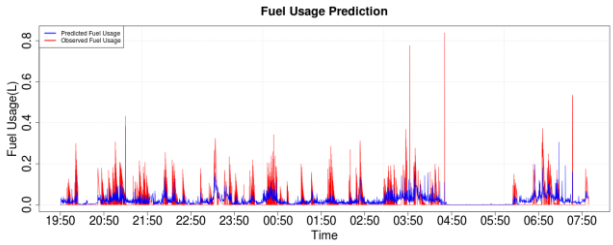

Fig. 6. Predicted and observed instantenous fuel consumption using RF.

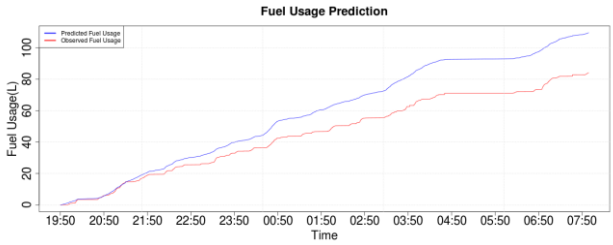

Fig. 7. Predicted and observed commulative fuel consumption using RF.

B. Prediction Using 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-

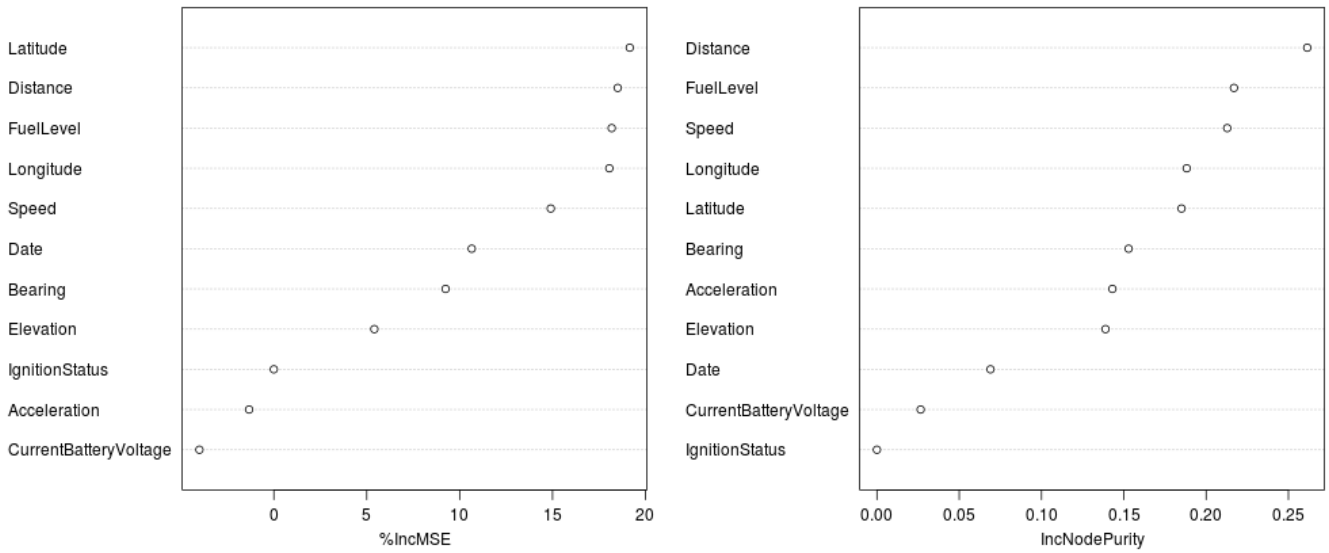

Fig. 5. Predicted and observed commulative fuel consumption using random forest.

dimensional data. In this study we used GAM to predict the fuel consumption as the relationship between given predictors and fuel usage is non-linear. GAM can be accessed from *gamboost* function 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 [6]. Therefore, based on the exploratory analysis in the section II, the model given for *gamboost* function was derived as follows:

$$Fuel_{consum} = bols(Long) + bbs(Lat) + bbs(Speed) + bols(Acc) + bols(Elev.Change) + bols(Dis) + bols(Day) \quad (2)$$

Fig. 8 and 9 illustrate the predicted and observed instantaneous and cumulative fuel consumption, respectively.

C. Prediction Using Neural Networks

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 [14]. Theoretically, *nuralnet* can handle an arbitrary number of predictors and responses, as well as hidden layers and hidden neurons [14]. In this analysis, we used a neural network 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. The graph of the trained neural network including trained synaptic weights and basic information about the training process can be found in Fig. 10. Overall error obtained for this network was 497.86, which is an indication that neural networks are not performing well in for this particular dataset. Fig. 11 and 12 depict the predicted and observed instantaneous and cumulative fuel consumption.

V. Evaluation of Predictive Accuracy

To assess the accuracy of each predictive model, its efficiency and error statistics were analyzed. Efficiency measure can be carried out using different methods. We use Nash-Sutcliffe efficiency coefficient to measure the predictive power of each model, which is defined as follows [17]:

$$Nash\text{-}Sutcliffe\ efficiency = 1 - \frac{\sum_{i=1}^{n}(EST_i - OBS_i)^2}{\sum_{i=1}^{n}(OBS_i - \overline{OBS})^2} \quad (3)$$

where EST_i and OBS_i denote the i-th estimated and observed fuel consumption values. 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:

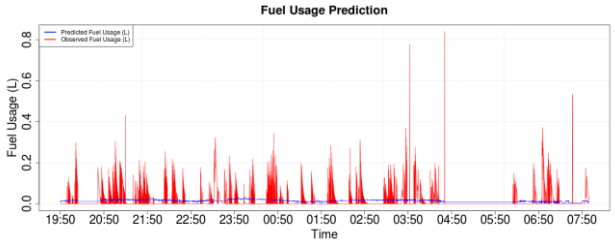

Fig. 8. Predicted and observed instantenous fuel consumption using GB.

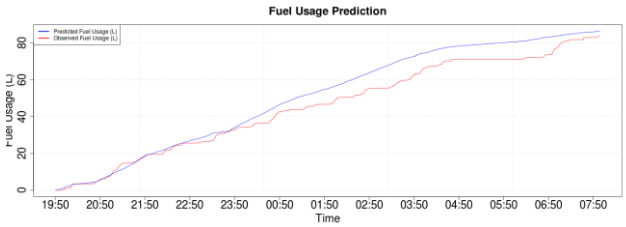

Fig. 9. Predicted and observed commulative fuel consumption using GB.

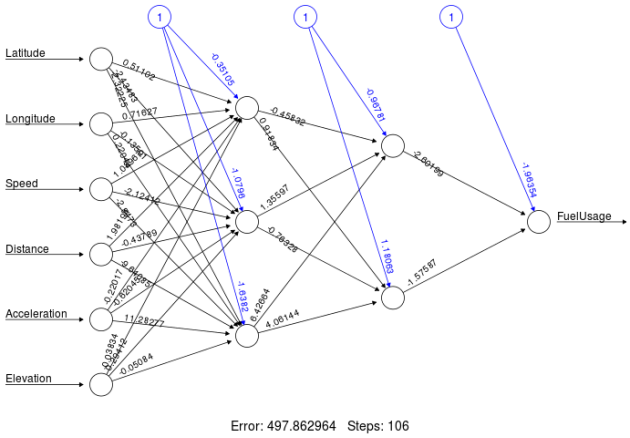

Fig. 10. Corrosponding neural network diagram.

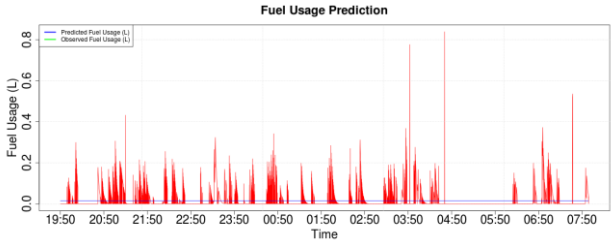

Fig. 11. Predicted and observed instantenous fuel consumption using ANN

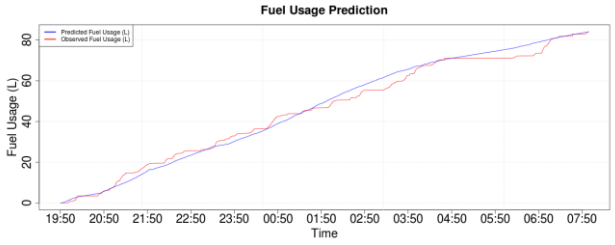

Fig. 12. Predicted and observed commulative fuel consumption using ANN.

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(EST_i - OBS_i) \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|EST_i - OBS_i\right| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(EST_i - OBS_i\right)^2} \quad (6)$$

As seen in Table I RF-based prediction produces the largest Nash-Sutcliffe efficiency coefficient. RF has the least difference between predicted and actual fuel consumption figures. It indicates that RF is more accurate in predicting the fuel consumption than the other two models. While both GB and ANN have negative values for Nash-Sutcliffe Efficiency measure, RF has a positive measure. That means the developed RF model is well explaining the relationship among predictors and response [17]. The prediction graphs in Fig. 6 and 7 also verify this. While the RF captures the trend more accurately, GB and ANN are only predicting the fuel consumption in a conservative manner. Moreover, RF also has the overall lowest error statistics. Therefore, one can conclude that a RF-based predictive model can be used to predict the fuel consumption with given parameters. This kind of a prediction can be used to detect fraudulent activities by drivers of fleet vehicles.

TABLE I. NUSH SUTCLIFFE EFFICIENCY.

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

TABLE II. ERROR STATISTICS OF THREE TEHNIQUES.

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

VI. SUMMARY

We evaluate the predictive ability of three machine-learning prediction models in predicting the fuel consumption of a long-distance public bus. While the selected dataset has several parameters that directly influence the fuel consumption, several other important parameters such as load, engine RPM, and traffic are not available. Even in the absence of such key parameters, we demonstrated that the RF model could predict the fuel consumption more accurately while capturing the trends in data. Such a model is useful detection of fuel fraud where the actual consumption of the vehicle can be compared against the predicted value based on other parameters like distance, location, elevation, speed, and day of the week. As future work we plan to integrate additional data influencing the fuel consumption such as traffic, weather, and load of the bus to enhance the predictive ability even further. We are also developing a module to guide the process reengineering steps to reduce the fuel consumption via improved fleet scheduling and better driving habits.

ACKNOWLEDGMENT

Authors are grateful for Nimbus Venture (Pvt) Ltd. for proving the dataset for the analysis.

REFERENCES

[1] L. Breiman, "Statistical Modeling: The Two Cultures," *Statistical Science,* vol. 16, pp. 199-231, 2001.

[2] A. Viswanathan, "Data driven analysis of usage and driving parameters that affect fuel consumption of heavy vehicles," Master's thesis, Linköping University, Sweden, 2013.

[3] J. Lindberg, "Fuel consumption prediction for heavy vehicles using machine learning on log data," Master's thesis, KTH, School of Computer Science and Communications (CSC), 2014.

[4] J. Gondar, M. Earleywine, and W. Sparks "Analyzing vehicle fuel saving opportunities through intelligent driver feedback," *SAE World Congr.*, Detroit, Michigan, 2012.

[5] J. S. Stichter, "Investigation of vehicle and driver aggressivity and relation to fuel economy testing," Master's thesis, University of Iowa, Iowa, 2012.

[6] I. M. Berry, "The effects of driving style and vehicle performance on the real-world fuel consumption of U.S. light-duty vehicles," Master's thesis, Massachusetts Institute of Technology, Cambridge, 2010.

[7] L. Rokach, (2009, Nov, 19), *Ensemble-based classifiers,* [Online], Available: http://www.ise.bgu.ac.il/faculty/liorr/AI.pdf

[8] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, Dec. 2002.

[9] 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, pp. 141-150, June 2010.

[10] 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, Sep. 2012.

[11] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, vol. 29, no. 5, pp. 1189-1232, 2001.

[12] B. Hofner, A. Mayr, N. Robinzonov, and M. Schmid, "Model-based boosting in R: A hands-on tutorial using the R package mboost," Comput. Stat., vol. 29, no. 1-2, pp. 3-35, Dec. 2012.

[13] P. Bühlmann and B. Yu, "Boosting with the L 2 loss: Regression and classification," *Journal of American Statistical Association*, vol. 98, no. 462, pp. 324-339, 2003.

[14] V. Jack, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *Journal of Clinical Epidemiology,*vol. 49, pp. 1225-1231, 1996.

[15] A. N. Refenes, A. Zapranis, and G. Francis, "Stock performance modeling using neural networks: A comparative study with regression models," *Neural Network,* vol. 7, pp. 375-388, 1994.

[16] B. Ghobadian et al., "Diesel engine performance and exhaust emission analysis using waste cooking biodiesel fuel with an artificial neural network," *Journal of Clinical Epidemiology,* vol. 34, pp. 976-982, 2009.

[17] J.E. Nash, J.V Sutcliffe, "River flow forecasting through conceptual models: Part 1- A discussion of principles", *Journal of Hydrology,* vol. 10, pp. 282-290, 197