Performance of Motor Vehicle based on Driving and Vehicle Data using Machine Learning

MSc Research Project
Data Analytics

Punith Kumar Nagaraje Gowda
Student ID: x18130771

School of Computing
National College of Ireland

Supervisor: Dr. Cristina Muntean

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Punith Kumar Nagaraje Gowda
Student ID:	x18130771
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Cristina Muntean
Submission Due Date:	12/12/2019
Project Title:	Performance of Motor Vehicle based on Driving and Vehicle Data using Machine Learning
Word Count:	6442
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	29th January 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Performance of Motor Vehicle based on Driving and Vehicle Data using Machine Learning

Punith Kumar Nagaraje Gowda
x18130771

Abstract

With the increasing population demographics and the dependency of man on motor vehicles as the primary source of transportation, the number of motor vehicles being registered for commercial as well as non-commercial activities on a daily basis is massive and yet continues to increase at an alarming rate. This has a direct and an unambiguous effect on the amount of fossil fuels being utilized globally and its subsequent environmental effects, which is of great concern in the present situation. Several attempts from various research sectors are ongoing in order to overcome this global issue and promising results are expected. This project is one such attempt at identifying the performance of small passenger cars in terms of fuel efficiency and map them with factors affecting it using machine learning techniques. The commencing activity while carrying out any such research activity will be the identification of the problem and all its possible sources. In this case, two potential sources can be identified and they are; the vehicle characteristics and the driver/driving behaviour. The relevant data for this analysis was taken from the public source, Kaggle which is the data collected from the OBD of the car and models are built using techniques like Multiple Linear Regression, XGBoost, Support Vector Machine and Artificial Neural Network and their performance is compared to discover the first-rate technique in predicting the fuel efficiency and to propose the optimum driving behaviour in terms of throttle position to achieve better fuel efficiency. The results reveal that XGBoost model outperforms all other models developed in predicting the fuel efficiency for the different split ratios evaluated and comparing the throttle position with the predicted fuel efficiency explains that to achieve better fuel efficiency the throttle position must be around 70 to 80 on a scale of 100, referred to as full throttle position. The knowledge discovered from the research could be used by car manufacturers to design cars in future to mitigate the fuel consumption.

1 Introduction

Since the advent of the industrial revolution, transportation facilities has played a vital role in all areas of livelihood such as travel, trade and exchange. Despite the invention of multiple transportation modes, roadways transportation is the most commonly preferred course by people to carry out day to day activities. Irrespective of the type of activity being carried out i.e., commercial or non-commercial purpose, roadways take up a major share in the global transportation statistics. Although the growth of the automobile industry has contributed to the luxury of commuting to the communities and boosted economic growth, it certainly has had unfavourable effect on environment (Predić et al.

(2016)). The result of using fossil fuels as the primary source for the operation of these motor vehicles is already well known to man and extensive efforts are being carried out to at least subside them. Global temperature rise increased anthropogenic CO_2 emissions, depletion of fossil fuels and hazardous health risks are most concerning.

Several legislative principles and policies are being drafted proactively by the government to help mitigate the emission of CO_2 emissions. Here it becomes important to understand that the levels of CO_2 emission by a particular motor vehicle is parallel to the amount of fuel being consumed by the same motor vehicle during operation. This in turn corresponds to the performance characteristics and the efficiency of the vehicle. Periodic maintenance of these motor vehicles becomes very essential in maintaining its fuel efficiency. Yet, with the assistance of the new and updated developments in technology, monitoring of the fuel consumption of these motor vehicles can be achieved easily today. Several researchers have attempted to analyse the data in order to figure out the factors affecting the fuel efficiencies of these motor vehicle, which led to the formulation of another source of cause; the driving/driver behaviour (Çapraz et al. (2016)).

The purpose of the study is to address the research question "How well the machine learning technique XGBoost predict the fuel efficiency of a car by considering both vehicle characteristics and driving data like speed, throttle position, air intake temperature and pressure?"

The modern vehicles comes equipped with many sensors offering lot of data to be analysed to make maximum use of it and contains information about engine characteristics as well as the driving behaviour. The information about vehicle characteristics include mass air flow, manifold absolute pressure, air intake temperature and pressure and the driving data includes the throttle position. Machine learning techniques like Multiple Linear Regression, Artificial Neural Network, Support Vector Machine and XGBoost are implemented and models are developed to predict the fuel efficiency of small passenger car. Speed and engine RPM are also considered to build the model.

Objectives of this research are -

- Analyse the correlation between fuel efficiency of a car and its characteristics and driving behaviour.
- Develop machine learning models using Multiple Linear Regression, Support Vector Machine, Artificial Neural Network and XGBoost considering vehicle characteristics and driving behaviour as input data to predict the fuel efficiency of a small passenger car.
- Propose the optimum throttle position and other characteristics that would help in achieving better fuel efficiency and thereby reduce fuel consumption and emissions.

This document is structured as follows: Section 2 Critique on the literature highlighting the purpose, method and the limitations, section 3 Methodology this research follows and the steps involved, section 4 Design Specification and the architecture of the techniques used in this research, section 5 Implementation details of the techniques used to develop models, section 6 Results/Evaluation of the models developed, section 7 Conclusion and future scope, Acknowledgement and References.

2 Related Work

Vehicle's fuel consumption is influenced by external and internal factors. Road conditions, weather and traffic are considered as the external factors and vehicle characteristics, driving behavior and load are considered to be the internal factors. Wickramanayake and Bandara (2016) in their work predicted the fuel consumption of fleet vehicles using Machine Learning techniques like Random Forest, Gradient Boosting and Artificial Neural Network. The dataset consisted of a number of parameters of which few are speed, fuel level, fuel consumption and acceleration, the data was collected from a public bus in Sri Lanka. Random Forest outperformed the other two models built. Several factors which directly influences the fuel consumption was considered in the study but other main factors like the engine RPM, traffic conditions and load were not considered in the study.

Although engine and drive technology, vehicle type and condition influences vehicle's fuel consumption, personal driving style is an eminent factor and change in the style can minimize the fuel consumption. Thitipatanapong and Luangnarutai (2011) conducted a study in Thailand to analyse the relationship between driving behaviour and fuel economy of a car. Based on the acceleration, driving behaviour was classified as 'moderate', 'aggressive' and 'claim'. The study concluded that fuel consumption rate and vehicle driver index (VDI), measure of driving behaviour, were deeply related and hence VDI can be used to estimate driving behaviour. Driver is said to be aggressive when there is a higher VDI resulting in lower fuel economy.

2.1 Influence of vehicle characteristics on performance

Çapraz et al. (2016) studied the fuel consumption of automobiles using the real-time data of three vehicles driven by different drivers on three different routes in Turkey. The data was collected in real-time from the vehicles using the on-board diagnostics (OBD2), a smartphone and a bluetooth interface. Multiple Linear Regression, Artificial Neural Networks and Support Vector Machine models were developed and compared to analyse the data. Results revealed that the Support Vector Machine model outperformed the other two models under the two tests conducted where in one test only speed, acceleration and slope were considered and in the other test all the variables where considered. The broader goal of the study is to take advantage of the road conditions to predict fuel consumption for long distance.

Perrotta et al. (2018) also conducted a similar study by applying Machine Learning techniques to model fuel consumption of trucks. Support Vector Machine (SVM), Random Forest (RF) and Artificial Neural Network (ANN) models were developed and compared for this purpose. Telematic data was collected for the vehicle characteristics and data from the Highways Agency Pavement Management System (HAPMS) of Highways England was collected for road characteristics. The study shows that all three models had good precision, but RF slightly outperformed the other two models. As a future work driver/driving behaviour and air temperature could also be considered to predict the fuel consumption. Similarly Yin et al. (2015) modelled and predicted the fuel efficiency based on an informative vehicle database for common automobiles. Mutual Information Index (MII) was employed to determine the characteristics which influences the fuel efficiency. Five machine learning models were developed like the Quantile Regression (QR), Support Vector Regression (SVR), Partial Least Squares (PLS), Ordinary

Least Squares (OLS) and Gaussian Process Regression (GPR). QR was found to be the better technique than other adopted techniques.

Vehicle manufacturing includes a number of factors to be considered to provide the best vehicle with better safety, fuel efficiency and environmental performance. A lot of research and development goes before coming up with the final specification of the vehicle. Torrao et al. (2016) designed and developed a tool which produces a Safety, Efficiency and Green (SEG) score to evaluate the vehicle performance not only for policy makers but also for the public. This score is the combined score of safety of a car, fuel efficiency and green emissions of a car. The crash data was collected in Oporto, one of the highest crash severity districts in Portugal. This data also included the vehicle characteristics required for calculating the score for fuel efficiency and green emission. The prediction models were developed using the advanced logistic regression models. The results revealed that cars with lesser engine size had better fuel efficiency while cars with bigger engine size had better safety.

Byerly et al. (2019) developed a machine learning model using Artificial Neural Network to predict the average fuel consumption for heavy vehicles. The type of Artificial Neural Network used is the Feed Forward Neural Network and the model is evaluated using Root Mean Squared Value, Mean Absolute Error and Mean Absolute Percentage Error. The average fuel consumption is predicted considering the distance travelled rather than traditional time period. Lee and Choi (2016) conducted a similar research by evaluating vehicle survival patterns and fuel efficiency to predict the total energy consumption. Vehicle characteristics like fuel type, model year, distance travelled every year, size of vehicle were studied to predict the energy consumption and analyse the efficiency.

Kanarachos et al. (2019) focused on building a model that predicts the instantaneous fuel consumption rather than average consumption that most of the researches are based on. Two different type of Recurrent Neural Networks were explored, using the smartphone's speed, position, acceleration and altitude to process and build the model. Comparison between the LSTM and RNN revealed that the later one is better in predicting instantaneous fuel consumption.

For more than a decade now constant effort is witnessed to reduce the energy consumption. Munyon et al. (2018) investigates whether the Jevon's paradox remains good when all the factors are controlled which are available that could affect the consumption of efficient goods or services. The general assumption is that, ceteris paribus, increase in the energy efficiency leads to decrease in the consumption of goods or services rendered efficiently which is opposed by Jevon's paradox. Regression analysis was carried out on the data collected and it was found that with 1% increase in the fuel efficiency there was 1.2% increase in the vehicle miles travelled thus proving the Jevon's paradox true.

The limitation observed in the above literature survey is that the main concentration was utilising the vehicle characteristics to build a model that predicts the fuel consumption either average consumption or on-the-go consumption, and the vehicle characteristics considered are the external characteristics like the weight, power and speed, not the sensor data of the engine or other factors like driving style or road condition combined with vehicle characteristics.

2.2 Influence of driving behaviour on performance

Along with the vehicle characteristics, driver/driving behavior plays a massive role in determining vehicle fuel efficiency and safety. Fugiglando et al. (2019) proposed a new methodology for classification of driver and near-real-time analysis using the selected subset of CAN bus signals. The sensors on the modern vehicles provide a huge amount of information about control and safety to the drivers and can be used for analyzing fuel consumption and emissions. Unsupervised learning techniques were used to cluster the drivers among different groups. Vehicle characteristics or the road conditions were not considered in the study and more sophisticated concepts like "nervousness" and "aggressiveness" could be fully characterized which helps in better analysis of driving/driver behaviour.

Further to understand the impact of driver behaviour on fuel consumption Ping et al. (2019) proposed two machine learning techniques, one was the unsupervised spectral clustering technique to study the relation between driving behaviour and fuel consumption in the first stage and to model the relation between them including the driving environment in the second stage. Second technique was the Long Short-Term Memory (LSTM) used to predict the fuel efficiency for short-term while the spectral clustering method was used to predict for long-term. The proposed techniques showed noteworthy relationship and hence better prediction of fuel consumption which can be used in Advanced Driver Assistance Systems (ADAS). Vehicle characteristics and group personality feature could be considered in future to make more generic fuel consumption model. Similarly Xu et al. (2018) proposed two approaches to predict the fuel consumption of trucks and to model the relationship between fuel consumption and driving behaviour using Internet of Vehicles data. The first approach was the energy consumption index and the other approach was the Generalized Neural Network model to establish the relation between driver behaviour and fuel consumption. The two proposed models were compared with the three existing state-of-the-art models like the Vehicle Specific Power (VSP), Virginia Tech microscopic (VT-Micro) model and Comprehensive Modal Emission Model (CMEM) and found that the proposed models had excelling performance in predicting fuel consumption.

Gilman et al. (2015) proposed and developed a prototype called Driving coach, a driver assistance system architecture for fuel efficient driving. Three models were developed Linear Regression model, Decision Tree model and Neural Network model to solve the regression task. The data is collected from the city of Oulu, Finland. The relearned models performed better, and this prototype provides in-depth information for the driver regarding the previous trip and the system audits its performance and provides feedback to the driver to make better prediction and learn itself better in future to serve the driver.

Artificial Neural Network model to predict passenger car fuel consumption was developed by Predić et al. (2016) a case study in NIŠ city. The investigation revealed that the Neural Network with a single hidden layer and ten neurons can be trained well with Levenberg-Marquardt algorithm to predict the car fuel consumption. The model is not restricted to one particular car type but can be applied to other similar compact car types but can expect an approximate prediction.

In this section of the literature survey the driving style is considered mainly for building the model and predicting the fuel consumption and other factors which highly influence the fuel consumption is not considered.

From the review of the researches done in this field it is noticed that the focus is mainly on one factor and the fuel efficiency or the consumption is analysed, but it is important to address the factors which hugely influences the fuel consumption, hence this research tries to use both driving behaviour and vehicle characteristics (sensor data) together to predict the fuel consumption. To have a better understanding of fuel consumption of a vehicle and to predict fuel efficiency it is necessary to consider both driving behaviour and vehicle characteristics data.

3 Methodology

The Knowledge Discovery in Databases (KDD) methodology is used for the research and analysis of huge data collected by instruments given by Fayyad et al. (1996) and points out that KDD applications in areas like business, finance are difficult than in the field of science in general because science end-users know their data in detail.

It consists of the following steps:

- Data Selection
- Data Pre-processing
- Data Transformation
- Data Mining
- Evaluation/Interpretation

The steps of KDD are shown in Figure 1 with respect to the research and are explained in detail below

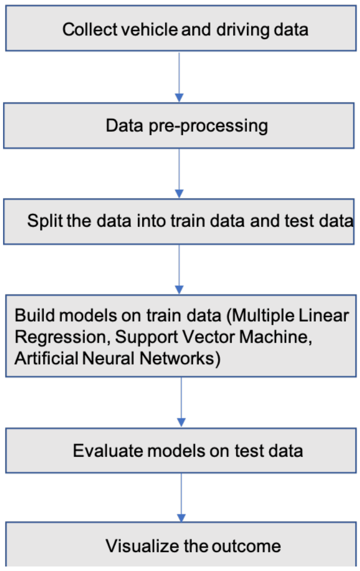
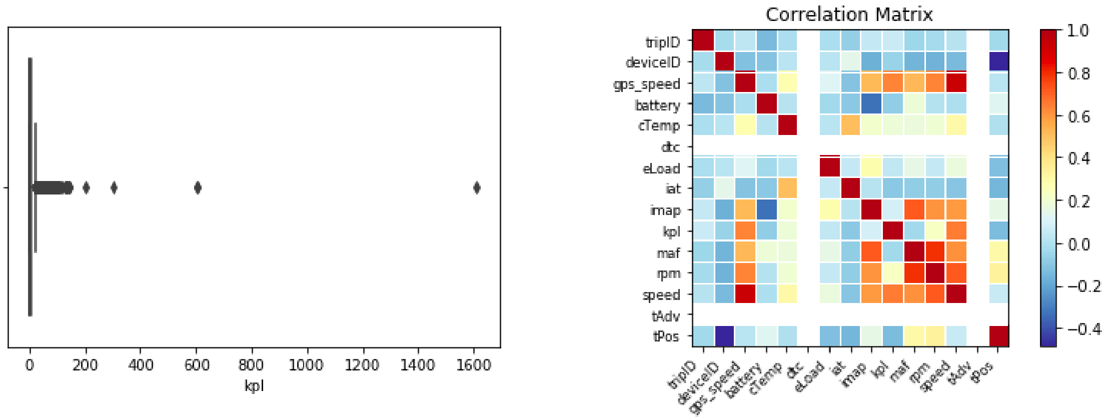

Collect vehicle and driving data
Data pre-processing
Split the data into train data and test data
Build models on train data (Multiple Linear Regression, Support Vector Machine, Artificial Neural Networks)
Evaluate models on test data
Visualize the outcome

Figure 1: KDD steps followed for the research

3.1 Data Selection

The research required dataset with both vehicle characteristics and driving data. All modern vehicles are well equipped with many sensors which gives lot of data, but the data lies with the owner of the vehicle or with the company of the vehicle. One such



Boxplot

Correlation Matrix

Figure 2: Data Pre-processing

dataset is publicly available on Kaggle[1] which has 17 features, some of them are speed, engine RPM, mass air flow rate, intake air temperature, kpl (fuel efficiency in kilometre per litre) and around 3 million records. The dataset contains the telematic data of the vehicle. The air intake temperature and pressure, battery temperature, load on the engine, vehicle speed and engine RPM are the data for vehicle characteristics and the throttle position is considered for the driving data.

3.2 Data Pre-processing

The dataset consists of 17 features of which 'kpl' is the dependent variable indicating the fuel efficiency in kilometre per litre. Null check was run, the outliers were detected using the boxplot shown in Figure 2 and were removed, resulting in the sample size to be around 1.5 million. To analyse the relationship between fuel efficiency and vehicle characteristics and driving behaviour, which is one of our objectives, Spearman correlation was applied and found that speed of the vehicle highly influenced the fuel efficiency followed by factors like the throttle position, air intake temperature and pressure, engine RPM and others. The correlation matrix is shown in Figure 2. The remaining factors like deviceID,tripID were not considered during the development of models. The models were built and validated for the split 70/30, 75/25 and 80/20 ratio.

3.3 Data Transformation

As major transformation was not required for the data, normalization and reshape are the only transformation techniques used to make sure the model developed does not overfit/underfit and predicts well on the test set. Normalization and reshape is applied for Multiple Linear Regression model, Artificial Neural Network model (Multilayer Perceptron and LSTM) and Support Vector Regression model. No normalization was done on XGBoost as trees do not require normalization of data.

[1] `https://www.kaggle.com/yunlevin/levin-vehicle-telematics#v2.csv`

3.4 Data Mining

The models built to predict the fuel efficiency are Multiple Linear Regression, Support Vector Machine, XGBoost and Artificial Neural Network. The models were compared to come up with the best which predicts fuel efficiency accurately. The parameters for these techniques was chosen after a series of tests for few algorithms and GridSearch was also performed to come up with best parameters for building the model. Based on the literature review XGBoost is yet to be implemented and see if there is any better result compared to the existing models. All the models developed are validated for the split 70/30, 75/25 and 80/20 ratio.

3.5 Interpretation/Evaluation

The evaluation metrics used to analyse the models built are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R^2.

- **Root Mean Squared Error (RMSE)** – One of the standard evaluation metrics used in regression analysis is Root Mean Squared Error (RMSE). RMSE is the squared root of Mean Squared Error (MSE), MSE is defined as the square of the distance between the expected value and predicted value for the sample data. The formula to calculate RMSE[2] is shown in Equation (1)

  $$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - p_i)^2} \tag{1}$$

  where p represents the i$^{\text{th}}$ predicted value , y represents the i$^{\text{th}}$ actual value and n represents the number of sample data. Lesser the value of RMSE, explains that the model built is better as the error in prediction will be minimum.

- **Mean Absolute Error (MAE)** – Mean Absolute Error is the average of the absolute difference between the actual value and the predicted value. MAE does not consider the direction of error and measures the average of magnitude error in prediction. The formula to calculate MAE[3] is given by the Equation (2)

  $$MAE = \frac{1}{n}\Sigma_{i=1}^{n}|y_i - p_i| \tag{2}$$

  where n is the sample size, y represents the i$^{\text{th}}$ actual value and p represents the i$^{\text{th}}$ predicted value.

- **R^2** – R^2 is a measure to indicate how near the data values are to the fitted regression line. The higher the R^2 better the model fits the data. The accuracy of the model is indicated by R^2 and is the derivative of Relative Squared Error (RSE)[4]. The formula to calculate RSE[5] is given by Equation (3)

[2] https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

[3] https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

[4] https://www.saedsayad.com/model_evaluation_r.htm

[5] https://www.saedsayad.com/model_evaluation_r.htm

$$RSE = \frac{\Sigma_{i=1}^{n}(y_i - p_i)^2}{\Sigma_{i=1}^{n}(y_i - \mu)^2} \tag{3}$$

where y is the i^{th} actual value, p is the i^{th} predicted value, μis the mean value and n is the number of sample data. R^2 is given by Equation (4)

$$R^2 = 1 - RSE \tag{4}$$

The accuracy measure is given by R^2 and not the error measure.

4 Design Specification

As shown in Figure 3 the 2-tier design is adopted as the dataset is not created explicitly for this research, a publicly available dataset is used and is in structured format. The first and major step for any research is data cleaning and pre-processing without which the results obtained might be inappropriate. Python is used for cleaning and also to develop models. In the computational layer or the business logic tier the data is cleaned, pre-processed, transformed and then the regression models are built using techniques like Multiple Linear Regression, SVM, XGBoost and ANN, the models built are evaluated and in the presentation layer using relevant software and tools the results are visualized. Using the matplotlib python library the results are visualized in the presentation layer or client tier.

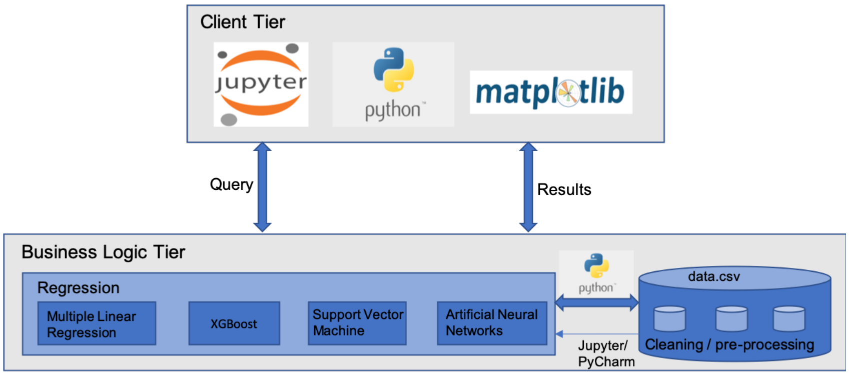


Figure 3: 2-tier design architecture

The architecture of the techniques used for this research are described below:

4.1 Multiple Linear Regression

Linear Regression assumes a linear relationship between predictor (independent) variable and response (dependent) variable, the case of one independent variable is referred to as simple linear regression and more than one independent variable is multiple linear regression (Freedman (2009)) and is the first type of technique to be studied and analysed critically, also used in prediction models extensively. Equation (5) gives the mathematical intuition of multiple linear regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{5}$$

where y represents the dependent or response variable, (x1. . . .xn) are predictor or the independent variables and (β1...βn) represent their coefficients and β0 is the constant.

4.2 Artificial Neural Network

Inspired by the biological neural networks that empowers a brain, Artificial Neural Networks were designed. It is based on a group of nodes or connected units called artificial neurons. Similar to the biological brain an artificial brain receives the signal, processes it and sends it to other connected artificial neurons. The input layer is the first layer with n-dimensional input vector, hidden layer is the middle layer and output layer is the last layer with output vector y. Along the connection between the layers there is a connection weight, the output of any layer is the function of connection weight and input neuron. In the final output layer the weighted sum of its inputs is taken and non-linear function is applied, the result of this is the output of the neural network[6]. The commonly used basic ANN architecture is the multilayer perceptron. Another class of artificial neural network is the Recurrent Neural Network (RNN) which uses the internal state, also known as the memory, to handle the sequence of inputs. Long-Short Term Memory (LSTM) is based on the RNN architecture which is well suited for making predictions with sequence to sequence data. A generic LSTM unit consists of a cell, input gate, output gate and a forget gate, the three gates control the flow of information into and out of the cell while the cell holds the values. The forget gate regulates to what extent the value remains in the cell.

4.3 Support Vector Regression

Support Vector Machines or the Support Vector Networks (Cortes and Vapnik (1995)) belong to the supervised learning method in machine learning that evaluate data for regression and classification analysis. SVM divides the data into classes and tries to find a line or a hyperplane between the data of the classes. The data points closest to the hyperplane are called support vectors, the distance is called the margin and the goal is to maximize the margin. Along with the linear classification SVMs can methodically perform non-linear classification using the kernel tricks. Some of the kernel functions are linear, polynomial, radial basis and sigmoid. Non-linear kernel functions can be used when the dataset is less than 50000 records.

4.4 XGBoost

XGBoost is an ensemble Machine Learning algorithm based on decision trees that uses Gradient Boosting architecture. Boosting refers to collection of algorithms that make use of weighted averages to make a strong learner from weak learners. Through parallel processing, tree pruning and regularization to avoid overfitting/bias, optimized gradient boosting algorithm was developed which is Extreme Gradient Boosting (XGBoost)[7]. XG-

[6] https://www.researchgate.net/figure/Artificial-neural-network-ANN-architecture-ANNs-consist-of-artificial-neurons-Each_fig1_5411405

[7] https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Boost tackles one of the major incapability of gradient boost of potential loss in creating a new branch for all splits by considering the feature distribution in a leaf across all data points to reduce search space for feature splits.

5 Implementation

This section describes in detail the implementation of the research project of predicting the fuel efficiency using the vehicle characteristics like speed, engine RPM, air intake (temperature and pressure), load on the engine and driving data like the throttle position.

Python 3.7.3 was used with PyCharm and Jupyter notebook to implement the project. R is better for statistical analysis while Python is better for analysis and development of machine learning models because of the vast availability of the libraries, hence Python was chosen for implementation. The necessary packages for the project was downloaded and installed using the pip install command. The parameters used for each model is described below:

5.1 Multiple Linear Regression

The sklearn Linear Regression is used to implement the model. The parameters that are available in building Linear Regression model are fit_intercept, normalize, copy_X and n_jobs. The parameters fit_intercept expects a bool value based on which it decides whether to calculate the intercept for the model, normalize also expects a bool value, True indicates the regressors will be normalized before the regression and if set to False fit_intercept will be ignored, copy_X expects a bool value and if set to True X will be copied else will be overwritten and n_jobs indicates the number of jobs to use for the computation. The data is normalized before building the model hence all the parameters are ignored.

5.2 LinearSVR

The sklearn svm LinearSVR is used to implement the model. Since the sample data is huge LinearSVR is implemented, for non-linear kernels sample size must be less than 50000 records to perform well. The parameters tuned for better results in LinearSVR model are the tol, C and max_iter. GridSearchCV was tried on the model and the values passed for the parameter grid are tol = 1e-4, 1e-5 and C = 1, 10, 100 and max_iter = 500, 1000. The data is normalized before building the model. The values found to be optimum for the research are tol = 1e-4, C = 1 and max_iter = 500.

5.3 Multilayer Perceptron

The sklearn neural network MultiLayer Perceptron (MLP) Regressor is used to develop the model. The parameters like hidden_layer_sizes, activation, solver, learning_rate, max_iter, learning_rate_init and alpha were tuned and others the default is used. The data is normalized before building the model. After conducting a number of runs the optimum values found were hidden_layer_size = (100,50,50,50), activation = 'tanh', solver = 'adam', learning_rate = 'adaptive', max_iter = 2000, learning_rate_init = 0.0001 and alpha = 0.001.

5.4 Long-Short Term Memory

Keras is installed using the command pip install keras from which LSTM, Dense, Sequential and Dropout is imported to develop the model. The epochs and batch_size are the parameters which are tuned to get the best model. A series of tests was conducted with different values for epochs and batch_size, the optimum value of Root Mean Squared Value was observed with 4 hidden layers and 50 neurons in each layer with the dropout value of 0.2, adam optimizer, 15 epochs and 10000 batch_size.

5.5 XGBoost

The XGBoost is installed using the command pip install xgboost, XGBoost Regressor has three types of parameters namely general parameters, booster parameters and task parameters. GridSearch was run on the parameters to come up with the optimum values to develop the best model. The values returned by the GridSearch are alpha = 5, learning_rate = 0.1, nthread = 2, max_depth = 5 and colsample_bytree = 0.8. The n_estimators chosen was 500 and objective was set to 'reg:squarederror'.

The data is split to train and test in the ratio of 70:30, 75:25 and 80:20, all the models are run for each split and the results are analysed.

6 Evaluation

This section presents the evaluation of the results achieved towards the objectives of the research. The parameters chosen for the models built are after a series of tests and for few models parameters were chosen with the help of GridSearch. As discussed in the previous section the models were run for different split ratios like 70:30, 75:25 and 80:20. RMSE, MAE and R^2 is used as the evaluation metrics. Since the sample size is large, for the visualisation purpose average of the samples are taken. Results obtained from each model built is discussed below.

6.1 Case Study with the different train and test split ratio

The following results are obtained for the split ratio 70:30, 70% train data and 30% test data:

6.1.1 Multiple Linear Regression

Figure 4 represents the graph of actual fuel efficiency versus the fuel efficiency predicted of the Multiple Linear Regression model. The y-axis represents the fuel efficiency in kilometre per litre and x-axis represents the average of the recorded instances, due to high volume of data the average is considered for better visualisation.

Figure 5 shows the error of prediction in percentage, the difference between predicted fuel efficiency and actual fuel efficiency for the Multiple Linear Regression model. RMSE of 4.16853, MAE of 2.90606 and R^2 of 0.05250 is obtained for the Multiple Linear Regression model. At few points the error is above 100 percentage indicating that the model does not predict well when there is a sudden raise and drop in fuel efficiency with huge difference.

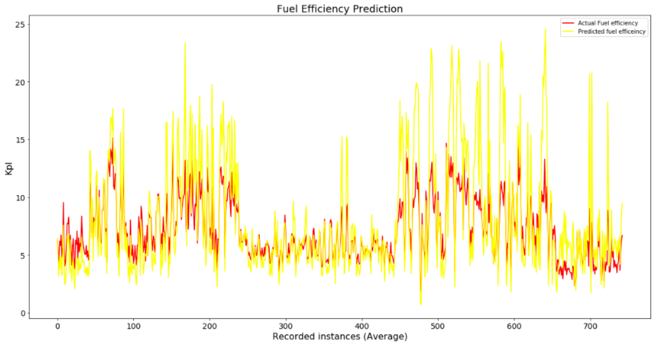


Figure 4: Fuel efficiency prediction of Multiple Linear Regression model

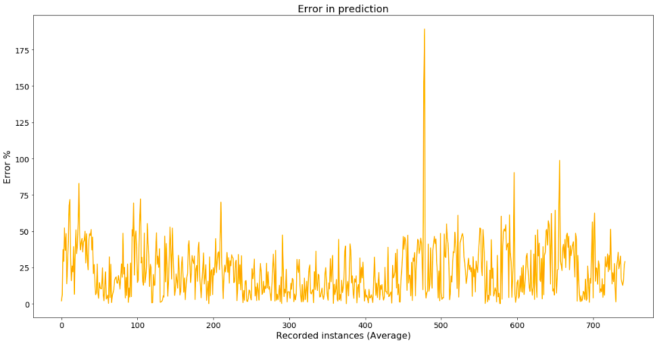


Figure 5: Error in prediction of Multiple Linear Regression model

6.1.2 LinearSVR

Actual fuel efficiency versus the fuel efficiency predicted of the Linear Support Vector Regression model is shown in Figure 6. The y-axis represents the fuel efficiency in kilometre per litre and x-axis represents the average of the recorded instances, due to high volume of data the average is considered for better visualisation.

Figure 7 shows the error of prediction in percentage, the difference between predicted fuel efficiency and actual fuel efficiency for the Linear Support Vector Regression model. RMSE of 4.28632, MAE of 3.31346 and R^2 of -0.00180 is obtained for the Linear Support Vector Regression model. At few points the error is above 200 percentage indicating that the model's performance is poor compared to the Multiple Linear Regression model. The RMSE and MAE values are also not impressive compared to the previous model.

6.1.3 Multilayer Perceptron

Figure 8 represents the graph of actual fuel efficiency versus the fuel efficiency predicted of the Multilayer Perceptron model. The y-axis represents the fuel efficiency in kilometre per litre and x-axis represents the average of the recorded instances, due to high volume of data the average is considered for better visualisation.

Figure 9 shows the error of prediction in percentage, the difference between predicted fuel efficiency and actual fuel efficiency for the Multilayer Perceptron model. RMSE of 2.76280, MAE of 2.18959 and R^2 of 0.58379 is obtained for the Multilayer Perceptron model. The error percentage is less than 100, so in comparison to the other two models

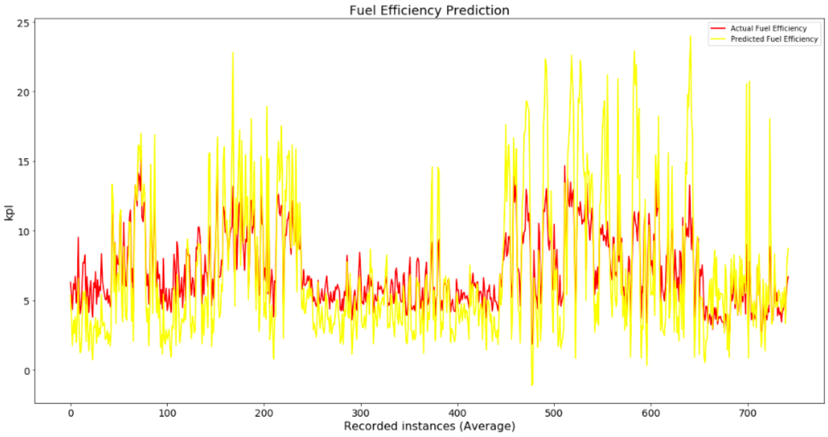


Figure 6: Fuel efficiency prediction of Linear Support Vector Regression model

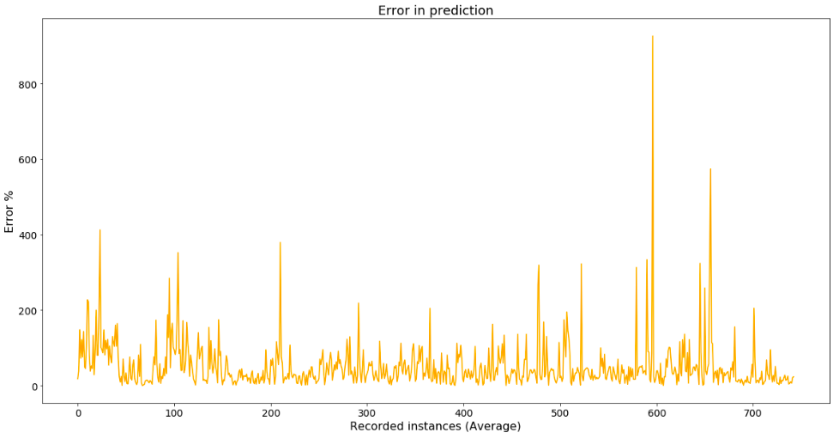


Figure 7: Error in prediction of Linear Support Vector Regression model

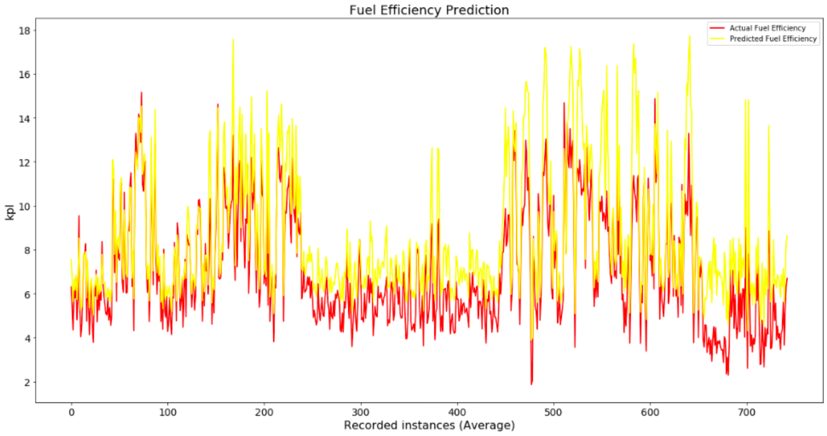


Figure 8: Fuel efficiency prediction of Multilayer Perceptron model

developed the Multilayer Perceptron is better in predicting the fuel consumption. The RMSE and MAE values are also impressive compared to the previous models.

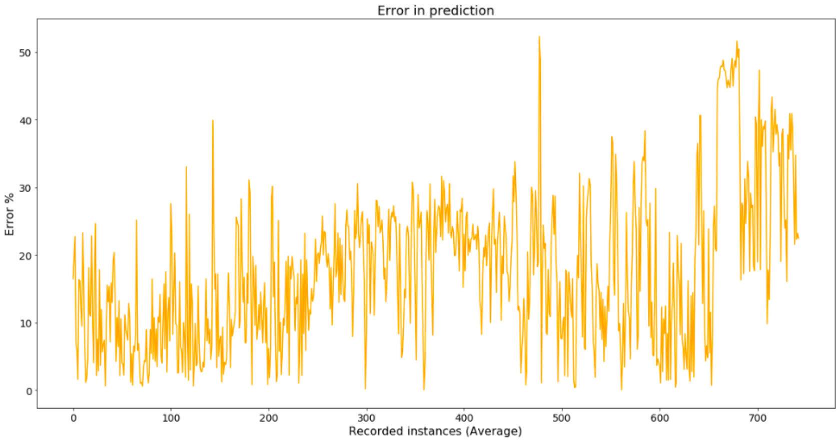


Figure 9: Error in prediction of Multilayer Perceptron model

6.1.4 Long-Short Term Memory

Actual fuel efficiency versus the fuel efficiency predicted of the Long-Short Term Memory model is shown in Figure 10. The y-axis represents the fuel efficiency in kilometre per litre and x-axis represents the average of the recorded instances, due to high volume of data the average is considered for better visualisation.

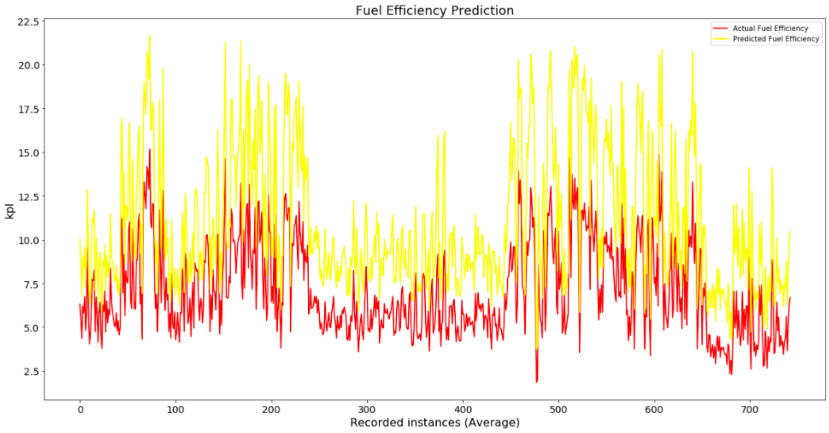


Figure 10: Fuel efficiency prediction of Long-Short Term Memory model

Figure 11 shows the error of prediction in percentage, the difference between predicted fuel efficiency and actual fuel efficiency for the Long-Short Term Memory model. RMSE of 4.36321, MAE of 3.85350 and R^2 of -0.03807 is obtained for the Long-Short Term Memory model. The error percentage is around 50 in this model, so in comparison to the other three models developed the Long-Short Term Memory model is better in terms of error of prediction in percentage but the RMSE and MAE values of the Multilayer Perceptron model is better than this model.

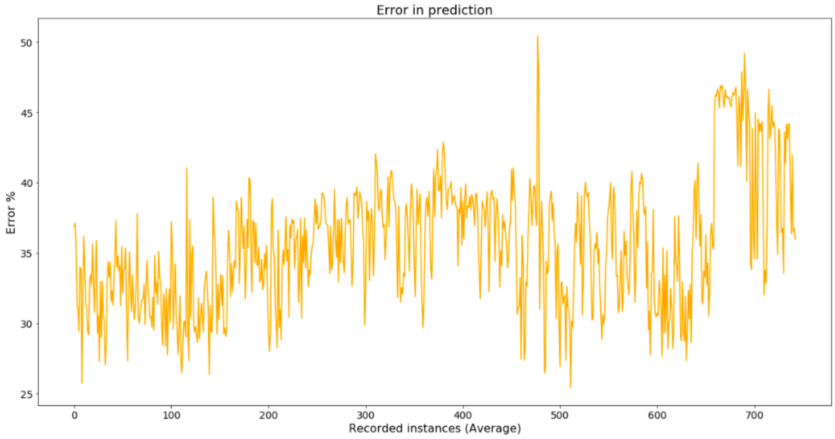


Figure 11: Error in prediction of Long-Short Term Memory model

6.1.5 XGBoost

Figure 12 represents the graph of actual fuel efficiency versus the fuel efficiency predicted of the XGBoost model. The y-axis represents the fuel efficiency in kilometre per litre and x-axis represents the average of the recorded instances, due to high volume of data the average is considered for better visualisation.

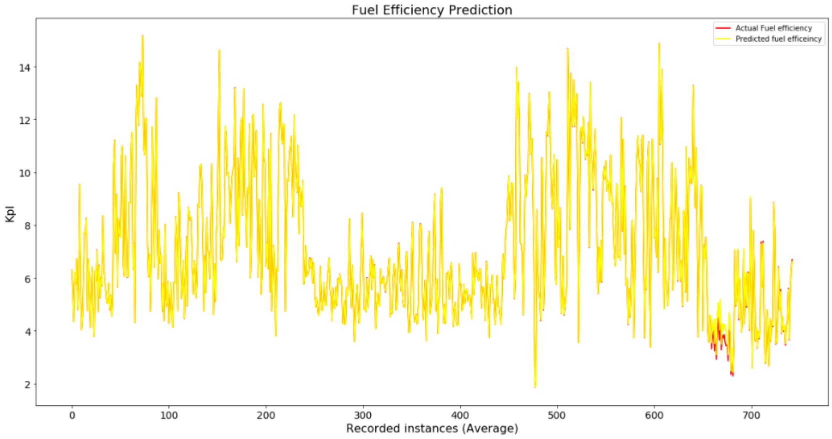


Figure 12: Fuel efficiency prediction of XGBoost model

Figure 13 shows the error of prediction in percentage, the difference between predicted fuel efficiency and actual fuel efficiency for the XGBoost model. RMSE of 0.17107, MAE of 0.09378 and R^2 of 0.99840 is obtained for the XGBoost model. The error percentage is around 15 in this model, so in comparison to all the other models developed the XGBoost model is better in terms of error of prediction in percentage and also the RMSE and MAE values of this model is better than all other models developed. From the graph it is evident that the XGBoost model predicts well even when there is sudden raise and drop with the massive difference in fuel efficiency. At few points it is observed the error percentage to be around 15 indicating that the XGBoost model's prediction is not up to the mark when the fuel efficiency is continuously in low range, except those points XGBoost model outperforms other models in prediction of fuel efficiency.

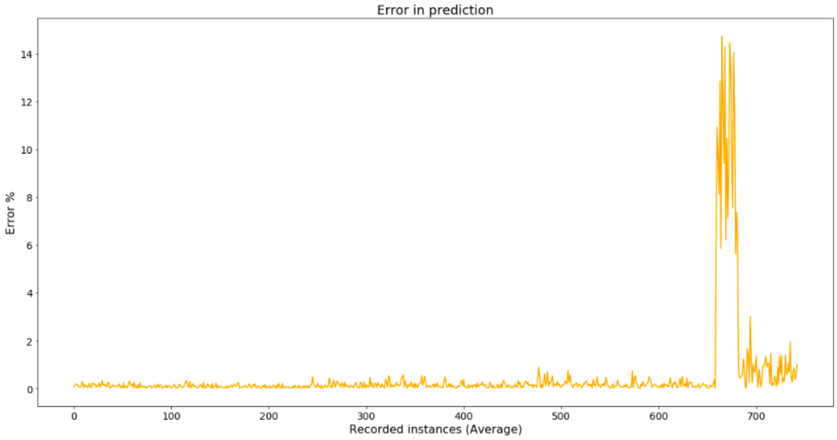


Figure 13: Error in prediction of XGBoost model

Similar case study was done for the split ratio 75:25 and 80:20 the results are shown in the Table 1 and Table 2 respectively.

Table 1: Results for the split ratio 75:25

Model	**RMSE**	**MAE**	**R^2**
Multiple Linear Regression	4.32890	3.05285	-0.06931
LinearSVR	4.46226	3.37958	-0.10838
Multilayer Perceptron	4.13546	2.71837	0.04802
LSTM	4.59520	4.15971	-0.17541
XGBoost	0.14048	0.08958	0.99890

Table 2: Results for the split ratio 80:20

Model	**RMSE**	**MAE**	**R^2**
Multiple Linear Regression	4.33148	2.95356	-0.16300
LinearSVR	4.38962	3.33760	-0.19443
Multilayer Perceptron	3.74051	2.49914	0.13270
LSTM	4.02409	3.59022	-0.00379
XGBoost	0.16319	0.09583	0.99835

XGBoost model has consistently performed good for all the split ratios and the values in the table are evident that XGBoost outperforms the other models and with the RMSE value of around 0.15 the model performs well in prediction of fuel efficiency.

6.2 Recommended characteristics for better fuel efficiency

The other objective of this research is to recommend the throttle position and other vehicle characteristics to achieve better fuel efficiency. The two main factors, speed and throttle position are analysed and Figure 14 shows the graph of throttle position versus predicted fuel efficiency and Figure 15 shows the graph of speed vs predicted fuel efficiency. The predicted fuel efficiency of XGBoost is chosen. From the Figure 14 it is clear that the fuel efficiency is highest when the throttle position is maintained between 70 and 80, 100 refers to full pressed throttle, and when there is huge difference in the throttle position repeatedly the fuel efficiency falls down.

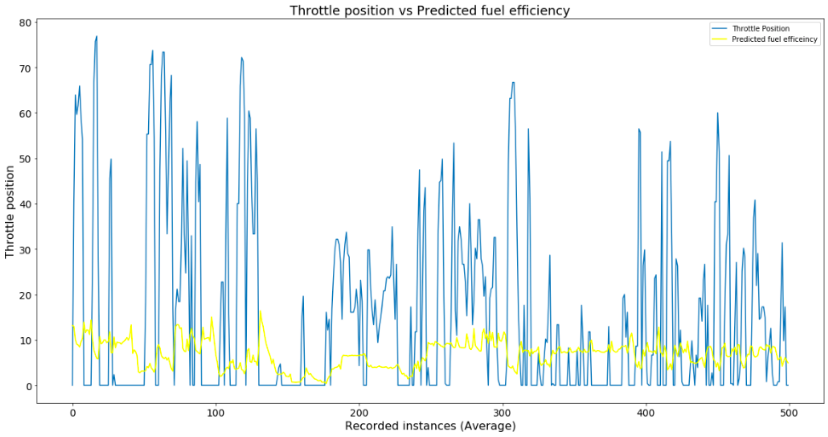


Figure 14: Comparison between throttle position and predicted fuel efficiency

Similarly from the Figure 15 it is evident that when the vehicle's speed is between 40 km/h and 60 km/h maximum fuel efficiency is achieved, this is the economy speed and necessary actions are to be taken to make sure less fuel is consumed either by giving a warning to the driver to stay in the economy speed or the manufacturing companies should limit the maximum speed to ensure better fuel efficiency.

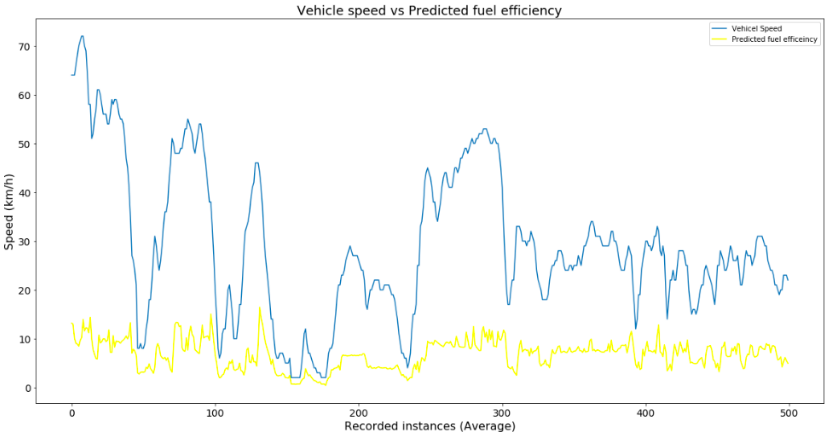


Figure 15: Comparison between speed and predicted fuel efficiency

6.3 Discussion

The primary objective of the research was to develop a model using machine learning techniques which precisely predicts the fuel efficiency and to propose the optimum driving style and vehicle characteristics to achieve better fuel efficiency. A review on the literature related to the research exposed the techniques that could be used to build the model and the analysis to be carried out to meet the objectives. Machine learning techniques like Multiple Linear Regression, Support Vector Machine, Artificial Neural Network and XGBoost was chosen to develop the model. Data pre-processing and data transformation like normalisation was carried out before building the model. Five models were built using the machine learning techniques Multiple Linear Regression, Support Vector Machine, Artificial Neural Network and XGBoost.

The parameter was selected by running GridSearch for few algorithms like Support Vector Machine and XGBoost and for the other algorithms through parametric analysis. The developed models were evaluated with the help of standard evaluation metrics like RMSE, MAE and R^2. Throttle position and speed are examined with the predicted fuel efficiency to evaluate their relationship with the fuel consumption. Analysis on mass air flow rate, intake air temperature and other vehicle characteristics with the predicted fuel efficiency is also carried out which gives deeper insight and better recommendations to mitigate fuel consumption. One limitation was the hardware resource available because of which the GridSearch was unable to run on all algorithms for all parameters. Another limitation was the availability of the data required, although the car manufacturing companies collect all the data related to the car, it is limited.

7 Conclusion and Future Work

As discussed in section 6 the models developed have promising results in predicting the fuel efficiency with the XGBoost model outperforming all other models by constantly predicting better for all the experiments conducted with different train and test split ratio. The XGBoost model's performance is low only when there is low fuel efficiency repeatedly but in comparison with other models developed XGBoost model's performance is exceptional and the values obtained for RMSE, MAE and R^2 is also acceptable. Although this model was run on the data collected from small passenger car, the model is not limited only to that class and can be generalised for any vehicle with the driving data and vehicle characteristics available.

There is more scope in future for research and analysis of fuel efficiency by including other factors like the road condition and real-time traffic with the help of google maps, this would help in analysing much deeper. The knowledge discovered from the research and future work can be used by the car manufacturing companies to improve the fuel economy by considering the characteristics that substantially influence the fuel efficiency.

8 Acknowledgement

My most heartfelt thanks to my supervisor Dr. Cristina Muntean for her constant support and encouragement throughout the research work. I thank her immensely for the guidance she has offered since the very beginning. All the ideas she was willing to share with me were of great help in shaping this research work. I would like to thank National College of Ireland and the School of Computing for providing an opportunity for me to carry out this research work.

Special thanks to my friends and family for supporting me and boosting my morale all through the research period.

References

Byerly, A., Hendrix, B., Bagwe, R., Santos, E. and Ben-Miled, Z. (2019). A machine learning model for average fuel consumption in heavy vehicles, *IEEE Transactions on Vehicular Technology,* 68(7)*,* 6343-6351*,* doi: 10.1109/TVT.2019.2916299 .

Çapraz, A. G., Özel, P., Şevkli, M. and Beyca, Ö. F. (2016). Fuel Consumption Models Applied to Automobiles Using Real-time Data: A Comparison of Statistical Models, 83, pp. 774-781, doi: 10.1016/j.procs.2016.04.166.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine learning,* 20(3)*,* pp. 273-297 .

Fayyad, U. M., Haussler, D. and Stolorz, P. E. (1996). Kdd for science data analysis: Issues and examples., *KDD* pp. 50-56.

Freedman, D. A. (2009). *Statistical models: theory and practice*, cambridge university press.

Fugiglando, U., Massaro, E., Santi, P., Milardo, S., Abida, K., Stahlmann, R., Netter, F. and Ratti, C. (2019). Driving behavior analysis through can bus data in an uncontrolled environment, *IEEE Transactions on Intelligent Transportation Systems.* 20(2)*,* pp. 737-748*,* doi: 10.1109/TITS.2018.2836308 .

Gilman, E., Keskinarkaus, A., Tamminen, S., Pirttikangas, S., Röning, J. and Riekki, J. (2015). Personalised assistance for fuel-efficient driving, *Transportation Research Part C: Emerging Technologies,* 58*,* pp. 681-705 doi: 10.1016/j.trc.2015.02.007 .

Kanarachos, S., Mathew, J. and Fitzpatrick, M. E. (2019). Instantaneous vehicle fuel consumption estimation using smartphones and recurrent neural networks, *Expert Systems with Applications,* 120*,* pp. 436-447*,* doi: 10.1016/j.eswa.2018.12.006 .

Lee, H. and Choi, H. (2016). Analysis of vehicle fuel efficiency and survival patterns for the prediction of total energy consumption from ground transportation in Korea, *International Journal of Automotive Technology,* 17(4)*,* pp. 605-616*,* doi: 10.1007/s12239-016-0060-7 .

Munyon, V. V., Bowen, W. M. and Holcombe, J. (2018). Vehicle fuel economy and vehicle miles traveled: An empirical investigation of Jevon's Paradox, *Energy Research and Social Science* 38*,* pp. 19-27*,* doi: 10.1016/j.erss.2018.01.007 .

Perrotta, F., Parry, T. and Neves, L. C. (2018). Application of machine learning for fuel consumption modelling of trucks, *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017,* doi: 10.1109/BigData.2017.8258382.

Ping, P., Qin, W., Xu, Y., Miyajima, C. and Takeda, K. (2019). Impact of Driver Behavior on Fuel Consumption: Classification, Evaluation and Prediction Using Machine Learning, *IEEE Access,* doi: 10.1109/access.2019.2920489 .

Predić, B., Madić, M., Roganović, M., Kovačević, M. and Stojanović, D. (2016). PREDICTION OF PASSENGER CAR FUEL CONSUMPTION USING ARTIFICIAL NEURAL NETWORK : A CASE STUDY IN THE CITY OF NIŠ, *FACTA UNIVERSITATIS, Series: Automatic Control and Robotics* .

Thitipatanapong, R. and Luangnarutai, T. (2011). Effects of a vehicle's driver behavior to the fuel economy.

Torrao, G., Fontes, T., Coelho, M. and Rouphail, N. (2016). Integrated indicator to evaluate vehicle performance across: Safety, fuel efficiency and green domains, *Accident Analysis and Prevention,* 92*,* pp. 153-167*,* doi: 10.1016/j.aap.2016.03.008 .

Wickramanayake, S. and Bandara, D. H. (2016). Fuel consumption prediction of fleet vehicles using Machine Learning: A comparative study, *2nd International Moratuwa Engineering Research Conference, MERCon 2016,* doi: 10.1109/MERCon.2016.7480121.

Xu, Z., Wei, T., Easa, S., Zhao, X. and Qu, X. (2018). Modeling Relationship between Truck Fuel Consumption and Driving Behavior Using Data from Internet of Vehicles, *Computer-Aided Civil and Infrastructure Engineering,* doi: 10.1111/mice.12344 .

Yin, X., Li, Z., Shah, S. L., Zhang, L. and Wang, C. (2015). Fuel efficiency modeling and prediction for automotive vehicles: A data-driven approach, *2015 IEEE International Conference on Systems, Man, and Cybernetics,* pp. 2527-2532*,* doi: 10.1109/SMC.2015.442.