Master Thesis in Statistics and Data Mining

Data driven analysis of usage and driving parameters that affect fuel consumption of heavy vehicles

Aiswaryaa Viswanathan

Abstract

The aim of this thesis is to develop data mining models able to identify and classify usage and driving parameters that affect fuel consumption of heavy vehicles. Heavy vehicles open up for a new range of systems aimed at improving the efficiency and intelligence of road transports. The most important feature of all these systems and services is the fuel efficiency of these vehicles. For developing energy optimized autonomous vehicles and for helping drivers in eco-driving training, there is a need to understand the usage parameters of these vehicles. One part of this is to understand the factors that affect fuel consumption. In this thesis, comparison of usage and driving pattern parameters has been done to analyze fuel consumption of heavy vehicles. The importance of the parameters has been evaluated depending on its contribution in predicting fuel consumption of heavy vehicles. This particular idea is of huge interest for the company and it will be used in building optimal control strategies for fuel efficiency.

Data mining techniques random forest and gradient boosting were used for the analyses because of their good predictive power and simplicity. Driving parameters like speed, distance with cruise control, distance with trailer, maximum speed and coasting were found to be important factors that affect fuel consumption of heavy vehicles. Evaluation of performance of each of these models based on Nash- Sutcliffe measure demonstrated that random forest (with an accuracy of 0.808) could do a better prediction of the fuel consumption using the input variables compared to gradient boosting method (accuracy=0.698). The results proved that it is wise to rely on predictive efficiency of random forest model for future analysis.

Acknowledgements

I would first like to express my deep appreciation and sincere gratitude to Linköping University for encouraging me and giving me the opportunity to be a part of the Master's program in statistics and data mining. Special thanks to Oleg Sysoev for being such an able program director and guiding us throughout the course. I would also like to thank Mattias Villani, for being so patient, for always trying to explain everything really clearly during our lessons and also during consulting sessions which made me gain lot of confidence. It would be incomplete if I don't mention special thanks to Anders Nordgaard for spending his time giving me valuable ideas about project work, report writing which made motivated me during my studies and work.

I would like to extend my gratitude to Patrik Waldmann, my supervisor at Linköping University, for giving me valuable tips about the project work, suggesting me new techniques to implement and for taking time to review my work and sharing his views about my project work.

I thank Scania for its acceptance to carry out this project and Carl Svärd, my supervisor at Scania who gave me the chance and confidence to work on this project. Thank you Carl for helping me obtain lot of knowledge about Scania, about my project work and for suggesting nice ideas and very politely listening to my ideas as well. Thank you for all your help and effort. I would also like to thank people in Scania who were in one way or another involved in my project work.

Thank you Magnus Adolfson, Gunnar Ledfelt and Henrik Petterson for their valuable suggestions and advices during my thesis work.

Last but not least, I would like to thank my beloved family: my father, for your endless love, support and blessings for successful completion of my project work. Thank you for always believing in me and for always being there for me. My mother, for always supporting me in all stages of my life. Thank you for your immense love and faith in me, which made me gain lot more confidence. My sisters Swarna and Bhu for being such nice friend to me and helping me with your love in all walks of my life. Thanks for being there for me to share my happiness and worries. My friend Prateek Sharma, for all your support and help, for always encouraging me, for being there for me during my difficult times and showing me the positive side of every situation. My grandparents, for their blessings and love.

Table of contents

1 Introduction..2
1.1 Scania..2
1.2 Background..2
1.3 Related Work..4
1.4 Objective..5
2 Data..6
2.1 Raw data..6
2.2 Secondary data..7
3 Methods..9
3.1 Supervised learning methods..9
3.2 Methodology..9
3.3. Random forests..10
3.4 Gradient boosting..12
3.5 GAM boost..15
3.6 Efficiency measure..16
4 Results..17
4.1 Descriptive statistics..17
4.2 Random forests..21
4.3 Gradient boosting..28
5 Discussion and conclusions..34
5.1 Random forests..34
5.2 Gradient boosting..35
5.3 Evaluation..36
6 References..38
7 Appendix..40

1 Introduction

1.1 Scania

Scania is a major Swedish automotive industry manufacturer of commercial vehicles. The company has ten production facilities and over 35000 employees around the world. The company's objective is to deliver optimized trucks and buses, enabling customer with very less total operating economy, making the company one of leading companies in the industry. This master thesis has been done at research, pre-development and concept studies, department of intelligent transport systems and services. The mission of this group is to perform pre-studies within the area intelligent transport systems and services, to achieve transport efficiency and more profitable customers through information and communication technology (Scania Inline, 2013).

1.2 Background

Recently, increased fuel consumption of heavy vehicles has become one of major challenges of automotive industries. For controlling fuel consumption and maintaining it at an average level, lot of new techniques have been introduced. These techniques range from energy optimized route optimization, driver support and training, autonomous or semi-autonomous vehicles, intelligent fleet management and platooning. Hence, there is a need for understanding the parameters that affect fuel consumption of heavy vehicles for developing energy optimized autonomous vehicles. This also provides a great help in assisting drivers in eco-driving training. The fuel consumption of heavy vehicles can vary depending on various different reasons which can be categorized as,

1. due to vehicles,
2. due to roads,
3. due to the usage (drivers),
4. due to ambient conditions (temperature, wind, etc.)

This project will mainly focus on the fleet management data source. Fleet management system interface (Scania Inline, 2013), is a medium to vehicle data of heavy vehicles. Various vehicle parameters like speed of the vehicle, total fuel consumed, total distance travelled, total runtime of the vehicle, etc. is broadcasted in Fleet management system interface. Due to the Fleet management system interface data, it is possible to have manufacturer independent application, usage and evaluation of data.

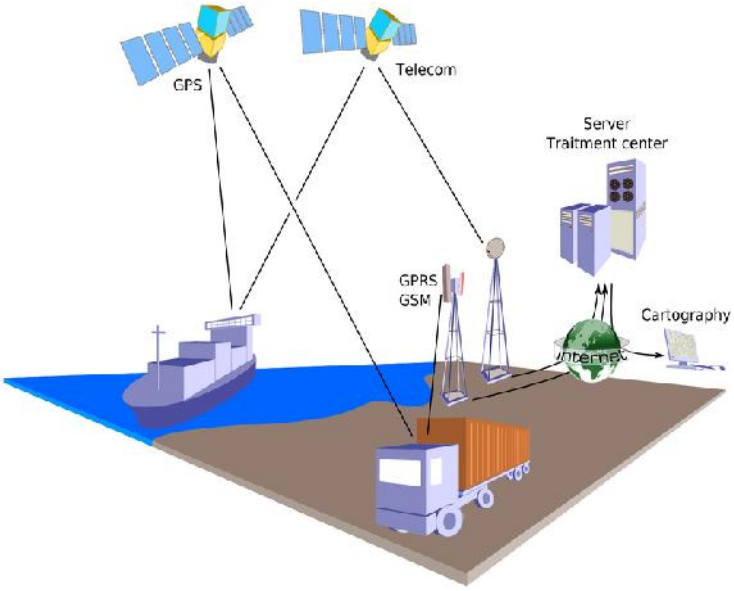

Figure 1: Principle of geo-location for the position determination and GSM/GPRS for the data transmission (http://en.wikipedia.org/wiki/Fleet_management)

Figure 1 gives information about how the data is collected in fleet management system. Each vehicle's position is regularly monitored and the data is transmitted using GPRS and GSM technologies. The web based fleet management system consists of two parts: an on-board computer in the trucks and a fleet management portal (FMP) at the office. The on board computer consists of various information about drivers and usage parameters and at the office a haulier supervises vehicle and transport management.

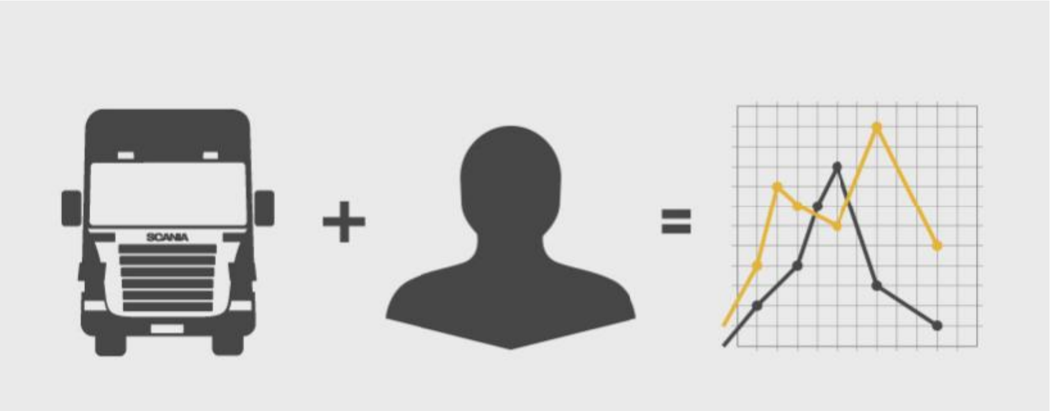

Figure 2: Analysis package connects data on vehicles and drivers, which gives more precise understanding of the fleet management system (Scania.se)

Figure 2 depicts the functioning of the fleet management system. The stored information in the FMP from the vehicle's on board computer containing information about drivers and vehicles is very useful for any kind of analysis.

Given the idea of the project, selecting a group of variables that are potential enough in predicting the fuel consumption of heavy vehicles has been an exciting task.

1.3 Related Work

Fuel efficiency of road transports is one of the major challenges faced all over the world. Hence, there are lot of literature works that deal with this problem.

Bratt and Ericsson (1998) investigate fuel consumption of heavy vehicles, differences in driving patterns and characteristics and emissions using various types of measuring equipment. The authors have considered two important measures speed and acceleration profiles as the driving behavior and investigated on the fuel consumption and emissions (Bratt and Ericsson,1998).

Constantinescu *et al.* (2010) analyze driving style using data mining techniques. Each of the driver's driving style has been modeled using various driving parameters. The idea behind the project is to classify drivers to enable traffic safety. Cluster analysis and principal component analysis have been carried out to cluster driver according to their driving behavior.

Yun *et al.* (2010) deal with application of data mining techniques on commercial vehicles. This analysis has been carried out for reducing the carbon dioxide emission due to commercial vehicles. GPS based data and data mining techniques together helped in the successful completion of the work.

1.4 Objective

The main objective of this thesis is to develop a predictive model using data mining methods for identifying and classifying usage and driving parameters that affect fuel consumption of heavy vehicles. The importance of each of the parameters that helps in predicting fuel consumption should be analyzed and evaluated for future use.

2 Data

The analysis will be concentrated on long haulage vehicles, for which, Scania has followed many years of strong presence in the market. The data is selected from the databases- Fleet management and Vehicle specification database (Scania Inline, 2013).

Vehicle operational data from the fleet management database has been used for the analysis of vehicles and their properties. The data contains various parameters including vehicle id, odometer reading, total runtime, speed of the vehicles, harsh accelerations, number of brakes applied, over speeding, over revving, maximum speed, position message id, etc.

The vehicle specification database contains variant codes which contains information about physical attributes of vehicles like length, weight, color, gearbox, engine type, type of the vehicle, etc.

2.1 Raw data

The dataset contains information about 550 vehicles and their usage parameters. The selected dataset represents one of the largest databases which contain accumulated data about the vehicle parameters.

In general, many variables are needed for accurate prediction of the target variable. The FMP contains various different parameters of heavy vehicles ranging from usage parameters, physical parameters, vehicle ids, driver ids, vehicle position ids, etc. But only few variables have been chosen for the analysis. The chosen variables include usage and driving parameters- mean speed, total fuel, total fuel idle, total fuel power take off (PTO), odometer reading, total runtime, runtime idle, runtime PTO, harsh brakes, number of brake applications, over speeding, over revving, total distance with cruise control (CC) active, total distance with trailer, maximum speed, green band driving, distance while out of gear and coasting. The data stored in the FMP has a repetition rate between 20ms and 10 seconds.

2.2 Secondary data

The FMP contained information about both buses and trucks. Since this project focusses on trucks, all the data regarding buses has been removed.

Since the data stored in FMP has a very low repetition rate, the data has been remodeled in such a way that it is not very large and easier to work on. For simplicity reasons, the raw data has been transformed to create new variables, with each measurement taken at a time difference of 10 minutes.

Due to the presence of three different fuel parameters, runtime parameters in the FMP, it was important to decide which of the three parameters would contribute in the predictive analysis. Hence a wholesome measure of fuel and runtime was calculated using the three parameters

$$Fuel_{driven} = Total_{fuel} - (Fuel_{PTO} + Fuel_{idle})$$

$$Runtime_{driven} = Total_{runtime} - (runtime_{PTO} + runtime_{idle}).$$

The above formulas were used to calculate a single parameter of fuel and runtime.

Fuel consumption of each vehicles was calculated as a measure of liters per 100 kilometers following

$$Fuel\ consumption = \frac{Total\ fuel\ driven}{Total\ distance} * 100.$$

The above formula has been used for calculating the fuel consumption of the vehicles.

Speed was also not available in FMP, therefore the speed of each vehicle has also been calculated using the distance and runtime. Speed is measured in km/hour units

$$Speed = \frac{Total\ distance}{Runtime_{driven}} * 3.6.$$

The dataset was reorganized so the final data contained 40000 observations with 12 descriptive input variables which were considered variables able to predict the output variable fuel consumption of heavy vehicles.

After preprocessing, the data was further split into training and test data. Training data is required for carrying out any kind of analysis on the data, whereas test data evaluates the analysis done. Data was split so that both training and test data had 20,000 observations each.

A quantitative analysis of the input and output variables is given in section 4.1. Summary of all variables can be the first step to understand the whole data set. Further information on how well the variables have helped in predicting the fuel consumption of heavy vehicles, which variables are important in interpreting and predicting the fuel consumption is given in detail in Section 4.

3 Methods

Two main methods random forests and gradient boosting are used in this report. Both the methods are considered to have good predictive properties and all analysis have been carried out in R.

3.1 Supervised learning methods

The supervised learning methods are helpful in describing properties of the dataset and in predicting the value of the output variable knowing the values of input variables. In supervised learning, a training set is given and aim is to achieve prediction of previously unseen examples. The training set is a representative of the real world use of function. Using the training set, a model is built to predict the output variable.

Supervised learning is a machine learning technique of inferring an information or function from a training set. The procedure of supervised learning starts with a training set of the data$\{(x_1 ,y_1),(x_2,y_2) \ldots.,(x_n,y_n)\}$; and a function g: $X \rightarrow Y$, where X is the input and Y is the output and g is a function describing relationship between X and Y. To determine how well the model fits the training data, a loss function is introduced.

3.2 Methodology

- Firstly, quantitative analysis of all input variables and output variable was carried out for a simple understanding of the variables. Graphical representation of variables was done to find inconsistent values and outliers.
- Partitioning of the dataset was required for training and testing of the models, because training data teaches the model whereas test data tests the model performance for future predictions.
- For predicting the output variable using all input variables supervised learning methods was used. Random forests and gradient boosting was suggested to have high predictive performance, and was therefore the choice for the analyses.

- The chosen supervised learning methods was carried out on the training data. The performance of prediction was then measured and optimized using the test data. Forward and backward model selection methods was also carried out for selecting the variables that were more accurate in predicting the output variable.
- The performance of the models was analyzed using various residual squared error and Nash-Sutcliffe efficiency measure. The best model was chosen to predict the fuel consumption for a whole new set of data.

3.3. Random forests

Random forests is an ensemble learning method for classification and regression. It operates by constructing a group of decision trees for training data. Random forests carry out bagging of unpruned decision tree learners with a random selection of importance features at each split. In particular, generalization error of random forest is directly related to the strength of each tree in the forest and how they are correlated to each other. Various internal estimates and correlation are related to the increasing number of features at each split and also help in measuring variable importance of random forests (Breiman, 2001).

A collection of many trees is random forest. A large number of trees has been considered important when variable importance is of interest. In random forests each tree is a random set of data and each split is created based on a random subset of candidate variables. Average predictions over all individual trees will be the overall prediction for the random forest. The prediction of the random forest will be a smoother function compared to the prediction of single tree because it is averaged over a large number of trees (Genuer, 2010; Gromping, 2012).

In classification, variable importance is measured using the Gini coefficient and in regression, average impurity reduction is used as a measure of variable importance. Variable importance is typically used for variable selection in random forests. Variable selection in random forests can be helpful in two ways (Gromping, 2012):

1. Identify variables that are important in predicting the output variable for interpretation purposes,

2. Identify a small number of input variables that are sufficient for good prediction of the response variable.

Figure 3 illustrates the working of random forest model.

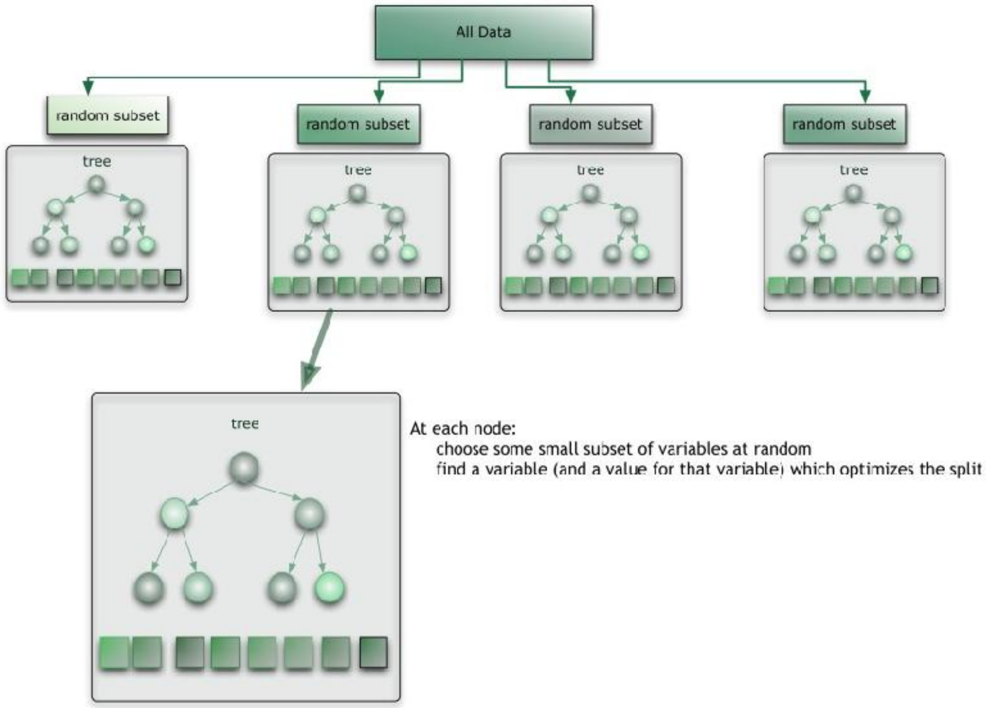
All Data
random subset
random subset
random subset
random subset
tree
tree
tree
tree
tree
At each node:
choose some small subset of variables at random
find a variable (and a value for that variable) which optimizes the split

Figure 3: Illustration of a random forest

Random forests work as follows (Montilo, 2009):

Let N_{trees} be the number of trees to be built.

1. Select new bootstrap sample from the training set.
2. An unpruned tree is grown on these bootstrap sample
3. Select $\mathrm{m}_{\mathrm{try}}$ random predictor variables and find the best split using only these predictor variables at each internal node.
4. Repeat the above three steps for the whole dataset.

The random forest has various useful features (Breimann, 2001):

1. produces accurate and precise results,
2. produces useful estimates of error, correlation, variable importance measure and strength,
3. comparatively robust to noise, inconsistent data and outliers and overcomes over fitting problem,
4. easy to use, easy to interpret and easy to understand,
5. very simple and easily parallelized,
6. relatively faster and efficient than bagging and boosting (Breimann, 2001),
7. levels of measurement don't affect the results of random forest, for example, random forest of fuel consumption will produce same results regardless of how fuel consumption has been measured, (in thousands or hundreds or even as discrete range of values).

The random forest also has some limitations (Montilo, 2009):

1. The computing time is high. In case of huge amount of data, fitting a random forest model to the data is time consuming,
2. Interpretation is difficult in some cases.

The random forest method is implemented in R using the built in package *randomForest* (Breiman, 2012).

3.4 Gradient boosting

Boosting is a machine learning algorithm introduced to reduce bias in supervised learning. Gradient boosting is also a machine learning method for regression problems that produce a predictive model in the form of an ensemble of weak predictive models. Gradient boosting method generalizes all boosting methods by optimizing the differentiable loss function.

Various regression techniques have emerged over the years. The classical methods of predictive model building and variable selection were not considered very reliable and some of the models were even considered biased. This led to a high progress in statistical methodologies and innovations. In this thesis, component wise gradient boosting (Breimann, 1998) has been used to predict the output variables. It is a machine learning method which obtains estimates using gradient decent approaches and produces optimized prediction accuracy (Hofner *et al.*, 2012). The main feature of this method is that it does variables selection during the fitting process. It doesn't rely on any heuristic techniques.

Consider a set of input variables $x_1, x_2 \dots . x_p$ and an output variable y. The aim of this method is to model the relationship between the input and the output variables and to produce an optimized prediction of the output variable given the input variables. This can be achieved by a minimal loss function $\rho(y, f) \in R$ over a prediction function$f$ that depends on input variable$x$. The aim of gradient boosting is to determine optimal prediction function f^* which is given by the formula,

$f^* := argmin_f E_{Y,X}[\rho(y, f(x^T))]$, where the loss function is differentiable with respect to prediction function (Hofner *et al.*, 2012).

Gradient boosting method functions as follows:

1. A set of base learners is first specified. Base learners are regression estimators with set of input variables and single response variables. The base learners are denoted as p and number of iterations m is set to zero. The n-dimensional vector $\hat{f}^{[0]}$ is first intialized to some offset value
2. The number of iterations is increased to one and negative gradient is computed using the formula, $-\frac{\partial}{\partial f}\rho(Y, f)$ and is evaluvated at $\hat{f}^{[m-1]}(X_i)$ where *i=1, 2...n.* A negative gradient vector is yielded from$U_i^{[m-1]} = -\frac{\partial}{\partial f}\rho(Y, f)$ where *Y*= Y_i and *f* =$\hat{f}^{[m-1]}(X_i)$
3. Negative gradient $U^{[m-1]}$ is estimated by using the base learners (regression estimators *p)*, leading to p vectors, each vector acts as an estimate for negative gradient vector $U^{[m-1]}$.

4. The best base learner is selected using the fact that the selected base learner produces minimum SSE. Then $\hat{U}^{[m-1]}$ is set to the fitted value of the selected best base learner.
5. $\hat{f}^{[m]}= \hat{f}^{[m-1]} + v\hat{U}^{[m-1]}$ is updated, where v is a step length factor.
6. Steps 2 -5 is repeated until m=m_{stop}

As seen from the above steps, only one predictor variable is selected in each iteration. For the additive update that the method does, the final estimate can be interpreted as the additive prediction function.

In working with gradient boosting method, various tasks are considered important for the selection of parameters. The most crucial task is choosing number of stopping iterations m_{stop}. Due to the problem of over fitting, boosting algorithms shouldn't be run until convergence. So a finite stopping point has to be given to optimize the prediction accuracy. Also, a choice has to be made for step factor length *v*, which is of minor importance when predicting the output variable using boosting algorithms. Smaller values of v would be preferred to prevent the problem of overshooting the minimum of the empirical risk.

Gradient boosting is useful in several ways (Hofner *et al.*, 2012):

1. At each iteration of gradient boosting, the base learners are selected using a built in function which carries out variable selection between the input variables,
2. It can be applied for data that contains more predictor variables than number of observations,
3. One can fit a wide range models using boosting methods including (generalized) linear models, (generalized) additive models, survival models, etc.,
4. It doesn't rely on users to predefine complex features, instead, it uses function approximation to induce the same,
5. It also addresses multicollinearity problem and optimizes the prediction accuracy,

3.5 GAM boost

An additive model using covariates x=$(x_1, x_2 \dots \dots . x_p)^T$ takes a form (Hofner *et al.*, 2012),

$$g(\mu) = \beta_0 + f_1 + f_2 \dots \dots . . f_p$$

with conditional expectation **μ**=$E(y|x)$, the link function *g* and arbitrary functions $f_1, f_2, \dots \dots . f_p$ of the covariates. Both linear and nonlinear functions can be included in this model. It has been showm that boosting is a gradient descent algorithm (Breiman, 1998), where at each iteration a base learner (e.g. tree) is fitted to the negative gradient of the loss function (Friedman *et al.*, 2000).

Additive regression models has been considered using least square base learners in addition to the L_2 loss function (Buhlmann and Yu, 2003).

In this project, fitting an additive model with a continuous outcome and smooth functions of input variables was considered. In this case smooth splines were used as base learners. Due to the fact that smoothing splines are less efficient than other base learners, the natural approximation of smoothing P-splines was used (Eilers and Marx, 1996).

The mboost() package in R provides efficient and helpful tools for fitting generalized additive boosting models. The type of effects (linear or nonlinear) can be specified using different base learners. *Bols()* base learner allows to specify ordinary least square base learners with following properties: 1. Linear effects, 2. Categorical effects, 3. Linear effects of group of variables. The logical variable *intercept*=FALSE make continuous covariates mean centered.

The best modeling tool for smooth effects is P-splines. Smooth effects are defined in gradient boosting by the *bbs()* base learner for all those covariates which are needed to be comparatively smooth. Examples of *bbs()* type are: 1.smooth effects, 2.spatial effects, 3.cyclic effects. The logical variable *center*=TRUE produces smooth deviation (Hofner *et al.*, 2012).

Minimum description length (MDL) is one important technique for model selection in statistics (Hansen and Yu, 2001). MDL is based on finding the model that produces the shortest description of the data. It is related to Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), but has better statistical properties than these. As it has been mentioned in section 3.4, it is important to stop the boosting algorithm at an optimal iteration to ensure the maximal predictive accuracy.

3.6 Efficiency measure

Each of the supervised learning methods used, should be evaluated on a basis of its predictions. This evaluation can be done by many different ways. Nash-Sutcliffe model efficiency coefficient was used in this project to observe the predictive power of the random forest and gradient boosting methods (Nash and Sutcliffe, 1970).

Nash-Sutcliffe measure gives the difference between observed and predicted values, normalized over the variance of the observed values. The formula for calculating the efficiency measure is given below.

$$E = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The predictive power of the methods is considered good, if the efficiency measure is closer to one.

4 Results

4.1 Descriptive statistics

This chapter deals with the descriptive analysis of input and output data. It is clear that the distribution of fuel consumption is highly skewed and has heavier right tail. The average fuel consumption = 35.3026 litres/ 100km. Figure 4 and Figure 5 explain the distribution of target variable-fuel consumption of heavy vehicles.

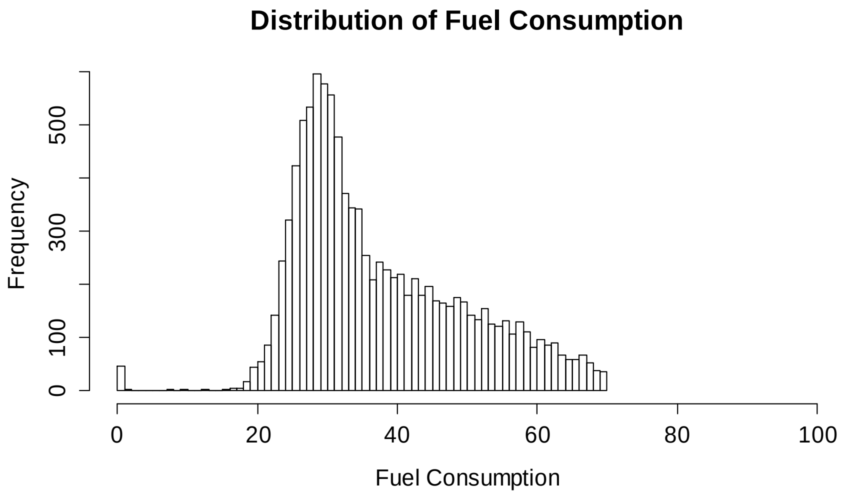

Figure 4: Histogram of output variable fuel consumption

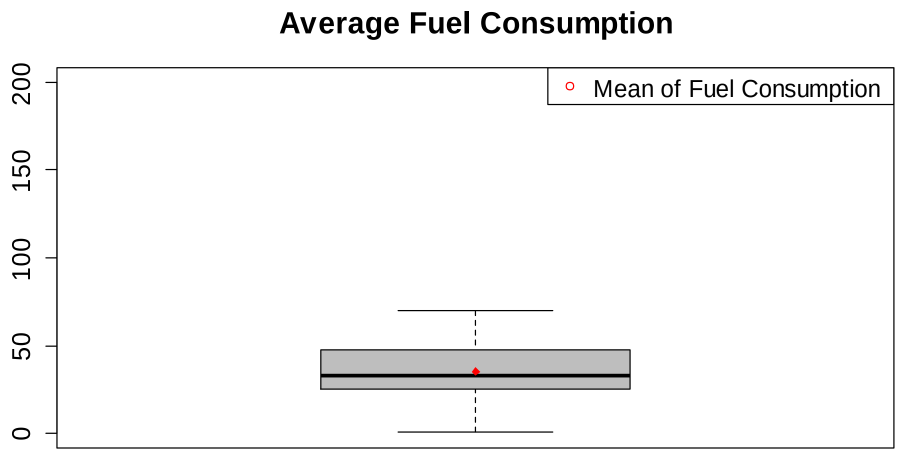

Figure 5: Box-plot of fuel consumption of heavy vehicles

Figure 6 shows the distribution of input variable coasting. It clearly shows a normal distribution. Figure 7 shows the distribution of variable speed. The speed of heavy vehicles is distributed between 0 and 200km/hr.

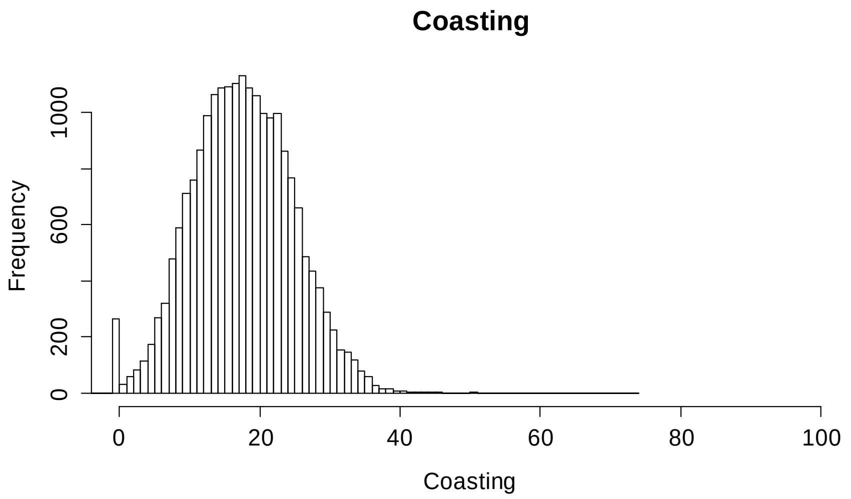

Figure 6: Histogram of coasting

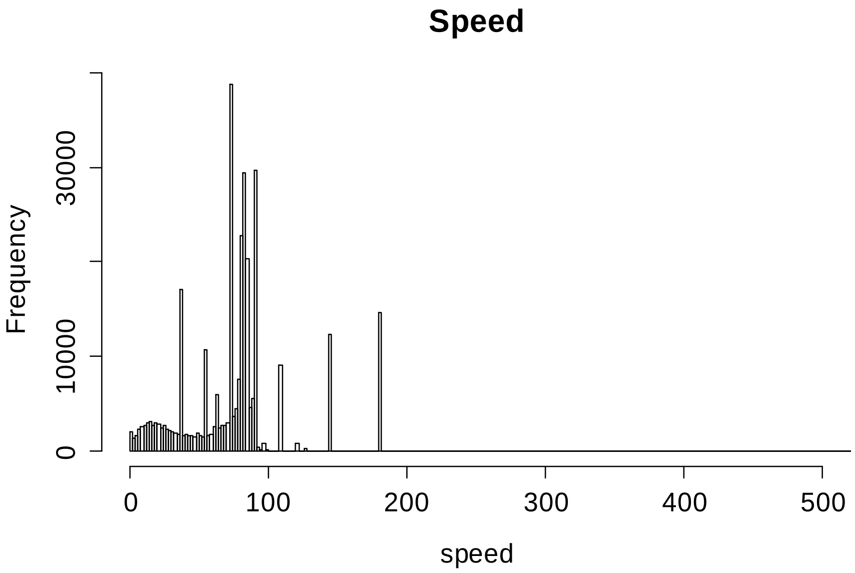

Figure 7: Histogram of input variable speed

Table 1 gives a detailed description of mean, median and standard deviation values for each of the months between March, 2012 and February, 2013. From the table, it could be concluded that average fuel consumption is highest in May month. It's not reasonable due to good ambient conditions during summers. Standard deviation during the month gives an answer to this doubt. Standard deviation is very high due to the fact that there are lot of outliers, which makes the analysis unreasonable.

Table 1: Statistical summary of fuel consumption in each month between 2012 and 2013

Month, Year	**Mean**	**Standard deviation**	**Median**
March, 2012	35.44239	32.29124	29.74972
April, 2012	35.92697	43.02022	30.29177
May, 2012	41.41261	154.8616	30.39451
June, 2012	33.40775	23.89525	29.78114
July, 2012	32.90204	32.0759	29.96268
August, 2012	31.02009	26.21233	29.57123
September, 2012	31.6247	47.40319	30.06165
October, 2012	32.40807	21.06322	30.64224
November, 2012	33.15107	21.36647	31.00583
December, 2012	31.70924	23.17233	31.25363
January, 2013	30.75369	23.26382	30.68329
February, 2013	30.11127	24.40953	29.91467

The graphical description of the fuel consumption over months is given in Figure 8.

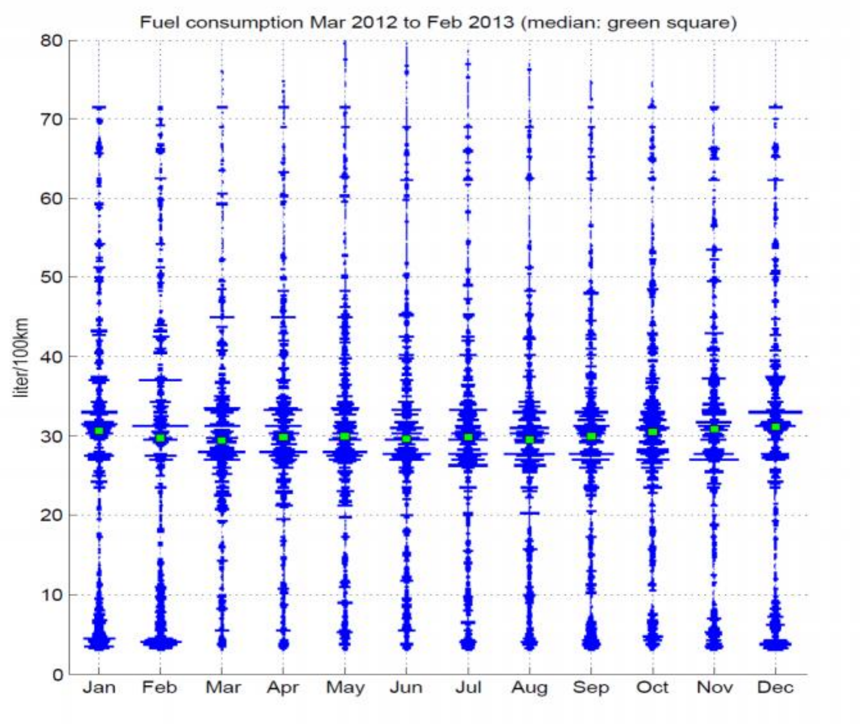

Figure 8: Distribution of fuel consumption of all heavy vehicles from March 2012- February 2013

Table 2 gives a statistical summary of all input and output variables. Reading and understanding the statistical summaries of all predictors and targets is the preliminary task in building a data mining model.

Table 2: Statistical summary of all input and output variables

Variable	Mean	Standard deviation
Fuel consumption	35.25	13.63
Speed	75.62	37.50
Coasting	64.12	35.67
Harsh brakes	0.014	22.80
Distance out of gear	0.686	21.21
Distance with trailer	387.5	20.90
Green band driving	0.115	37.61
Brake applications	1.508	34.57
Harsh accelerations	0.668	14.04
Distance with CC	289.2	15.43

Overreving	1.811	27.09
Over speeding	10.78	35.13
Maximum speed	77.92	17.12

Table 3 gives the extent of correlation of all input variables with the target variable, fuel consumption. When building a predictive model, multicollinearity is one minor problem. Hence correlation between the output variable and each of the input variables is calculated and enlisted below. From Table 3, it is evident that the input variables are not much correlated with the target variable.

Table 3: Correlation between output and all input variables

Variables	Correlation with fuel consumption
Speed	0.003547088
Coasting	-0.01485554
Harsh brakes	0.03466233
Distance out of gear	0.01901613
Distance with trailer	0.02694345
Green band driving	0.01470636
Brake applns	0.02168884
Harsh acceleration	0.01777529
Distance with CC	-0.0574294
Overreving	-0.0001246
Over speeding	0.00034450

4.2 Random forests

Different techniques work well with different datasets but random forest is a technique which works well on almost all kinds of data.

The random forest method was performed, producing a series of important variables in predicting the output variable. Selecting the tuning parameters of random forest was a crucial task.

The random forest regression was built using all input variables and the response variable. The output variable is continuous hence squared error rate was used as one of the measures of efficiency of the model. The results obtained are shown below.

The results in Table 4 explain that the predictive capability of random forests was not improved greatly by increasing the number of split variables. It shows a minute increase in the quality of prediction.

Table 4: Mean squared error and percentage of variance explained by the random forest

No. of trees	No. of variables	Variance explained %	MSE
500	2	45.3	0.1177
500	4	43.2	0.1126
700	2	45.1	0.1179
700	4	49.2	0.1125
1000	2	45.4	0.1177
1000	4	52.3	0.1124

The importance of variables in predicting the fuel consumption can be estimated from Figure 9 and Figure 10. In Figure 10, the prediction accuracy of each variable in predicting the output variable gives a detailed description of which variables are important in this predictive model.

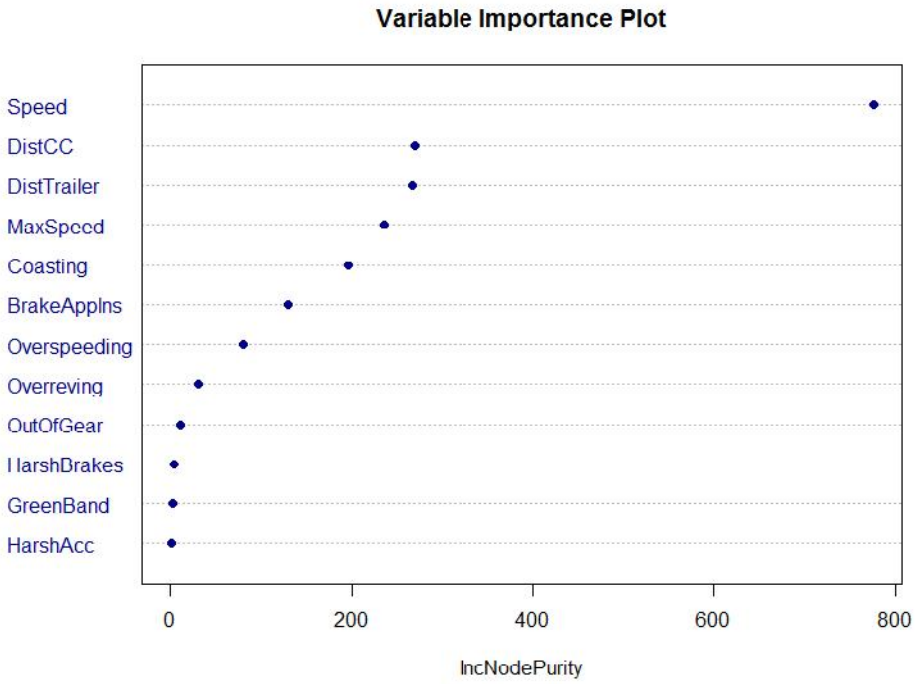

Figure 9: Variable importance plot – increase in node impurity

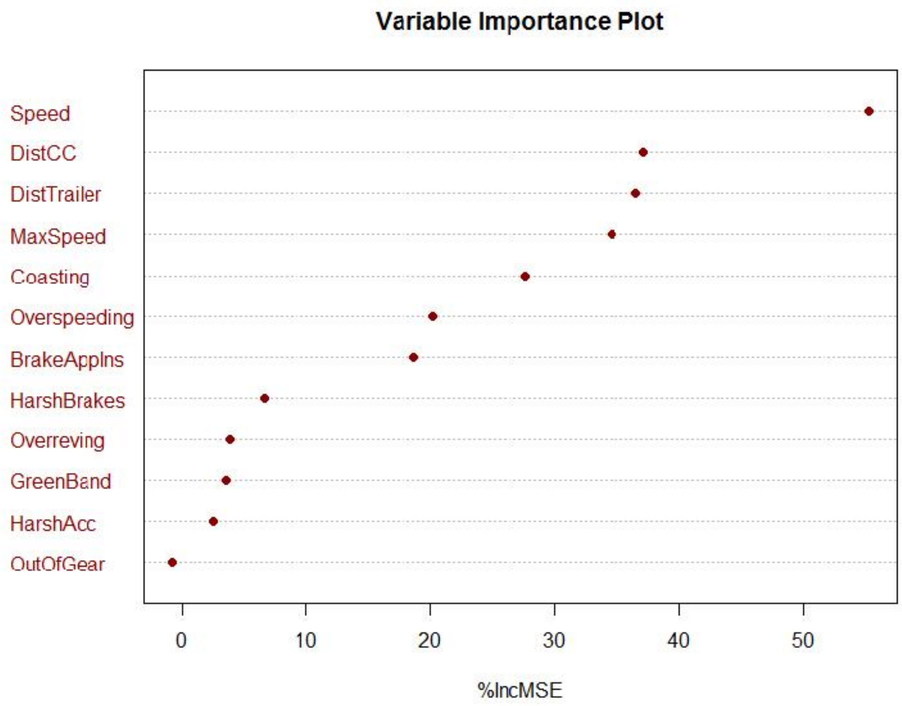

Figure 10: Scaled average of the prediction accuracy of each variable

The variable importance index shows the increase in the mean error of a tree (MSE in regression) in the forest, when the observed value of any input variable is randomly permuted in the out of bag samples (Genuer *et al.*, 2010).

Table 5 explains the importance measure of each variable in predicting the target variable. The variables have been arranged in descending order to show the importance of the variables more clearly. The higher the importance measure, the better the input variable predicts the target variable. Therefore stronger splits are represented by the variables with high importance measure.

Table 5: Importance measure of input variables

Variables	Importance measure
Speed	75.575
Distance with trailer	13.944
Distance with CC	13.092
Max speed	12.374
Coasting	9.717
Brake applns	9.974
Over speeding	3.163
Distance out of gear	0.260
Over reving	0.230
Harsh brakes	0.098
Green band driving	0.023
Harsh accelerations	0.000

The following conclusions can be made from the plots and tables above:

1. The random forest produces good results in predicting the variable that affect the fuel consumption of heavy vehicles.
2. In Table 4, the results confirm that smaller values of *mtry* - number of split variables, could give better prediction. Increase in *ntree* - number of trees, increases the predictive capability of random forests.

One of the main objective of variable selection is to find the important variables highly related to the target variable, for which, variables with very low importance measure can be removed. It is evident from Table 5, the variable Speed is the one that affects fuel consumption to the maximum extent compared to other variables.

After fitting a random forest model to the training data, the model is evaluated using the test data. The predicted values and the observed values are plotted to check the efficiency of the model. The density plot of observed and predicted values in Figure 11 shows that the predictive distribution has peaks and troughs for the same values as the original data.

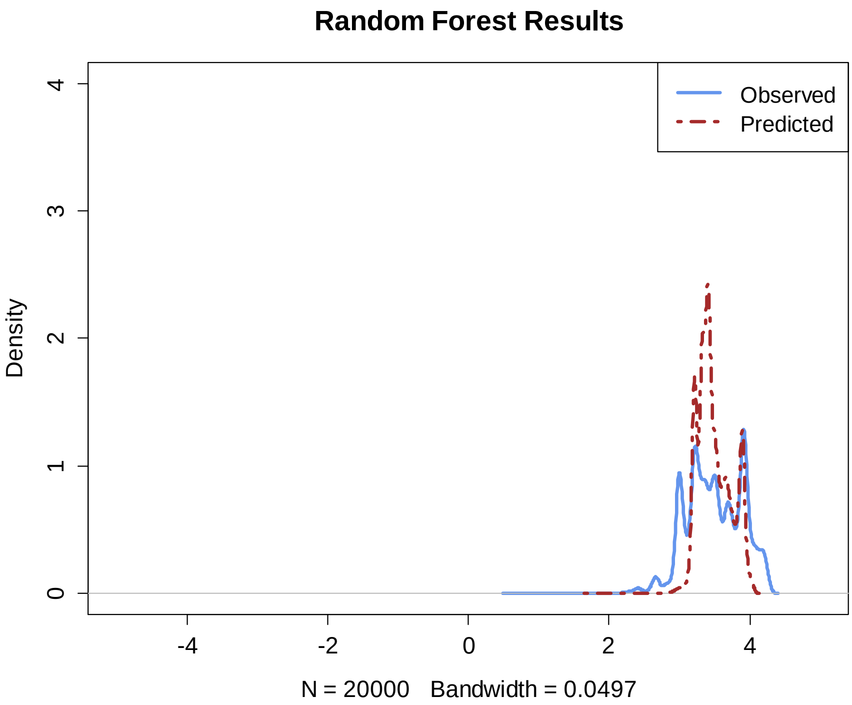

Figure 11: Density plot of predicted vs. observed values

When building a predictive model, it is hard to judge the model comparing the density plot of predicted and observed values of the data. Figure 12 shows the scatter plot of observed values vs. predicted values.

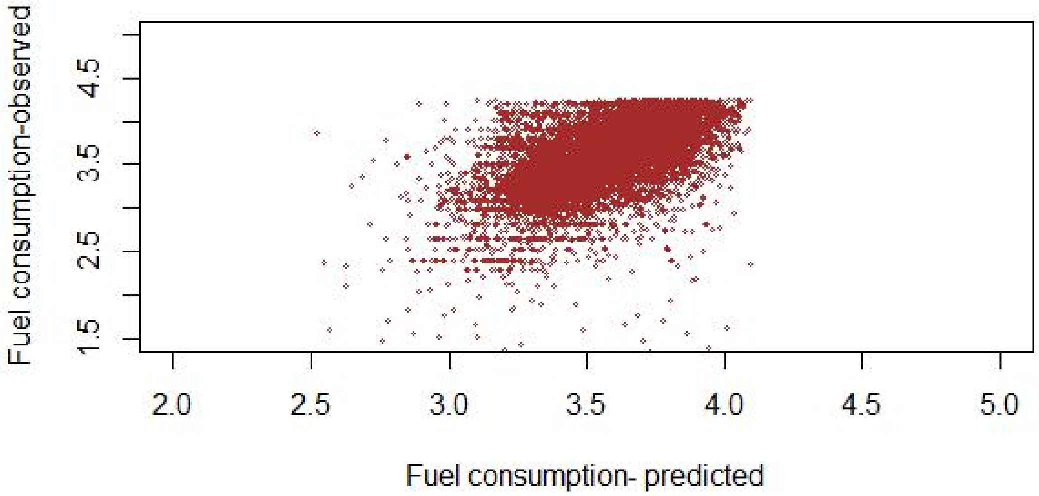

Figure 12: Scatterplot of observed vs. predicted values

The histogram of residuals in Figure 13, show that the residuals are almost normally distributed. A bell shaped plot of residuals shows normal distribution is a good approximation of distribution of residuals in the method.

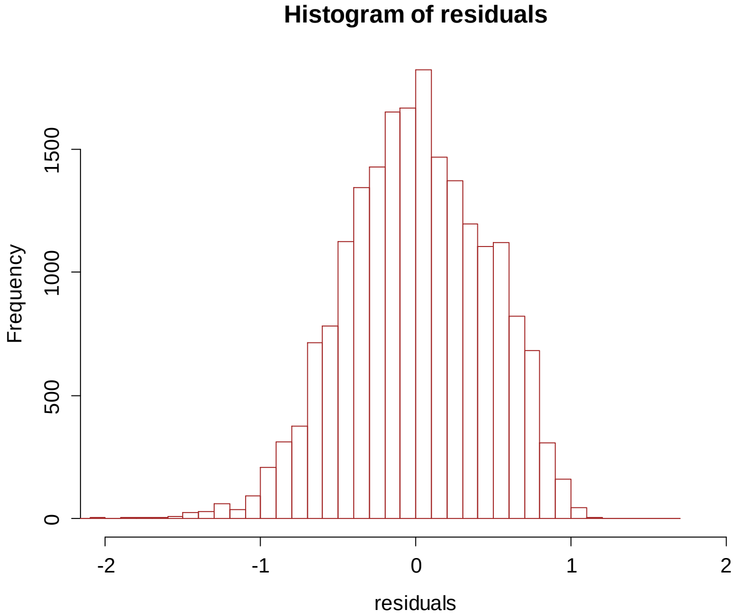

Figure 13: Histogram of residuals- Random forest

As a final step to estimate the efficiency of the model, Nash- Sutcliffe model efficiency is calculated for the random forests model. The value obtained is 0.808, which is relatively closer to 1. Hence random forests could be considered a good model for prediction of the fuel consumption of heavy vehicles.

Nash Sutcliffe efficiency of random forest method = 0.808

Figure 14 shows the assessment plot of random forests. MSE values of training data and test data have been plotted against the number of trees. It is clearly evident that MSE values of random forest are very low. Training data achieves lower MSE earlier than the test data. Overall assessment of random forests is very good.

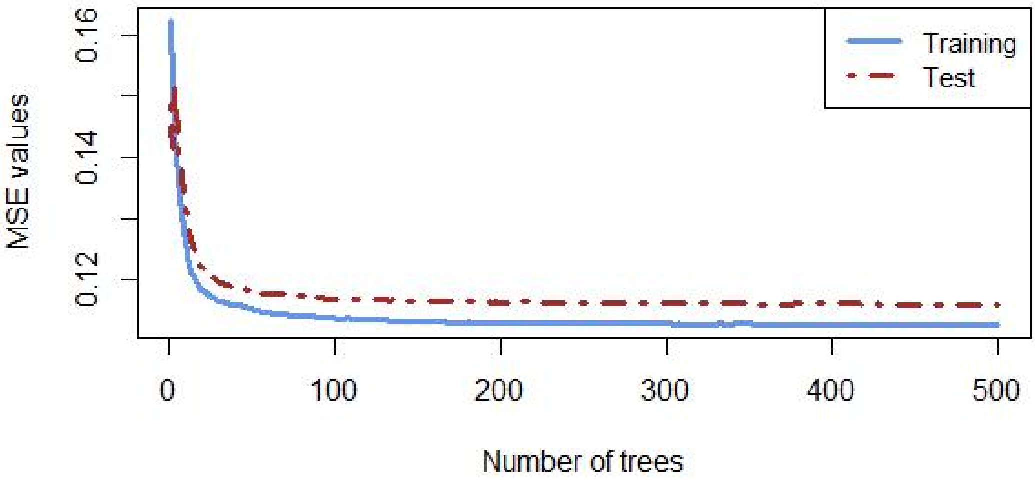

Figure 14: Assessment plot of random forests

4.3 Gradient boosting

Component-wise gradient boosting technique is a method which is helpful in optimizing prediction accuracy and provide good variable selection of input variables that could predict target variable with highest accuracy.

The main advantage of gradient boosting over other machine learning techniques such as decision trees and random forests is that it produces prediction rules that have same interpretation as in the statistical models. The gradient boosting is performed in R using the *mboost* package, which can implement methods to do regression, classification and fit generalized linear models, generalized additive models, etc. (Hofner *et al.*, 2012)

The component-wise boosting technique involves various steps as mentioned in section 3. The technique follows an additive update procedure, which makes it very closer to generalized additive models (Hastie and Tibshirani, 1990).

In this project, *gamboost()* is performed, which provides an interface to fit a generalized additive model to the data.

The interpretation of *gamboost()* is very similar to *gam() e*xcept that, in *gamboost()* variable selection is performed in each step. At each iteration *gamboost()* fits an additive model to all the input variables and calculates a negative gradient measure. Only the model that best fits the data is used in the update step.

Once the model is fitted to the data, it is important to calculate the appropriate number of iterations via out of sample empirical risk. One important task while fitting *gamboost()* is to select the tuning parameters. Number of iterations is one of the important parameters to be chosen before fitting a model. The minimum description length value is considered as a value for measuring the efficiency of the model. The value doesn’t change dramatically either with the change in number of trees or with the step control factor. But it is required that value of v is small (Schmid and Hothorn, 2008) to make sure that the algorithm doesn’t go beyond the minimum empirical risk R (Hofner *et al.,* 2012).

Table 6 shows the difference in MDL measure when the tuning parameters are changed. Lesser the MDL value better the predictive model, hence number of trees = 100 and step control factor= 0.05 is chosen for the fitting the predictive model.

Table 6: GAM boosting- MDL measure

Number of iterations	Step control factor		MDL measure
10	0.1		5.15
100	0.1		5.20
100	0.05		5.12
150	0.05		5.20
200	0.05		5.12

The plots in Figure 15 and Figure 16, explain the partial effects of the input variables- speed, maximum speed, distance with trailer on fuel consumption. The plot depict linear and nonlinear effects of the input variables. In this case, decision can be more rigorous between perfectly linear and nonlinear effects if the algorithm is not stopped at an appropriate iteration *mstop* (Hofner *et al.*, 2012).

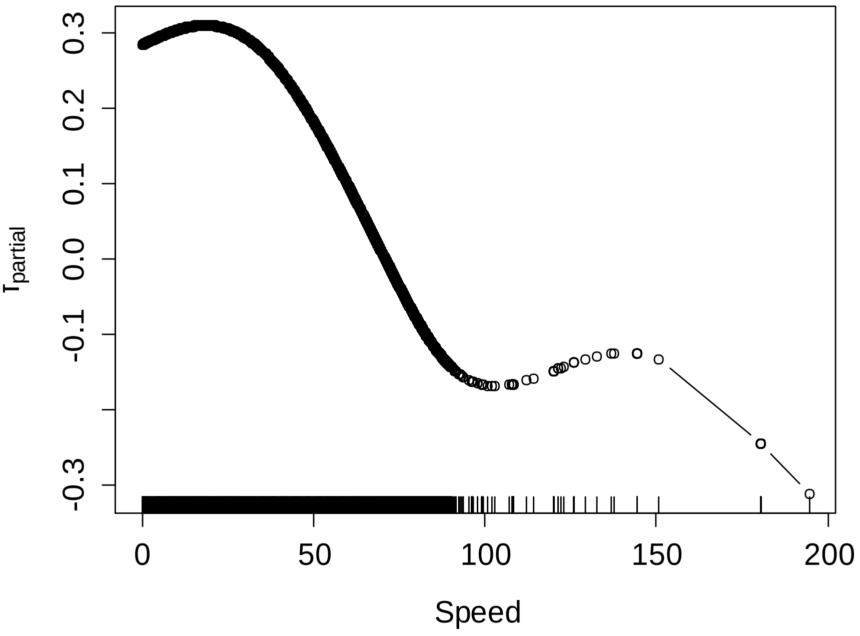

Figure 15: Partial Effects of speed on fuel consumption

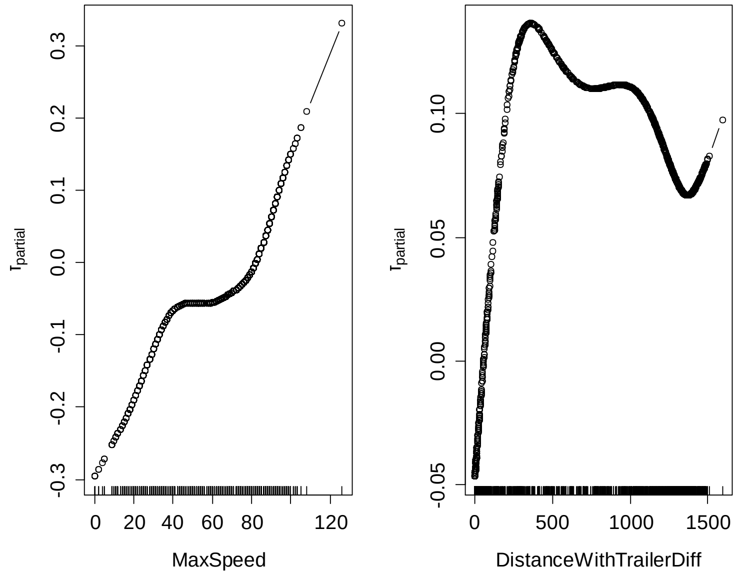

Figure 16: Partial effects of Maximum speed and distance with trailer on Fuel Consumption

Figure 17 also interprets the partial effect of input variables green band driving and overreving on the target variable fuel consumption. As it is evident from the graph that for higher values of the input variable overreving there is existence of a nonlinear relationship with the target variable

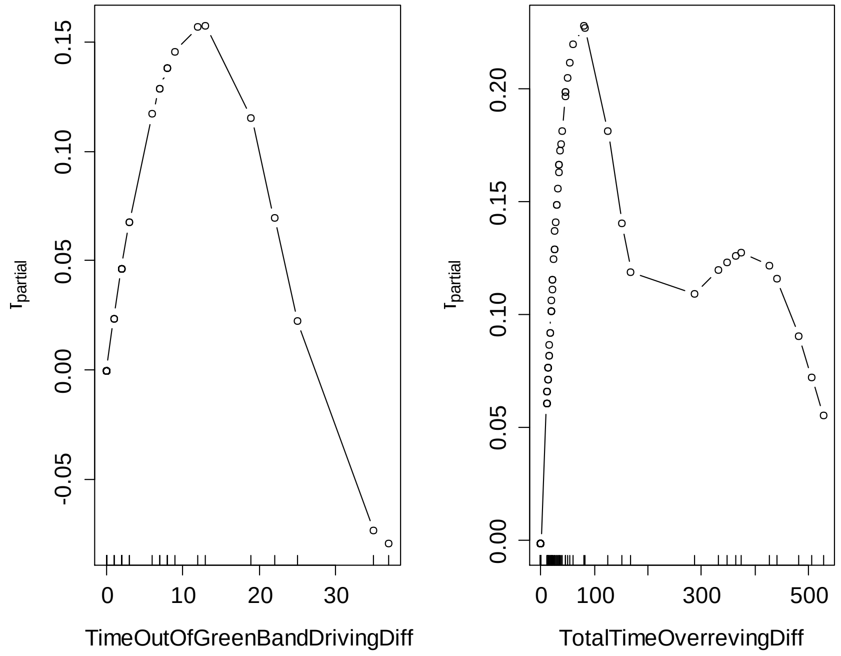

Figure 17: Partial effects of Green band driving and Overreving on Fuel Consumption

It’s hard to interpret the importance of variable from the plot, hence each variable and its selection frequency has been mentioned in Table 7. The gradient boosting method chooses only the variables mentioned in Table 7 for predicting the fuel consumption of heavy vehicles. These variables are given higher importance in random forests method as well.

Table 7: GAM boosting- Selection frequencies

Variables	Selection frequencies
Speed	0.365
Coasting	0.185
Max speed	0.170
Distance with trailer	0.125
Distance with CC	0.065
Overreving	0.050
Over speeding	0.040

Figure 18, shows a density plot of observed vs. predicted values of the *gamboost* method. As seen in Figure 11 the predictive distribution has peaks and troughs almost at the same values as the original data. The gradient boosting method predicts the high and low values of fuel consumption using the input variables clearly, but when compared with random forest prediction model in Figure 11, the latter one seems better.

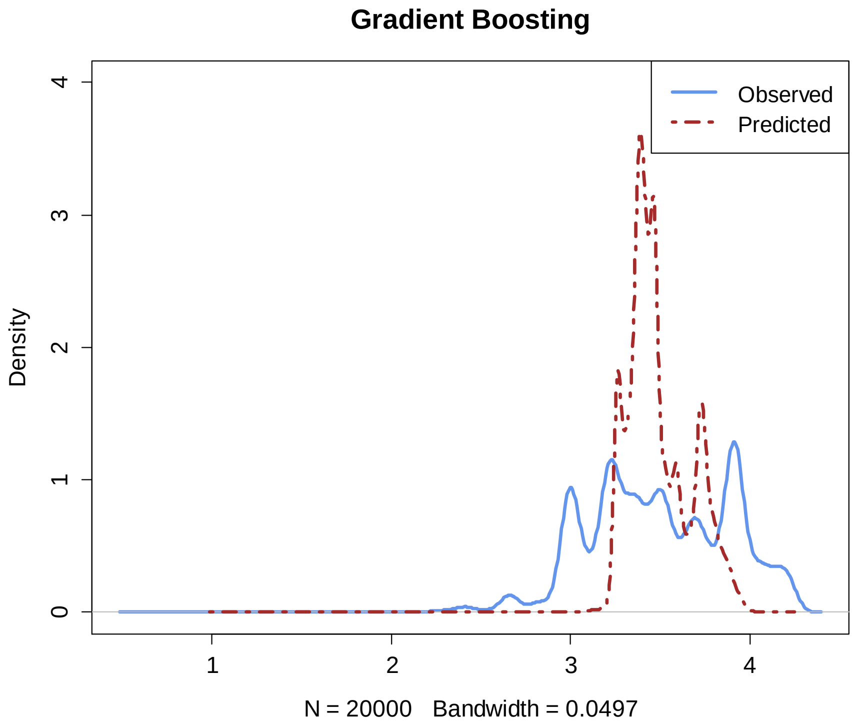

Figure 18: Density plot of observed vs. predicted values

For knowing in detail about the reason why the predictive fit of gam boost is not as good as random forest model, histogram of residuals was plotted. Figure 19 shows the distribution of residuals of gradient boosting method. The residuals are higher in this method compared to the random forest method. The residual distribution has a heavier left tail.

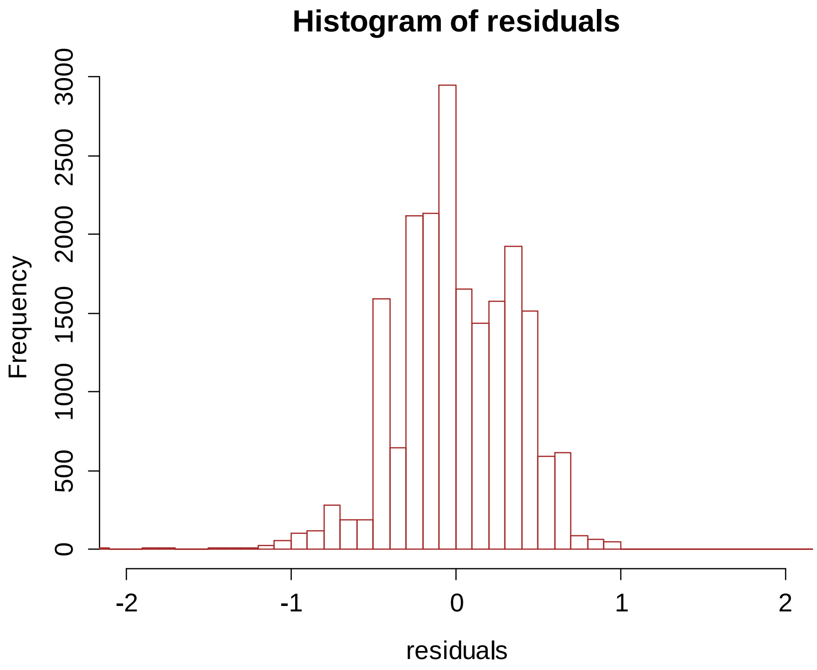

Figure 19: Histogram of residuals

Finally, predictive efficiency of the model is estimated using Nash- Sutcliffe model efficiency. The value obtained is 0.698, which is lesser than the efficiency of random forest method.

Nash Sutcliffe efficiency of random forest method =0.698

5 Discussion and conclusions

For achieving energy optimized autonomous vehicles with high intelligence and efficiency, it is important to study the reasons that contribute to the fuel consumption of heavy vehicles. The supervised learning methods, random forest and gradient boosting were used throughout this thesis work to build a predictive model that predicts the target variable, fuel consumption.

Some problems were encountered due to high dimensional data. There were huge number of observations and input variables. The original distribution of the input variables was not known. Since the data is accumulated there were relationships between predictor and target variables. Increased number of observations exhibited problems with pattern recognition systems.

Splitting the whole data into training and test set, made the analysis easier. Training set came in handy to perform analysis of data using different data mining techniques. Test set was used to measure the efficiency of the techniques used. Random forests technique was first performed since it is easier to use understand and interpret. It produced very good results, but to check if a better prediction can be achieved, gradient boosting was also performed.

5.1 Random forests

Importance of input variables in predicting the target variable was measured for the random forests. Speed was considered as the most important variable that affects the fuel consumption of heavy vehicle, followed by distance with trailer, distance with cruise control active, maximum speed and coasting parameters exhibiting relatively higher importance in predicting fuel consumption.

The parameters number of harsh accelerations and green band driving had the lowest importance in predicting the target variable fuel consumption. This concludes that, the harsh acceleration or the green band driving doesn’t affect the fuel consumption greatly.

The efficiency of the performed model was analyzed using squared error rate. The mean squared error of the model is consistently very low all through the process (even when the tuning parameters has been changed), which proves that the efficiency of the random forest model is very good.

The most important advantage of random forest model is that pruning of each tree is not done, because for prediction, pruning eliminates the negative effects of over fitting. An added advantage of the random forest is that it selects *mtry* variables at first and then update the variables, which makes the model more accurate.

5.2 Gradient boosting

Gradient boosting is also a good prediction method that optimizes the prediction accuracy to produce good variable selection of input variables that could predict the target variable. Since gradient boosting doesn’t need users to predefine complex functions, it is user friendly.

Gradient boosting also produces same results as the random forests. The important variables that affects fuel consumption of heavy vehicles according to gradient boosting method are speed, coasting, maximum speed, distance with trailer, distance with cruise control, over speeding and overreving. The other parameters green band driving, harsh accelerations and brake applications have not been considered in predicting fuel consumption of heavy vehicles. Apparently, these variables don’t affect the fuel consumption to a great extent.

The best fit of the gradient boosting model was analyzed using minimum description length (MDL) (Hofner *et al.*, 2012). The method is supposed to be reliable as long as MDL is very low. MDL of the designed gradient boosting model is between 5.10 and 5.20 for different number of iterations, which concludes that prediction done by gradient boosting is also good.

Partial effects of the input variables on the target variables show that the smooth effects for variables is reasonable, which means, there exists nonlinear relationship between input and target variables.

5.3 Evaluation

Each method has been evaluated based on the Nash-Sutcliffe efficiency measure. Both the methods proved good for this data. The predictive efficiency of random forest was 0.808 which showed that it could provide a better prediction of fuel consumption using the variables selected by the model compared to gradient boosting method, which produced a predictive efficiency of 0.698.

The residual is the difference between predicted and observed value. A model is considered to be very good for a particular data if it produces predicted values very close to observed values leading to very small values of residuals.

Both the suggested models produced low residuals, but still, the residuals of random forest were comparatively lower than the residuals of gradient boosting model. The lower predictive accuracy of gradient boosting model over random forest model is the main reason behind the higher residual values.

Despite the fact that gradient boosting produced lesser efficiency compared to random forest, it produces more interpretable results over random forest. The main disadvantage with random forest method is its “black box” nature, where the internal working of the model is invisible and is just defined by its input and output variables, whereas gradient boosting produces prediction rules very similar to other statistical models making its working understandable and interpretable.

It is evident from the above results and discussions, that parameters- speed, coasting, distance with trailer, distance with cruise control and maximum speed are the factors that affect the fuel consumption of heavy vehicles. Controlling these parameters can ensure lower fuel consumption of heavy vehicles.

6 References

Berry M.I. (2010). The Effects of Driving Style and Vehicle Performance on the Real-World Fuel Consumption of U.S. Light-Duty Vehicles.

Bratt H., Ericsson E. (1998). Measuring vehicle driving patterns- estimating the influence of different measuring intervals.

Breiman L. (1998). Arcing classifiers (with discussion). The Annals of Statistics 26, 801-849.

Breiman L. (2001). Random Forests. Machine Learning, Volume 45, Issue 1, pp 5-32.

Buhlmann P., Yu B. (2003). Boosting with the L2 loss: Regression and classification. Journal of the American Statistical Association 98, 324-339.

Constantinescu Z., Marinoiu C., Vladoiu M. (2010). Driving style analysis using data mining techniques. Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844 Vol. V, No. 5, pp.654-663

Friedman J.H., Hastie T., Tibshirani R. (2000). Additive Logistic Regression: A statistical view of boosting (with discussion). The Annals of statistics 28, 337-407.

Geng X., Arimura H., Uno T. (2012). Pattern mining using trajectory GPS data. IIAI International Conference on Advanced Applied Informatics.

Genuer R., Poggi M.J., Malot T.C. (2010). Variable Selection using Random Forests.

Hansen H.M., Yu B. (2011). Model selection and the principle of minimum description length. Journal of the American Statistical Association.

Hastie T., Tibshirani R. (1990). Generalized Additive Models.

Ho T. K. (1995). Random Decision Forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC.

Ho T. K. (1998). The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 8.

Hofner B., Mayr A., Robinzonov N., Schmid M. (2012). Model-based Boosting in R: A Hands-on Tutorial Using the R Package mboost. Computational Statistics, Volume 27, Issue 4.

Hothorn T., Bühlmann P., Kneib T., Schmid M., Hofner B. (2010). Model-based Boosting 2.0. Journal of Machine Learning Research, Volume 11.

Hothorn T., Schmid M. (2008). Boosting additive models using component-wise P-Splines. Computational Statistics and Data Analysis 53 (2008) 298-311.

Krause P., Boyle DP. (2005). Comparison of different efficiency criteria for hydrological model assessment. Advance in geosciences; 5:89-97.

Liaw A., Wiener M. (2002). Classification and Regression by randomForest, web education in chemistry.

Lin Y., Jeon Y. (2002). Random Forests and Adaptive Nearest Neighbors. Journal of the American Statistical Association.

Maimon O., Rokach L. (2010). Data Mining and Knowledge Discovery Handbook. Second edition, Springer, 1305 p.

Montillo A. (2009). Random Forests. Guest lecture at Temple University.

Vapnik, V. N (2000). The Nature of Statistical Learning Theory (2nd Ed.), Springer Verlag.
Yun C., Fu C., Hung F., Jie Lin D (2010). The applications of data mining technologies on commercial vehicle GPS data- A case study on Formosa Plastics transport Corp., Taiwan.
Scania Inline, retrieved in 2013 from www.scania.inline.com

7 Appendix

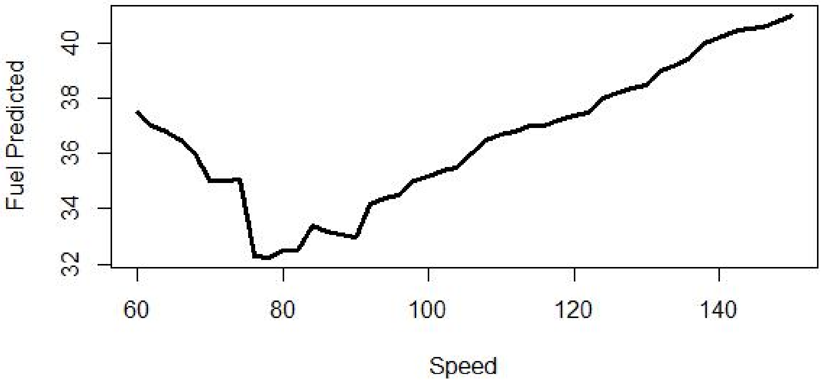

Figure 20: Plot of speed against predicted fuel consumption

The plots in figure 20 and figure 21 show the dependence between input variables speed, coasting and the predicted fuel consumption of heavy vehicles. The plots interpret that speed of heavy vehicles should be optimized and maintained in a range between 78km/hr and 82km/hr and also, increased coasting can result in lesser fuel consumption.

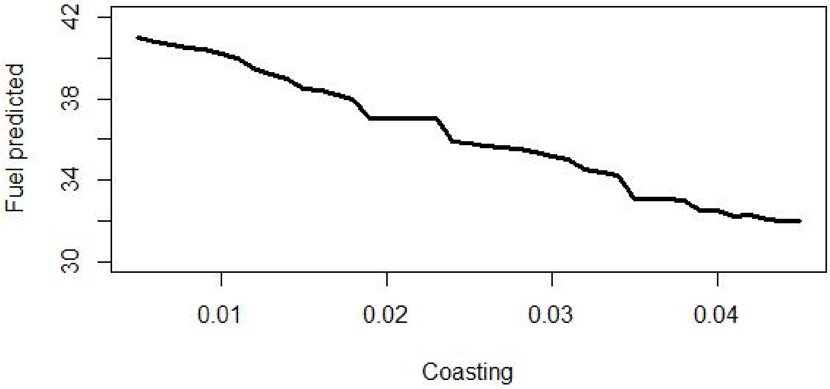

Figure 21: Plot of coasting against predicted fuel consumption

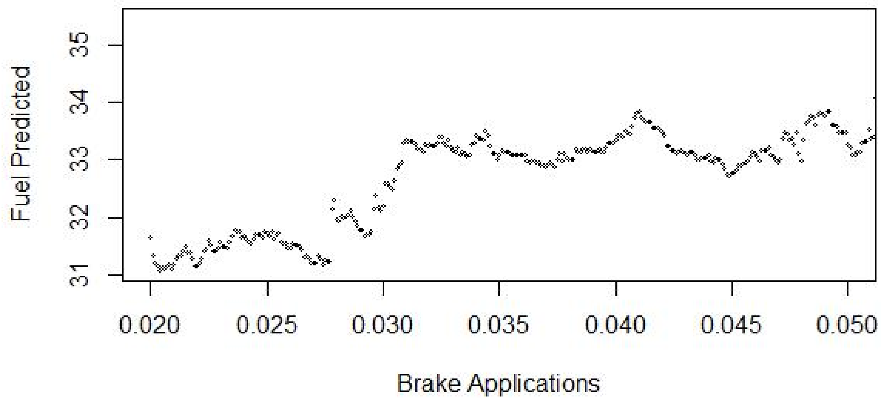

Figure 22: Plot of brake applications against predicted fuel consumption

Figure 22 shows the plot of brake applications against predicted fuel consumption for new set of values of brake applications. Figure 23 shows the partial effects of input variables distance out of gear and distance with cruise control on fuel consumption.

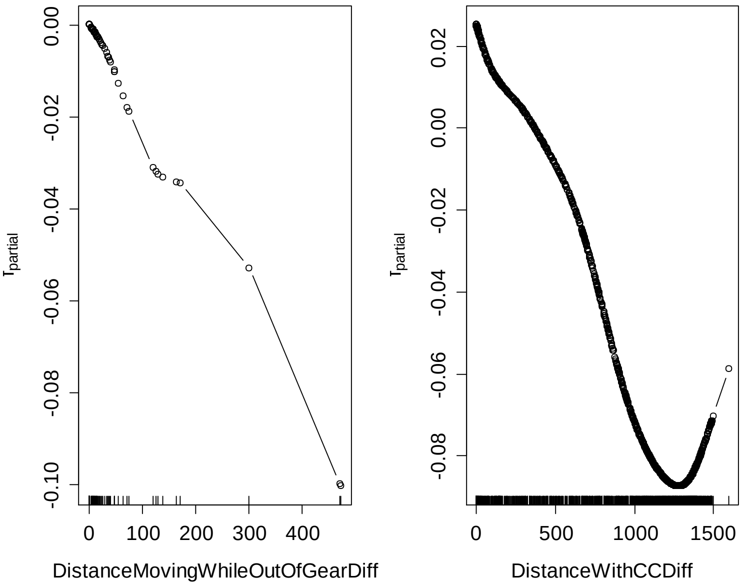

Figure 23: Partial effects of distance out of gear and distance with cruise control on fuel consumption

LiU-IDA-008-SE