Analysis of Variable Importance Measurement Techniques for Classification of Road Surfaces

Anupama Jawale(✉) and Ganesh Magar

Post Graduate Department of Computer Science, SNDT Women's University, Mumbai 400049, India
anupama.jawale26@gmail.com

Abstract. The term variable importance refers to the role of an attribute in making accurate predictions. A particular model, when relies majorly on multiple variables, increases variable importance of those variables in positive direction. Variable importance is applied to various classification and regression models using different methods. For example, in regression model, higher value Root Mean Squared Error (RMSE) is the indicator of high importance to that variable, whereas in classification model, higher number of splits associated with a variable determines its importance in the model. In this research study, we have considered a problem of road surface classification depending upon 17 variables associated with vehicle parameters. This is a multiclass classification problem. Different classification and regression models are used, and variable importance of each model is evaluated on the metrics like RMSE, Goodness of fit model. Outcome of this research study shows all models define a common set of 5 to 7 higher importance variable rankings to predict dependant variable.

Keywords: Classification · Decision Trees · Regression · Variable Importance

1 Introduction

Variable importance and its various measures are essential outcome of exploratory data analysis. Variable importance assesses role of a variable in prediction. It helps in improvement of overall performance of predictive model, classification or regression. Variable importance analysis is performed on the dataset. This analysis gives some useful insights about data, for example;

1. To know which variables in the dataset are important for the model – Variables those are not important for model prediction performance can be excluded.
2. For a particular variable, to learn how does it influence model's prediction – Assessing influence of the variable is helpful for examining validity of the model for specific domain
3. To learn if any specific combination of variables or set of observations cause incorrect prediction or is responsible for decrease in accuracy – This may result to generation of new factors or new models.

R. Manza et al. (Eds.): ACVAIT 2022, AISR 176, pp. 521–537, 2023.
https://doi.org/10.2991/978-94-6463-196-8_40

4. Comparison of different models based on variable importance – Useful for performance benchmarking of various models

Methods of variable importance measurement are divided into two groups, viz; model-specific and model-agnostic.

1.1 Model Specific Metrics

1.1.1. Linear Models – Linear models describe y as a continuous variable function of x_i predictor variables. The most common measure for variable importance is t-statistics test. T-statistics is defined as a ratio of the difference of the estimated value of a parameter from its hypothesized value to its standard error. Equation 1 provides formula for computation of t-test.

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{s.e.\left(\hat{\beta}\right)} \tag{1}$$

where $\widehat{\beta}$*is estimator for parameter* β *and* $s.e.\ (\widehat{\beta})$ *is standanrd error of estimation*

In linear models, absolute value of t-statistics is used for each model parameter.

1.1.2. Random Forests and decision tress - For each tree, prediction accuracy for out of bag data which is left out observation form the bootstrap training set. For regression, Mean Square Error (MSE) is computed on this data for each tree and same is computed by permuting a variable. The difference is normalized with standard error. Equation 2 shows computation of MSE in regression or classification trees

$$\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y_i} - Y_i\right)^2 \tag{2}$$

where $\widehat{Y}$ *is predicted value*, Y *is true value label and n is number of observations*

1.1.3. Partial Least Squares (PLS) – PLS models use weighted sum of regression coefficients as metric of variable importance. Weights are multipliers used to decrease sum of squares across various PLS variables. To find most important variables can be seen as binary classification problem. According to researcher [1] sensitivity and specificity are considered basic measures of accuracy for a classification task are obtained from confusion matrix.

1.1.4. Recursive Partitioning – Reduction in mean square error is a loss function of recursive partitioning. This function is returns certain value at each partition of tree and then summed up. Along with loss function value, upper competing variables are also added at each split and recorded. Total of these values are used to compute variable importance.

1.1.5. Bagged Trees and Boosted Trees – These models implement same technique as in recursive partitioning method. In bagged trees, total variable importance is computed for all bootstrapped trees, whereas in boosted trees, total variable importance is returned for each boosting iteration.

1.1.6. Multivariate Adaptive Regression Splines (MARS) – MARS models proposed by Friedman, use backward elimination for reduction in cross validation error estimate. This algorithm creates piece wise linear model, sums up GCV error of estimation for each predictor variable. Total loss function given by total reduction of GCV values is used for assignment of variable importance.

1.2 Model – Agnostic Methods

1.2.1. Intuition [2] – The idea behind this method is to calculate change in a models' performance with effect of removal of selected variables. Resampling or permutations are used to remove effect of removal. This is another version of the idea of variable importance measure for random forest [3]. If the important variable is removed, performance of the model will be decreased.

Loss function computation is given by Eq. 3.

$$L^{*j} = L\left(\hat{y}^{*j}, X^{*j}, y\right) \tag{3}$$

Where $\hat{y}^{*j}$ Model Prediction based on X^{*j}, y is the true label value.

In this research study, we have discussed and compared various variable importance measurement methods to generate different metric. Methods understudy are RandomForest, Partial Least Square, Bagging and Multivariate Adaptive Regression Splines (MARS).

2 Background

The latest study for Variable importance proving it as a useful tool for larger datasets and multi-objective optimization [4], implements differential evolution algorithm on importance ranking basis. Reduction in variable dimensionality has been tried by researchers including various feature selection and feature extraction techniques [5–9]. However, variable importance measurement is the technique that can be applied to dataset before any feature engineering. The advantage is to reduce the pre-processing complexity. As the machine learning accuracy highly dependant upon this stage, for imbalance datasets, accuracy performance becomes a challenging task. Variable importance measurement techniques described in [10], highlights usage of imbalance dataset and shows out performance of proposed method. To achieve high dimensional selection consistency in decision tree algorithm, researchers have presented model selection algorithm named DSTUMP that outperforms in nonlinear additive model settings [10].

There are few research studies carried out on variable importance measurement metrics. Basically model specific variable importance is carried out for various models. A popular model among all is a Random Forest model [11, 12]. The reason behind is only random forests model with conditional inference trees provide unbiased variable importance [13]. In the research study [14], researchers have worked on variable ranking using Mean Decrease in Accuracy (MDA) and Mean Decrease in Gini (MDG) using *random forest*. They have concluded that both the measures are different even if same model is used. They have suggested to use *randomforest* model only once in order to

select variable importance based on ranking. Another popular model for classification is Partial Least Square Regression (PLS). To select relevant important predictors, PLS regression coefficients are investigated using Receiver Operating Characteristic (ROC) analysis. The comparison of various variable selection methods for PLS have shown that variable importance methods have outperformed other methods [1].

Variable contribution can be computed at part stages of a model also. For example, research study [15] suggests variable contribution computation at each stage of a multistage process using random forest regression and a new measure of conditional permutation metric. This combination is further used to quantify local contribution of each variable, which is then integrated to calculate global quantification. Even in muti objective optimization problem, variable importance play role in enhancing accuracy of the model [16]. However, it is also observed that if number of ranks in overall observations is small then importance of variable is not predicted accurately. There are many experiments in defining variable importance mechanism. Variable importance are computed on the basis of similarity between margin distributions prior to random permutation and after that [17]. Researchers have noticed more stability in computation of variable importance with this methodology. Based on variable importance calculation, some new feature selection approaches are also presented. Importance combined with prior knowledge parameters to select features, when applied to soft measuring model, these features have shown increase in performance, as stated in [18]. Similarly permutation based framework, a dissimilarities based algorithm is proposed by [10] researchers that computes variable importance using distribution of misclassification errors. In the area of image classification, researchers have proposed method of quantifying of variable importance, employing concept of game theory and metric of Shapely value which is applicable to any type of model [19]. Researchers in research study [14] has presented systematic approach that computes variable importance using optimal number of runs. Shapely method refers to the concept of treating every variable as a player in collaborative game, where the objective of maximum accuracy is followed.

As observed in above theoretical survey, we can conclude that variable importance is not a stable input and is very much model specific. There is a scope of further research in model specific behaviour of variables importance measure. This research study focuses on computation of variable importance for different models like regression trees, random forest, bagged trees and multivariate adaptive regression splines – MARS models on the basis of relative metric of accuracy of these metrics. Next sections describe different algorithmic models and experimental setup of this research study.

3 Methodology

This research study focuses on five different methods of classification/regression and tries to analyse different variable importance ranking for all theses methods. Methods under consideration are explained as below-

i. Recursive partitioning and regression tree model- this model performs successive binary partitions on the basis of various predictor variables. Once the partition is made, prediction on the basis of average of depended variable y in each partition can be made.

This technique is applicable to both classification and regression [20]. For classification, predicted value is based on the formula

$$max(Pr(Y = s|XA_k)) \tag{4}$$

Here, A is a class of road surface conditions (Y = roadSurface), having 3 different values (1 = SmoothCondition, 2 = FullOfHolesCondition, 3 = UnevenCondition)

Pr is probability value, s is a split or partition of a tree.

ii. Random forest model – The idea behind this model is to grow many classification trees on the basis of probabilistic scheme. Classify the new observation from predictor variable by putting it down each of the tree. The tree votes for that observation thus giving its classification [21].

iii. Bagging – The name bagging stands for bootstrap aggregating. Bagging generated multiple versions of predictors and aggregate them in later phase. These multiple versions are formed by replicating bootstrap of learning sets and using them as new learning sets [22]

We can formulate bagging as numerical equation, when used for classification. Consider Eq. 5 given below for a predictor $\varphi(x, L)$ predicts the class label j $\epsilon(1, 2, \ldots\ldots..J)$.

$$Q(j|x) = P(\emptyset(x, L) = j) \tag{5}$$

where P is the probability, $\varnothing$ is the function to predict class label and Q is one of the many replicates of learning set.

iv. Partial Least Squares regression model (PLS) – This model is popular in the situations wherever we have multiple, possibly correlated predictor variables [23]. As shown in the Table 1, we have a dataset of 17 variables, possibly correlated with each other for prediction of road surface condition class. PLS model could be the ideal model to solve this type of problem.

v. Multivariate Adaptive Regression Splines (MARS) model – This approach adopts non linearity of polynomial regression by evaluating cut points that are similar to step functions. Mathematical function of MARS model is shown in below equation

$$\hat{f}(x) = \sum_{i=1}^{k} C_i B_i(x) \tag{6}$$

All of the above models have their own set of variable importance ranking. In this research study we have computed variable importance for all the models and compared them. The methodology used in this research study is described in Fig. 1.

i. Data Cleaning and Preparation

The dataset [24] is collected using various sensors using vehicle running in different road conditions. There are missing values due to hardware failure and/or bad road or other conditions. The very first step includes removal of missing values.

ii. Dataset Exploration

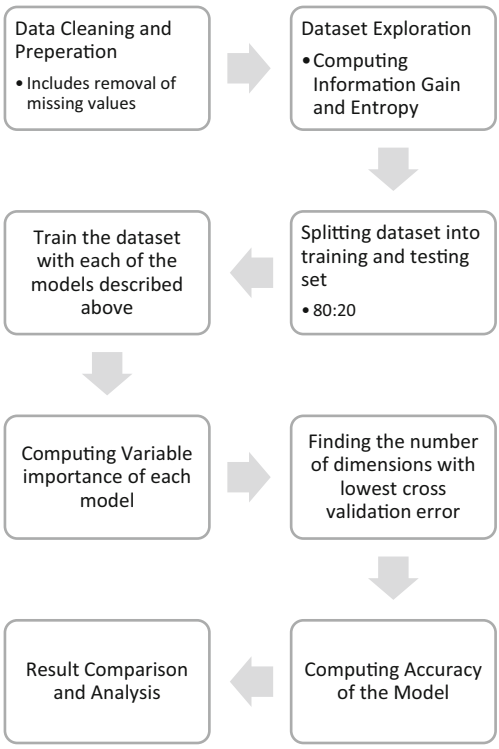

Fig. 1. Flow of Methodology

The dataset used in this research study has dataset with 17 variables. We have computed Entropy and Information gain from the dataset. This step explores intricate nature of dataset. The concept of entropy introduced by Shannon in 1948 is used to quantify the information contained inside the variable. For the datasets used for classification task, entropy is used to determine how balanced the dataset is. In this research study, we have obtained the entropy value 1.547135. The dataset has 3 classes defining road surface condition. Decision Tree algorithms use entropy to calculate information gain at each split to decide variable importance. Table 1 shows information gain for every variable.

iii. Splitting dataset
With 80:20 training – testing proportion, dataset is split for machine learning. Dataset is shuffled to maintain adequate weightage to every class.

iv. Training Dataset with Models
The dataset is trained with five different methods as described earlier in this section.

v. Variable Importance computation
Variable importance calculation for Recursive partitioning is computed using reduction in loss function, like mean squared error. The reduction value is calculated for

Table 1. Information Gain Computation

Variable	Information Gain
IntakeAirTemperature	0.843938644
EngineCoolantTemperature	0.694697585
ManifoldAbsolutePressure	0.369650062
MassAirFlow	0.30801103
VehicleSpeedAverage	0.265671043
EngineRPM	0.194717111
VehicleSpeedInstantaneous	0.165999861
FuelConsumptionAverage	0.149185424
EngineLoad	0.110725003
AltitudeVariation	0.038622989
LongitudinalAcceleration	0.031212503
VehicleSpeedVariation	0.017425414
VerticalAcceleration	0.015265343
VehicleSpeedVariance	0.009659753

each variable and each split and then tabulated to add up. For Random Forest, Gini Importance (mean decrease in impurity) is used to calculate variable importance. Higher value of node probability (calculated as Number of samples to reach that node / Total number of samples) indicates high importance of the feature. In bagged tree models, the variable that appears frequently in splitting function and decrease in squared error is considered for importance. In Partial Least Squares, the variable importance is calculated based on weighted sum of absolute regression coefficients. Weight is computed as reduction of sum of squares across all Partial Least Squares Variables. In MARS models, reduction in generalized cross validation estimate of error is considered to calculate variable importance. The total reduction is used as variable importance.

vi. Finding number of dimensions with lowest cross validation error
Minimum of Root mean squared error of prediction (RMSEP) is used to find out best dimensions. From different estimators, minimum value is taken t consider best predictor variable.

vii. Comparison of Accuracy of various models
Accuracy on test dataset is computed for all models with set of best dimensions

viii. Result Comparison and analysis
Variable Importance obtained from all models understudy are presented in Sect. 5.

Experimental setup and dataset description is covered in the below section.

Table 2. (a): Summary of Dataset

Altitude Variation	**Vehicle Speed Instantaneous**	**Vehicle Speed Average**
Min.:-9.900e + 09	Min.:0.000e + 00	Min.:0.000e + 00
1st Qu.:-1.100e + 09	1st Qu.:4.050e + 02	1st Qu.:1.250e + 09
Median:-2.000e + 00	Median:1.890e + 09	Median:2.788e + 09
Mean:-1.405e + 08	Mean:2.617e + 09	Mean:3.452e + 09
3rd Qu.: 6.000e + 08	3rd Qu.:4.140e + 09	3rd Qu.:4.656e + 09
Max.: 9.600e + 09	Max.:9.900e + 09	Max.:9.999e + 09
NA's:63	NA's:9	NA's:415
VehicleSpeedVariance:	**VehicleSpeedVariation**	**LongitudinalAcceleration**
Min.:0.000e + 00	Min.:-9.900e + 09	Min.:-14576
1st Qu.:1.455e + 09	1st Qu.:-9.000e + 08	1st Qu.: 2972
Median:2.510e + 09	Median: 0.000e + 00	Median: 10187
Mean:3.424e + 09	Mean: 3.613e + 07	Mean: 9971
3rd Qu.:5.005e + 09	3rd Qu.: 9.000e + 08	3rd Qu.: 15703
Max.:9.999e + 09	Max.: 9.900e + 09	Max.: 39798
NA's:415	NA's:78	
EngineLoad	**EngineCoolantTemperature**	**ManifoldAbsolutePressure**
Min.:0.000e + 00	Min.: 8.00	Min.: 88.0
1st Qu.:1.098e + 09	1st Qu.:51.00	1st Qu.:103.0
Median:3.412e + 09	Median:79.00	Median:106.0
Mean:4.134e + 09	Mean:65.89	Mean:114.6
3rd Qu.:7.294e + 09	3rd Qu.:79.00	3rd Qu.:124.0
Max.:9.961e + 09	Max.:86.00	Max.:170.0
NA's:5 NA's:5	NA's:5	
EngineRPM	**MassAirFlow**	**IntakeAirTemperature**
Min.: 0	Min.:3.490e + 08	Min.: 7.00
1st Qu.: 1476	1st Qu.:1.699e + 09	1st Qu.:21.00
Median: 7295	Median:2.438e + 09	Median:35.00
Mean: 8392	Mean:2.970e + 09	Mean:32.82
3rd Qu.:14965	3rd Qu.:4.020e + 09	3rd Qu.:41.00
Max.:28025	Max.:9.990e + 09	Max.:65.00
NA's:5	NA's:5	NA's:5
FuelConsumptionAverage	**roadSurface**	**traffic**

(*continued*)

Table 2. (*continued*)

Altitude Variation	**Vehicle Speed Instantaneous**	**Vehicle Speed Average**
Min.:1.230e + 06	Min.:1.000	Length:8614
1st Qu.:1.212e + 09	1st Qu.:1.000	Class:character
Median:1.454e + 09	Median:2.000	Mode:character
Mean:2.997e + 09	Mean:1.925	
3rd Qu.:2.035e + 09	3rd Qu.:3.000	
Max.:9.999e + 09	Max.:3.000	
NA's:96		
VerticalAcceleration	**drivingStyle**	
Min.:-27631.0	Length:8614	
1st Qu.: -9796.5	Class:character	
Median: -5357.0	Mode:character	
Mean: -5721.8		
3rd Qu.: -761.2		
Max.: 9999.0		

4 Experimental Setup

In this research study we have taken a dataset of accelerometer sensor from Kaggle [25] for study of safety driving and road condition analysis. There are 17 variables and 8614 objects in the dataset. Table 2 (a) shows summary of the dataset.

We have used VIP package from R Library for variable importance analysis and graphical representation. These computations are model specific. However, all computations can not be compared on the basis of same parameter like accuracy. For example RMSE based importance calculations can not be directly compared with tree models' accuracy score or t-statistics values of linear model [26]. Findings and interpretation of these findings are described in below section.

5 Results and Interpretation

Experimental results for computation of Variable importance are sorted from largest to smallest order and are presented in below Tables. All the models have shown variable sets of important variables, keeping few common. Graphical representation of the same is also shown in Fig. 2.

Fig. 2. (a) – Top 6 feature variables for.Recursive Partitioning and Regression Tree – IntakeAirTemperature, ManifoldAbsolutePressure, EngineCoolantTemperature, VehicleSpeedAverage, EngineRPM, FuelConsumptionAverage

Table 3. (b): Variable importance rankings obtained for different models

a.) Recursive Partitioning and Regression Tree	
Variable	**Importance value**
IntakeAirTemperature	4847.469
ManifoldAbsolutePressure	3293.519
EngineCoolantTemperature	3063.557
VehicleSpeedAverage	1744.039
EngineRPM	1534.012
FuelConsumptionAverage	1012.668
MassAirFlow	537.7126
VehicleSpeedInstantaneous	314.968
VerticalAcceleration	290.1684
AltitudeVariation	102.9991
VehicleSpeedVariance	84.99847
LongitudinalAcceleration	44.21902
VehicleSpeedVariation	0
EngineLoad	0
b) Random Forest Model	
Variable	**Importance value**
FuelConsumptionAverage	47.71845
EngineCoolantTemperature	42.13896
IntakeAirTemperature	38.52865
ManifoldAbsolutePressure	36.4526
VehicleSpeedAverage	35.49989
VehicleSpeedVariance	28.15275
VerticalAcceleration	28.09975
LongitudinalAcceleration	25.96866
VehicleSpeedInstantaneous	24.78347
AltitudeVariation	21.00622
MassAirFlow	19.18455
EngineRPM	17.17326
VehicleSpeedVariation	14.79859
EngineLoad	13.42231
c) Partial Least Square Regression Model	

(*continued*)

Table 3. (*continued*)

a.) Recursive Partitioning and Regression Tree	
Variable	**Importance value**
Variable	**Importance Value**
ManifoldAbsolutePressure	0.0185
EngineCoolantTemperature	0.0133
IntakeAirTemperature	0.00161
EngineRPM	0.00001
LongitudinalAcceleration	0.00001
VerticalAcceleration	0.00001
EngineLoad	0
VehicleSpeedAverage	0
FuelConsumptionAverage	0
AltitudeVariation	0
VehicleSpeedInstantaneous	0
VehicleSpeedVariation	0
MassAirFlow	0
VehicleSpeedVariance	0
d) Bagging Model	
Variable	**Importance Value**
FuelConsumptionAverage	2.361472
IntakeAirTemperature	2.232337
VehicleSpeedAverage	1.571968
ManifoldAbsolutePressure	1.332739
VerticalAcceleration	1.190778
EngineCoolantTemperature	1.178852
MassAirFlow	0.846832
VehicleSpeedVariation	0.534877
VehicleSpeedInstantaneous	0.511967
LongitudinalAcceleration	0.477734
AltitudeVariation	0.3921
EngineRPM	0.391357
VehicleSpeedVariance	0.195058

(*continued*)

Table 3. (*continued*)

a.) Recursive Partitioning and Regression Tree	
Variable	**Importance value**
EngineLoad	0.04444
e) Multivariate Adaptive Regression Splines (MARS) Model	
Variable	**Importance Value**
ManifoldAbsolutePressure	18
EngineCoolantTemperature	18
IntakeAirTemperature	18
FuelConsumptionAverage	16
VehicleSpeedAverage	15
VerticalAcceleration	12
MassAirFlow	7
AltitudeVariation	5
VehicleSpeedInstantaneous	3
VehicleSpeedVariance	0

Table 4. Performance Parameter

	Decision Tree	Random Forest	Partial Least Square Regression	Bagging	Multivariate Adaptive Recursive Spline
Accuracy (All)	93.20%	97.30%	75.80%	22.20%	25.20%
Accuracy (Top 6)	93.83%	98.29%	81.17%	19.28%	24.83%

Fig. 2. (b) – Top 6 feature variables for Random Forest – FuelConsumptionAverage, EngineCoolantTemperature, IntakeAirTemperature, ManifoldAbsolutePressure, VehicleSpeedAverage, VehicleSpeedVariance

Fig. 2. (c) – Top 6 feature variables for Partial Least Square Regression Model – ManifoldAbsolutePressure, EngineCoolantTemperature, IntakeAirTemperature, EngineRPM, LongitudinalAcceleration

Fig. 2. (d) – Top 6 feature variables for Bagging Tree Model – FuelConsumptionAverage, IntakeAirTemperature, VehicleSpeedAverage, ManifoldAbsolutePressure, VerticalAcceleration, EngineCoolantTemperature

Figure 2 (e) – Top 6 feature variables for Multivariate Adaptive Regression Splines (MARS) Model – ManifoldAbsolutePressure, EngineCoolantTemperature, IntakeAirTemperature, FuelConsumptionAverage, VehicleSpeedAverage, VerticalAcceleration.

From Table 1, we can observe there are some variables with low information gain, like EngineLoad, VehicleSpeedVarience etc. We can drop such variables from our machine learning models to save computational resources. Similarly, from Table 2 (b), we can observe that certain variables with very low importance can be omitted from the machine

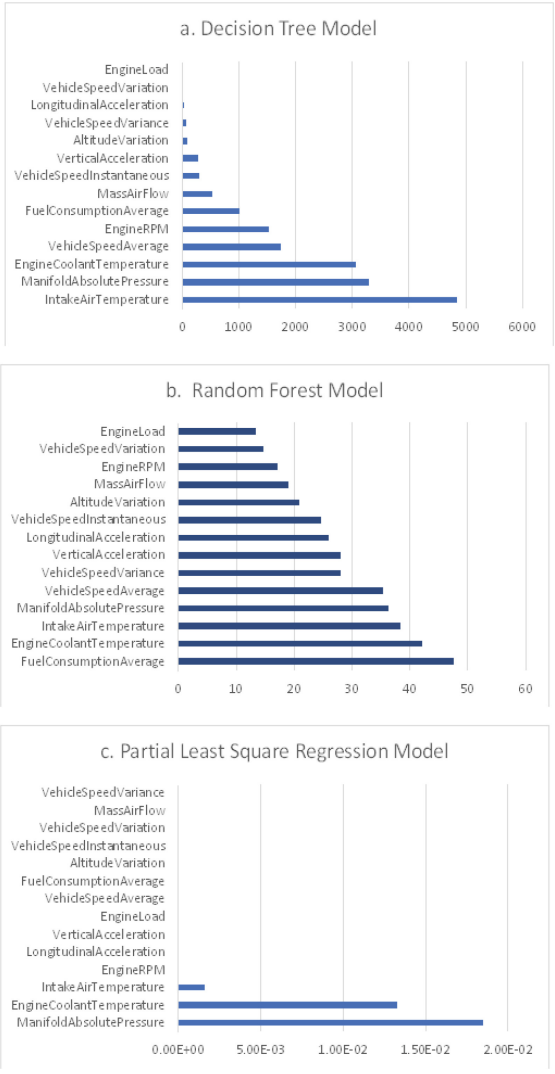

Fig. 2. Variable Importance graphs for different model

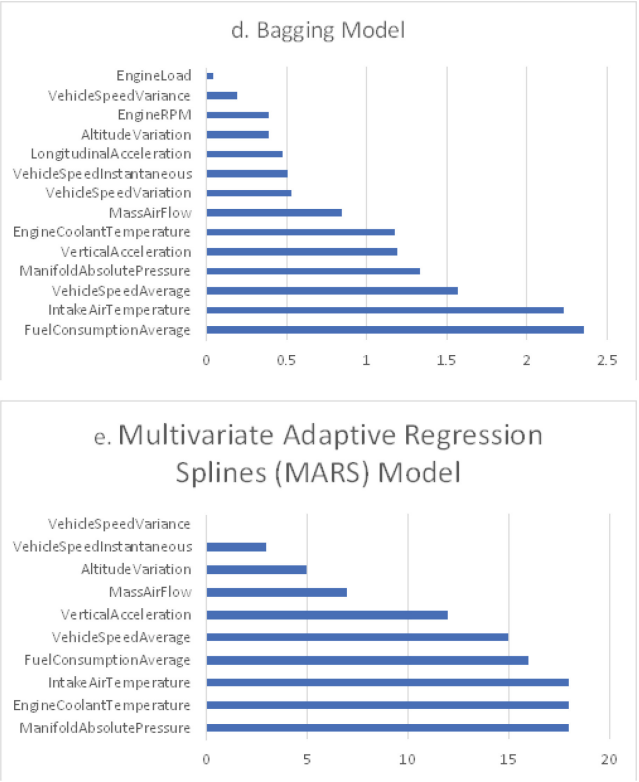

Fig. 2. (*continued*)

learning model. We also observe that, even though variable analysis ranking is model specific, all models have ranked *FuelConsumptionAverage, EngineCoolantTemperature, IntakeAirTemperature, ManifoldAbsolutePressure and VehicleSpeedAverage* as of higher importance. On the basis of these results, we can reform our predictive model formula as Y ~ 6 most important variables, rather than using Y ~ 17 variables. Table 4 shows performance of various models on the basis of above ranked variables.

It is observed that instead of all 17 variables formula in Decision Tree, Random forest, PLS, 6 most important variables formula works better with increased accuracy. However, it is also noticed that the performance of Bagging and Multivariate Adaptive Recursive Spline models are very poor. Bagging has a drawback of sensitivity to variance. A very little addition in number of observations highly impact prediction accuracy. MARS model, when unable to fit the spline function, results in poor accuracy of predictive model.

6 Conclusion

In this research study we have studied a problem of variable importance ranking based on different models. We have studied five different models; decision trees, random forests, PLS, bagging and MARS. Variable importance computation specific to these models can extract a set of 5 to 6 common variables like FuelConsumptionAverage, EngineCoolantTemperature, IntakeAirTemperature, ManifoldAbsolutePressure and VehicleSpeedAverage and rank them as the highest importance. Accuracy comparison on the basis of these important variables gives us good results for Decision Tree (93.83%), Random Forest (98.29% and PLS models (81.17%) indicates instead of all 17 variables, we can use these ranked variables for computational cost reduction. In future we would like to work on model-independent variable importance ranking methods and quantification of the same.

References

1. G. Palermo, P. Piraino, and H.-D. Zucht, "in selecting variables for each response of a multivariate PLS for omics-type data," Advances and Applications in Bioinformatics and Chemistry, p. 14.
2. A. Fisher, C. Rudin, and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," p. 81.
3. T. M. Therneau, E. J. Atkinson, and M. Foundation, "An Introduction to Recursive Partitioning Using the RPART Routines," p. 60.
4. S. Liu, Q. Lin, Y. Tian, and K. C. Tan, "A Variable Importance-Based Differential : https://doi.org/10.1109/TCYB.2021.3098186.
5. N. Mlambo, W. K. Cheruiyot, and M. W. Kimwele, "A Survey and Comparative Study of Filter and Wrapper Feature Selection Techniques," p. 11.
6. A. S. Ashour, M. K. A. Nour, K. Polat, Y. Guo, W. Alsaggaf, and A. El-Attar, "A Novel Framework of Two Successive Feature Selection Levels Using Weight-Based Procedure for Voice-Loss Detection in Parkinson's Disease," IEEE Access, vol. 8, pp. 76193–76203, 2020, https://doi.org/10.1109/ACCESS.2020.2989032.
7. J. Cao, G. Lv, C. Chang, and H. Li, "A Feature Selection Based Serial SVM Ensemble Classifier," IEEE Access, p. 1, 2019, https://doi.org/10.1109/ACCESS.2019.2917310.
8. F. Feng, K.-C. Li, J. Shen, Q. Zhou, and X. Yang, "Using Cost-Sensitive Learning and Feature Selection Algorithms to Improve the Performance of Imbalanced Classification," IEEE Access, vol. 8, pp. 69979–69996, 2020, https://doi.org/10.1109/ACCESS.2020.2987364.
9. A. Jawale and G. Magar, "Study of Feature Extraction Techniques for Sensor Data Classification:," International Journal of Information Communication Technologies and Human Development, vol. 13, no. 1, pp. 33–46, 2021, https://doi.org/10.4018/IJICTHD.2021010103.
10. I. Ahrazem Dfuf, J. Forte Perez-Minayo, J. M. Mira Mcwilliams, and C. Gonzalez Fernandez, "Variable Importance Analysis in Imbalanced Datasets: A New Approach," IEEE Access, vol. 8, pp. 127404–127430, 2020. https://doi.org/10.1109/ACCESS.2020.3008416.
11. J. Ehrlinger, "ggRandomForests: Exploring Random Forest Survival," arXiv:1612.08974 [stat], Dec. 2016, Accessed: 21, 2022. [Online]. Available: http://arxiv.org/abs/1612.08974
12. R. Diaz-Uriarte and S. A. de Andres, "Variable selection from random forests: application to gene expression data," arXiv:q-bio/0503025, Jun. 2005, Accessed: 21, 2022. [Online]. Available: http://arxiv.org/abs/q-bio/0503025

13. D. Ollech and K. Webel, "A Random Forest-Based Approach to Identifying the Most Informative Seasonality Tests," SSRN Journal, 2020, https://doi.org/10.2139/ssrn.3721055.
14. A. Behnamian, K. Millard, S. N. Banks, L. White, M. Richardson, and J. Pasher, "A Systematic Approach for Variable Selection With Random Forests: Achieving Stable Variable Importance Values," IEEE Geosci. Remote Sensing Lett., vol. 14, no. 11, pp. 1988–1992, 2017, https://doi.org/10.1109/LGRS.2017.2745049.
15. G. Gazzola et al., "Integrated Variable Importance Assessment in Multi-Stage Processes," IEEE Trans. Semicond. Manufact., vol. 31, no. 3, pp. 343–355, . 2018, https://doi.org/10.1109/TSM.2018.2853586.
16. M. Sagawa et al., "Learning Variable Importance to Guide Recombination on Many-Objective Optimization," in 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), Hamamatsu, 2017, pp. 874–879. https://doi.org/10.1109/IIAI-AAI.2017.158.
17. F. Yang, P. Piao, Y. Lai, and L. Pei, "Margin based permutation variable importance: A stable importance measure for random forest," in 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, 2017, pp. 1–8. https://doi.org/10.1109/ISKE.2017.8258842.
18. J. Tang, J. Qiao, and W. Yu, "Selective Ensemble Modeling Approach based on Variable Importance of Projection With its Application 1," in 2018 IEEE International Conference on Information and Automation (ICIA), Wuyishan, China, 2018, pp. 99–104. https://doi.org/10.1109/ICInfA.2018.8812566.
19. S. D. Nandlall and K. Millard, "Quantifying the Relative Importance of Variables and Groups of Variables in Remote Sensing Classifiers Using Shapley Values and Game Theory," IEEE Geosci. Remote Sensing Lett., vol. 17, no. 1, pp. 42–46, 2020, https://doi.org/10.1109/LGRS.2019.2914374.
20. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification And Regression Trees, 1st ed. Routledge, 2017. https://doi.org/10.1201/9781315139470.
21. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, https://doi.org/10.1023/A:1010933404324.
22. L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996, https://doi.org/10.1023/A:1018054314350.
23. B.-H. Mevik and R. Wehrens, "Introduction to the pls Package," p. 24.
24. "Absolutegaming, 'Road_prediction Dataset,' Kaggle, 08-Aug-2021. [Online]. Available: https://www.kaggle.com/code/absolutegaming/road-prediction/data. [Accessed: 02-Jun-2022].," Absolutegaming, "Road_prediction Dataset," Kaggle, 08-Aug-2021. [Online]. Available: https://www.kaggle.com/code/absolutegaming/road-prediction/data. [Accessed: 02-Jun-2022].
25. https://www.kaggle.com/gloseto/traffic-driving-style-road-surface-condition/download.
26. B. Greenwell M. and B. Boehmke C., "Variable Importance Plots—An Introduction to the vip Package," The R Journal, vol. 12, no. 1, p. 343, 2020. https://doi.org/10.32614/RJ-2020-013.