A framework for vehicle quality evaluation based on interpretable machine learning

Mohammad Alwadi1 · Girija Chetty2 · Mohammad Yamin3

Received: 14 June 2022 / Accepted: 13 October 2022 / Published online: 27 November 2022
© The Author(s), under exclusive licence to Bharati Vidyapeeth’s Institute of Computer Applications and Management 2022

Abstract Ensuring high quality of a vehicle will increase the lifetime and customer experience, in addition to the maintenance problems, and it is important that there are objective scientific methods available, for evaluating the quality of the vehicle. In this paper, we present a computational framework for evaluating the vehicle quality based on interpretable machine learning techniques. The validation of the proposed framework for a publicly available vehicle quality evaluation dataset has shown an objective machine learning based approach with improved interpretability and deep insight, by using several post-hoc model interpretability enhancement techniques.

Keywords Machine learning · Vehicle quality · Risk assessment · Interpretability · Explainability

1 Introduction

Artificial Intelligence (AI) in general and machine learning systems in particular are currently the subject of intense research, particularly in the domain of cyber physical systems, which involves monitoring large physical environments within homes, autos, and other indoor and outdoor places. The automobile and automotive industries have recently undergone a tectonic change, moving away from traditional techniques of marketing cars, towards data-driven solutions based on machine learning, big data, and artificial intelligence. As car and ride sharing has become more popular, quality assurance aspects, such as approaches for assessing vehicle quality and safety standards, using historical information and data-driven models is gaining increasing importance, and demands for a variety of new services have increased. As a result, the car is turning into a connected device on wheels, complete with a real-world browser for everyone. As a result, traditional vehicle manufacturing companies and dealers must share the market opportunity with the IT industry. This paper proposes a new interpretable machine learning framework for a use case involving evaluation of vehicle quality, in terms of safety standards as acceptable based on the quality of several component subsystems.

According to an automotive industry report [1], as of 2020, each vehicle has been generating up to 30 gigabytes of data per day. This vast amount of data presents enormous opportunities for manufacturers to leverage machine learning and data mining algorithms to improve vehicle safety and assess the quality and risks for failure of different component subsystems, and allow vehicle manufacturers and owners to do pro-active, predictive maintenance of vehicles. It is feasible, for example, to predict when an automobile is likely to break down using data processing and vehicle’s car management system block [20, 21]. With a single click, the automobile owner can be notified, and the manufacturer can call a service center and register for repairs [1, 24]. Furthermore, compared to selling automobiles as independent products, providing value-added connected car services, with sophisticated infotainment systems embedded...
with vehicle quality monitoring dashboards, might increase user experience and acceptability [23, 23]. Cars with linked predictive maintenance services, provide not only the value adding feature, but also provide a connected services vehicle ecosystem allowing human centered communication and interaction between the manufacturer and the client, resulting in improved customer experience and acceptability.

These services can be built by leveraging the massive data that gets collected in the vehicle computer systems, and using data driven approaches based on machine learning and AI for assessing the vehicle performance and quality [20–23]. However, these approaches, particularly based on sophisticated machine learning algorithms are often perceived as “black boxes”, due to their advanced algorithm architectures, and though are excellent in terms of their prediction and analytics capabilities, they suffer from problem of being able to explain its' predictions and analysis outcomes. Due to the difficulties associated with extracting useful insights from the models, there is a lack of trust in deploying the sophisticated machine learning solutions for real world problems for critical application settings.

In this paper, we propose a novel computation framework based on interpretable machine learning and explainable AI, for vehicle quality evaluation settings. The proposed framework allows machine learning models to provide interpretations and explanations by different post-hoc processing techniques, after the model evaluation is done. The experimental evaluation of the proposed interpretable machine learning framework allows a deeper insight into machine learning predictions using various objective metrics and can help create increased trust and accountability. Rest of the paper is organized as follows. Next Section describes the background on interpretable machine learning and explainable AI techniques and the related work, and details of computational framework proposed is described in Sect. 3. The experimental work for validating the proposed computational framework is described in Sect. 4, and the paper concludes with conclusions and plan for further work in Sect. 5.

2 Background and related work

In recent years, machine learning technologies have been used widely in several application areas, including vehicle manufacturing, production, sales supply chain, health management, privacy and security, crowd management, and many other domains of our day-to-day life [2–11]. Table 1 shows a brief review of background literature and related work.

Motivated with the prior related work, in this paper, a similar investigative strategy was used in developing an innovative framework for vehicle quality evaluation based on machine learning and data analytics, with a goal to provide the stakeholders with a set of interpretable and explainable tools at various points in the supply chain routing, and clarify the workings being machine learning model, and whether you can trust the decision support provided by the model on several tasks related to predictive maintenance activity, such as evaluating the overall vehicle quality for insurance purposes, and scheduling, maintenance, safety, and risk assessment checks [12]. This can help the customers, who can ask for an explanation of an algorithmic decision that was made by the computer-based systems about the vehicle quality checks. By providing model interpretability capability, it is possible to extract deep insight on the vehicle quality, and its relationship between various subsystems, and can garner trust with the machine learning based decision making.

Some of the insights that these explainable tools can provide could include:

- Which of data features does the model think is most important?
- What is the impact of each feature on local prediction, for a single instance?
- What is the impact of each feature on predictions at the global level, when considered over a large set of possible predictions?
- What is the value of these local and global insights?

These insights might help in getting an in-depth understanding of the model behavior and performance, for different purposes, such as debugging, information about feature engineering, and need for a human in the loop for certain decision-making activities, directing collection of future data, and building trust and accountability. In general, the data that is public and accessible for building machine learning models, and is often unreliable, disorganized, and dirty. Use of appropriate pre-processing techniques can help data preparation and cleaning, and it is one of the important tasks, before any modelling activity can begin. Insight on effectiveness of pre-processing and subsequent feature engineering step is another important stage for building machine learning models. When there are large set of features to be used, it is important to understand the relative importance of each feature, and whether it is a redundant feature, so that it can be eliminated.

Sometimes you can go through this process using nothing but intuition about the underlying features, and their impact on predictions. But you will more insight and understanding when you have 100 s of raw features, or when you do not intend to have a human in the loop and want a computer-based model to provide predictions and explain the reasoning behind the predictions made. Model based insights can provide good understanding of the value of features used, and it is also important to know whether there is a human in the loop needed for decision
| Authors/Reference | Title | Findings |
|-------------------|-------|----------|
| [2] Almutairi, M.M., Yamin, M., Halikias, G. and Abi Sen, A.A. (2021) https://doi.org/10.3390/su14010303 | A Framework for Crowd Management during COVID-19 with Artificial Intelligence. Sustainability | In this work, authors provided validation details for different components and features for extending the framework for different application settings |
| [3] Yamin, M., Ahmed Abi Sen, A., Mahmoud AlKubaisy, Z. and Almarzouki, R. (2021) https://doi.org/10.32604/cmc.2021.017433 | A Novel Technique for Early Detection of COVID-19 | An automatic early detection technique for COVID-19 detection based on CT images and a deep neural architecture called the AIRRCNN was reported, and a detection performance of 99% was achieved |
| [4] Abdi, A.I., Eassa, F.E., Jambi, K., Almarhabi, K., Khemakhem, M., Basuhail, A. and Yamin, M. (2022) https://doi.org/10.3390/electronics1050711 | Hierarchical Blockchain-Based Multi-Chain code Access Control for Securing IoT Systems | In this work, the authors developed a light-weight hierarchical blockchain-based multi-chain code access control to protect the security and privacy of IoT systems, and some of the key findings include, enhanced scalability of the system based on a novel clustering concept based on block chain managers |
| [5] Basahel, S., Alsabban, A. and Yamin, M. (2021) https://doi.org/10.1007/s41870-021-00812-w | Hajj and Umrah management during COVID-19. International Journal of Information Technology | This work provided a detailed review of several crowded events which have taken place during the COVID-19 pandemic, with an in-depth investigation of the impact of pandemic done, and a framework was developed for effectively organizing crowded events during similar emergency and disaster scenarios |
| [6] Yamin, M. and Abi Sen, A.A. (2020) https://doi.org/10.1109/access.2020.3038825 | A New Method with Swapping of Peers and Fogs to Protect User Privacy in IoT Applications | In this work, the authors provided a new method to protect users’ privacy in IoT systems], based on a technique known as the Swapping of Peers and Fogs (SPF), and by exploiting the features of fogs and smart dummies. Some of the key findings from this work include, remarkable improvement offered by SPF approach towards the level of protection of users’ identity |
| [7] Yamin, M., Alsaawy, Y., B. Alkhodre, A. and Abi Sen, A.A. (2019) https://doi.org/10.3390/s19153355 | An Innovative Method for Preserving Privacy in Internet of Things. Sensors | Another innovative framework for preserving the privacy of users in IoT systems was proposed in this work reported, and two methods, namely: Blind Third Party (BTP) and Blind Peers (BLP), and a combination of them to form a new one to be known as the Blind Approach (BLA). The method was shown to be superior to other methods available in the field, and extensible to other similar settings in E-Health, smart transportation and smart-home systems |
| [8] Yamin, M. and Sen, A.A.A. (2018) https://doi.org/10.4018/ijaci.2018010102 | Improving Privacy and Security of User Data in Location Based Services | The authors in this work proposed an approach involving an integration of peer-to-peer (P2P) protocol with the caching technique and dummies from real queries. The findings from this work were that this approach has led to increased efficiency and improved performance, and provides solutions to many problems, and offers an improved way of managing cache |

Table 1 Car evaluation dataset and class distribution
making, or a fully autonomous decision-making model is needed for solving a problem. Sometimes, insights can be more valuable than actual predictions to build the trust in a decision provided by machine learning model, whether it has human-in-the loop or fully autonomous. Further, having an insight not only allows deeper understanding of the problem, it can elicit trust, and improved user acceptance for safety critical cyber physical systems.

Although, interpretability is an important requirement for building a machine learning model for real-world applications, for having a better insight, for most of the newly developed sophisticated machine learning approaches based on shallow and deep machine learning and its variants [13, 14], the decisions seem like coming out of a black box, with no clear explanation and reasoning provided on how the algorithm has made the decisions, and is hardly interpretable from evaluation metrics used for assessing the model performance, causing lack of trust accountability on the decisions provided by the machine.

One of the most common, simple, and inherently interpretable machine learning approaches are those based on linear regression [15], as it just involves analysis of the magnitude and sign of the coefficients, to get a good understanding of how a variable contributes to the model’s prediction. However, for linear regression models, when the number of variables is large, it is still hard for people to understand the meaning behind this linear model. Therefore, certain variants of linear regression called sparse regression were proposed to find a model with few variables [16, 17]. Other similar linear models have also been widely used, such as the Logistic regression models, which is one of the most widely used binary classification methods and known to scale well for large-scale datasets [18]. The model is based on a logistic function to change a linear function to the probability of a response belonging to a certain class. Support vector machines are another widely used models [19], which can use linear or nonlinear kernel function to maximize the separation between different classes. Decision trees are other widely used techniques in machine learning. Humans can easily follow the condition on each split to understand how the model makes the final prediction.

Most of the linear models outlined above and their variants based on shallow machine learning are white box models, which are easy to interpret and can be categorized as inherently interpretable models, but they do not result in high performance. Use of sophisticated non-linear models, such as those based on neural networks, random forests and ensemble models, and their variants, can result in better performance, but they are hard to interpret on how the model makes the decision, and often categorized as black box models, due to the complex algorithm architectures used by these models to achieve improved performance and robustness.
By using some of novel post processing techniques, it is possible to make the black box models more interpretable and capacity to provide better insight. Some of these post processing techniques are discussed in the next section.

3 Techniques for interpreting black box models

Some of the post processing techniques for interpreting black box models include used of Permutation Importance plots [20], SHAP value plots [21] and SHAP tree plots [22], and are as described below:

- **Permutation importance plot**
  
  Permutation importance plot provides an insight into the features which the model thinks are important, and which of them can have biggest impact on predictions? This concept is also referred feature importance, and there are multiple ways to measure feature importance metric. Permutation importance is calculated after a model has been fitted and tries to address the main question—how the prediction accuracy of a model would be impacted, with a random shuffle of a single attribute or feature of the validation dataset, with all other attributes/features and target variable/class label left intact. If a random re-ordering of a single attribute or feature causes less accurate predictions, that feature is highly significant and an important feature. The process for computing permutation importance for a model involves following steps:
  
  - Train the model with one of the machine learning algorithms.
  - By shuffling the values for a single feature/attribute column, make predictions using the resulting dataset.
  - By using these predictions and true target values, compute the impact on the loss function.
  - The performance deterioration as assessed by the loss function indicates the importance of the attribute/feature that was just shuffled.
  - Undo the shuffle and return the data in the original order.
  - Repeat step 2–5 for every attribute/feature column until the completion of feature importance computation for every column.
  
  The permutation importance values towards the top of the permutation importance plot are the most important features, and those towards the bottom are the least important features. For a Random Forest model, which is essentially a tree-based learner, the permutation feature importance measure can be obtained as shown in Fig. 1.

- **SHAP values**
  
  SHAP is an abbreviation for “SHapley Additive exPlanations”, and computation of Shapley values is drawn from the concepts of cooperative game theory. The Shapley value measures the contributions from each player separately on the outcome (model prediction) from the coalition (interaction between input features). By using SHAP values for model interpretation and insight, it is possible to measure the contribution of input features to individual predictions. Using SHAP values has more benefits over other approaches, such as:

  - Global interpretability: SHAP values not only show the importance of features but also show if a particular feature has a positive or negative impact on predictions.
  - Local interpretability: By calculating SHAP values for each individual prediction, it is possible to get an insight on how the features contribute to that single prediction. Other approaches often show aggregated results over the whole dataset.
  - SHAP values are suitable to interpret a large set of black box models including linear and non-linear models, such as linear regression, tree-based models (e.g. XGBoost), random forests, SVM, and deep neural networks.

Next section shows the model evaluation and model interpretability for some of the popular black box machine learning models for a vehicle evaluation use case study.

4 Performance evaluation of black box models

The details of the publicly available dataset used for the vehicle evaluation study is described first, before presenting the studies for comparing black box models, and their interpretability.

4.1 Car evaluation dataset

This publicly available multivariate data set [19, 23] contains information about car evaluation, labelled with the car acceptability metric as the target class label. Table 2 below summarizes the statistical summary for the data set. The car acceptability metric is care safety assessed as being acceptable, unacceptable, good, or very good. The car quality
variable was used as the target variable or class label. The other attributes were used as the predictors for building the machine learning model.

Table 2 displays the structure of the car evaluation dataset and target class distribution. The pie chart shown in Table 2 also presents the distribution of class label with respect to other attributes from the dataset. As can be seen for class distribution, this dataset is significantly imbalanced.

4.2 Performance matrix for black box models

The performance comparison of different black box models in terms of % train and test accuracy are shown in Table 3, the confusion matrix for each model for the test subset in Fig. 2 and classification report for the test subset in Table 4. The details of each of machine learning model are described in [24, 25].

Six different black box models were considered for the comparative study, viz., Decision Tree with Gradient Search, Random Forest, SVM, KNN, Bernoulli Naïve Bayes, and Gaussian Naïve Bayes models. As each of the black model must do multiclass classification (unacc, acc, good and v-good) to predict the vehicle quality in terms of safety assessment, and due to highly imbalanced class distribution, as shown in Table 2, the prediction accuracy metric alone cannot be used to assess the model performance. For imbalanced class distribution, F1 measure is more indicative performance metric, and as can be seen in the classification report matrix shown in Table 4, black box models based on complex algorithms such as SVM and Random Forest are better in terms of F1-score performance metric, with an F1-score of 0.98 for SVM model, and 0.91 for Random Forest Model.

However, the algorithms are not as interpretable as the other low performing shallow learning models based on simple algorithms such as the Bernoulli and Gaussian Naïve Bayes, and Nearest Neighbor (KNN) classifier, which are more interpretable and explainable. Hence the explainability and interpretability is lost if model performance improves, and it becomes less trustworthy. Figure 3 show the model interpretability plots, for one of the well performing model (random classifier model) with permutation importance, SHAP plot (single prediction) for local interpretability, and SHAP tree plot for global interpretability.

Table 2 Car evaluation dataset and class distribution

| Class | N   | N [%] |
|-------|-----|-------|
| unacc | 1210| (70.023 %) |
| acc   | 384 | (22.222 %) |
| good  | 69  | (3.993 %)  |
| v-good| 65  | (3.762 %)  |

Table 3 Accuracy table—white box models
The insight into model performance for any of the black box model (only random forest model interpretability shown here) can be explained better and more interpretable by using permutation importance plots, SHAP value plot and SHAP tree plots for a local and global interpretability.

**Fig. 2** Confusion matrix—white box models (test dataset)

**Table 4** Classification report—white box models (test dataset)

| Model          | Accuracy | Precision | Recall | F1-score |
|----------------|----------|-----------|--------|----------|
| Decision—GS    | 0.93     | 0.85      | 0.91   | 0.88     |
| Random forest  | 0.95     | 0.92      | 0.90   | 0.91     |
| SVM            | 0.99     | 0.98      | 0.98   | 0.98     |
| KNN—GS         | 0.93     | 0.78      | 0.93   | 0.83     |
| Bernoulli NB   | 0.69     | 0.25      | 0.17   | 0.20     |
| Gaussian NB    | 0.73     | 0.62      | 0.55   | 0.52     |

**Fig. 3** Model interpretability—permutation importance, SHAP Value (local interpretability and SHAP tree plot (global interpretability) for random forest model
5 Conclusion

In this paper, we have offered a novel computational framework for vehicle quality evaluation based on interpretable and explainable machine learning. The vehicle evaluation dataset was used for experimental evaluation and assessed with various black box models, enhanced with model interpretability techniques resulted in promising results. The future directions for this research includes developing highly interpretable and explainable machine learning techniques without compromised model performance, that can elicit trust and accountability provided by sophisticated black box models for real world problem solving.

References

1. Automotive. Decisions fueled by insight. Online. Accessed 1/11/2017. https://www.ihs.com/industry/automotive.html
2. Almutairi MM, Yamin M, Halikias G, Abi Sen AA (2021) A framework for crowd management during COVID-19 with artificial intelligence. Sustainability 14(1):303. https://doi.org/10.3390/su14010303
3. Yamin M, Ahmed Abi Sen A, Mahmoud AlKubaisy Z, Almarzouki R (2021) A novel technique for early detection of COVID-19. Comp Mater Continua 68(2):2283–2298. https://doi.org/10.32604/cmc.2021.017433
4. Abdí AI, Eassa FE, Jamni K, Almarhabi K, Khemakhem M, Basuhail A, Yamin M (2022) Hierarchical blockchain-based multi-chain code access control for securing IoT systems. Electronics 11(5):711. https://doi.org/10.3390/electronics11050711
5. Basheh S, Alsabban A, Yamin M (2021) Hajj and Umrah management during COVID-19. Int J Inf Technol. https://doi.org/10.1007/s41870-021-00812-w
6. Yamin M, Abi Sen AA (2020) A method with swapping of peers and fogs to protect user privacy in IoT applications. IEEE Access 8:210206–210224. https://doi.org/10.1109/access.2020.3038825
7. Yamin M, Alsawwy Y, Alkhodre BA, Abi Sen AA (2019) An innovative method for preserving privacy in internet of things. Sensors 19(15):3355. https://doi.org/10.3390/s19153355
8. Yamin M, Sen AAA (2018) Improving privacy and security of user data in location based services. Int J Ambient Comput Intell 9(1):19–42. https://doi.org/10.4018/ijaci.2018010102
9. Chetty G, Yamin M, White M (2022) A low resource 3D U-Net based deep learning model for medical image analysis. Int J Inf Technol 14:95–103. https://doi.org/10.1007/s41870-021-00850-4
10. Sen AAA, Yamin M (2021) Advantages of using fog in IoT applications. Int J Inf Technol 13:829–837. https://doi.org/10.1007/s41870-020-00514-9
11. Yamin M (2020) Counting the cost of COVID-19. Int J Inf Technol 12:311–317. https://doi.org/10.1007/s41870-020-00466-0
12. Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a “right to explanation.” AI Mag 38(3):50–57
13. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
14. Chen T, Guestrin C (2008) Xgboost: a scalable tree boosting system. in KDD ’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining August 2016 Pages 785–794. https://doi.org/10.1145/2939672.2939785
15. Kenney JF, Keeping ES (1962) Mathematics of statistics Pt 1, 3rd edn. Van Nostrand Co, New Jersey
16. Friedman JH (2012) Fast sparse regression and classification. Int J Forecast 28(3):722–738
17. Bertsimas D, Van Parys B et al (2020) Sparse high-dimensional regression: exact scalable algorithms and phase transitions. Ann Stat 48(1):300–323
18. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1
19. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
20. Molnar C (2022) Interpretable machine learning: a guide for making black box models explainable, 2nd edn. https://christophm.github.io/interpretable-ml-book/
21. https://dev.to/mage_ai/how-to-interpret-machine-learning-models-with-shap-values-54jf
22. https://github.com/slundberg/shap
23. Zupan B, Bohanec M, Bratko I, Demsar J (1997) Machine learning by function decomposition. ICML-97, Nashville, TN
24. Alwadi MD, Chetty G, Yamin M (2018) A virtual sensor network framework for vehicle quality evaluation. In: Hoda MV (ed) 12th INDIACom-2018; 2018 5th international conference on computing for sustainable global development, pp 1416–1420. BVICAM. http://bvicam.ac.in/news/INDIACom%202018%20Proceedings/Main/papers/134.pdf
25. Alwadi M, Chetty G (2015) Energy efficient data mining scheme for high dimensional data, biodiversity environment. Proc Comp Sci 46:483–490

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.