IS EXPLAINABILITY ALWAYS NECESSARY?
DISCUSSION ON EXPLAINABLE AI

Gayane Grigoryan

Department of and Eng. Management and
Systems Engineering
Old Dominion University
5115 Hampton Blvd
Norfolk, VA, USA
ggrigory@odu.edu

Andrew J. Collins

Department of and Eng. Management and
Systems Engineering
Old Dominion University
5115 Hampton Blvd
Norfolk, VA, USA
ajcollin@odu.edu

ABSTRACT

The explainability of a model has been a topic of debate. Some research states explainability is unnecessary, and some ”white-box” models, such as regression models or decision trees, are inherently explainable. This paper conducts a multiple regression model analysis with highly correlated features to illustrate how the model's explainability fails when dealing with complex data. In this case, trusting the model explanations can be problematic. The Shapley net effect technique, which helps determine the marginal contribution of the features, is employed to improve the model explainability and reveal more information about the prediction. The work concludes that explainability is necessary to avoid biased and erroneous conclusions in all circumstances, including simple models or even more apparent cases.

Keywords: machine learning model, explainability, explainable AI, feature importance

1 INTRODUCTION

An analyst or the machine learning expert may have good knowledge about the inner workings of the algorithm; however, it is vital to communicate algorithm findings with non-experts clearly. This includes providing a transparent explanation about how the model reached a particular solution and justification of why one should accept that result. The models that are hard to comprehend usually are described as ”black box” models, referring to an increased level of uncertainty to understand the algorithm outcomes. Simpler machine learning models, also known as ”white box” models, can be easily understood by humans due to their lack of rules that design the model and generate the outcome. An example of a less complicated ”white box” machine learning model is the regression model, and convolutional neural networks are considered "black-box." Many analysts blindly ‘accept’ the outcome of the ”black-box” model, whether by necessity or by choice (Doran et al., 2017).

The explainability of the model outcomes becomes more complex when the output predictions are not clearly erroneous. Consider a legal system where understanding why a particular outcome was achieved is vital to confirm that the investigation was not faulty and the conclusion is accurate. Otherwise, incorrect decisions can be devastating. For an example of a model incorrectly labeling the outcome, see LeCun, Bengio, and

MSVSCC, April 14, 2022, Suffolk, VA, USA; ©2022 Society for Modeling & Simulation International (SCS)

Hinton (LeCun et al., 2015). In that work, a deep learning model incorrectly labels an image of a dog lying on a floor as "a dog standing on a hardwood floor."

Therefore it may seem explainability should be necessary for any model we use to analyze the real-world phenomena. A field of Artificial Intelligence called explainable Artificial Intelligence (XAI) proposes creating models and techniques with high explainability whilst maintaining high accuracy. However, some research debates this stance (Adadi and Berrada, 2018). Specifically, these sources suggest that explainability is not necessary when using simple interpretable models, such as regression models (Čík et al., 2021).

This paper employs a regression modeling approach to evaluate the argument for model explainability. Our analysis illustrates that even when using simple "white-box" models, explanations are necessary to fully comprehend the problem as well as the model outcome. The next section provides background on explainability and explainable AI (XAI). Section three describes a regression model, and section four presents explainable results for the regression model. This is followed by the discussion and future work on why explainability is necessary. The paper is concluded in section six.

2 BACKGROUND

Explainability is generally described as the ability for the human user to understand the model's logic. Gregor and Benbasat (Gregor and Benbasat, 1999) describe explainability as a “declaration of the meaning of words spoken, actions, motives, etc., with a view to adjusting a misunderstanding or reconciling differences.” Explanations help to understand the system's malfunctions or anomalies (Schank, 1986). The explanation is assumed to be provided by some source of information, and it usually supplies data, knowledge, and evidence or resolves a disagreement. Explainability in machine learning means that a model can be explained from input to output. This requires knowledge about the underlying ML algorithm.

In 2017, the Defense Advanced Research Projects Agency (DARPA) launched the Explainable AI (XAI) program (Gunning, 2017). The program highlights the importance to “produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately, trust, and effectively manage the emerging generation of artificially intelligent partners.” DARPA defines explainable AI as follows: “AI systems that can explain their rationale to human user, characterise their strengths and weaknesses, and convey an understanding of how they will behave in the future” (see Figure 1).

Model explainability and XAI have been gaining increasing attention in the last few years (Gunning and Aha, 2019). The rapid growth is associated with having more trustworthy AI-powered systems in practical settings. Adadi and Berrada (Adadi and Berrada, 2018) identify the following reasons for explainability: (i) justification, (ii) control, (iii) improvement and (iv) discovery. The authors (Adadi and Berrada, 2018) classify XAI methods based on the following three criteria: (i) the complexity, (ii) the scope, and (iii) the level of dependency from the used ML model . Complexity and explainability are inversely related. The more complex the model is, the less explainable the model outcomes are. One solution is to design an inherently explainable model, such as Bayesian Rule Lists (Letham et al., 2015). Based on the scope, two types of explainability are categorized, i.e., global and local. Global scope refers to the explainability of the whole logic of the model, while local explainability tries to explain a specific instance or an individual prediction. Finally, based on the model used, explainability is model-specific and model-agnostic. Model-specific refers to the explainability techniques that are specifically designed for a particular modeling paradigm. Model-agnostic approaches do not require any information about how the model makes predictions. Figure 2 shows the Google trends of XAI and model explainability terms.

We can notice the increasing trend of explainable models and XAI after 2016. This is perhaps due to the rapid technological development that generated a large amount of data. However, even when observing the

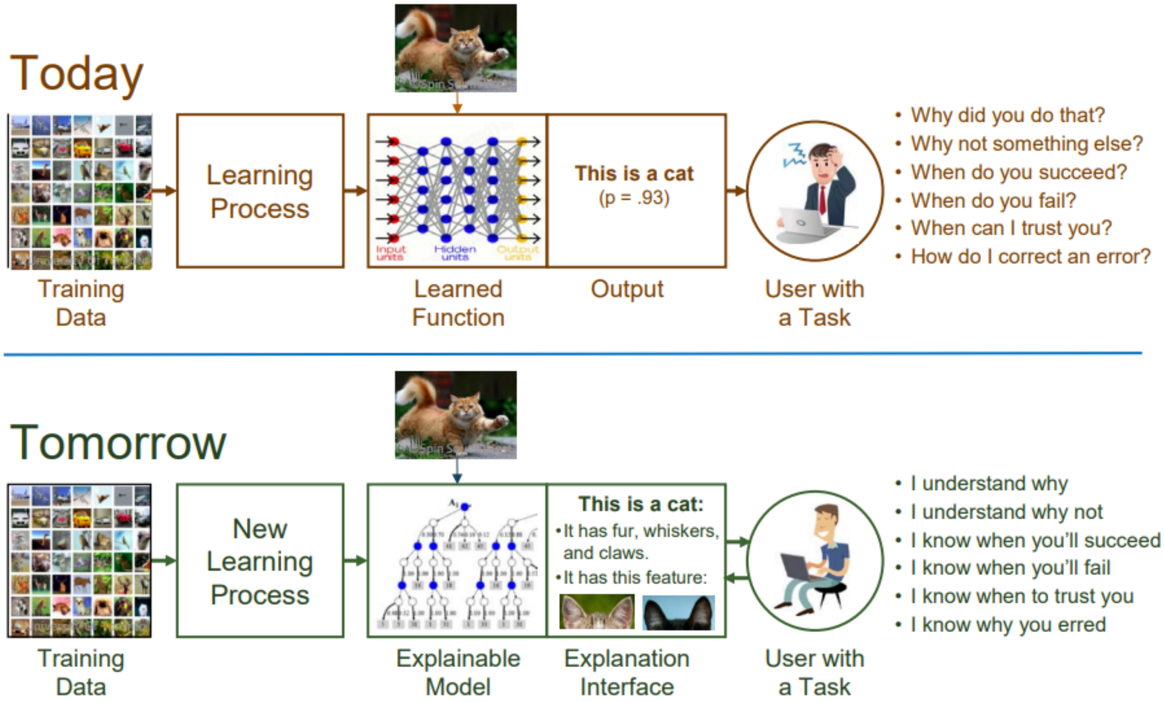


Figure 1: Explainable AI (XAI) concept: DARPA

Source: Gunning and Aha (Gunning and Aha, 2019)

benefits explainability generates, some groups of research question and debate whether explainability is always necessary. Adadi and Berrada (Adadi and Berrada, 2018) state *"explainability is an essential property however, it is not always a necessity."* The authors further elaborate that requiring models to explain every decision could result in less efficient systems. Wang, Kaushal, and Khullar (Wang et al., 2020) suggest *"lower explainability expectations and regulatory requirements for models that address discrete and known tasks."* Čík, Rasamoelina, Mach, and Sinčák (Čík et al., 2021) further highlight that for some applications, such as online translation services, object recognition in the image explainability is not necessary. A significant argument used in most of these works is that attention to the explanation is unnecessary when working with simple and inherently explainable models.

The following section presents a regression model considered for our analysis. The regression model was selected since it is usually identified as a simple, explainable model (Adadi and Berrada, 2018). However, our study highlights how the explainability of the regression model fails when working with complex data.

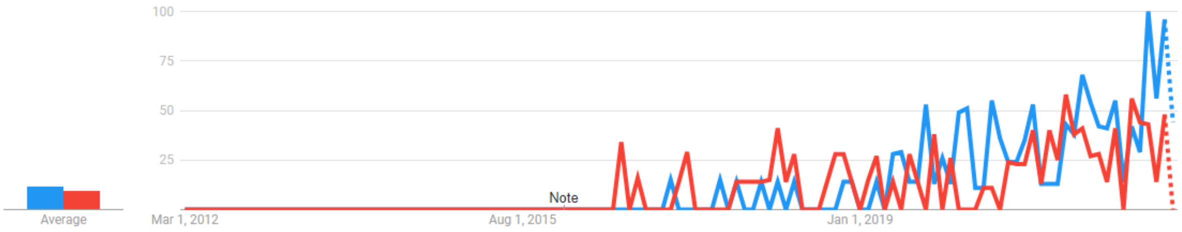


Figure 2: Google trend results for Model Explainability (blue) and XAI (red) terms

3 REGRESSION MODEL

Multiple linear regression models are learning paradigms that can accommodate multiple features to explain more of the variation in the predicted variable (Hastie et al., 2009). These models are employed to describe more realistic modeling situations and find association between the target variable and features. Mathematical form of a regression model is as follows:

$$y_i = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \varepsilon \tag{1}$$

Where, y_i are the observations of the target variable, $x_1, x_2, \dots, x_n$ are the features, ε is the regression error term, that is assumed to be normally distributed, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

3.1 Regression model background

This section explain what the regression model explanation is and what outcome it supplies.

Regression models are considered inherently explainable models due to their generated output. It is easy to understand how the mathematical computations came to a particular result. When dealing with multiple linear regression model it simply determines the feature relationship with the target by determining partial differential equations. Regression models are ideal for contexts to predict a numeric value like the price of the house, and this prediction usually is an approximation, i.e., the closeness of the predictions to the expected values. A regression models is usually evaluated based on its performance with regard to how much error the model makes generally when making predictions. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) are example matrices used to characterize the regression model performance.

Regression outputs β_0 intercept, and $\beta_1, \beta_2, \dots, \beta_n$ are regression coefficients that can be found through the Ordinary Least Square approach (OLS). Regression coefficient values explain the change of the mean of the target variable given a one-unit shift in the independent variable while holding other variables in the model constant. Regression coefficients are regarded unstandardized effect sizes as they indicate the strength of the relationship between features using values that retain the natural units of the target variable. Effect sizes tell us how important the findings are in a practical setting. For example if the effect size is negligible, then we interpret that the variation to the feature has almost no effect on the target variable. The objective of the OLS estimator is to minimize the difference between actual and fitted data points i.e. error sum of squares (ESS). In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased. Residual plots usually are used to reveal residual patterns that indicate bias, irregularities, omitted variable or other model construct issues. For more information about OLS estimators and residual plots, reader can refer to (Montgomery et al., 2021).

Regression model produces multiple determination R^2 coefficient, which measures the model performance. R^2 coefficient is a statistical measure to show how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. Sum of squares values are used as indicators to present the dispersion of data and suggest how well the data fits the regression model. The three main sum of squared indicators to determine the R^2 value are total sum of squares (TSS), regression sum of squares (RSS), error sum of squares (ESS). The formula of multiple determination R^2 coefficient is as follows:

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} \tag{2}$$

Regression also supplies statistical significance values, also known as *p*-values. *P*-values evaluate the null hypothesis that the regression coefficient is equal to zero. A feature has no effect on the regression model if the regression coefficient equals 0 or is not statistically significant. A low *p*-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a feature with a low *p*-value is likely to be a meaningful addition to the model. Larger *p*-values suggest the features are not appropriate for the model to predict the target variable.

So, a user employs a regression model to explain a particular estimation based on the following outputs: (a) regression coefficients, (b) *p*-values, and (c) R^2 and Adj. R^2 values. However, some of these results can be problematic if generated from a regression model with highly correlated features. Regression may supply output with statistically insignificant *p*-values, implying some features are irrelevant for the model and should be removed. Better explanation techniques and careful analysis are essential to deal with similar cases. An example below discusses a regression model with a multicollinearity issue and highlights how the model outcome becomes deceitful and unreliable.

3.2 ”Seatpos” model

This subsection describes the dataset and the regression model construct generated using the "seatpos" dataset. The dataset collected by HuMoSim laboratory researchers at the University of Michigan is intended to study a car seat position given the demographic attributes of 38 drivers. The dataset is used to demonstrate the association between the car seat design given various characteristics of a driver. A detailed description of the dataset and an example of its use can be found in Faraway (Faraway, 2014).

Human body ratios are described to be symmetric. Features included in the dataset to model the car seat position are age in years (Age), weight in lbs (Weight), height in shoes in cm (HtShoes), height bare foot in cm (Ht), seated height in cm (Seated), lower arm length in cm (Arm), thigh length in cm (Thigh), lower leg length in cm (Leg), horizontal distance of the midpoint of the hips from a fixed location in the car in mm (hipcenter). To study the car seat position given different demographic characteristics of the driver, the following regression model was estimated:

$$y_i = \beta_0 + \beta_1 Age + \beta_2 Weight + \beta_3 HtShoes + \beta_4 Ht + \beta_5 Seated + \beta_6 Arm + \beta_7 Thigh + \beta_8 Leg + \varepsilon \qquad (3)$$

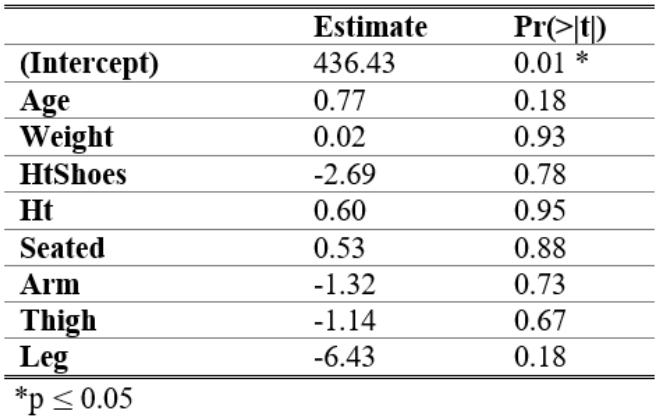
The output of the regression model is presented in the subsection below.

3.3 ”Seatpos” results

The initial results of the model are presented in Table 1. As we can see, the multiple regression results are insignificant. An important question here arises is "Should we trust the model? Does the model provide an accurate explanation about the data." The model seems to be unsatisfactory. If we study the car seat design requirements we learn that the data is intended to have a substantial role in predicting the driver’s car seat position. The R^2 value of the model is about 0.7.

From these results, we cannot determine whether the predictions are biased. Perhaps we cannot trust the *p*-values, respectively, we cannot trust the coefficients and finally cannot trust the prediction. This will suggest specifying a better model and justify the model predictions. Therefore, further explanations are needed to have a better understanding of the features employed as well as the model prediction. These results illustrate

Table 1: Initial results of the regression model



	Estimate	Pr(>|t|)
(Intercept)	436.43	0.01 *
Age	0.77	0.18
Weight	0.02	0.93
HtShoes	-2.69	0.78
Ht	0.60	0.95
Seated	0.53	0.88
Arm	-1.32	0.73
Thigh	-1.14	0.67
Leg	-6.43	0.18

*p $\leq$ 0.05

how a model performance reduces and the explanation changes with twisted data. In the above-discussed example, the regression model could no longer provide correct predictions about the data.

We have also checked the relationships between the features for multicollinearity, and identified some highly correlated features. Feture Age does not have a close association with the rest of the variables. Hipcenter has a negative correlation with most of the variables, except the Age. The negative relationship is especially strong with variables HtShoes (-0.8), Ht (-0.8), and Leg (-0.79). Note, that HT and HtShoes have perfect positive correlation; with correlation coefficient equal to 1. The remaining variables mostly have positive associations with one another.

After the analysis, we should determine whether or not to follow the model suggests that none of the features are relevant to designing the car seat. Other actions we can do are: (i) remove some of the correlated features, (ii) linearly combine the features, or (iii) apply LASSO and Ridge regression models that are advanced forms of regression analysis and can handle multicollinearity. However, for our model, none of these options will improve the performance or explainability of the model. We can also do nothing and use the model with hardly any significant predictors. In that case, statistical insignificance will mean - lost explainable power about our estimate.

4 REGRESSION MODEL EXPLAINABILITY

To improve the explainability of the regression model we suggest to use Shapley value net effects proposed by Lipovetsky and Conklin (Lipovetsky and Conklin, 2001). Shapley value net effects determines the feature importance of regression model with multicollinearity issue. The proposed approach employs a cooperative game theory solution concept called Shapely value (Shapley, 1953) (Eq. 4)

$$\phi_i(v) = \sum_{s\subseteq N\setminus\{i\}} \frac{|S|!(|N|-|S|-1)!)}{|N|!}(v(S\cup\{i\})-v(s)) \tag{4}$$

Feature importance evaluation consists of comparing the model performance, measuring the multiple determination R^2 value, with and without particular feature i using Shapley value (Eq. 5).

$$U_i = R^2 - R^2_{-i} \tag{5}$$

Shapley value considers a coalition S, where i is not part of the coalition. The first part of the equation randomly chooses a set size $|S|$ out of $\{0,1,2,\ldots,|N|-1\}$, each having probability $\frac{1}{|N|}$ to be drawn. Afterwards, a subset of $N\{i\}$ of size $|S|$ is chosen. Marginal contribution of coalition member i is computed subsequently $v(S \cup \{i\})) - v(S)$. For more information about cooperative game theory and Shapley value calculation please see Grigoryan and Collins (Grigoryan and Collins, 2021). The computation of the regression Shapley values estimation is summarized in Algorithm 1.

Figure 3: Algorithm to explain regression feature marginal contributions

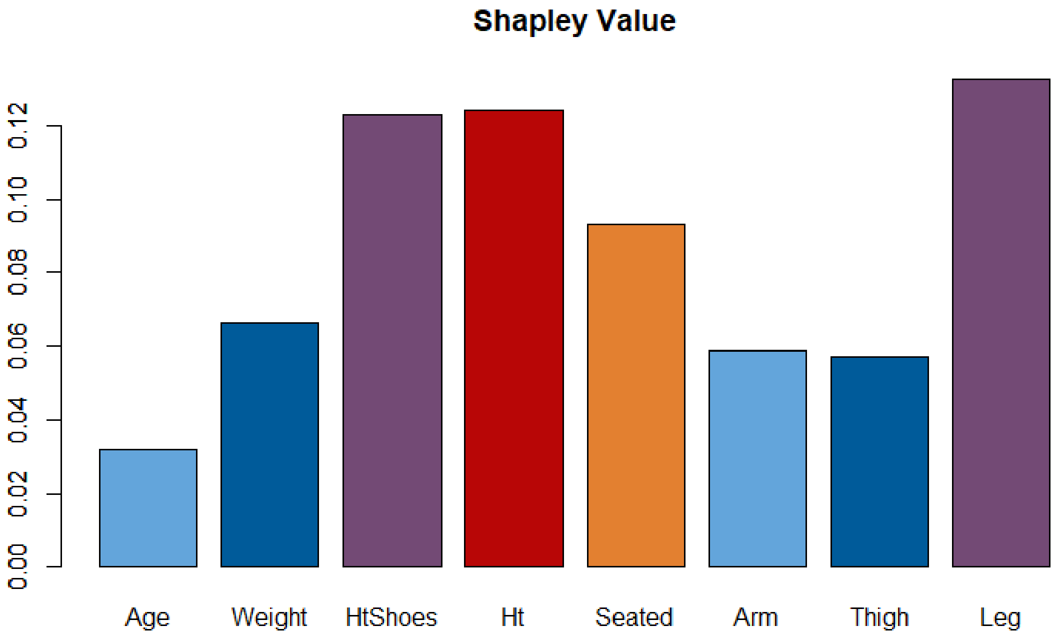
```
Algorithm 1 Regression Shapley Value Estimation
Initial_input: $D(X_1, X_2, …, Xn)$                        # a training dataset with $n$ features
r_squared ← []                                              # initialize r_squared value
combinations C ← []                                         # initialize combinations of features
for i in  range$(1 : n)$
    determine combinations of features
        for a  combination C in the list of combinations
        run regression models (Eq. 1)
        get model summaries to extract r_squared values
        end for
end for
New_input: $(v(\{0\}), v(\{1\}), …, v\ (all\ features))$   # use r_squared values as coalition value $V$
begin defining the game$(n, V)$ in characteristic form      # coalition values $V$ are in lexicographic order
calculate Shapley value φ(i) (Eq 4 & Eq 5)
end when all permutations of coalitions are evaluated
Return: Shapley values of features φ(i)                     # returns feature marginal contributions
```

The algorithm starts determining different combinations of features and runs regression models for each combination of the features. It ends the run by returning the model summaries and extracting R^2 values. The newly obtained R^2 values are used as the new data to define the game in characteristic form. The next step of the algorithm defines the game for n features given the respective characteristic values, i.e., coalition values V in lexicographic order. The algorithm ends when all permutations are evaluated and returns the regression Shapley values, which are the features' marginal contributions. The "seatpos" dataset and the Python Jupyter Notebook code could be obtained online from: https://github.com/grigoryangayane/RegressionShapley

4.1 Explainable results

This subsection presents the computational results of features marginal contributions for the "seatpos" dataset following the Algorithm 1. With 8 variables, it is possible to construct 256 regression models. These models include 8 single variables models, 28 models with two variables, 56 three-variable models, 70 four-variable models, 56 five-variable models, 28 six-variable models, 8 seven-variable models and one model that contains all the 8 variables. Note that the model that does not contain any features will have the characteristic value $v(0) = 0$, because the regression model with intercept only does not yield any R^2 result. The model that has the lowest multiple determination value is a single variable model and contains the variable "Age". As expected, the model that includes all the variables has the highest characteristic function with R^2 equivalent to 0.686553. The outcome of Shapley value computation demonstrates the marginal

Figure 4: Algorithm to explain regression feature marginal contributions



contributions of all 8 features to predict the target variable. These results explain the feature contribution when designing the prediction model based on their respective importance levels.

The Figure 4 reveals that the most relevant variables to predict the driver's car seat position are height in shoes (HtShoes), height bare foot (Ht), and lower leg length (Leg). The marginal contributions of the variables Age, Weight, HtShoes, Ht, Seated, Arm, Thigh, Leg respectively are $\phi_i = (0.032, 0.0661, 0.122, 0.124, 0.093, 0.058, 0.0572, 0.132)$. Surprisingly, the arm and thigh length is one of the least contributors in explaining the predictor variable. As expected, the variable Age is not very relevant to predict the car seat position.

5 DISCUSSION

Some machine learning models or simulation models are not always as intuitive as other models. Understanding the model outcome and explaining the model becomes more challenging when features are correlated, or there are some interactions. The ”seatpos” dataset we discussed in our model is an appropriate dataset for a multiple regression model. Other models that may be applicable for this dataset are LASSO, Ridge, Elastic Net regressions, stepwise regression, or best subsets regression. However, these models are not generally accepted as inherently explainable models, and they come with some disadvantages. For instance, LASSO cannot make a group selection and arbitrarily selects one of the highly correlated variables. In future work, we plan to compare the performances of these models to determine which one has better explainable power. For detailed information about the machine learning explanation evaluation methods one can review the following survey (Zhou et al., 2021) The best explainable model is the intuitive model that clearly describes the relationship between variables as accurately as possible.

Explainable model concepts can be further explored when designing simulation models. Verification and validation of the simulation models can be challenging. Employing explainability techniques like SHAP or LIME to verify that the model designed is correct or to explain the results from a global or local scope, can provide more insights about each iteration of the simulated model. For example, the following simulation model developed by Vernon-Bido et al. (Vernon-Bido et al., 2018) studies the relationship between cyberloafing behavior and cyber risk. The simulation analysis reveals interesting patterns about cyber risk. However, it does not explain why the phenomenon occurs. Paying attention to the explainability of the model could disclose more knowledge about the effect of cyberloafing on cyber vulnerabilities and cyber risk.

6 CONCLUSION

Models are used to make highly crucial decisions, varying from medical diagnosis to cyber-physical systems analysis. Understanding the decision-making process of systems-powered outcomes will provide more knowledge and confidence about our conclusions. However, when dealing with complex data, the explanation of the model may yield biased results, such as the example we discussed in this paper. Having a trustworthy model that we can rely on is crucial.

This work discusses a multiple regression model that is inherently considered explainable and shows how the explanations of the model change when data is twisted. We conducted a multiple regression model analysis and notice that features that have high correlations drastically affect the output of the model. The p-values are statistically insignificant, suggesting the data used is not appropriate or the model construct is not optimal. The data used to run the regression model describe the anthropometric characteristics of 38 drivers. These characteristics are symmetric in most humans and are expected to be highly correlated. Therefore, removing some of the features from the model will not help the situation. This example highlights how vulnerable the output of the explainable model can be to some data twist, including influence from outliers, data interactions, and correlations.

To improve the explainability of the regression model, we have employed the Shapley net effects technique. This helps to determine the feature importance of regression models with multicollinearity issue. This technique assigns scores to input features of the predictive model. Based on the case analysis of the car seat study, we could determine that the feature "lower leg length" was the most important feature when designing the car seat. Applying this technique generates more trust in the model's suggestions and makes it easier to interpret the model output by identifying the marginal contributions of each feature. From this viewpoint, we can also simplify the complex model representation by simply determining the most crucial factors necessary for the prediction.

This work shows how vital the explainability of a model is, no matter the case. Explainability is essential to make sure the message of the model is clear and accurately delivered. We believe explainability is necessary, especially because of the rapid advancement of intelligent systems and the expansion of digital engineering that generates massive data. We think explainability will contribute to gaining more reliable knowledge about the behavior of the models and their underlying processes.

REFERENCES

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

Čík, I., Rasamoelina, A. D., Mach, M., and Sinčák, P. (2021). Explaining deep neural network using layer-wise relevance propagation and integrated gradients. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000381–000386. IEEE.

Doran, D., Schulz, S., and Besold, T. R. (2017). What does explainable ai really mean? a new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.

Faraway, J. (2014). Linear models with r. crc press. *Boca Raton, Florida*.

Gregor, S. and Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, pages 497–530.

Grigoryan, G. and Collins, A. J. (2021). Game theory for systems engineering: a survey. *International Journal of System of Systems Engineering*, 11(2):121–158.

Gunning, D. (2017). Explainable artificial intelligence (xai)(2017). *Seen on*, 1.

Gunning, D. and Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371.

Lipovetsky, S. and Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Schank, R. C. (1986). Explanation: A first pass. *Experience, memory, and reasoning*, pages 139–165.

Shapley, L. (1953). A value for n-person games. *Contributions to the Theory of Games*, 28(2):307–317.

Vernon-Bido, D., Grigoryan, G., Kavak, H., and Padilla, J. (2018). Assessing the impact of cyberloafing on cyber risk. In *Proceedings of the Annual Simulation Symposium*, pages 1–9.

Wang, F., Kaushal, R., and Khullar, D. (2020). Should health care demand interpretable artificial intelligence or accept "black box" medicine?

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593.