A machine learning research template for binary classification problems and shapley values integration

Matthew Smith a,b,∗, Francisco Alvarez a,2

a Universidad Complutense Madrid, Madrid, Spain
b Barcelona Supercomputing Center, Barcelona, Spain

A R T I C L E I N F O

Keywords:
Machine Learning
Binary classification
COVID19
Shapley values

A B S T R A C T

This paper documents published code which can help facilitate researchers with binary classification problems and interpret the results from a number of Machine Learning models. The original paper was published in Expert Systems with Applications and this paper documents the code and work-flow with a special interest being paid to Shapley values as a means to interpret Machine Learning predictions. The Machine Learning models used are, Naive Bayes, Logistic Regression, Random Forest, adaBoost, Classification Tree, Light GBM and XGBoost.

1. Introduction

The COVID19 pandemic has affected every corner of the world and research efforts have increased exponentially since the start of it. We provide a basic framework for researchers who wish to apply Machine Learning models to predict COVID19 mortality at a patient level. This will allow researchers to save time in writing their own models, it is by no means an exhaustive resource but provides a starting point for other investigations. Predicting the probability of mortality from patient characteristics also may help hospitals better treat patients with personalized treatments upon arrival at the hospital. This paper documents the models used in [1] and how they can be used on future datasets. Obtaining high quality prediction performance is an important aspect and a vital one when predicting which patient is likely to survive or not. However, in our code we aim to go into more detail on the interpretation of the Machine Learning models results since it provides greater insight into how practitioners can better treat patients and learn about important features in a dataset. In order to better interpret the model we make use of Shapley values which is a concept first developed...
in cooperative game theory, see [2] and has recently been applied to Machine Learning models, see [3] and [4].

2. Contribution to the scientific community

The code was written in RMarkdown [5] and can be used as a template for Machine Learning applications to other problems. RMarkdown is a file format for making dynamic documents with R, a document can be created using plain text and LaTeX formatting along with chunks of R code embedded throughout the document. This allows researchers to save time and facilitates reproducible results by directly replicating the whole research paper. In the Code Ocean capsule we provide the full RMarkdown script along with a standalone R script such that others can reproduce the exact same results as in the original paper, see [6].

We apply a series of Machine Learning models, notably, Naive Bayes, Logistic Regression, Random Forest, adaBoost, Classification Tree, Light GBM and XGBoost. We then compute a number of relevant statistics based on a confusion matrix for each model. The results are visualized using the Receiver Operating Characteristics (ROC) and Precision Recall (PR) Curves. Individual predictions and case studies can also be analyzed from the XGBoost model which may help practitioners make more informed decisions based on the Machine Learning models predictions. We make use of Shapley values in order to analyze the model agnostic methods for interpreting Machine Learning models. This allows us to analyze the models predictions at a local and global level along with seeing how different variable values change with Shapley contributions. Finally, we apply what-if analysis to show how (holding constant all other variables) changes in one variable impacts a prediction. This is important since we can see how the probability of mortality changes with increases or decreases in variable values, i.e. how does keeping a patient in hospital a day longer impact patients probability of survival. In this paper we highlight the main results focusing on the aforementioned analysis, this time applied to a new dataset, to show its reproducibility.

3. Shapley values

Shapley values [2] from coalition game theory allows us to interpret the predictions of Machine Learning models by treating each variable as a player in a game with the prediction being the payoff [7] and [8]. A cooperative game can be considered as the following. A characteristic function game $G$ is given by a pair $(N, v)$ where $N$ is the number of players and $v : v^N \rightarrow \mathbb{R}$ is a characteristic function which maps every coalition of players to a payoff.

Consider the following example which has three variables and the output is a prediction from a Machine Learning model. Each variable contributes differently to the prediction (some variables contribute more, others less). The characteristics function $v(c)$ tells us that if the variable $X_1$ was the only variable in the model then it would make a contribution of 80, the variables $X_2$ and $X_3$ on their own make a contribution of 56 and 70 respectively. The value of the whole coalition is 90 with all three variables $X_1$, $X_2$, and $X_3$ contributing differently to the prediction (some variables contribute more). The Shapley value is defined as $\phi_i(G) = \frac{1}{n!} \sum\limits_{c \in 2^N} v^c(i)/|c|$. The table on the right considers every permutation of players, that is, the first line considers the permutation of $X_1$, $X_2$, and $X_3$ with a score of $(80, 0, 10)$. Looking at the characteristic function $v(c)$ on the left we can see that $X_1$ on its own contributes the value of 80, additionally, we can see that when $X_1$ and $X_2$ are in a coalition the payout is also 80, therefore $X_1$ contributes to the prediction and $X_2$ contributes nothing, finally, when $X_1$, $X_2$, and $X_3$ are in a coalition, the score is 90, therefore, $X_1$ must contribute 10. Consider now when we swap the order of the variables, $X_1$ still contributes 80 but the coalition of $X_1$ and $X_2$ is 85, therefore $X_2$ contributes 5. Still, the coalition of $X_1$, $X_2$, and $X_3$ is 90, therefore $X_3$ must contribute 5 also. Finally, if the order is as in line three, then $X_2$ contributes 56 on its own with the coalition of $X_1$ and $X_3$ being 80, $X_1$ must contribute 24 and the coalition of all variables is 90, therefore $X_3$ must contribute 10. This is done for all permutations of the variables and then the average of these values are taken, given as $\phi$ which gives us the average marginal contribution for each variable over all of the possible subsets.

| $c$ | $v(c)$ |
|-----|--------|
| $\{X_1\}$ | 80 |
| $\{X_2\}$ | 56 |
| $\{X_3\}$ | 70 |
| $\{X_1, X_2\}$ | 80 |
| $\{X_1, X_3\}$ | 85 |
| $\{X_2, X_3\}$ | 72 |
| $\{X_1, X_2, X_3\}$ | 90 |

4. Reproducible results: Application to new data

[9-12] and [13] used clinical features to predict the survival probabilities of patients with COVID19. [10] released their patient characteristic dataset and thus in this paper we replicate the results from our first study [1] using this new data. The code for the models and figures can be found on CodeOcean [6].

The data contains COVID19 patient characteristics along with the patients mortality outcome for people admitted to hospital with COVID19. The replication of the code onto a new dataset are left to Appendix with the purpose of illustrating the outcome of the code. The data contains only 9 features with some of these features likely to be related to each other, additionally two of the features are binary features and can only be used a single time in each tree, thus reducing
Fig. 1. Number of daily cases in the Netherlands. Shaded region is the region understudy. The coronavirus package in R was used and it collects data for 192 countries worldwide.

Fig. 2. XGBoost most important variables. Age and daysInHospital appear to be important and consistent with the original paper.

Fig. 3. Shapley additive explanations for variables. Low (yellow) values of age have a negative impact on the probability of mortality, measured by negative Shapley values. The opposite occurs for high (blue) values of age. The highest impacts, that is, highest Shapley values in absolute value are for Age, days in hospital and LD, in line with the previous figure (information gain). In contrast to the previous figure, here we observe the sign of the impact and how it changes with the feature values. Notice that Shapley values do not detect causality. For instance, consider admissions to ICU, which is binary. When it occurs (coded as 1, or blue in the figure), it has a positive Shapley value, that is, it increases the probability of death, though we know the causality is the other way around: a patient enters the ICU because they have a high probability of death. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
the complexity and non-linearity of the tree. Therefore, some of the more complex tree based algorithms are unable to distinguish between the noise and the correct functional form in the data.

The code, plots and interpretation can be applied to any binary classification problem. [14] applied Machine Learning to the classification of bankrupt versus non-bankrupt firms with interpretation on case study levels. [15] applied Shapley values to credit risk management using a number of financial ratios to show the contribution each variable made to the models prediction. [16] applied Shapley values to the detection of acute myocardial infarction (AMI). [17] applied Shapley values to the detection of relevant Alzheimer’s disease characteristics. [18] used Shapley values as a feature selection model using data on mental health. [19] measured Shapley values to the input of images of a Convolutional Neural Network to find pixels which contribute most to the classification of input images. [20] studied the early prediction of a teams performance in software engineering. They use a Deep Neural Network to make a classification of the teams and Shapley values for feature importance. [21] discusses some of the problems with using Shapley values as feature importance measures.

5. Limitations and improvements

Whereas the XGBoost and LightGBM models both contain cross-validation code (commented out in the scripts), other models were
optimized away from the main script. Therefore, users should be mind-ful that cross-validation should also be integrated for the adaBoost and Random Forest models also. Additionally, whereas the main focus of the paper was on the use of Shapley values for model-agnostic inter-pretability, other model-agnostic methods can also be implemented, such as partial dependence plots, individual conditional expectations, accumulated local effects, permutation feature importance, local inter-pretable model agnostic explanations (LIME) and others.

6. Conclusion

We believe that researchers may benefit from the models and code used to generate the results and figures in our paper. The code can be applied to binary classification problems, we have shown two examples of use cases applied to COVID19 mortality prediction, one in [1] and the other here. The reproducibility of our paper may help researchers apply the same models and code to their own research problems. Adding Machine Learning models, improving and building on top of the current code should be a next step.

Table 2

| Metric         | Naive.Bayes | Logistic.Regression | Random.Forest | adaBoost | Classification.Tree | XGBoost |
|----------------|-------------|---------------------|---------------|----------|---------------------|---------|
| Accuracy       | 0.62        | 0.77                | 0.77          | 0.79     | 0.79                | 0.78    |
| Sensitivity    | 0.56        | 0.87                | 0.89          | 0.85     | 0.92                | 0.90    |
| Specificity    | 0.87        | 0.40                | 0.33          | 0.53     | 0.27                | 0.27    |
| Precision      | 0.95        | 0.84                | 0.83          | 0.88     | 0.84                | 0.84    |
| F1             | 0.71        | 0.85                | 0.86          | 0.87     | 0.88                | 0.87    |
| MCC            | 0.34        | 0.29                | 0.25          | 0.37     | 0.23                | 0.20    |
| AUC            | 0.72        | 0.64                | 0.61          | 0.69     | 0.59                | 0.58    |
| AUPRC          | 0.31        | 0.35                | 0.33          | 0.38     | 0.30                | 0.28    |
| TP             | 35.00       | 47.00               | 48.00         | 53.00    | 57.00               | 56.00   |
| FP             | 27.00       | 7.00                | 6.00          | 9.00     | 5.00                | 6.00    |
| FN             | 2.00        | 9.00                | 10.00         | 7.00     | 11.00               | 11.00   |
| TN             | 13.00       | 6.00                | 5.00          | 8.00     | 4.00                | 4.00    |

* Note: The Logistic Regression and Random Forest model removes missing values from its final results and cannot be adequately compared with the other results.

MCC: Matthew’s Correlation Coefficient
AUC: Area Under the Curve
AUPRC: Area Under the Precision Recall Curve
TP: True Positive — FP: False Positive — FN: False Negative — TN: True Negative.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

See Figs. 1–6 and Tables 1 and 2.

References

[1] M. Smith, F. Alvarez, Identifying mortality factors from Machine Learning using Shapley values - a case of COVID19, Expert Syst. Appl. (2021) 114832, http://dx.doi.org/10.1016/j.eswa.2021.114832, URL: https://www.sciencedirect.com/science/article/pii/S0957417421002736.

[2] L.S. Shapley, A value for n-person games, Contrib. Theory Games 2 (28) (1953) 307–317.
[3] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.

[4] S.M. Lundberg, G.G. Erlon, S.-I. Lee, Consistent individualized feature attribution for tree ensembles, 2018, arXiv preprint arXiv:1802.03898.

[5] J. Allaire, Y. Xie, J. McPherson, J. Luraschi, K. Ushey, A. Atkins, H. Wickham, J. Cheng, W. Chang, R. Iannone, rmarkdown: Dynamic Documents for R, 2021, R package version 2.7, URL: https://github.com/rstudio/rmarkdown.

[6] M. Smith, F. Alvarez, Identifying mortality factors from Machine learning using Shapley values - a case of COVID19, Expert Syst. Appl. (2021) http://dx.doi.org/10.24433/CO.1107431.v1, https://www.codeocean.com/.

[7] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, pp. 4765–4774, URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[8] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, M.J. Quanjel, T. Adams, D.E. Liston, D.K.-W. Low, S.-F. Newman, J. Kim, et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (10) (2018) 749.

[9] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang, et al., An interpretable mortality prediction model for COVID-19 patients, Nat. Mach. Intell. 2 (5) (2020) 283–288.

[10] M.J. Quanjel, T.C. van Holten, P.C. Gunst-van der Vliet, J. Wielaard, B. Karakaya, M. Söhne, H.S. Moeniralam, J.C. Grutters, Replication of a mortality prediction model in Dutch patients with COVID-19, Nat. Mach. Intell. 3 (1) (2021) 23–24.

[11] A.S. Yadav, Y.-c. Li, S. Bose, R. Iyengar, S. Bunyavanich, G. Pandey, Clinical features of COVID-19 mortality: development and validation of a clinical prediction model, Lancet Digit. Health 2 (10) (2020) e516–e525.

[12] A.S. Yadav, Y.-c. Li, S. Bose, R. Iyengar, S. Bunyavanich, G. Pandey, Clinical predictors of COVID-19 mortality, 2020, medRxiv, Cold Spring Harbor Laboratory Preprints.

[13] D. Berndas, G. Lukin, L. Mingardi, O. Nohadani, A. Orfanoudaki, B. Stellato, H. Wiberg, S. Gonzalez-Garcia, C.L. Parra-Calderon, K. Robinson, et al., COVID-19 mortality risk assessment: An international multi-center study, PLoS One 15 (12) (2020) e0243262.

[14] M. Smith, F. Alvarez, Predicting firm-level bankruptcy in the Spanish economy using extreme gradient boosting, Comput. Econ. (2021) 1–33.

[15] N. Bussmann, P. Giudici, D. Marinelli, J. Papenbrock, Explainable machine learning in credit risk management, Comput. Econ. 57 (1) (2021) 203–216.

[16] L. Ibrahim, M. Mesinovic, K.-W. Yang, M.A. Eid, Explainable prediction of acute myocardial infarction using machine learning and Shapley values, IEEE Access 8 (2020) 210410–210417.

[17] L. Bloch, C.M. Friedrich, Developing a machine learning workflow to explain black-box models for Alzheimer’s disease classification, in: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2021-Volume, Vol. 5, 2021.

[18] N. Jothi, W. Ilusain, et al., Predicting generalized anxiety disorder among women using Shapley value, J. Infect. Public Health 14 (1) (2021) 103–108.

[19] M. Rashed-Al-Mahfuz, M.A. Moni, S.M.S. Islam, S. Berkovsky, M. Khushi, J.M. Quinn, et al., Deep convolutional neural networks based ECG beats classification to diagnose cardiovascular conditions, Biomed. Eng. Lett. (2021) 1–16.

[20] F. Giannakas, C. Troussas, I. Voiyatizis, C. Sgouropoulou, A deep learning classification framework for early prediction of team-based academic performance, Appl. Soft Comput. (2021) 107355.

[21] I.E. Kumar, S. Venkatasubramanian, C. Scheidegger, S. Friedler, Problems with Shapley-value-based explanations as feature importance measures, in: International Conference on Machine Learning, PMLR, 2020, pp. 5491–5500.