Meeting on Codes, Lattices, and Information
June 15th and 16th, 2023, Campinas/SP

Extraction of interpretable information in models based on symptoms

Lucas Tramonte, Cristiano Torezzan

Introduction

A problem of great interest in the field of machine learning is that the results obtained by algorithms are auditable and interpretable [3]. This study presents preliminary results on the effectiveness of local interpretability methods, such as Lime and SHAP, in supervised machine learning applications for medical diagnostics. To this end, public datasets were used, with particular interest in COVID-19 diagnoses based on symptoms.

In this context, linear models have been widely used, as they are interpretable based on the weights of each attribute [2]. Such models will be used as a reference for comparison.

In general terms, the efficiency property of linear models allows for estimating the contribution, ϕ_j, of the j-th attribute in the prediction by calculating:

$$\sum_{j=1}^{p} \phi_j(f) = f(x) - E(f(x)), \tag{1}$$

where $E(f(x))$ is the average predicted value for the instance x..

Materials and methods

As a solution, post-hoc interpretability methods can be used to estimate the importance of the main attributes for classification. The post-hoc interpretability method SHapely Additive exPlanations (SHAP), for example, aims to explain local predictions of a given model by randomly forming various coalitions of attributes and determining their respective effects on the model. The associated theoretical foundation is based on game theory, where marginal costs are assigned to each attribute, and their contribution is analyzed across all possible coalitions, resulting in a value known as the Shapley value [3].

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!(p-|S|-1)!}{p!} \cdot (val_x(S \cup \{j\}) - val_x(S)) \tag{2}$$

where S is the subset of attributes used in the model, x the vector of attribute values of the instance to be explained, and val_x the prediction of the attributes in set S that have been marginalized over the attributes not belonging to this set:

$$val_x(S) = \int f(x_1, \ldots, x_p) dP_{x \notin S} - E_X[f(X)]. \tag{3}$$

In this work, we present the results of an empirical case study based on anonymized real-world data, as presented in [1], in which the authors use machine learning methods to predict SARS-CoV-2 test outcomes in the Brazilian population based on reported symptoms and socio-demographic characteristics.

Results

In the tests conducted based on the data made available in [1], the aim is to understand the relationship between the symptoms Cough, Fever, Sore Throat, Runny Nose, Muscle Pain, Nausea, Diarrhea, Loss of Smell, and Shortness of Breath, with the binary outcome of the COVID-19 test.

Figure 1 presents a comparison of the ROC curve for the 3 models used. It can be observed that the RF and XG-Boost models outperform LR. However, these are models whose results are not easily interpretable, unlike LR, whose parameters are directly related to attribute importance and allow for the calculation of odds ratios.

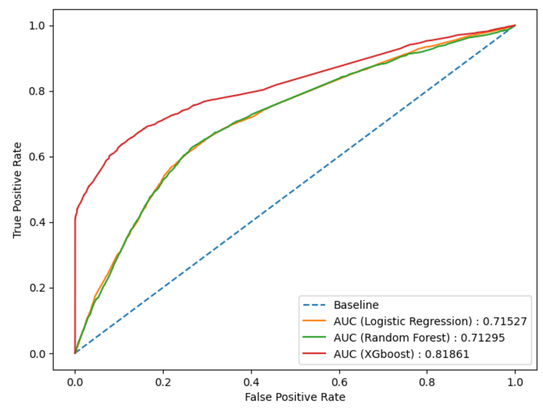

Figure 1: ROC curve for each model used.

In this sense, we used the SHAP method to calculate the distribution of Shapley values, by attribute, for the LR and XG-Boost methods, which are represented in Figure 2. Figure 3 presents the odds ratio values obtained from the LR model, along with their respective 95% confidence intervals.

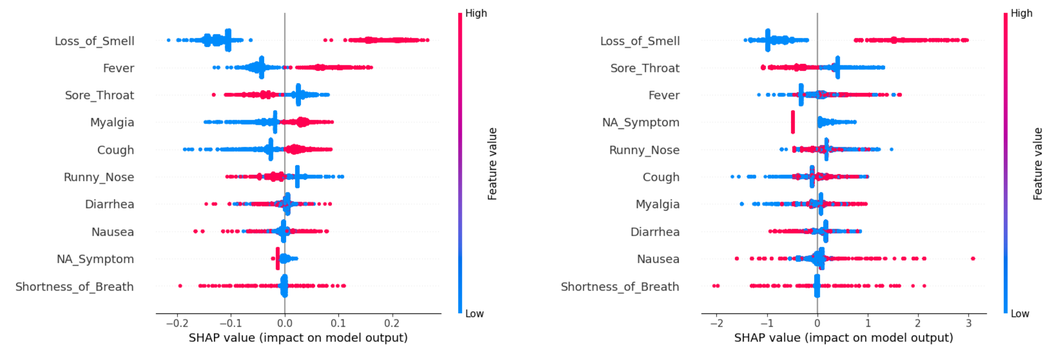

Figure 2: Summary Plot Random Forest e Xgboost.

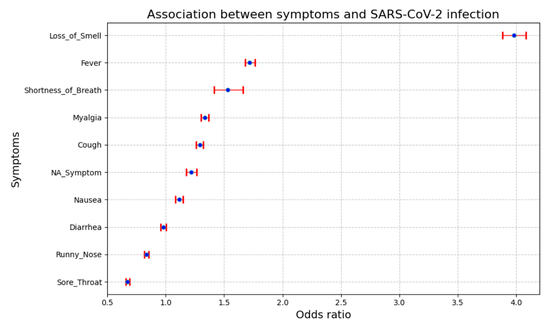

Figure 3: Association between symptoms and SARS-CoV-2 infection using the Logistic Regression model.

Conclusion

It can be observed that the symptom related to Loss of Smell showed the highest importance in the interpretation across all models used, indicating that the presence of loss of smell in patients significantly increases the predicted probability of having the SARS-CoV-2 pathogen.

When analyzing the symptom "Sore Throat," it's clear that the interpretability results align. Although this attribute ranks high in importance for both SHAP methods, the color distribution along the x-axis indicates that it reduces the likelihood of testing positive for COVID-19. This finding is consistent with the result shown in Figure 3, where the odds ratio for this symptom is approximately 0.67, suggesting a 33% reduction in the chance of a positive test when this symptom is present. Given that the data were collected in 2020, this finding aligns with medical literature on symptom prevalence during the first wave of COVID-19.

References

[1] Leila F Dantas and et al. App-based symptom tracking to optimize sars-cov-2 testing strategy using machine learning. *PloS one*, 16(3):e0248920, 2021.

[2] et al. Loftus, T. J. Ideal algorithms in healthcare: Explainable, dynamic, precise, autonomous, fair, and reproducible. *PLOS Digital Health*, 2022.

[3] Christoph Molnar. *Interpretable Machine Learning*. Github, 2 edition, 2022.