MEDINFO 2025 — Healthcare Smart × Medicine Deep
M.S. Househ et al. (Eds.)

doi:10.3233/SHTI250908

Explaining Machine Learning: A Deeper Look into Admission Prediction

Bernardo CONSOLI [a,1], Vinícius PEDROSO [a], Artur KNIEST [a], Renata VIEIRA [b], Rafael H. BORDINI [a], Isabel H. MANSSOUR [a]
[a] *Pontifical Catholic University of Rio Grande do Sul*
[b] *University of Évora*
ORCiD ID: Bernardo Consoli https://orcid.org/0000-0003-0656-511X

Abstract. The popularization of artificial intelligence solutions in both research and industry that has been occurring due to the rise of tools such as the GPT, Gemini and Claude large language models has revitalized research in the area. There are many possible uses within the medical field, but a key determinant of the adoption of new tools by medical professionals is trust. To augment tool trust, the tool must be made understandable and explainable, but this is a problem for "black box" machine learning models. In an effort to promote transparency, we have performed a deep study of the reasoning behind an XGBoost machine learning model that performed well in the task of inpatient admission prediction.

Keywords. Explainable AI, SHAP, XGBoost, Inpatient admission prediction

1. Introduction

Many new solutions for computational medicine suggest using artificial intelligence and machine learning tools [1]. Many of these solutions do not tackle one of the most serious issues that faces the implementation of such solutions in real scenarios and adoption by practicing medical professionals: explainability [2].

In order to establish the trust of professionals in AI tools and solutions, it is vital that the users are able to understand the reasons behind a model's answers. Enhancing the explainability and interpretability of often opaque "black box" machine learning models is often seen as the best way to bridge this "understanding gap" [2].

The tasks in computational medicine that could benefit from the use of AI include nursing assistance, diagnostic assistance, managerial assistance, patient engagement, reduced cost, etc [3].

Many studies have shown that machine learning solutions are effective for these tasks [4]. As such, it is paramount to build these systems to be trustworthy and accountable so that they may find real world usage. Explainability can be achieved in many ways for different architectures [5]. Partial Dependence Plots (PDP) can explain how a feature influences the predicted outcome by maintaining the rest as constants; Accumulated Local Effects is similar, but based on differences rather than PDP's averages and, as such, considered less biased; Local Interpretable Model-Agnostic Explanations performs multi-feature perturbations around a prediction and measures the

[1] Corresponding Author: Bernardo Consoli, bernardo.consoli@edu.pucrs.br.

differences; Shapley Additive Explanations (SHAP) uses game theory to assign feature importance; and quite a few others [5].

We have chosen to tackle the task of inpatient admission prediction through a lens of explainability using the BRATECA collection [6], a collection of tertiary care hospital data from Brazilian hospitals. We used an XGBoost model to perform this task, and then used SHAP to explain the results achieved.

2. Methods

2.1. Data

For our data, we used the BRATECA collection [6]. We extracted 72 structured features and free-text clinical notes from it. These 72 features can be divided into three categories: Administrative, Exam, and Prescription. Administrative features include age, sex, and skin color. Exam and prescription features are the most frequent exam results and the details of the most frequent prescriptions in the training data, respectively. These data are sparse.

In addition to these data, clinical notes were compiled for each patient. The clinical notes were vectorized using a Term Frequency Inverse Document Frequency featurizer. Figure 1 describes the data pipeline, and a more in-depth explanation can be found in Consoli et al. 2024 [7].

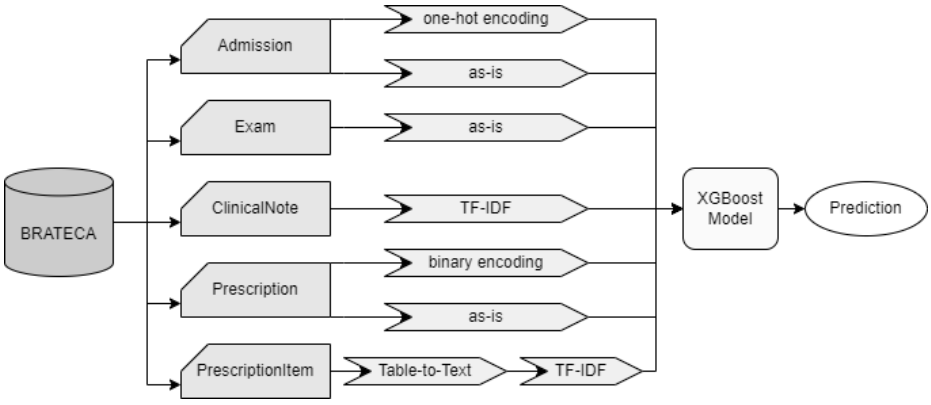

Figure 1. Diagram of the data pipeline.

2.2. Machine Learning Model and Task

The machine learning model used is a simple XGBoost [8] trained to perform binary predictions. The task it was trained to perform was “Inpatient Admission Prediction”. This task is to predict whether a patient will remain in the hospital for more than 24 hours once they are admitted or if they will leave before this timeframe. The task was performed twice, using patients at the 1-hour and 8-hour marks of their hospital stay.

The best scores achieved by the model for each task were as follows: weighted F1 of 0.76 with 1-hour patients; and weighted F1 of 0.88 with 8-hour patients, as detailed in Consoli et al. [7].

2.3. SHapley Additive exPlanations (SHAP)

SHAP [9] is a game theory-based approach to explain the output of machine learning models. It can be employed to evaluate all features used in a model for their contribution and is able to provide insight into potential biases and the reliability of the predictions.

We used SHAP to determine which of the 72 structured features used by our model were the most and least impactful for the two scenarios we tested. This resulted in a few plots we relied on to examine feature relevance.

3. Results

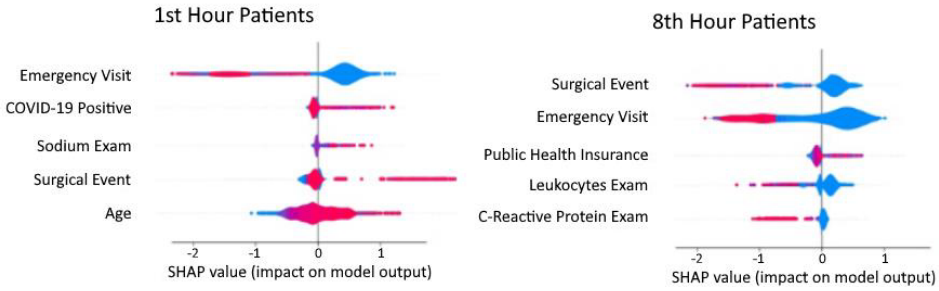

Figure 1. SHAP beeswarm plots for the five most important features for each kind of patient. Blue indicates a low value or False (i.e. blue Surgical Event means the patient did not have a Surgical Event while a blue age means a lower age. Red indicates a high value or True. A widening of the line means more patients at that level of effect, and the closer to the edges, the more significant the impact. Negative impact indicates that the patient is more likely to leave, while positive impact indicates that the patient is more likely to be admitted.

3.1. Scenario 1: Patients at the end of their 1st hour in hospital

There is a dearth of information about the current situation of patients who have been in hospital for a very short time. Usually, only a few basic tests have been run on them, and mostly administrative information relating to the nature of their visit, their age and sex, and their health insurance is known at this point. This is shown in the importance metrics discovered with our SHAP test, as seen in Table 1.

Table 1. Most important structured variables.

Variable	Importance (%)
Emergency Visit	55.85
COVID-19 Positive	10.97
Sodium Exam	6.89
Surgical Event	5.10
Age	3.25
Leukocytes Exam	2.07
C-Reactive Protein Exam	1.91
CI Exam	1.74

As can be seen, the “Emergency Visit” feature, a binary feature that determines whether a patient required emergency care upon entry to the hospital, is extremely relevant for predicting the stay of recent arrivals. The test reveals that most emergency patients will not stay in hospital, leaving before 24 hours.

The remaining of the five most relevant variables are: COVID-19 Positive status; a sodium blood exam to check for hyponatremia; the advent of a surgical event; and patient age. These are sensible variables for the model to focus on. Given a shortage of information, COVID-19 positive status, especially in 2020/2021, when this data was taken, would often result in hospitalization, as would many surgical events slated for as little as an hour after the patient's arrival. The sodium test results could indicate that symptoms of hyponatremia were observed during triage, which can be serious even if it is caused by something else. Finally, it stands to reason that older patients would have a higher likelihood of staying in the hospital for longer periods.

3.2. Scenario 2: Patients at the end of their 8th hour in hospital

At 8 hours, a patient is likely to have had more exams performed, so their immediate circumstances upon arrival are weighted less by the model. The Emergency Visit variable, for example, has dropped from 55% importance to 20%, and the Surgical Event variable has increased in importance from 5% to 28%. The importance of the patient's insurance type has also increased, with it having been shown that public insurance patients are more likely to stay in hospital for longer periods. This is shown in the importance metrics discovered with our SHAP test, as seen in Table 1.

Table 2. Most important structured variables.

Variable	Importance (%)
Surgical Event	28.50
Emergency Visit	20.37
Public Health Insurance	6.75
Leukocytes Exam	5.19
C-Reactive Protein Exam	4.67
Hemoglobin Exam	2.61
Fluoridated Plasma Exam	2.50
Platelets Exam	2.17

Here, we can observe that, after 8 hours, exams become much more relevant since more of them are likely to have been conducted. It is very interesting to note, however, that the top three most important variables have nothing to do with exams. Whether a patient has experienced a surgical event and whether the patient is in emergency care are by far the most relevant variables when attempting to predict longer patient stays.

It is notable that now the model considers surgical events to be indicative that a patient will soon leave the hospital rather than be admitted. This reflects the higher prevalence of benign, pre-scheduled surgeries over emergency surgeries, which have not been differentiated in this dataset.

4. Discussion

For both scenarios, we can see that the machine learning algorithms choose sensible variables to base their predictions on. The most important variables and how they are used are explainable and understandable enough that, with perhaps some tuning, they could reach a state of usefulness in real-world scenarios.

It also shows that models treat variables from patients at different times of their treatment differently. This indicates that further dividing the task into more scenarios,

such as using 1 hour of information to predict whether a patient will leave before 8 hours, could be a valid approach to improve model scores and distinguish clearly different kinds of patients earlier on in their stay for improved efficacy and trust, since accuracy of prediction is the best way to build trust in tools, alongside explainability and interpretability.

5. Conclusion

In this study, we have shown that the use of pre-established methodology is vital in increasing medical professionals' understanding and trust in machine learning tools. As such, these studies must be made a priority for computer scientists who wish for their tools to be implemented in real-world scenarios.

The examination of variable use by models for the admission prediction task specifically proved helpful in showing that the model's decisions are not completely unintelligible to human operators. In future work, we hope to use similar methods to explain neural network style models, as well as to help tune existing models to achieve higher accuracy of predictions. We also intend to expand the list of examined tasks, including length-of-stay prediction and re-admission prediction, to form a complete examination of the patient flow dynamics of hospitals.

Acknowledgements

We gratefully acknowledge partial financial support by CNPq Scholarship - Brazil (projects PIBIC, 303208/2023-6 and 25/2020), FAPERGS 22/2551-0000390-7 (RITE CIARS), Capes, the FCT under projects CEECIND/01997/2017 and UIDB/00057/2020 (Portugal), and the Brazilian Ministry of Science, Technology and Innovation, with resources from Law no 8.248, 23 of October 1991, in the scope of PPI-SOFTEX, coordinated by Softex.

References

[1] Sijie Y, Zhu Fei, Buckwalter JG, Xinghong L. Intelligent Health Care: Applications of Deep Learning in Computational Medicine. Frontiers in Genetics. 2021; Vol. 12, doi: 10.3389/fgene.2021.607471

[2] Combi C, Amico B, Bellazzi R, et al. A manifesto on explainability for artificial intelligence in medicine. Artificial Intelligence in Medicine. 2022; Vol 133, doi: 10.1016/j.artmed.2022.102423

[3] Lee D, Yoon SN. Application of Artificial Intelligence-Based Technologies in the Healthcare Industry: Opportunities and Challenges. International Journal of Environmental Research and Public Health. 2021; 18(1):271. doi:10.3390/ijerph18010271

[4] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthcare Journal. 2019; 6(2):94-8. doi: 10.7861/futurehosp.6-2-94

[5] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy. 2021; 23(1):18. doi: 10.3390/e23010018

[6] Consoli BS, dos Santos HD, Ulbrich AH, Vieira R, Bordini RH. Brateca (brazilian tertiary care dataset): a clinical information dataset for the portuguese language. In: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille, 2022 Jun.

[7] Consoli BS, Viera R, Bordini RH, Manssour IH. Predicting Inpatient Admissions in Brazilian Hospitals. In: Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS). 2024 Jun 25; pp. 284-295.

[8] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13; pp. 785-794.

[9] Scott M, Su-In L. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017 Dec 4; 4765-74.