MODEL-AGNOSTIC EXPLANATIONS: A CASE STUDY FOR DRUG SERIOUSNESS PREDICTIONS

Arman Grigoryan
Ph.D. in Engineering (Computer Science)
EUA, Chair of Information Technology and Applied Mathematics
armgrigoryan@gmail.com

Abstract

In practice, complex machine learning models are commonly outperforming traditional models, however, it is significantly difficult for clinicians to understand and trust these complex models due to the lack of intuition and explanation of their predictions. This paper aims to study and demonstrate the use of various model-agnostic explanation techniques of machine learning models with a case study for explaining drug seriousness predictions based on the FDA Adverse Event Reporting System dataset.

The experiments and results in this paper show that different interpretability techniques can vary in the explanations of the model behavior leading to better and more meaningful predictions and decisions. While global interpretations can generalize the entire population of the model-generated results and help the clinicians to understand the entire behavior of the model, the local interpretations enable the clinicians to understand the explanations at the level of individual instances. All these open interesting insights and provide new opportunities for AI adoption in the pharma domain.

Keywords: Model-agnostic explanation, model behavior, drug seriousness prediction, explainable artificial intelligence, XAI, LIME, SHAP.

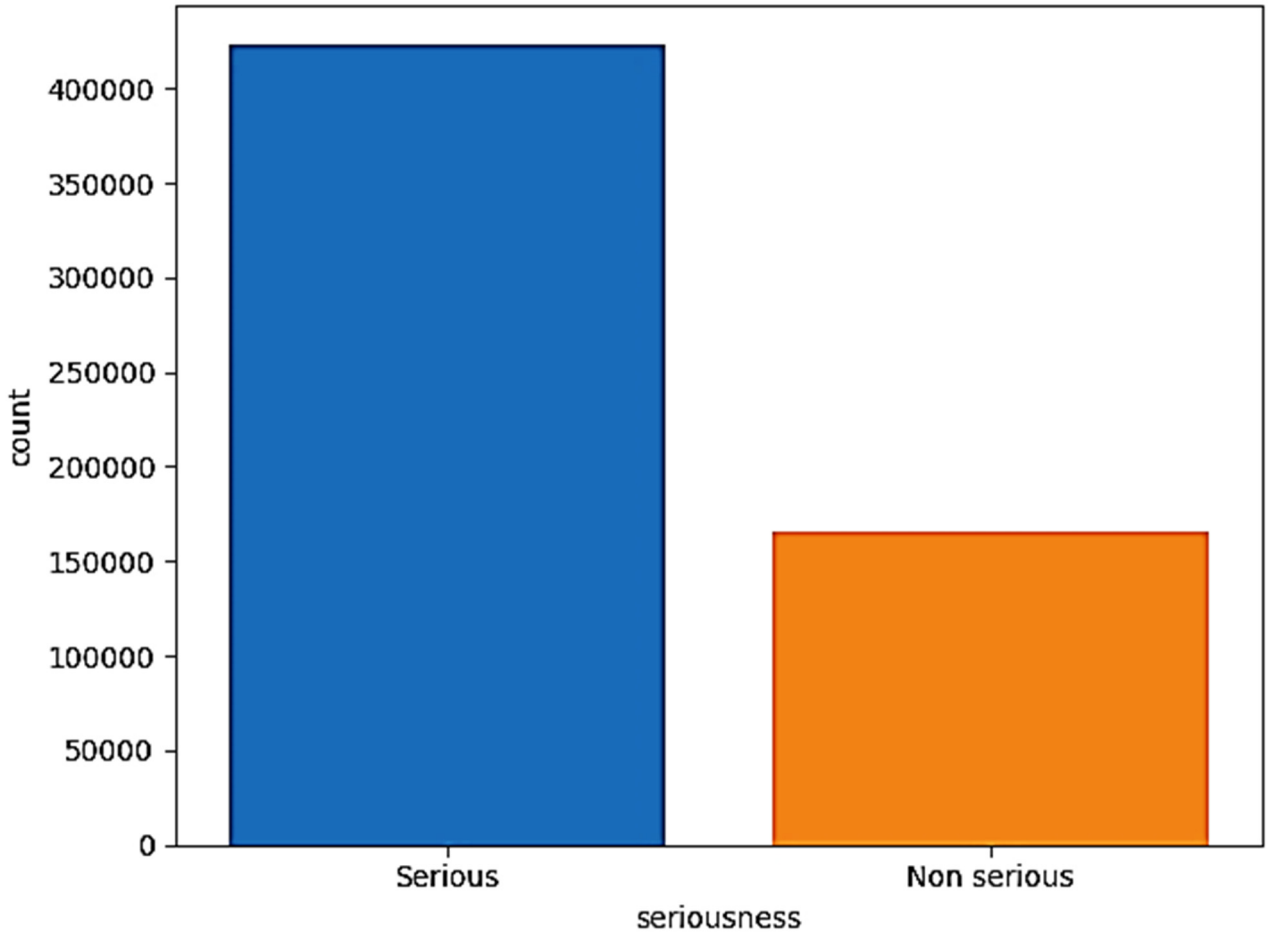
Introduction

Although, there are numerous models developed to deal with adverse event analysis and predict the level of severity, however, there is a lack of frameworks that establish trust and confidence in these predictions. Generally, the medical/pharma domain was always adopted a preservative approach to innovations, and thus, there was always some criticism for using machine learning models, even in terms of the high accuracy of the models. On the one hand, such an issue is critical because the models that are being used in practice, especially black-box models, should provide trust and safety in terms of the right predictions and decisions, on the other hand, meaningful explanations will be useful to improve model performance and lead to the business value [Chen, Asch 2507].

Data Background

In this paper, the FDA Adverse Event Reporting System (FAERS) database which contains adverse event reports, medication error reports, etc. was used. The database is designed to support the FDA's post-marketing safety surveillance program for drug and therapeutic biologic products. The information in these reports has not been scientifically or otherwise verified as to a cause-and-effect relationship and cannot be used to estimate the incidence of these events. For any given report, there is no certainty that a suspected drug caused the reaction. This is because physicians are encouraged to report suspected reactions; however, the event may have been related to the underlying disease being treated or caused by some other drug being taken concurrently, or simply occurred by chance at that time. Accumulated reports cannot be used to calculate incidence (occurrence rates) or to estimate drug risk. Comparisons between drugs cannot be made from these data [FDA FAERS Database]. The dataset that was created based on the FAERS database during data preprocessing contains 13 features and 2 target classes.

Fig. 1.



Relationship between serious and non-serious cases.

Machine Learning Model

A Voting Classifier by Sklearn was used to create a machine learning model for training on an ensemble of numerous models, predicting an output (class) based on the highest probability of a chosen class, and aggregating the findings of each classifier passed into Voting Classifier according to the results of the highest majority of voting. The voting classifier aims to create a single model which trains the other models and predicts the output based on their combined majority of voting for each output class. In this case, RandomForestClassifier, LogisticRegression, and SVC were included in the Voting Classifier to create the ensemble model which allows ensuring the error of one model is resolved by the other. The model performs binary classification that predicts whether or not the case in the data set has a serious effect [Machine Learning in Python].

Feature Selection

Features used in the model:

- F1: *indi_c*: Medical terminology describing the Indication for use,
- F2: *gender_c*: Gender of the patient,
- F3: *drug_seq*: Unique number for identifying a drug for a case,
- F4: *indi_drug_seq*: Drug sequence number for identifying a drug for a case,
- F5: *ai_nm*: Product active ingredient,
- F6: *age*: Age of the patient.

The following features presented in the figure below were selected based on the calculations on the feature importance matrix.

Fig. 2.

	drug_seq	caseid	age	indi_drug_seq	indi_c	gender_c	seri_c
drug_seq	1.000000	-0.058478	-0.027464	0.239476	0.002970	0.005671	0.106472
caseid	-0.058478	1.000000	-0.066927	-0.227828	-0.075802	-0.000836	-0.152175
age	-0.027464	-0.066927	1.000000	0.083388	-0.023464	0.005075	0.081143
indi_drug_seq	0.239476	-0.227828	0.083388	1.000000	0.033140	0.022143	0.182091
indi_c	0.002970	-0.075802	-0.023464	0.033140	1.000000	0.031622	0.259666
gender_c	0.005671	-0.000836	0.005075	0.022143	0.031622	1.000000	0.157918
seri_c	0.106472	-0.152175	0.081143	0.182091	0.259666	0.157918	1.000000

Feature importance matrix.

Classification Report

The whole data set is split into 67% train and 33% test data sets. The classification report reveals that the macro average of the F1 score is about 0.78, which indicates that the trained model has a classification strength of 78%.

Fig. 3.

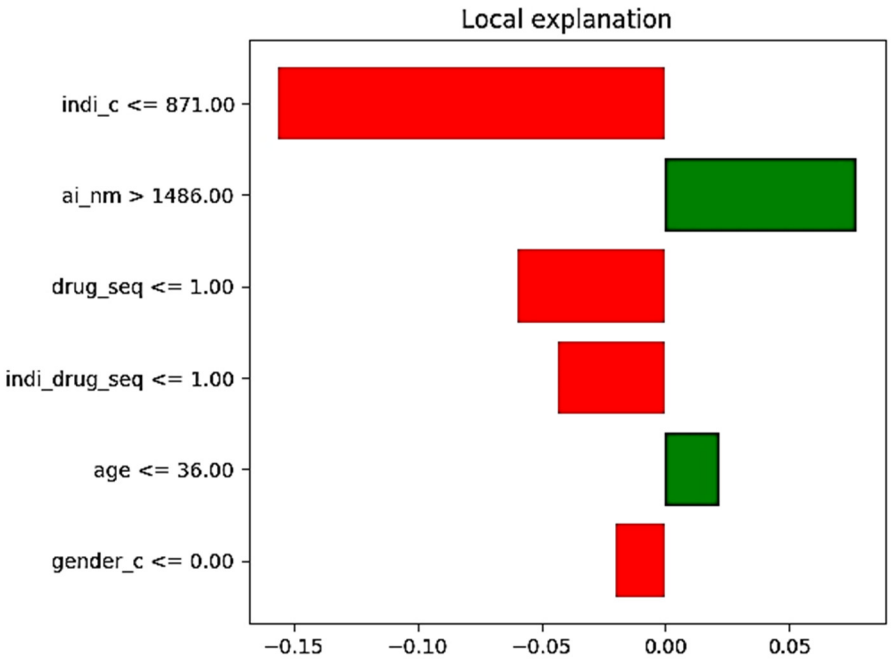
	precision	recall	f1-score	support
0	0.96	0.50	0.66	54624
1	0.83	0.99	0.91	139567
accuracy			0.85	194191
macro avg	0.90	0.74	0.78	194191
weighted avg	0.87	0.85	0.84	194191

Classification report for Voting Classifier model.

Explanations by LIME

LIME (Local Interpretable Model-Agnostic Explanations) is based on the concept of surrogate models. Local refers to local fidelity which means that explanation must reflect the behavior of the classifier around the instance being predicted. Lime is model-agnostic and supports explanations for individual predictions from a wide range of classifiers. Lime can be used to explain predictions of tabular, text, and image data. You can find drug seriousness prediction explanations performed by Lime tabular explainer [Ribeiro et al. 1139; Garreau, von Luxburg 1289; LIME: Explaining the Predictions of any Classifier].

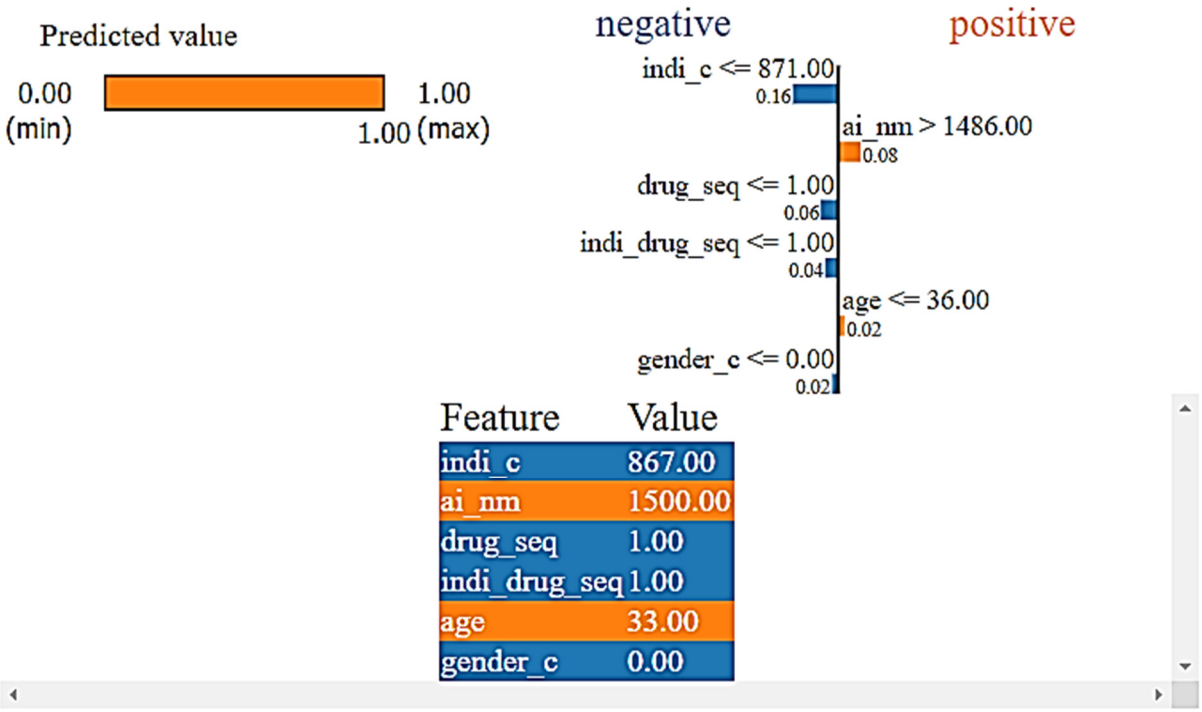
Fig. 4.



Local explanation by LIME.

The abovementioned plot represents the coefficients generated by the Lime tabular explainable model. Green color indicates the features that have positive and the red one has negative correlations with the target. In this regard, high values for active ingredients and age positively correlate with seriousness. Accordingly, low values for gender and indication, as well as drug sequence negatively correlate with seriousness. The same logic works for the rest of the features. More detailed explanations for this specific observation can be found also in the plot presented below.

Fig. 5.



LIME tabular explanation.

The following plots illustrate the results of Lime text explainer for the serious and non-serious classes by emphasizing the words found in the narrative. As a result, though the case report was classified as serious, however, the explanation indicates that only "Breast" positively affects classifying the case report as serious, while "cancer" supports the non-serious class.

Fig. 6.

```
Document id: 256
Probability(Non serious) = 0.08983317431962472
Probability(Serious) = 0.910166825680375 2
True class: Serious
```

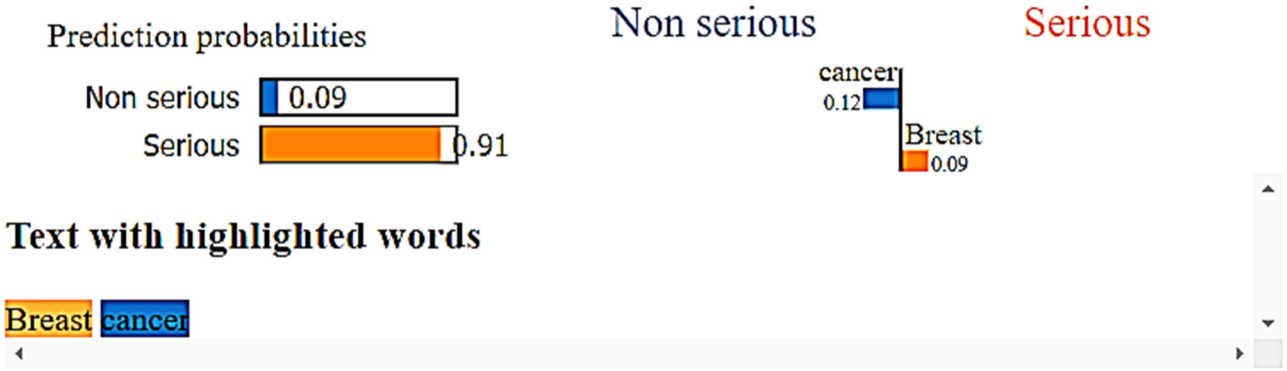
Model generated prediction for text explanation.

Fig. 7.

	precision	recall	f1-score	support
Non serious	0.88	0.59	0.70	54624
Serious	0.86	0.97	0.91	139567
accuracy			0.86	194191
macro avg	0.87	0.78	0.81	194191
weighted avg	0.86	0.86	0.85	194191

Classification report for LIME text explanation model

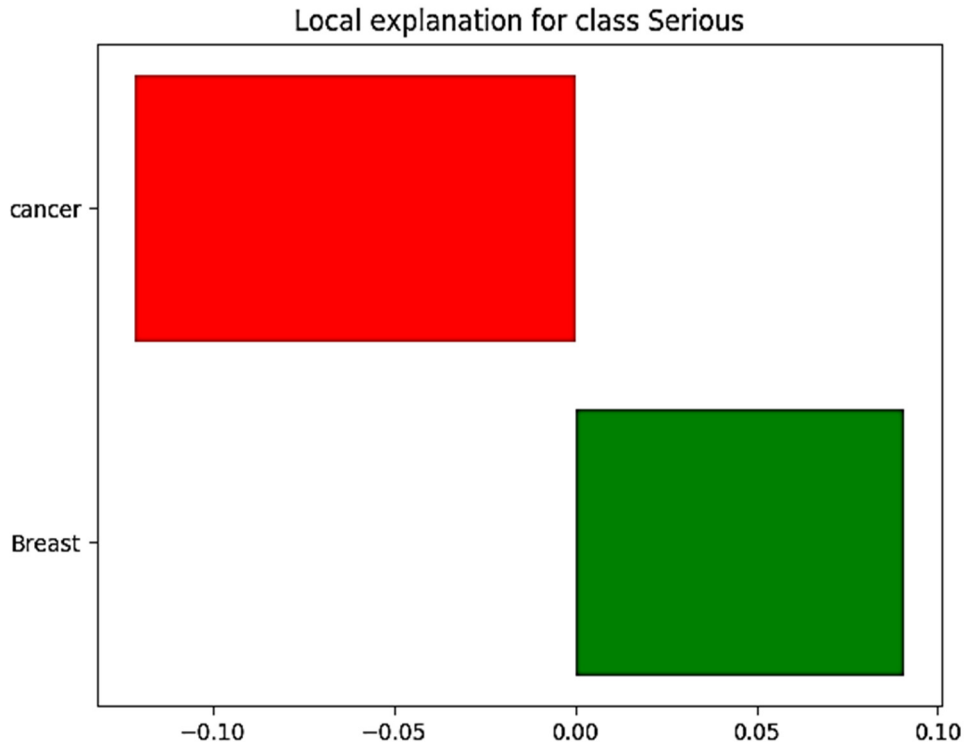
Fig. 8.



LIME text explanations.

The chart below presents the local explanation for the class serious, where the supporting words for the serious class are in green, while red color indicates the words that are supporting the non-serious class.

Fig. 9.

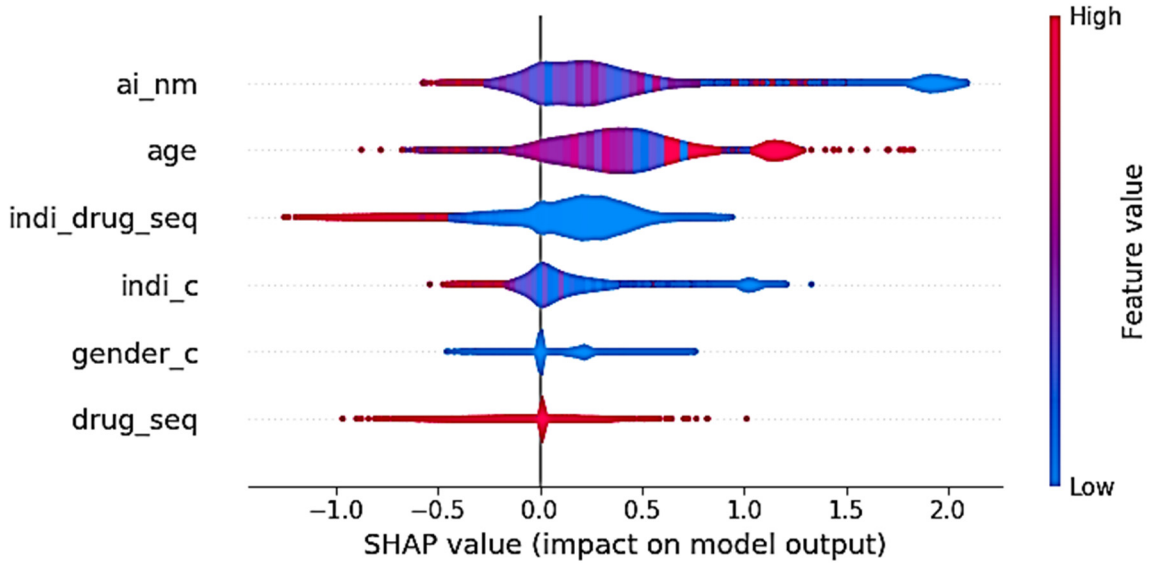


LIME explanation for the "Serious" class.

Explanations by SHAP

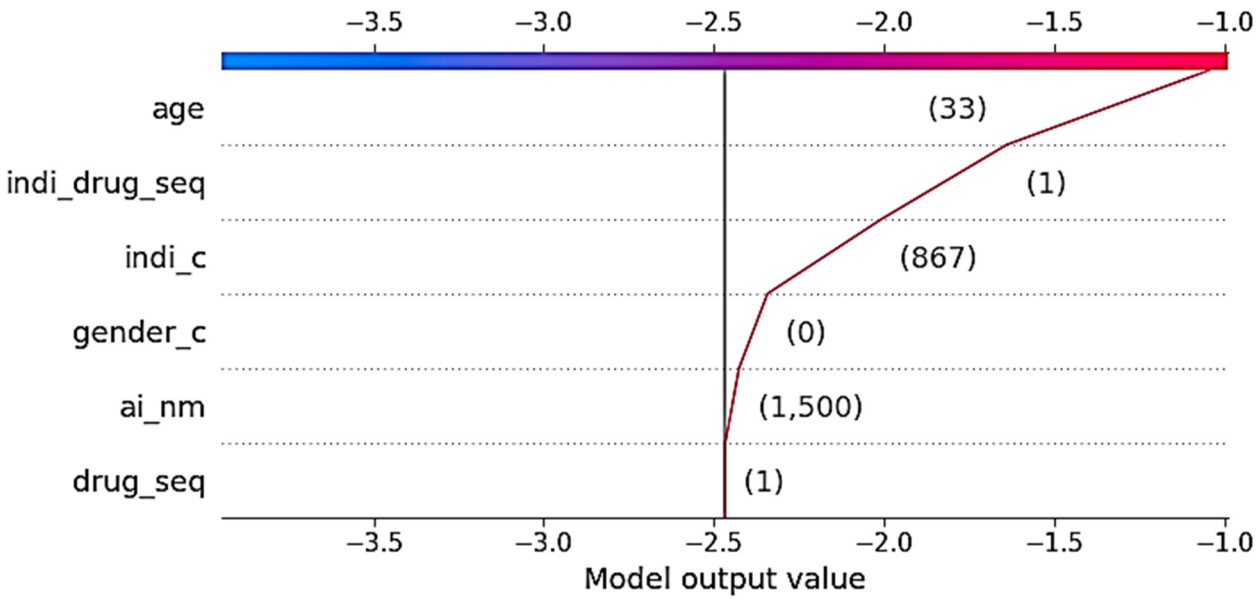
SHAP (SHapley Additive exPlanations) unifies all available frameworks for interpreting predictions. SHAP assigns each feature an importance value for a particular prediction. Based on insights from this unification, it presents new methods that show improved computational performance and/or better consistency with human intuition than previous approaches [Strumbelj, Kononenko 654; Lundberg et al.72-76; SHAP: A Game Theoretic Approach to Explain the Output of Any Machine Learning Model].

Fig. 10.



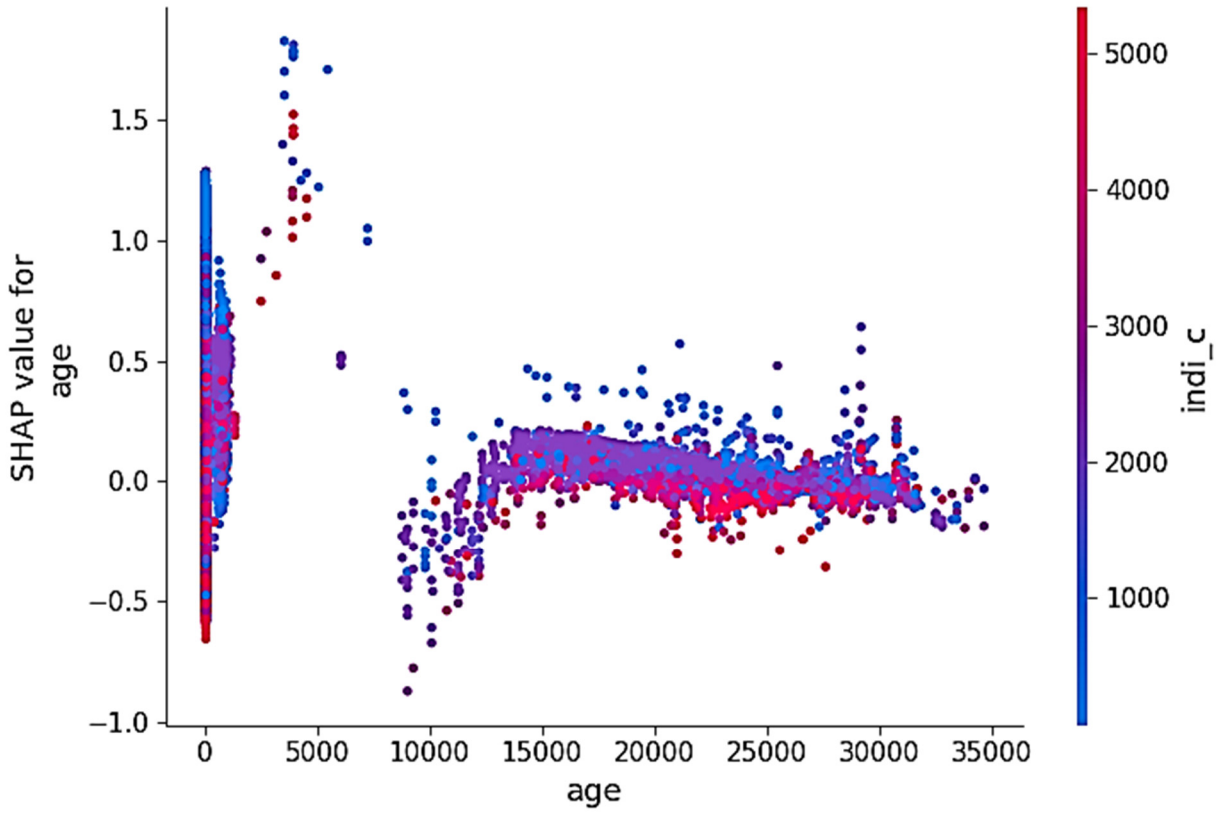
SHAP explanation with the summary plot.

Fig. 11.



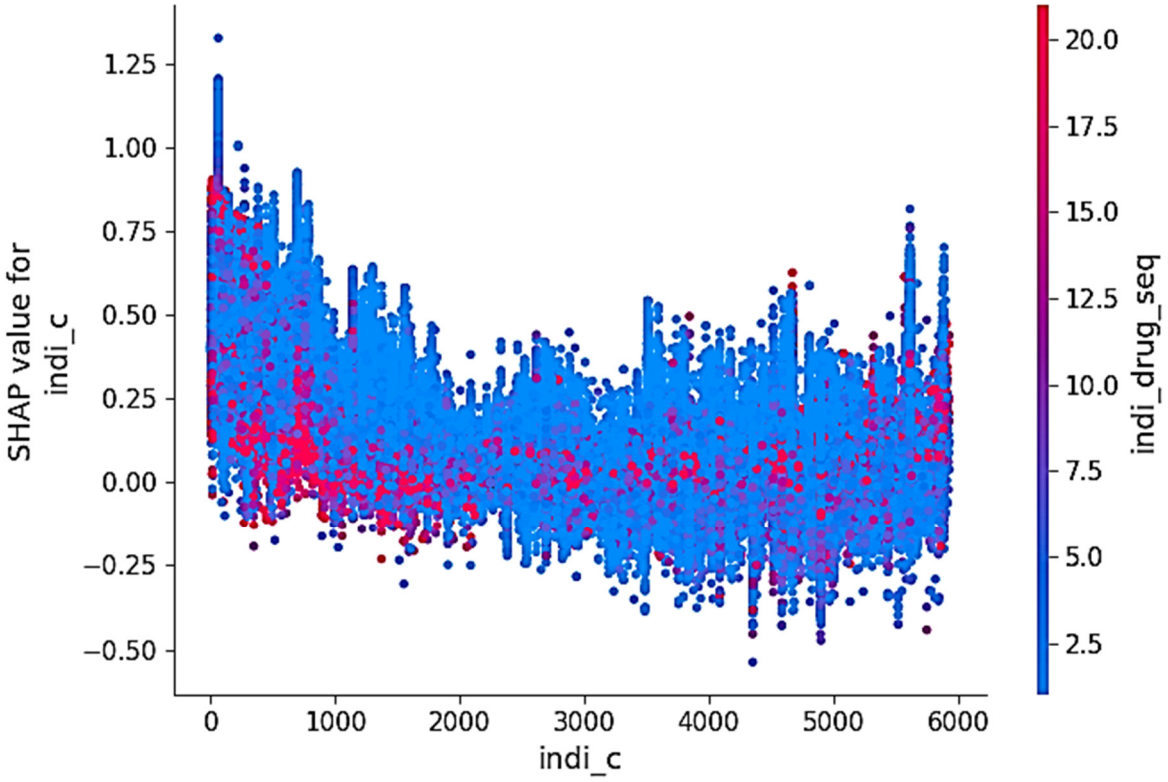
SHAP explanation with decision plot.

Fig. 12.



SHAP dependence plot (age/indication).

Fig. 13.



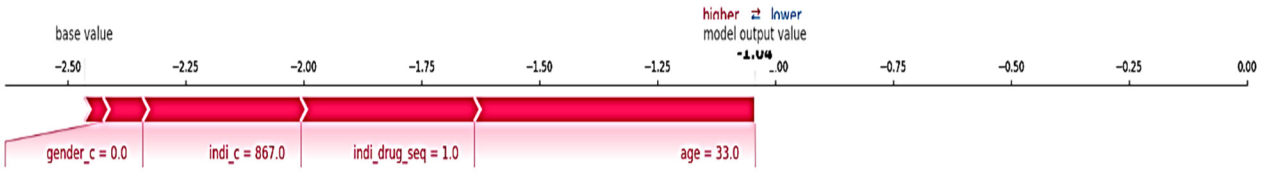
SHAP dependence plot (indication/indication drug sequence).

SHAP summary plot shows top feature contributions and data point distribution. It provides visual indicators of feature values that affect predictions, where red and blue colors indicate accordingly higher and lower feature values.

The decision plot shows the model's inner workings and the way it makes decisions. The vertical line is the model base value. The colored line is the prediction. Feature values are printed next to the prediction line indicating feature effects.

The dependence plots illustrate the relationship between feature values and predicted outcomes. SHAP values show the role of the feature in changing the model output. The model automatically selects another feature for coloring. In this case, coloring by indication drug sequence highlights that the average number of indications has less impact on serious cases in terms of high sequence value.

Fig. 14.



SHAP explanation with force plot.

The above-considered explanation shows the features, each of which is contributing to pushing the model output from the base value to the model output. Features pushing the prediction higher are shown in red. Accordingly, pushing the prediction lower is in blue. Fig. 14 illustrates the explanation for the individual case. The same explanations can be done for the entire dataset.

Conclusion

Although the provided methods support predictions with explanations, it's up to the domain experts to accept or reject the outcome of the explainable model based on their domain expertise. To compare LIME with SHAP for this specific drug seriousness classification task, it is recommended to choose LIME over SHAP, as explanations in this context are mainly being provided for individual cases. Besides, LIME is faster than SHAP and will be more suitable to utilize in the text explaining scenarios. On the other hand, SHAP can be applied in parallel with LIME, if there is a need to perform global model interpretations, as well as double-check the results of the explanations of one or more methods.

Indeed, FAERS is a pretty good resource to study drug effects. However, structured data does not incorporate confounding factors including concomitant medications and patient medical histories found in narratives. In this regard, a combination of adverse drug reactions in social media with FAERS and drug information databases, as well as developing more sophisticated explainable models for named entity recognition (NER), etc. will bring the studies in drug seriousness predictions to a new level.

The abovementioned can be achieved by adopting well-defined validation and qualification mechanisms, which will be projected into the user-friendly interfaces, as well as more detailed and meaningful documentation.

References

1. Chen, J.H.; Asch, S.M. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med. 2017; 376(26):2507-2509.
2. FDA Adverse Event Reporting System (FAERS): Latest Quarterly Data Files. Available at: https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html
3. Machine Learning in Python (scikit-learn). Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html
4. Ribeiro, M.T.; Singh, S.; Guestrin, C. 2016. Why Should I Trust You? Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD. ACM, 1135–1144.
5. Garreau, D.; Von Luxburg, U. Explaining the Explainer: A First Theoretical Analysis of LIME. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS). Palermo, Italy. 2020; Volume 108: 1287-1296. Available at: https://arxiv.org/pdf/2001.03447.pdf
6. LIME: Explaining the Predictions of any Classifier. Available at: https://github.com/marcotcr/lime
7. Strumbelj, E.; Kononenko, I. "Explaining Prediction Models and Individual Predictions with Feature Contributions." Knowledge and Information Systems, 41.3 (2014): 647-665.
8. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. Explainable AI for Trees: From Local Explanations to Global Understanding. arXiv preprint arXiv:1905.04610, 2019; 72p.
9. SHAP: A Game Theoretic Approach to Explain the Output of Any Machine Learning Model. Available at: https://github.com/slundberg/shap

Ներկայացվել է՝ 01.10.2022թ.
Ուղարկվել է գրախոսման՝ 01.11.2022թ.