Abstract

The field of Explanatory artificial intelligence is showing promising growth in recent years, thus giving researchers the option of deeply exploring the benefits and drawbacks of many different proposed models for solving the enigma of interpretability and explainability regarding machine learning models and their predictions. There are currently a number of techniques that can help to assist researchers in understanding the logic behind decisions made by various models, but the focus of this paper will mostly be on discussing and comparing two strong options, LIME (Local Interpretable Model Agnostic Explanations) and SHAP (Shapley Additive explanations). The proposed pipeline for the comparison will be given in a form of an Orange Data Mining workflow. Secondly, the paper aims to give a proposal of how a custom widget encapsulating the functionality of the LIME library can be integrated into the graphical interface, making it’s usability more appropriate towards less experienced users.

Keywords: Explainable artificial intelligence (XAI), interpretability, explainability, LIME, SHAP, Orange Data Mining

Introduction

The rapid rise of machine learning and artificial intelligence based techniques in recent years has provided many opportunities for enabling researchers from different branches to develop a variety of solutions for real world problems. These problems are distributed through various domains such as healthcare, education, business, scientific research and many more.

Terms such as interpretability and explainability are often used in the Explanatory artificial intelligence (XAI) domain interchangeably and currently there are is not a clear consensus for their meaning, thus they appear underspecified. On the other hand, as argued in [2], Gilpin, Bau et al. give a view in which these terms should be viewed on their own. They take a stance that interpretability is not enough by itself, rather that explainability is needed for people to trust black-box models. In their view, the goal of interpretability is “to describe the internals of a system in a way that is understandable to humans”. They also add that “in order for a system to be interpretable, it must produce descriptions that are simple enough for a person to understand using a vocabulary that is meaningful to the user”.

As stated in [3], the desired characteristics for explainers, defined by Marco Tulio Ribeiro et al. are that explainers have to be interpretable, i.e. “to provide qualitative understanding between the input variables and the response”. The meaning behind this definition also acknowledges that explanations need to be easy to understand. It is also stated that interpretability is not a term that can be used in the same manner for different types of target audiences with various levels of expertise.

Some of the machine learning models are considered to have a black box approach which can be a problem when working with sensitive matters involving human lives. Black box models can’t give much of the necessary explanations that would be greatly needed and beneficial for researchers.

In order to make a leap towards fair and ethical decision making, there is a need for discussing the transparency notion of interpretability as shown in [1]. In order to confer interpretability, Lipton proposes that there are two categories in which techniques and model properties fall, first being transparency and the second post hoc explanations. The two terms are defined as “how does the model work” and “what else can a model tell me”, respectively. It is also stated that transparency is on the opposite side of the spectrum from the black-box approach mentioned earlier.

Knowing all of that, there was a need to develop solutions that would solve this particular problem and provide some additional explanations for the inner logic behind machine learning models and predictions. A number of solutions were proposed, giving a variety of options to choose from. Linardatos et al give a comprehensive review of interpretability methods to explain any black-box models [4].

The main focus of this paper is not to show all of the available solutions, rather to give an understanding how two different models, LIME (Local Interpretable Model-Agnostic Explanations) [3] and SHAP(Shapley Additive explanations) [5] work in order to explain the reasoning behind a machine learning model on a given dataset. Secondly, we propose an example of an alternative way of using LIME through Orange Data Mining as a custom Add-on widget.

We believe that this way of using LIME could make the library more accessible to a greater amount of people that don’t have the necessary technical skills, as Orange Data Mining provides an visual programming experience.

This paper is organized in ... different sections...(dodacu kad rad bude gotov)

Related work

In this section a brief overview of the current work regarding the use of SHAP and LIME libraries will be shown. In his work, Ignatiev raises the question of trustable explainable AI (XAI) as he states that the majority of approaches to XAI are of heuristic nature [7]. The alternative would be the use of rigorous logic-based approaches which he states “is indispensible if trustable XAI is of concern”. Later in the paper, he gives a comparison of the Anchor, LIME and SHAP (criteria used for comparison is the percentage of incorrect, rendudant and correct explanations provided by these libraries). The paper by Palacio, Lucieri et al. that aims to give a theoretical framework that can be used to enable a comparioson of LIME, SHAP and MDNet [8]. Comparing LIME and SHAP has also been the focus of Misheva, Hirsa et al. where they do a deep dive with evaluting both models on a credit risk menagement dataset [9]. There is a number of papers discussing the use of Explainable AI in healthcare using SHAP and LIME among other methods [10][11][12][13].

Explainable AI Libaries - Deo sa detaljnijim objasnjenjem LIME-a i SHAP-a

Methodology

The purpose of this section will be to compare two different approaches in the Explanatory artificial intelligence area. First, an overview of the proposed pipeline for the comparison of before mentioned LIME and SHAP libraries will be shown. Following that, a brief overview of the possible implementation of a custom Orange Data Mining Add-on widget will be given. The widget in question encapsulates some of the functionality given in LIME, more specific, the explain instance part.

It is important to acknowledge that Orange Data Mining supports an Explain Add-on in their widget catalog. It consists of Feature Importance, Explain Model, Explain Prediction, Explain Predictions and ICE widgets. The inner core of Explain Prediction widget is the SHAP model of explainability, while we provide an alternative using LIME. Both SHAP and LIME are widely used but they have some distinct strengths and weaknesses, so it is beneficial to use them both in order to be more certain of the given results.

A. Information about the dataset

The dataset used in order to give a comparison of the results is the Pima Indians Diabetes Database [6]. It consists of eight input variables and one target variable. The input variables are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigree and Age. The target variable is Outcome (eventualno da dodam detaljna objasnjenja kolona, nalaze se u Table1 tabeli u radu pod referencom 6). The dataset has no values that are missing, nor any null values, so the data pre-processing step was not the main priority. The total number of instances is 768 and it is important to notice that the dataset is not balanced, as 268 instances are diabetic instances, while 500 were non-diabetic instances (35% to 65% ratio).

B. Comparison pipeline

In order to give an adequate comparison of the technologies, the pipeline in Figure x is proposed. The following phases are considered:

(ovde je deo gde bih trebala da objasnim workflow)

Results – Tabelarni prikaz dobijenih skorova za feature kod LIME-a i SHAP-a

Conclusion – Tu bih uporedila dobijene rezultate za ova dva modela

References (sredicu tako da isti tip notacije bude svuda, npr ieee)

[1] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 3 (May-June 2018), 31–57. https://doi.org/10.1145/3236386.3241340

[2] Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter and Lalana Kagal. “Explaining Explanations: An Overview of Interpretability of Machine Learning.” *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (2018): 80-89.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should i trust you?’: Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining. New York, NY, USA: Association Computing Machinery, Aug. 2016, pp. 1135–1144.

[4] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods,” *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.

[5] Lundberg, Scott M. and Su-In Lee. “A Unified Approach to Interpreting Model Predictions.” *NIPS* (2017).

[6] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). [Using the ADAP learning algorithm to forecast the onset of diabetes mellitus](http://rexa.info/paper/04587c10a7c92baa01948f71f2513d5928fe8e81). *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.

[7] Ignatiev, Alexey. "Towards Trustable Explainable AI." In *IJCAI*, pp. 5154-5158. 2020.

[8] *Sebastian Palacio, Adriano Lucieri, Mohsin Munir, Sheraz Ahmed, Jörn Hees, Andreas Dengel*; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021, pp. 3766-3775

[9] Misheva, Branka Hadji, Joerg Osterrieder, Ali Hirsa, Onkar Kulkarni, and Stephen Fung Lin. "Explainable AI in credit risk management." *arXiv preprint arXiv:2103.00949* (2021).

[10] Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, 107161.

[11] Knapič, S., Malhi, A., Saluja, R., & Främling, K. (2021). Explainable artificial intelligence for human decision support system in the medical domain. *Machine Learning and Knowledge Extraction*, *3*(3), 740-770.

[12] J. Duell, X. Fan, B. Burnett, G. Aarts and S. -M. Zhou, "A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records," 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, 2021, pp. 1-4, doi: 10.1109/BHI50953.2021.9508618.

[13] Dave, D., Naik, H., Singhal, S., & Patel, P. (2020). Explainable ai meets healthcare: A study on heart disease dataset. *arXiv preprint arXiv:2011.03195*.

Delovi rada za extended abstract:

Abstract

Motivation

Methodology

Solution/Discussion

References