Методы объяснимого искусственного интеллекта на основе анализа пространства признаков

Н. В. Попов[1], Н. В. Шевская[2]
Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)
[1] nvpopov@stud.eltech.ru, [2] nvrazmochaeva@etu.ru

***Аннотация*. В 21 веке человечество активно внедряет машинное обучение и искусственный интеллект во все сферы жизни. Но большинство современных алгоритмов выводит конечный итог вычислений, не раскрывая подробностей получения результата, что является причиной некоторого скептицизма к нему. Чтобы исправить данную ситуацию, появляется необходимость в использовании методов объяснимого машинного обучения, которые повышают прозрачность использования и уровень доверия людей. В работе проводится обзор существующих решений этой задачи, а также составляется вывод об эффективности того или иного алгоритма. По итогам статьи предлагаются пути дальнейшего развития работы.**

Ключевые слова: объяснимое машинное обучение; машинное обучение

I. Введение

На протяжении уже нескольких десятилетий машинное обучение внедряется во все сферы деятельности человечества. Эти технологии используются во всевозможных областях, таких как, наука, образование, медицина, безопасность.

По данным IDC Worldwide Artificial Intelligence Spending Guide, российский рынок искусственного интеллекта в 2020 году достиг 291 млн долларов США [1].

Мировые доходы рынка искусственного интеллекта, включая программное обеспечение, аппаратное обеспечение и услуги, по прогнозам, вырастут на 16,4 % в годовом исчислении в 2021 году до 327,5 млрд долларов, согласно последнему выпуску International Data Corporation (IDC). Ожидается, что к 2024 году рынок преодолеет отметку в 500 млрд долларов с пятилетним совокупным годовым темпом роста в 17,5 %, а общая выручка достигнет 554,3 млрд долларов [2].

Часть этих средств используется для рекомендации музыки, фильмов, сериалов, продвижения товаров в интернете, а также для огромного количества других всевозможных применений этих алгоритмов, в которых машинное обучение, несмотря на несовершенный результат, приносит колоссальные прибыли. В то же время существуют отдельные сферы жизни, где любая даже самая маловероятная ошибка может привести к серьезным последствиям, начинающимися финансовыми потерями и заканчивающимися угрозой для жизни людей. При обсуждении таких вопросов остро встает проблема о применении алгоритмов машинного обучения и контроле над ними.

Большинство подобных современных технологий имеет высокий уровень эффективности, но за это требуется платить тем, что обычные люди, которые являются конечными пользователями и даже сами разработчики не до конца понимают каким именно образом компьютер вычисляет тот или иной результат. Алгоритмы машинного обучения имеют множество уровней абстракции, что делает понимание процесса работы невозможным по крайней мере для человеческого мозга.

Объяснимые методы машинного обучения позволяют решить ряд проблем, которые неизбежно возникают по мере их использования:

- социальные и этические нормы;
- доверие пользователей к модели;
- отладка моделей;
- состязательные примеры и атаки;
- исследование модели и получение из нее новых знаний.

II. Обзор литературы

В ходе исследования различных источников литературы было выявлено множество существующих подходов к так называемому Explainable Artificial Intelligence (объяснимый искусственный интеллект) или XAI. Информация по некоторым из них представлена в таблице [3–4].

ТАБЛИЦА I Техники объяснимого искусственного интеллекта

Техники	**Время применения***	**Масштаб ****	**Модель - агностик**	**Тип результата**
Partial Dependence Plot	П	Г/Л	+	Сумма признаков
Individual Condition Expectation	П	Г/Л	+	Сумма признаков
Accumulated Local Effects Plot	П	Г	+	Сумма признаков
Feature Interaction	П	Г	+	Сумма признаков
Feature Importance	П	Г/Л	+	Сумма

Техники	Время применения*	Масштаб**	Модель - агностик	Тип результата
				признаков
Local Surrogate Model (Lime)	П	Л	+	Суррогатная модель
Shapley Values	П	Л	+	Сумма признаков
BreakDown	П	Л	+	Сумма признаков
Anchors	П	Л	+	Сумма признаков
Counterfactual Explanations	П	Л	+	Точки данных
Prototypes and Criticisms	П	Г	+	Точки данных
Influence Functions	П	Г/Л	+	Точки данных
Decision trees	В	Г	-	Сумма признаков
Rule extraction	П	Г/Л	+	Точки данных
Model distillation	П	Г	+	Суррогатная модель
Sensitivity analysis	П	Г/Л	+	Сумма признаков
Layer-wise Relevance Propagation	П	Г/Л	+	Точки данных

*В столбце «Время применения» В – внутренняя интерпретируемость, П – интерпретируемость после тренировки.
**В столбце «Масштаб» Г – глобальный, Л – локальный

III. Алгоритмы

В качестве рассматриваемых в этой работе алгоритмов были выбраны Деревья решений, LIME и библиотека SHAP на основе векторов Шепли.

A. *Деревья решений*

Интерпретируемость реализуется в качестве определения у всех признаков важности Джини. Она вычисляется как общее уменьшение критерия, вызванного этим признаком.

B. *Локальная суррогатная модель-агностик техника интерпретируемости*

LIME объясняет конкретный прогноз выбранного классификатора. Модель выполняет проверку того, что происходит с предсказаниями, когда происходят изменения данных в модель машинного обучения. LIME генерирует новый набор данных, состоящий из варьирующихся выборок и соответствующих прогнозов модели черного ящика. Затем на этом новом наборе данных интерпретируемую модель обучается и взвешивается по близости выбранных экземпляров к интересующему экземпляру. Изученная модель должна быть хорошим приближением прогнозов модели машинного обучения локально, но она не обязательно должна быть хорошим глобальным приближением [5].

Преимуществом LIME является то, что этот подход может быть применен к любой модели и дает объяснение того, почему было сделано конкретное предсказание. LIME хорошо показывает себя при работе как с табличными данными, так и с картинками и текстом.

C. *Векторы Шепли*

Понятие векторов Шепли идет из теории игр и подразумевает, что наилучший результат может быть достигнут путем объединения игроков в коалиции. Таким же образом может быть вычислено влияние, которое оказывает признак на конечный результат прогнозирования. При расчёте вектора Шепли необходимо формировать коалиции из ограниченного набора признаков.

На практике в большинстве случаях точное вычисление значений Шепли невозможно в виду большого числа необходимых вычислений (2^k возможных коалиций признаков) и вынужденного заполнения отсутствующих признаков. Для аппроксимации решения число итераций ограничивается, что приводит к увеличению дисперсии значений Шепли.

IV. Сравнение алгоритмов

Сравнение вышеупомянутых алгоритмов было проведено на классификаторе деревьев решений на данных анализа сердечных приступов [6]. Точность прогнозирования модели составила 0.8525.

A. *Деревья решений*

С помощью библиотеки scikit-learn [7] на основе вычисленных важностей признаков была построена гистограмма, представленная на рис. 1.

Также было построено графическое отображение дерева решений, представленное на рис. 2. Как можно видеть интерпретируемость на дереве всего с четырьмя уровнями глубины уже является нечитаемой.

B. *Локальная суррогатная модель-агностик техника интерпретируемости*

Данный метод интерпретируемости был реализован библиотекой LIME [8]. На рис. 3 представлено графическое объяснение одного из результатов предсказания.

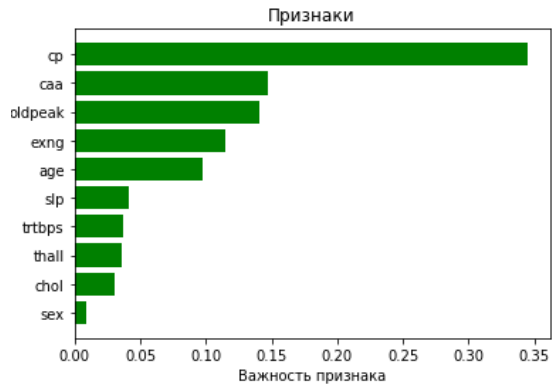

Рис. 1. Гистограмма важностей признаков

Рис. 2. Дерево решений

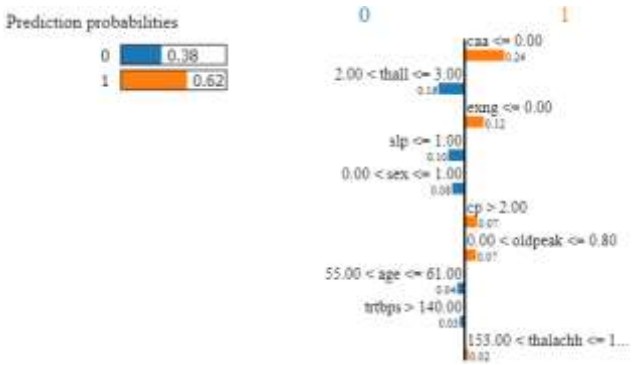

Рис. 3. Интерпретация единичного прогноза библиотекой LIME

C. Векторы Шепли

С помощью библиотеки SHAP [9] была получена интерпретируемость того же результата, что и в подразделе B, где использовался LIME. Объяснение представлено на рис. 4–5.

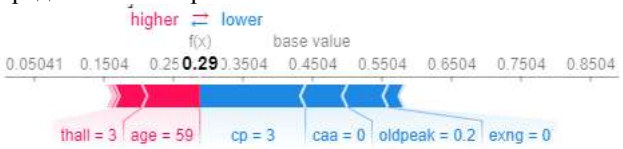

Рис. 4. Интерпретация единичного прогноза библиотекой SHAP

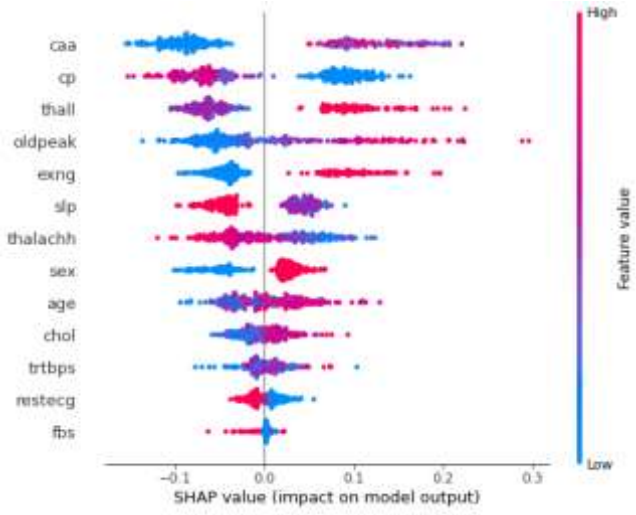

Рис. 5. Интерпретация влияния признаков на результат библиотекой SHAP

Можно сделать вывод о том, что интерпретация, имеющаяся в деревьях решений недостаточна удобна в силу не наглядности и чрезмерной загруженности. Локальная суррогатная интерпретация хорошо подходит для объяснения причин, по которым было сделано какое-то конкретное предсказание. В свою очередь SHAP предназначена для понимания сразу всей совокупности признаков. Однако подсчитать точное значение векторов Шепли не представляется возможным из-за большого числа вычислений.

V. Заключение

В статье были рассмотрены существующие подходы к интерпретации результатов машинного обучения. Более детально были разобраны такие подходы, как вычисление важности признаков при классификации на случайных деревьях, локальная суррогатная модель-агностик техника интерпретируемости и векторы Шепли. Был сделан вывод об эффективности последних двух по сравнению с первым. Для дальнейшей работы планируется расширить диапазон рассматриваемых подходов в определенных сферах, например, социальной или медицине.

Список литературы

[1] IDC: итоги развития рынка искусственного интеллекта в России. Available at: https://www.idc.com/getdoc.jsp?containerId=prEUR247642121 (accessed 25 July 2021).

[2] IDC Forecasts Improved Growth for Global AI Market in 2021. Available at: https://www.idc.com/getdoc.jsp?containerId=prUS47482321 (accessed 25 July 2021).

[3] Adadi A., Berrada M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2870052.

[4] Carvalho D.V., Pereira E.M., Cardoso J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 2019, 8, 832. Available at: https://doi.org/10.3390/electronics8080832 (accessed 25 July 2021).

[5] Molnar C. Interpretable machine learning. A Guide for Making Black Box Models Explainable, 2019. Available at: https://christophm.github.io/interpretable-ml-book/ (accessed 25 July 2021).

[6] Heart Attack Analysis & Prediction Dataset Available at: https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset (accessed 25 July 2021).

[7] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. Available at: https://scikit-learn.org/stable/index.html (accessed 25 July 2021).

[8] Local Interpretable Model-Agnostic Explanations library documentation. Available at: https://lime-ml.readthedocs.io/en/latest/index.html (accessed 25 July 2021).

[9] SHAP library documentation. Available at: https://shap.readthedocs.io/en/latest/index.html (accessed 25 July 2021)