УДК 004.891.2 DOI: 10.26102/2310-6018/2021.33.2.024

Объяснимый искусственный интеллект и методы интерпретации результатов

Н.В. Шевская

ФГАОУ ВО «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)», Санкт-Петербург, Российская Федерация

Резюме: Системы искусственного интеллекта используются во многих сферах обеспечения жизни человека, будь то, например, финансы или медицина. С каждым годом интеллектуальные системы обрабатывают все больше и больше данных и принимают все больше и больше решений. Эти решения оказывают большее влияние на судьбы людей. Краеугольным камнем становится недоверие к полностью нечеловеческим, автономным системам искусственного интеллекта. Ключ недоверия кроется в непонимании того, почему интеллектуальные системы принимают то или иное решение, исходя из каких убеждений такие системы действуют (и есть ли у них свои собственные убеждения или только те, что им передали разработчики). Для решения проблемы «недоверия» к таким системам стали применять методы объяснимого искусственного интеллекта. В этой статье представлен краткий обзор методов, получивших наибольшую популярность в академической среде (методы PDP, SHAP, LIME, DeepLIFT, permutation importance, ICE plots). На примере практических упражнений продемонстрирована легкость применения методов PDP и SHAP, а также удобство «чтения» графических результатов работы этих методов, которые объясняют построенные дерево решений и случайный лес на примере небольшого набора данных о продажах.

Ключевые слова: искусственный интеллект, объяснимый искусственный интеллект, интерпретируемый искусственный интеллект, объяснимость, интерпретируемость, XAI, PDP, SHAP.

Для цитирования: Шевская Н. В. Объяснимый искусственных интеллект и методы интерпретации результатов. *Моделирование, оптимизация и информационные технологии.* 2021;9(2). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=1005 DOI: 10.26102/2310-6018/2021.33.2.024

Explainable artificial intelligence and methods for interpreting results

N.V. Shevskaya

FSAEI OF HE Saint Petersburg State Electrotechnical University "LETI" named after V.I. Ulyanov (Lenin), Saint-Petersburg, Russian Federation

Abstract: Artificial intelligence systems are used in many areas of human life support, for example, finance or medicine. Every year intelligent systems process more and more data and make more and more decisions. All these decisions have an increasing impact on the fate of people. The cornerstone is a distrust of completely non-human, autonomous artificial intelligence systems. The key to not believing lies in the misunderstanding of why intelligent systems make this or that decision, based on what beliefs such systems operate (and whether they have their views or only those given them by the developers). The methods of explainable artificial intelligence have been used to solve the problem of distrust in such systems. This article provides a brief overview of the most popular technics in the academic environment, such as PDP, SHAP, LIME, DeepLIFT, permutation importance, ICE plots. Practical

exercises demonstrate the ease of application of PDP and SHAP methods, as well as the convenience of "reading" the graphical results of these methods, which explain the constructed decision tree model and random forest model on the example of a small set of sales data

***Keywords*:** artificial intelligence, explainable artificial intelligence, interpretable artificial intelligence, explainability, interpretability, XAI, PDP, SHAP

***For citation*:** Shevskaya N.V. Explainable artificial intelligence and methods for interpreting results. *Modeling, Optimization and Information Technology*. 2021;9(2). Available from: https://moitvivt.ru/ru/journal/pdf?id=1005 DOI: 10.26102/2310-6018/2021.33.2.024 (In Russ).

Введение

Актуальность данной работы основана на действительно новой и важной области искусственного интеллекта – объяснимых методах искусственного интеллекта. Необходимо понять, как и почему столь популярная система искусственного интеллекта может принимать решения и иногда даже изменять человеческие жизни. В важнейших сферах человеческой жизни, таких как медицина и финансы, система искусственного интеллекта используется для принятия решений. Чаще всего пациент или клиент доверяют врачу или банковскому работнику, но не машине.

В основе проблемы лежит широко используемый принцип модели черного ящика, который невозможно разложить на простые и понятные человеку части или подзадачи. И главная задача объяснимого искусственного интеллекта – разработать подходы, которые могут показать человеку, почему эта система работает именно так. Но эти подходы не должны снижать точность модели и не должны увеличивать время работы модели. В этом случае необходимо делать упор на те методы, которые не изменяют внутренней структуры моделей черного ящика.

В настоящей работе объектом исследования является искусственный интеллект в рамках объяснения того, как он работает. Предметом исследования являются несколько широко распространенных методов интерпретации искусственного интеллекта.

Определение рамок настоящего исследования.

Прежде всего, когда речь идет об искусственном интеллекте, необходимо удерживать в голове всю карту интеллектуальных методов, которая в общем случае содержит три большие области, как показано на Рисунке 1.

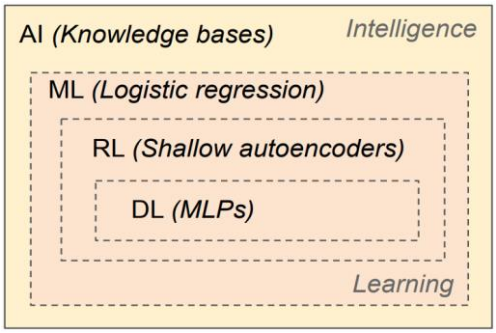

Рисунок 1 – Границы понимания ИИ
Figure 1 – Edges of understanding AI area

В этом случае можно определить два типа граней: интеллект и обучение, где первый тип включает второй тип (интеллект на основе обучения). Рассмотрим следующую иерархию в разрезе обучения. Самая большая часть обучения – это машинное обучение (*machine learning, ML*). В области машинного обучения можно выделить самые популярные примеры – логистическую регрессию и ее аналоги. Далее,

машинное обучение включает еще один, следующий тип, который представляет собой репрезентативное обучение (*representation learning, RL*), в качестве наглядного примера которого можно привести реализацию мелких автокодировщиков. Другими словами, RL – это обучение в соответствии с особенностями обучающей выборки.

С другой стороны, изучение характеристик (свойств или атрибутов, с точки зрения хранения данных – столбцов) (*feature learning*) данных включает в себя набор специальных подходов, которые позволяют автоматически обнаруживать представления, необходимые для обнаружения или классификации характеристик из необработанных данных. Изучение атрибутов помогает исследователям свести к минимуму разработку новых атрибутов вручную и позволяет атоматически как изучать характеристики данных, так и использовать их для решения проблем. Изучение атрибутов постоянно конкурирует с инструментами извлечения характеристик (*feature extraction*) и разработки характеристик (*feature engineering*). Скажем несколько слов об основном различии между этими тремя немаловажными подходами, работающими под одной целью – исследование и преобразование пространства признаков (пространства, в котором описаны наблюдения, строки таблицы).

Извлечение признаков – это простое преобразование необработанных данных в последовательность векторов признаков, с которыми можно работать. При изучении характеристик неизвестно, какую именно характеристику можно извлечь из своих данных. Фактически, вероятно, будут применяться методы машинного обучения только для того, чтобы узнать, какие полезные характеристики можно извлечь из набора данных.

Разработка и извлечение характеристик близки друг к другу из-за области применения. Они касаются преобразования данных обучения и добавления к ним дополнительных характеристик. Все эти методы используются для повышения точности алгоритмов машинного обучения.

И самая глубокая и последняя часть обучения – это глубокое обучение (*deep learning, DL*). Примером DL является многослойный персептрон (*multilayer perceptron, MLP*). MPL – это класс искусственной нейронной сети с прямой связью. Термин MLP используется неоднозначно, иногда свободно по отношению к любой искусственной нейронной сети прямого распространения, иногда строго для обозначения сетей, состоящих из нескольких уровней персептронов (с пороговой активацией). Многослойные персептроны иногда в просторечии называют простейшими нейронными сетями, особенно когда они имеют единственный скрытый слой.

Таким образом, когда речь идет об объяснимом искусственном интеллекте, имеются ввиду все существующие методы машинного обучения, включая репрезентативное обучение и глубокое обучение. Общая проблема, обсуждаемая в этой статье, заключается не в том, как построить очень точные, сложные и более быстрые модели, а в том, как объяснить, почему такие большие и сложные модели дают конкретный результат. Основной вид результата – прогноз (в финансах, медицине, продажах). Другой очень популярный результат – классификация (число ближайшего класса или вероятность принадлежности к классу). Один из слабо изученных результатов моделей машинного обучения – функциональные результаты (не точечные, как в предыдущих двух пунктах; функциональные результаты моделей особенно свойственны таким областям, как медицина). В этой работе ограничимся простейшими результатами моделей ИИ, которые описываются числовыми значениями.

Вопросы соотношения объяснимости и интерпретируемости.

Есть разные стороны проблем доверия к моделям ИИ. И эти стороны порождают разные подходы к обозначению на первый взгляд одних и тех же проблем. Но при ближайшем рассмотрении эти проблемы действительно будут иметь различный смысл.

Здесь не будет обсуждаться доверенный ИИ (*trusted AI*) и пути его реализации, цель этого этапа исследования – определить разницу между наиболее популярными терминами, такими как объяснимость и интерпретируемость.

Интерпретируемый ИИ способен описывать внутреннюю структуру системы понятным людям способом. Про интерпретируемый ИИ можно сказать так: «понятно как, но непонятно почему». Для такого вида ИИ характерно использование четких правил и метрик объяснения, но природа этих правил определяется значением термина «объясняемый».

Объяснимость моделей ИИ – это возможность кратко описать, почему модель работает (не вдаваясь в подробности). Про объясняемый ИИ можно сказать так: «непонятно как, но понятно почему», т. е. причина принятия того или иного решена понятна и может быть даже вполне обоснована, однако четкий алгоритм, описывающих переход от причины к конкретному действию остается неявным.

Можно рассматривать одновременно и объяснимость, и интерпретируемость, в некоторых источниках это называют верностью (*fidelity*), но определение этого термина выходит за рамки настоящей статьи.

Помимо верности, в источниках литературы можно встретить такие понятие как прозрачность (*transparency*) и ответственность (*responsibility*) [1]. Существует еще много терминов, всплывающих при работе с доверием к системам на основе ИИ. В данной статье придерживаемся идеологии, предложенной в [1] (на Рисунке 2), согласно которой есть две стороны ИИ: свойство интерпретируемости, которое можно представить количественными метриками, и свойство объяснимости, обладание которым определяется выполнением критериев (необходимых и достаточных) в интерпретируемости. Правила формирования необходимых и достаточных критериев интерпретируемости для достижения уровня объяснимого ИИ будут обсуждаться в последующих исследованиях.

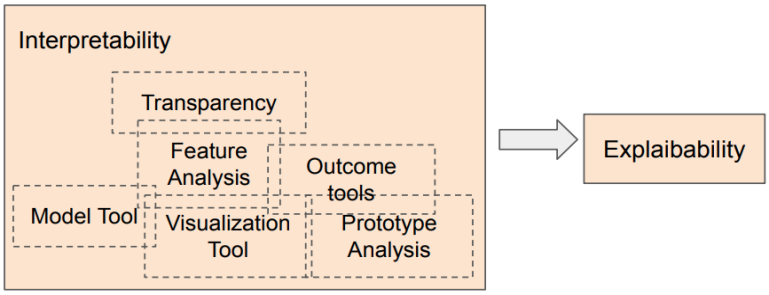

Рисунок 2 – Связь терминов «интерпретируемость» и «объяснимость», и «прозрачность»
Figure 2 – Connection between Interpretability and Explainabiliy and Transparency terms

В основе понимания, рассматриваемого в настоящей работе, лежит работа с моделями черного ящика, вне зависимости от того, какие типы алгоритмов обработки и обучения, какие типы моделей находятся внутри них. Точка контроля объяснимости для таких моделей – это входные значения и выходные данные. Управлять выходом в таких моделях невозможно, потому что это зависит от внутренней структуры модели, которую полагаем неизвестной. Но для понимания принципов работы внутренней структуры можно управлять входом модели и фиксировать, как изменяется выход. Этот тип исследования черного ящика модели включает в себя некоторый набор методов, которые будут обсуждаться в следующей главе.

Материалы и методы

В [2] можно найти наиболее широкую классификационную таксономическую модель для методов объяснимого искусственного интеллекта (*explainable artificial intelligence, XAI*) в зависимости от разных причин и предпосылок, здесь же кратко представляем один из наиболее общих аналогов (см. Рисунок 3).

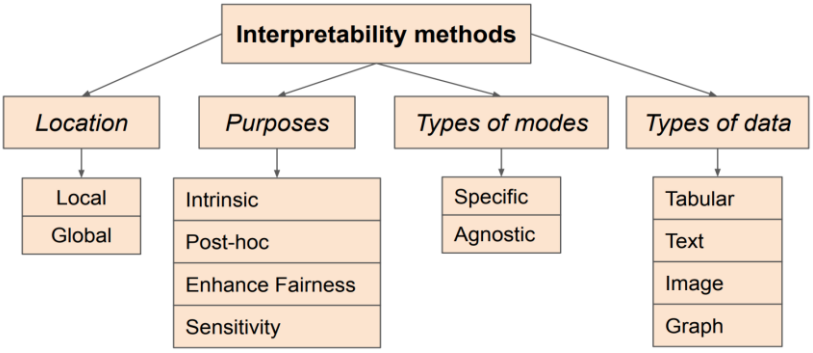

Рисунок 3 – Классификация методов интерпретируемости
Figure 3 – Classification of Interpretability methods

Различие методов интерпретации легко понять, если говорить об обрабатываемых моделями типах данных (*types of data*). Можно научить модель работать с очень популярными типами, такими как изображения и таблицы, но нужно понимать, что модели могут обрабатывать текстовые и графические данные. Нельзя игнорировать и тот факт, что каждый тип данных может быть преобразован в другой тип из представленного списка.

Когда речь идет о локализации метода (*Location*), то нужно понимать, что локальный метод – это метод, который может объяснить только одно-единственное предсказание. О глобальных методах можно сказать, что они должны объяснять общие принципы, по которым работает модель.

Классификация по типу моделей (*Types of models*) сложнее двух предыдущих, давайте это поясним. Специфическая модель (*Specific*) может быть применена к одной модели или нескольким моделям, которые имеют схожие структуры. Независимость (*Agnostic*) от модели – это свойство самой модели, когда ее объяснение может применяться к любым моделям (в таком случае полается, что модель – «черный ящик»).

И определяющая и важная часть этой классификации основана на целях интерпретации (*Purposes*). Можно определить четыре основные цели. Внутренняя цель (*Intinsic*) означает, что создается модель белого ящика или полностью интерпретируемая модель. Цель *Post-Hoc* (цель – объяснить модель постфактум) включает в себя объяснение существующих моделей черного ящика, которые очень сложны внутри (для таких моделей это единственный способ объяснения).

Исследование чувствительности моделей (*Sensitivity*) – важная часть в области интерпретируемого ИИ. Исследование чувствительности может помочь определить стабильность модели в зависимости от различных входных данных, когда эта модель будет использоваться в реальных условиях.

Проблемы справедливости (*enhance fairness*) являются частью этических проблем в XAI. Самый яркий пример – медицина, судебная система и аналогичные системы, где человек должен хранить в памяти большой набор правил и дел (историй болезни) и иметь возможность использовать их в любое время в своей работе (в условиях защиты персональных данных).

Следует отметить, что все интерпретируемые методы имеют возможность визуализации, что делает их более удобными для использования людьми, которые не являются разработчиками, а, например, экспертами в предметной области.

Прежде чем приступить к экспериментам, приведем краткое описание двух наиболее популярных методов интерпретируемого ИИ:

• Графики частичной зависимости (PDP) [3]

• SHAP-значения [4], [5].

Графики частичной зависимости (PDP) показывают незначительное влияние одной или двух функций на прогнозируемый результат модели машинного обучения. PDP может показать взаимосвязь между целевым параметром и другими выбранными параметрами, используя одномерные или двухмерные графики.

Метод SHAP расшифровывается как объяснение добавок Шепли (или векторы Шепли). В теории игр значение Шепли – это концепция решения справедливого распределения прибыли и затрат между несколькими участниками, работающими в коалиции. Значение Шепли применяется в первую очередь в ситуациях, когда вклады каждого актора неравны, но они работают в сотрудничестве друг с другом, чтобы получить суммарную отдачу. Этот метод помогает разбить прогноз так, чтобы раскрыть значение каждой характеристики.

Как правило, бывает сложно найти баланс между точностью модели и её интерпретируемостью, но значения SHAP могут обеспечить и то, и другое. SHAP – значения показывают, насколько данный конкретный признак изменил итоговый прогноз (по сравнению с тем, как был бы сделан этот прогноз при некотором базовом значении этого признака). Помимо классического метода SHAP, встречаются еще две его модификации KernelSHAP и DeepSHAP, но в рамках данной статьи они рассматриваться не будут.

Спектр методов интерпретации не ограничивается выше представленными. Среди популярных методов можно еще встретить LIME метод (*Local Interpretable Model-agnostic Explanations)* [6], позволяющий выполнять локальные объяснения для модей типа «черный ящик» [7]. Другой метод DeepLIFT [8] (*Layer-Wise Relevance Propagation*) относится к методам, модифицирующих модель нейронной сети изнутри (сравнивает активацию каждого нейрона с его «эталонной активацией» и присваивает баллы вклада в соответствии с разницей). Следующий метод, который относят к категории визуальных, графики индивидуальных условных ожиданий (*Individual Conditional Expectation (ICE) plots*) [9], которые отображают по одной строке для каждого экземпляра, которая показывает, как прогноз экземпляра изменяется при изменении функции. Если сравнивать с PDP, то второй подход для среднего эффекта функции является глобальным методом, т. к. фокусируется не на конкретных примерах, а на общем среднем.

Еще один популярный метод исследования модели, который здесь не обсуждался, – это важность признаков при перестановке (*permutation importance*) [10]. Какие характеристики модели считает важными? Эта концепция называется важностью характеристик, а важность перестановки – это отдельный метод, широко используемый для расчета важности атрибутов. Применение этого метода помогает определить, когда модель дает неожиданные результаты, и помогает убедиться, что наша модель работает именно так, как должна.

Важность перестановки работает для многих встроенных оценок в популярную библиотеку анализа данных *scikit-learn* (на языке программирования *Python*). Идея проста: произвольно переставьте или перемешайте один столбец в наборе данных проверки, оставив все остальные столбцы нетронутыми. Характеристика считается

«важной», если точность модели падает, а ее изменение вызывает увеличение ошибок. С другой стороны, характеристика считается «неважной», если изменение ее значений не влияет на точность модели. Важность перестановки велика, потому что она создает простые числовые меры, чтобы увидеть, какие функции имеют значение для модели. Это поможет легко сравнивать характеристики, и представлять полученные графики, например, нетехнической аудитории. Но данный метод не говорит, насколько важна каждая функция. Если функция имеет среднюю важность перестановки, это может означать, что она имеет большой эффект для нескольких прогнозов, но не имеет эффекта в целом, или средний эффект для всех прогнозов. Описание и варианты практического применения выше представленных и некоторых других методов, не рассмотренных в этой статье, можно найти в [2].

Эксперименты

Проведем исследование методов PDP и SHAP на примере нескольких упражнений с данными о продажах. В этом наборе данных каждая строка представляет собой наблюдение одного продукта, который описывает с помощью 10 характеристик (10 столбцов). Характеристики можно разделить на качественные (номенклатура, маркетинговая группа, свойства маркетинговой группы и т. д.) и количественные характеристики (себестоимость, выручка, маржа и т. д.). Размер набора данных сравнительно небольшой и составляет 164 строки (164 продукта).

Основная идея проведения экспериментального исследования на данном наборе – понять, какие количественные характеристики имеют наибольшее влияние на нашу величину маржи (*Margin*, целевую характеристику). Для построения моделей потребуется привести значения *Margin* в бинарный вид (0, 1). Для этого фиксируется пороговое значение 40 (такое допущение) и все значения, которые больше или равны 40, равны 1, а те, которые меньше 40, равны 0.

Первая модель для исследования – это дерево решений (будем использовать *DecisionTreeClassifier* из библиотеки *sklearn.tree* на Python 3.8). Посмотрим на результат интерпретации модели и выведем PDP [11] для дерева решений (см. Рисунок 4).

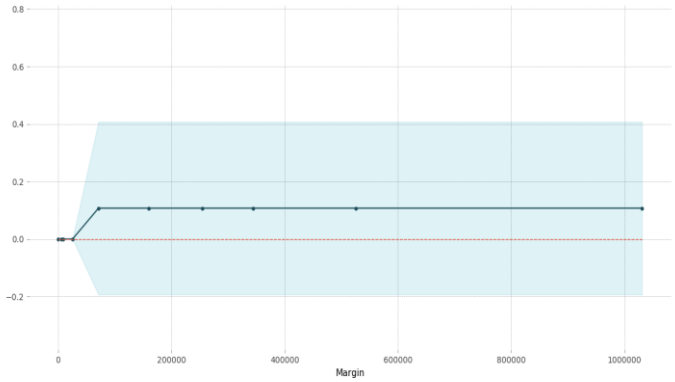

Рисунок 4 – Эксперимент со значением *Margin* для модели дерева решений
Figure 4 – Experiment with Margin value for DecisionTreeClassifier

Ось Y интерпретируется как изменение прогноза по сравнению с тем, что было бы предсказано на базовой линии или крайнем левом значении. Заштрихованная синим область указывает уровень уверенности.

Из этого конкретного графика видно, что чем больше товаров продано, тем больше конечная величина маржи (зависимость, практически, очевидная). При этом, влияние других характеристик незначительное.

Построим пример по второй модели – случайный лес (*RandomForestClassifier* из библиотеки *sklearn.tree* для *Python 3.8*) (см. Рисунок 5).

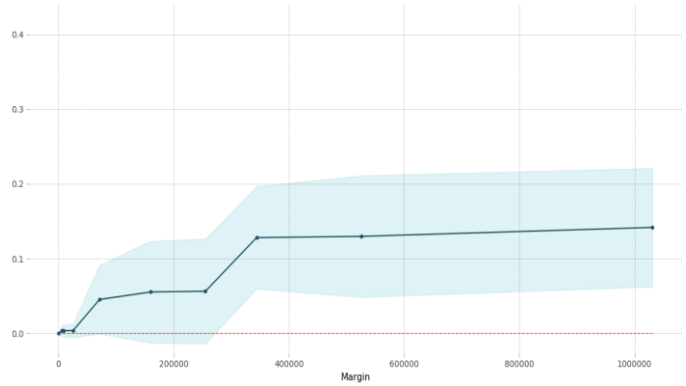

Рисунок 5 – Эксперимент со значением *Margin* для модели случайного леса
Figure 5 – Experiment with Margin value Random Forest model

Форма этой кривой кажется более правдоподобной, чем ступенчатая функция из интерпретации модели дерева решений. Хотя исследуемый набор данных достаточно мал, поэтому надо быть осторожными при интерпретации любой модели.

Также можно визуализировать частичную зависимость двух характеристик одновременно с помощью 2D-графиков (см. Рисунок 6).

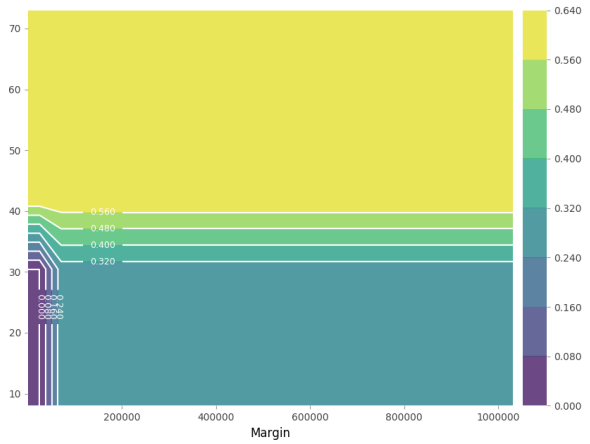

Рисунок 6 – Частичная зависимость *Margin* от другой переменной
Figure 6 – Partial Margin dependency on another variable

Давайте посмотрим на второй метод объяснения моделей ИИ. Сводные графики SHAP [12] дают нам представление о важности функции и ее факторах (см. Рисунок 7). Иллюстрация состоит из множества точек. Каждая точка имеет три характеристики:

(1) вертикальное расположение показывает, какой объект (характеристику) она изображает;

(2) цвет показывает, была ли эта характеристика высокой или низкой для этой строки набора данных;

(3) горизонтальное расположение показывает, повлияло ли это значения на более высокий или более низкий прогноз.

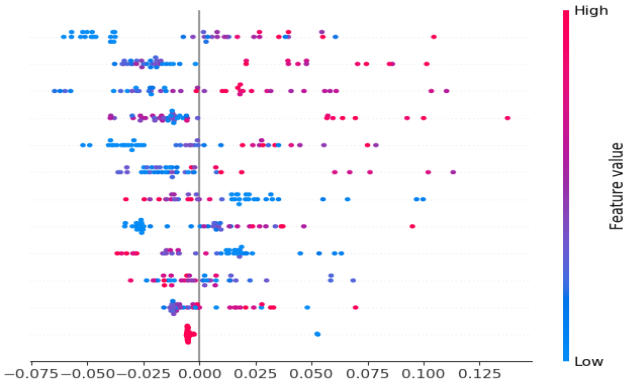

Рисунок 7 – Результат методов SHAP для анализа Margin
Figure 7 – SHAP methods result for Margin analysis

Для нашего целевого значения *Margin* из графика видно, что только первая половина (как минимум, первые 6 строк на этом графике) характеристик оказывают влияет на целевой параметр (истинные названия характеристик скрыты из-за коммерческой тайны). Чего нельзя сказать о второй половине. В сравнении с результатами PDP можем сделать вывод, что SHAP выделил большее число значащих характеристик (кол-во продаж, выявленное PDP, находится среди первых шести строк представленного результата).

Рассуждение и заключение

В данной статье кратко описаны основные идеи классификации подходов искусственного интеллекта в рамках проблемы объяснимости. Представлена иерархическая структура областей искусственного интеллекта. Была проиллюстрирована общая разница между объяснимостью, интерпретируемостью и прозрачностью. Два самых популярных метода для интерпретации ИИ (значения PDP и SHAP) были кратко проиллюстрированы на практике, а другие же популярные методы (permutation importance, LIME, DeepLIFT, и др.) обозначены для применения в дальнейшем исследовании.

Результаты проведенных экспериментов показывают, что два выбранных метода можно применить к неизвестному типу модели, построенной на данных о продажах, без особой подготовки. Способ объяснения результатов этих методов простым (в некоторых случаях очевидным) и не зависит от модели и типа данных. Конечно, эти гипотезы следует проверить более детально, но в этой статье был создан еще один прецедент, который показывает, что без определенной подготовки можно получить больше информации о том, как модели принимают решения. Но если так просто объяснять модели, почему до сих пор внедрение систем на основе искусственного интеллекта не происходит в областях, где это действительно необходимо (например, медицина или банки)? Краеугольным камнем в этой системе становится защита персональных данных, но этот вопрос уже выходит за рамки обсуждения данного исследования.

В самом широком смысле проблемы объяснимости искусственного интеллекта можно выделить больше сторон: теперь, помимо широкого внедрения системы искусственного интеллекта, необходимо реализовывать параллельно, в дополнение систему для объяснения результатов первой системы. Чем раньше начать это делать, тем больше поддержки методов искусственного интеллекта можно получить.

План будущей исследовательской работы состоит в том, чтобы расширить классификацию методов, представленных на Рисунке 3, провести технический сравнительный анализ методов и определить список необходимых требований к универсальному методу интерпретации; провести большее количество экспериментов с использованием методов разных типов, соответствующих расширенному дереву классификации.

Благодарности

Авторы благодарят Международный инновационный институт искусственного интеллекта, кибербезопасности и коммуникаций имени Александра Попова Санкт-Петербургского электротехнического университета «ЛЭТИ» за поддержку в работе и исследованиях.

ЛИТЕРАТУРА

1. Transparency and Responsibility in Artificial Intelligence. *Deloitte*. Доступно по: https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/innovatie/deloitte-nl-innovation-bringing-transparency-and-ethics-into-ai.pdf (дата обращения: 10.06.2021).
2. Linardatos P., Papastefanopoulos V., Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*. 2021:23(18). Доступно по: https://www.mdpi.com/1099-4300/23/1/18/pdf DOI: 10.3390/e23010018 (дата обращения: 10.06.2021).
3. Rosenfeld A., Richardson A. Explainability in human-agent systems. *Autonomous Agents and Multi-Agent Systems*. 2019;33(6):673-705. Доступно по: https://www.researchgate.net/publication/333084339_Explainability_in_human-agent_systems DOI: 10.1007/s10458-019-09408-y (дата обращения: 10.06.2021).
4. Lundberg S.M., Lee S.I. Consistent feature attribution for tree ensembles. *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning (WHI'2017)*. 2017. Доступно по: https://arxiv.org/abs/1706.06060 (дата обращения: 10.06.2021).
5. Lundberg S.M., Lee S.I. A Unified Approach to Interpreting Model Predictions. *Proceedings of the Neural Information Processing Systems (NIPS)*. 2017. Доступно по: https://arxiv.org/abs/1705.07874 (дата обращения: 10.06.2021).
6. Friedman J.H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*. 2001;29(5):1189-1232. Доступно по: https://statweb.stanford.edu/~jhf/ftp/trebst.pdf (дата обращения: 10.06.2021).
7. Ribeiro M.T., Singh S., Guestrin C. «Why should i trust you?» Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016;1135-1144. DOI: 10.1145/2939672.2939778.
8. Shrikumar A., Greenside P., Kundaje A. Learning important features through propagating activation differences. *Proceedings of the 34th International Conference on Machine Learning (PMLR)*. 2017;70:3145-3153.
9. Goldstein A. et al. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*. 2015;24(1):44-65. DOI: 10.1080/10618600.2014.907095.
10. Machine Learning Explainability. Permutation Importance. Доступно по: https://www.kaggle.com/dansbecker/permutation-importance (дата обращения: 10.06.2021).
11. Machine Learning Explainability Course. Partial Dependence Plots. Доступно по: https://www.kaggle.com/dansbecker/partial-plots (дата обращения: 10.06.2021).

12. Machine Learning Explainability. SHAP Values, Доступно по: https://www.kaggle.com/dansbecker/shap-values (дата обращения: 10.06.2021).

REFERENCES

1. Transparency and Responsibility in Artificial Intelligence. *Deloitte.* Available at: https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/innovatie/deloitte-nl-innovation-bringing-transparency-and-ethics-into-ai.pdf (accessed 10.06.2021).
2. Linardatos P., Papastefanopoulos V., Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy.* 2021:23(18). Available at: https://www.mdpi.com/1099-4300/23/1/18/pdf DOI: 10.3390/e23010018 (accessed: 10.06.2021).
3. Rosenfeld A., Richardson A. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*. 2019;33(6):673-705. Available at: https://www.researchgate.net/publication/333084339_Explainability_in_human-agent_systems DOI: 10.1007/s10458-019-09408-y (accessed 10.06.2021).
4. Lundberg S.M., Lee S.I. Consistent feature attribution for tree ensembles. *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning (WHI'2017)*. 2017. Available at: https://arxiv.org/abs/1706.06060 (accessed 10.06.2021).
5. Lundberg S.M., Lee S.I. A Unified Approach to Interpreting Model Predictions. *Proceedings of the Neural Information Processing Systems (NIPS).* 2017. Available at: https://arxiv.org/abs/1705.07874 (accessed 10.06.2021).
6. Friedman J.H. Greedy function approximation: a gradient boosting machine. *Annals of statistics.* 2001;29(5):1189-1232. Available at: https://statweb.stanford.edu/~jhf/ftp/trebst.pdf (accessed 10.06.2021).
7. Ribeiro M.T., Singh S., Guestrin C. «Why should I trust you?» Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016;1135-1144. DOI: 10.1145/2939672.2939778.
8. Shrikumar A., Greenside P., Kundaje A. Learning important features through propagating activation differences. *Proceedings of the 34th International Conference on Machine Learning (PMLR).* 2017;70:3145-3153.
9. Goldstein A. et al. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*. 2015;24(1):44-65. DOI: 10.1080/10618600.2014.907095.
10. Machine Learning Explainability. Permutation Importance. Available at: https://www.kaggle.com/dansbecker/permutation-importance (accessed 10.06.2021).
11. Machine Learning Explainability Course. Partial Dependence Plots. Available at: https://www.kaggle.com/dansbecker/partial-plots (accessed 10.06.2021).
12. Machine Learning Explainability. SHAP Values. Available at: https://www.kaggle.com/dansbecker/shap-values (accessed 10.06.2021).

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Шевская Наталья Владимировна, аспирант кафедры Математического Обеспечения и применения ЭВМ (МО ЭВМ), Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им.

Natalya V. Shevskaya, Postgraduate Student of the Software Engineering and Computer Applications Department of Saint Petersburg Electrotechnical University "LETI", Saint-Petersburg, Russia

В.И. Ульянова (Ленина), Санкт-Петербург, Российская Федерация
email: nvrazmochaeva@etu.ru
ORCID*:* 0000-0001-8936-8167