Интерпретация результатов машинного обучения для задачи регрессии

Н.К. Перминов, email: np.magistr@yandex.ru

Вятский государственный университет

Аннотация. *Проведен анализ современных методов интерпретации результатов машинного обучения для задачи регрессии. Предложена формализация понятия интерпретации, на основе которой описаны возможные задачи оценки результатов задачи регрессии. Продемонстрированы возможности библиотек для интерпретации ELI5, Lime и SHAP. Проанализированы результаты работы данных библиотек на двух датасетах – boston и diabetes, которые посвящены регрессионным задачам предсказания цен на жилье и оценке прогрессирования диабета соответственно. Даны рекомендации по решению задачи интерпретации результатов регрессионного анализа.*

Ключевые слова: *машинное обучение, регрессия, интерпретация, интерпретируемость, Lime, SHAP, ELI5.*

Введение

В последние годы происходит стремительное развитие в области машинного обучения, которое способствует новым открытиям во многих областях [2], а также используется во множестве сфер деятельности человека. Модели машинного обучения применяются для классификации, регрессии и кластеризации объектов. В данной статье рассматривается задача регрессии.

Вместе с развитием искусственного интеллекта растет запрос на объяснение результатов его деятельности. В связи с этим появляются различные работы по интерпретации результатов той или иной модели для решения задач медицины [6], химии [5] и т.д.

Крайне важным фактором в решении задач для медицины с помощью машинного обучения является объяснение результатов [1, с. 9]. Обосновывается это тем, что в данной сфере ошибка может повлиять на человеческую жизнь. Данное опасение является основной причиной недостаточной степени внедрения искусственного интеллекта при решении задач, связанных со здоровьем людей. Именно поэтому значительные усилия исследователей направлены на решение проблемы интерпретации результатов моделей машинного обучения.

В рамках бизнес-логики заказчик зачастую не доверяет полученному решению, так как ему непонятно, как именно получился тот или иной результат. Данную проблему также решает объяснение результатов машинного обучения.

Интерпретация модели – сложное для объяснения явление, которое имеет разные вариации определений и не имеет формально-технического представления [10]. В данной статье предложено формальное определение понятия интерпретации.

Цель настоящей статьи – выполнить обзор и сравнительный анализ наиболее популярных и цитируемых библиотек интерпретации результатов машинного обучения. По результатам исследования были выделены следующие библиотеки: SHAP [8], ELI5 [4] и LIME [7].

Данные библиотеки являются популярными, например, на GitHub SHAP имеет 14900 звезд, Lime – 9400, ELI5 – 2500. По актуальности лидирует SHAP, т.к. последнее изменение датировано 05.12.2021, у Lime 30.06.2021, а у ELI5 от 22.01.2020 года.

1. Понятие интерпретируемости результатов машинного обучения

Интерпретируемость – это свойство модели машинного обучения, которое измеряет, насколько легко отслеживать и объяснять ее процессы и/или результаты [2, с. 146].

Исходя из основополагающей статьи о интерпретируемости моделей машинного обучения [10], данное понятие имеет следующее дробление:

1. *Прозрачность* – касается описания модели, которое необходимо понимать, данные характеристики могут быть известны до момента обучения. В данный пункт входят следующие критерии: воспроизводимость, декомпозиция и алгоритмическая прозрачность. Первый критерий отвечает за вопрос: сможет ли человек повторить шаги алгоритма и получить тот же результат, что и используемая модель. Согласно второму критерию можно ответить на вопрос о том, можно ли интерпретировать промежуточные результаты модели или какие-то ее субкомпоненты. Последний дает ответ на вопрос о гарантии, которую предоставляет нам используемый алгоритм, например, результат модели всегда сходится к определенному типажу решений.

2. *Апостериорная интерпретируемость* – определяет то, что мы извлекаем из модели после проведения обучения. Данный пункт включает в себя следующие критерии: текстовое объяснение, визуализация и локальные объяснения и объяснения на примере. Первым делом задается вопрос о возможности объяснения результатов обучения модели на естественном языке. Визуализация и локальные объяснения дают ответ о том, имеется ли возможность обнаружения

особо важных признаков или критериев, на которые ссылалась модель при определении целевой переменной. Последний критерий объясняет может ли модель сопоставить новые данные с похожими из учебного множества и вывести приближенное значение целевой переменной, таким образом проводится аналогия.

2. Формализация задачи интерпретации результатов машинного обучения

Для формализации задачи интерпретации необходимо определить основные понятия машинного обучения.

Пусть входные данные для обучения и значения целевой переменной (формула 1) выражены следующим образом:

$$X_t = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \tag{1}$$

где n – количество примеров, m – количество признаков.

Таким образом, можно вывести формулу для целевой функции F, которая используется для предсказания целевой переменной Y на обучающих входных данных (формула 1):

$$F : X_t \rightarrow Y. \tag{2}$$

Следующим шагом будет вывод формулы интерпретации целевой функции:

$$I = inter\ (X, F)\ , \tag{3}$$

где *inter* – способ интерпретации, причем входной набор данных может быть следующим:

$$X \in \{ X_{test}\ ,\ X_{train}\ ,\ X_{new} \}\ , \tag{4}$$

то есть входной набор данных в настроенный интерпретатор может быть представлен: тестовой выборкой обозначение, обучающей выборкой обозначение или новыми сгенерированными данными обозначение.

Интерпретатор I может быть настроен несколькими способами, рассмотрим основные.

1. Задача – объяснение какой вклад был вложен в целевую переменную по каждому из признаков и для каждого примера, реализовано в библиотеке SHAP, Lime и ELI5. Значения, лежащие в I (формула 3) будут иметь следующий вид:

$$y_j = i_{j1} + i_{j2} + \ldots + i_{jm-1} + i_{jm} \ , \quad (5)$$

где i – показатель вклада по интерпретатору конкретного значения из датасета на значение целевой переменной, j – номер примера из входных данных.

2. Задача – демонстрация влияния каждого признака для получения целевой переменной на весь датасет, реализовано в библиотеке ELI5. Тогда значения, лежащие в I (формула 3) будут иметь следующий вид:

$$I(X, F) = [i_1, i_2, \ldots, i_{m-1}, i_m] \ , \quad (6)$$

где i – отображение влияния признака на результат.

3. Демонстрация работы библиотек

Для первой демонстрации работы был выбран набор данных из библиотеки sklearn – boston [9], в рамках него необходимо предсказать стоимость помещения по входным данным.

Признаки, которые хранятся в данном датасете имеют следующие определения:

1. CRIM – уровень преступности на душу населения по городам.
2. ZN – доля земли под жилую застройку, зонированная под участки более 25 000 кв. Футов.
3. INDUS – доля акров, не относящихся к розничной торговле, на город.
4. CHAS – фиктивная переменная реки Чарльз (1, если участок ограничивает реку; 0 в противном случае).
5. NOX – концентрация оксидов азота (частей на 10 миллионов).
6. RM – среднее количество комнат в доме.
7. AGE – доля занимаемых владельцами единиц, построенных до 1940 г.
8. DIS – взвешенные расстояния до пяти бостонских центров занятости
9. RAD – индекс доступности радиальных автомобильных дорог
10. TAX – полная ставка налога на имущество за 10 000 долларов США.
11. PTRATIO – соотношение учеников и учителей по городам
12. B – 1000 (Bk - 0,63) ^ 2, где Bk - доля черных по городам.
13. LSTAT – % более низкого статуса населения
14. MEDV – средняя стоимость домов, занимаемых владельцами, в 1000 долларов США

В рамках первой задачи – демонстрации вклада каждого показателя в целевую переменную, разобраны библиотеки SHAP, ELI5 и Lime.

Продемонстрированы возможности визуализации и даны объяснения к результатам их интерпретации.

После обучения модели можно выбрать случайным образом пример для интерпретации результатов, а можно взять конкретный экземпляр, целевая переменная которого нуждается в объяснении. Для сравнения библиотек наугад выбран индекс примера = 18.

Таблица 1

Соотношение имени признака из первого датасета и его значения для наугад выбранного примера

Name feature	Value
CRIM	0.80271
ZN	0.00000
INDUS	8.14000
CHAS	0.00000
NOX	0.53800
RM	5.45600
AGE	36.6000
DIS	3.79650
RAD	4.00000
TAX	307.000
PTRATIO	21.0000
B	288.990
LSTAT	11.6900

Для данных в табл. 1 логистическая регрессия предсказала значение = 16.18. Теперь посмотрим интерпретацию результата по библиотеке SHAP на рис. 1.

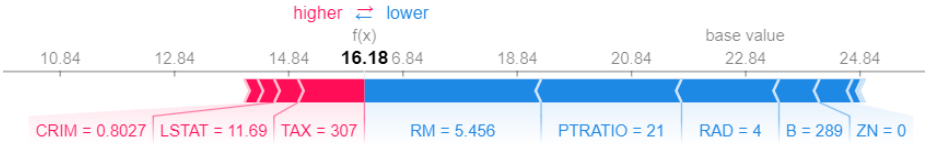

Рис. 1. Результат интерпретации результатов машинного обучения для предсказания целевой переменной у 18 примера с помощью библиотеки SHAP.

Как видно по рис. 1, сильно снизил стоимость дома признак *RM*, *PTRATIO* и *RAD*, повысило стоимость помещения значение признака *TAX*. Красные признаки влияют положительно, а синие – отрицательно.

Следующая библиотека – ELI5. Но рассмотрена первая и вторая задача. Для первой задачи выбран тот же пример – 18. Результат по двум задачам объединен на рис. 2.

y top features

Weight[?]	Feature
+36.459	<BIAS>
+3.810	RM
+2.687	CHAS
+0.306	RAD
+0.046	ZN
+0.021	INDUS
+0.009	B
+0.001	AGE
-0.012	TAX
-0.108	CRIM
-0.525	LSTAT
-0.953	PTRATIO
-1.476	DIS
-17.767	NOX

y (score **16.178**) top features

Contribution[?]	Feature	Value
+36.459	<BIAS>	1.000
+20.787	RM	5.456
+2.691	B	288.990
+1.224	RAD	4.000
+0.167	INDUS	8.140
+0.025	AGE	36.600
-0.087	CRIM	0.803
-3.787	TAX	307.000
-5.602	DIS	3.796
-6.134	LSTAT	11.690
-9.558	NOX	0.538
-20.008	PTRATIO	21.000

Рис. 2. Интерпретация результатов с помощью библиотеки ELI5: *левая таблица* – для второй задачи, *правая* – для первой задачи

По рис. 2 видно, что для первой задачи не хватает двух значений – интерпретации для признака *CHAS* и *ZN*, это обусловлено тем, что у данного признака значение = 0.

Для второй же задачи видно, что на целевую переменную наибольшее влияние в положительную сторону повлиял признак *RM*, а в отрицательную *PTRATIO*.

Библиотека Lime показала следующие результаты, отображенные на рис. 3.

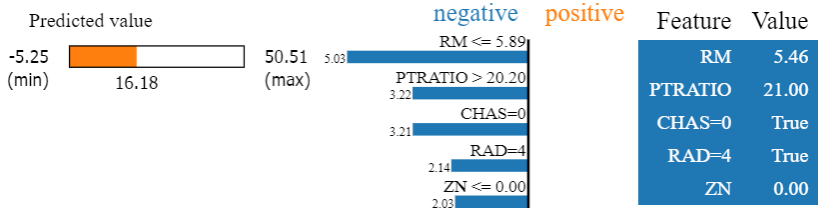

Рис. 3. Интерпретация результатов с помощью библиотеки Lime

На рис. 3 первый барплот отвечает за значение целевой переменной для выбранного примера. График посередине отвечает за влияние признаков на целевую переменную. Правая таблица демонстрирует исходные значения примера.

По рис. 3 можно сказать, что данная библиотека имеет все необходимые данные для визуализации результатов модели. Она включает все информативно-описательные характеристики.

Следующим примером данных для демонстрации работы был выбран датасет из библиотеки sklearn – diabetes [9], целевой переменной которого является количественный показатель прогрессирования заболевания через год после исходного уровня.

Признаки, которые хранятся в данном датасете имеют следующие определения:

1. age – возраст в годах.
2. sex – пол.
3. bmi – индекс массы тела.
4. bp – среднее артериальное давление.
5. s1 – общий холестерин сыворотки.
6. s2 – липопротеины низкой плотности.
7. s3 – липопротеины высокой плотности.
8. s4 – общий холестерин / ЛПВП.
9. s5 – логарифм уровня триглицеридов в сыворотке.
10. s6 – уровень сахара в крови.

Для данного датасета используется аналогичная трактовка первой задачи, что и для первого набора данных.

После обучения модели, для сравнения библиотек наугад выбран индекс примера = 5.

Таблица 2

Соотношение имени признака из второго датасета и его значения для наугад выбранного примера

Name feature	Value
age	-0.092695
sex	-0.044642
bmi	-0.040696
bp	-0.019442
s1	-0.068991
s2	-0.079288
s3	0.041277
s4	-0.076395

s5	-0.041180
s6	-0.096346

Для данных в табл. 2 логистическая регрессия предсказала значение = 106.35. Теперь посмотрим интерпретацию результата по библиотеке SHAP на рис. 4.

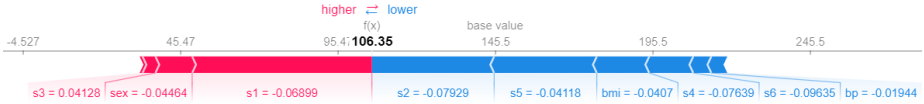

Рис. 4. Результат интерпретации результатов машинного обучения для предсказания целевой переменной y 5 примера с помощью библиотеки SHAP.

Как видно по рис. 4 сильно снизил целевую переменную признак *s2, s5, bmi*, повысило же значение признаков *sex, s1*. Красные признаки влияют положительно, а синие – отрицательно.

Для демонстрации работы библиотеки ELI5 был выбран тот же датасет – diabetes. Но рассмотрена первая и вторая задача. Для первой задачи выбран тот же пример – 5. Результат по двум задачам объединен на рис. 5.

y top features

Weight[?]	Feature
+751.279	s5
+519.840	bmi
+476.746	s2
+324.390	bp
+177.064	s4
+152.133	<BIAS>
+101.045	s3
+67.625	s6
-10.012	age
-239.819	sex
-792.184	s1

y (score **106.349**) top features

Contribution[?]	Feature	Value
+152.133	<BIAS>	1.000
+54.653	s1	-0.069
+10.706	sex	-0.045
+4.171	s3	0.041
+0.928	age	-0.093
-6.307	bp	-0.019
-6.515	s6	-0.096
-13.527	s4	-0.076
-21.155	bmi	-0.041
-30.938	s5	-0.041
-37.800	s2	-0.079

Рис. 5. Интерпретация результатов с помощью библиотеки ELI5: *левая таблица* – для второй задачи, *правая* – для первой задачи

По рис. 5 видно, что для второй задачи сильное негативное влияние имеет признак *s1*, а положительное влияние имеет признак *s5*. Для первой задачи наоборот получилось, что для примера под номером 5, признак *s1* имеет положительное влияние, а *s5* – отрицательное.

Результаты интерпретации, сформированные библиотекой Lime, для примера 5 отображены на рис. 6.

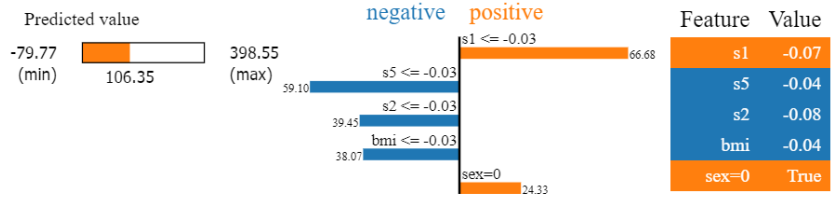

Рис. 6. Интерпретация результатов с помощью библиотеки Lime

На рис. 3 первый барплот отвечает за значение целевой переменной для выбранного примера. График посередине отвечает за влияние признаков на целевую переменную. Правая таблица демонстрирует исходные значения примера.

По рис. 6, можно сказать, что есть отличительный признак *sex*, который положительно повлиял на значение целевой переменной. В результатах по двум предыдущим библиотекам, данного признака нет.

4. Сравнение библиотек

Сравнение библиотек осуществлялось на стандартной задаче регрессии – об оценке помещения из Бостона и предсказании прогрессирования диабета через год после исходного уровня.

После получения результатов интерпретации модели для задачи регрессии с помощью библиотек ELI5, SHAP, Lime, было решено сравнить, какие самые влиятельные признаки выделили интерпретаторы. Для данного сравнения были выделены первые 5 интерпретируемых значений для признаков по модулю. Модуль взят для анализа и отрицательно влияющих признаков на результат целевой переменной. Данные для первого датасета занесены в табл. 3, а для второго в табл. 4.

Таблица 3

Сравнение результатов интерпретации по трем библиотекам для датасета об оценке помещения из Бостона

Name Feature	Lime	Name Feature	ELI5	Name Feature	SHAP

RM	-5,03	NOX	-17,76	**RM**	-3,09
PTRATIO	-3,22	*RM*	3,81	**PTRATIO**	-2,46
CHAS	-3,21	**CHAS**	2,68	*RAD*	-1,70
RAD	-2,14	DIS	-1,47	TAX	1,16
ZN	-2,03	*PTRATIO*	-0,95	B	-0,71

В табл. 3 курсивом выделены признаки, которые имеются в топ-5 хотя бы по двум библиотекам. Жирным выделены показатели с полным совпадением порядка и знака для интерпретированных данных.

Также для ELI5 есть “аномалия”, по сравнению с Lime и SHAP данная библиотека интерпретировала признак *RM* как положительно влияющий на цену, что является странным. Так как в анализируемом примере *RM* = 5,456 в то время, как среднее значение по данному признаку = 6,284. Таким образом, получается, что значение *RM* ниже среднего.

По второму датасету также были выделены топ-5 признаков и результаты занесены в табл. 4.

Таблица 4

Сравнение результатов интерпретации по трем библиотекам для датасета об оценке прогрессирования диабета

Name Feature	Lime	Name Feature	ELI5	Name Feature	SHAP
s1	66,68	**s1**	54,65	**s1**	57,08
s5	-59,10	**s2**	-37,80	**s2**	-38,83
s2	-39,40	**s5**	-30,93	**s5**	-32,60
bmi	-38,10	**bmi**	-21,15	**bmi**	-16,65
sex	24,33	**s4**	-13,52	**s4**	-13,99

Аналогично, как и для табл. 3 курсивом выделены признаки, которые имеются в топ 5 хотя бы по двум библиотекам. Жирным выделены показатели с полным совпадением порядка и знака для интерпретированных данных.

Получились противоположные результаты в сравнении с табл. 3. По табл. 4 видно, что произошло полное совпадение результатов интерпретации ELI5 и SHAP. Кроме того, признаки *s1, bmi* выделены на одинаковых уровнях важности для трех библиотек, то есть эти

показатели расположены на 1 и 4 месте в топ-5 самых важных признаков.

Что касательно визуального восприятия, то в библиотеке SHAP заложено намного больше функционала для отображения интерпретации результатов машинного обучения. Это обусловлено тем, что данная библиотека активно дорабатывается и развивается, разработчиками.

Заключение

Результаты интерпретации по библиотекам могут совпасть, а могут различаться, поэтому при решении задачи интерпретации результатов машинного обучения для задачи регрессии, необходимо пользоваться несколькими инструментами сразу.

Для выполнения данной задачи рекомендуется использовать библиотеки ELI5 и SHAP, так-как данные библиотеки активно разрабатываются и расширяют свой функционал. Также имеется подробная документация для обеих библиотек.

Однако, есть проблема с оценкой результатов интерпретаций по различным методам. Для каждой задачи в данный момент необходимо экспертное мнение, которое будет основываться на результатах интерпретации моделей машинного обучения.

Для того, чтобы облегчить эксперту работу или же объяснить для бизнес-логики тот или иной результат лучше всего использовать библиотеку SHAP. Данная библиотека использует JavaScript для визуализации интерпретации результатов машинного обучения, данный функционал позволяет просто и интерактивно подойти к объяснению ответа итоговой модели.

Список литературы

1. Атлас электрокардиографии. Интерпретация результатов: от простого к сложному / Н. А. Новикова [и др.] – Москва : Эксмо, 2022. – 128 с.

2. Будума, Основы глубокого обучения. Создание алгоритмов для искусственного интеллекта следующего поколения / Н. Будума, Н. Локашо ; пер. с англ. А. Коробейникова ; [науч. Ред. А. Созыкин]. – М. : Манн, Иванов и Фербер, 2020. – 304 с.

3. Бурков. Машинное обучение без лишних слов. – СПБ.: Питер, 2020. – 192 с..

4. Eli5 [Электронный ресурс] : источник библиотеки. – Режим доступа : https://github.com/TeamHG-Memex/eli5.

5. Interpretable and Explainable Machine Learning for Materials Science and Chemistry [Электронный ресурс] / F. Oviedo [and all] :

Computing Research Repository (CoRR), 2021. – Режим доступа : https://arxiv.org/abs/2111.01037.

6. Interpretable Machine Learning for Genomics [Электронный ресурс] / S. Watson : Computing Research Repository (CoRR), 2021. – Режим доступа : https://arxiv.org/abs/2110.03063.

7. Lime [Электронный ресурс] : источник библиотеки. – Режим доступа : https://github.com/marcotcr/lime.

8. SHAP [Электронный ресурс] : источник библиотеки. – Режим доступа : https://github.com/slundberg/shap.

9. Scikit-learn documentation [Электронный ресурс] : источник библиотеки. – Режим доступа : https://scikit-learn.org/stable/modules/classes.html#module-sklearn.datasets.

10. 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.