ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
http://siit.ugatu.su 2026. Т. 8, № 4(28)
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
Научная статья

УДК 004.89 DOI: 10.54708/SIIT-2026-no4-p75 EDN: UBUULW

Оценка объяснений моделей «чёрного ящика» посредством интеграции метода взвешенной суммы на основе Z-чисел

Л. А. Гардашова • П. И. Косов

Азербайджанский государственный университет нефти и промышленности

Рассматривается задача объективного сравнения качества объяснений, которые формируют методы объяснимого искусственного интеллекта (XAI). Предложена многометодная схема оценки. Она соединяет четыре автоматически вычислимые метрики, пять человеко-ориентированных критериев и процедуру интеграции в единый ранг. Верность объяснения вычисляется как апостериорная байесовская вероятность его согласованности с решением модели. Информативность оценивается взвешенным нечётким средним степеней принадлежности семантических свойств. Устойчивость измеряется коэффициентом Жаккара и Евклидовым расстоянием. Частные оценки агрегируются моделью взвешенной суммы на основе Z-чисел. Веса критериев вычисляются методом собственного вектора Z-значной матрицы парных сравнений. Компонента надёжности Z-числа отражает достоверность каждого измерения и снижает вклад вырожденных оценок. Схема апробирована при сравнении восьми подходов XAI на наборе данных German Credit Risk. В сравнение вошли SHAP, LIME, контрфактические объяснения, нечёткие правила, нечёткие деревья решений и три онтологических подхода. Количественное и человеко-ориентированное ранжирования согласованы. Ранговая корреляция Спирмена между ними составила 0,86. Наибольший интегральный ранг в обоих контурах получил нечёткий семантический подход, построенный на Fuzzy OWL2. Результат демонстрирует разрешающую способность схемы и её применимость к произвольным методам XAI

Объяснимый искусственный интеллект; оценка качества объяснений; нечёткая онтология; Fuzzy OWL2; Z-числа; многокритериальный анализ решений; модель взвешенной суммы.

Введение

Модели машинного обучения всё чаще участвуют в решениях с высокой ценой ошибки. К таким областям относятся кредитный скоринг, медицинская диагностика и информационная безопасность. Высокая точность современных моделей сочетается с непрозрачной внутренней логикой. Пользователь видит результат, но не видит его оснований. Для преодоления этого разрыва сформировалось направление объяснимого искусственного интеллекта [Ada18, Bar20]. Часть исследователей настаивает на применении изначально интерпретируемых моделей в критических задачах [Rud19]. На практике широкое распространение получили апостериорные методы объяснения. Метод LIME строит локальную линейную аппроксимацию решения [Rib16]. Метод SHAP распределяет вклады признаков на основе значений Шепли [Lun17]. Контрфактические объяснения указывают минимальное изменение входа, при котором вердикт модели меняется [Wex19].

Рост числа методов обострил вопрос сравнения качества самих объяснений. Эталонное «правильное» объяснение для решения модели обычно отсутствует. Поэтому прямая сверка с истиной невозможна [Yan19]. Систематические обзоры фиксируют, что значительная часть

Гардашова Л. А., Косов П. И. Оценка объяснений моделей «чёрного ящика» посредством интеграции метода взвешенной суммы на основе Z-чисел // СИИТ. 2026. Т. 8, № 4(28). С. 75-86. DOI: 10.54708/SIIT-2026-no4-p75. EDN: UBUULW.

Gardashova L. A., Kosov P. I. “Evaluation of black-box model explanations through the integration of a weighted sum method based on Z-numbers” // SIIT. 2026. Vol. 8, no. 4(28), pp. 75-86. DOI: 10.54708/SIIT-2026-no4-p75. EDN: UBUULW. (In Russian).

публикаций по XAI ограничивается демонстрацией удачных примеров без измеримых показателей [Zho21, Nau23]. Для человеко-ориентированной стороны качества предложены психометрические шкалы [Hof19]. Они охватывают восприятие объяснений людьми, но не заменяют объективных измерений. В работе [Ali24] задача сравнения методов XAI сведена к многокритериальному анализу решений на основе Z-чисел. Критериями там выступают субъективные суждения экспертов. Автоматически вычислимые метрики в интеграции не участвуют.

Таким образом в литературе сложился разрыв. Количественные метрики фрагментарны и не сводятся в единый показатель. Качественные критерии субъективны и редко сопоставляются с измерениями. Достоверность самих измерений при агрегировании обычно не учитывается. Целью настоящей работы является построение воспроизводимой многометодной схемы оценки объяснений, свободной от перечисленных ограничений. Научная новизна работы состоит в следующем:

1. Сформирован комплекс из четырёх количественных метрик, которые покрывают верность, информативность и два вида устойчивости объяснения. Метрики вычисляются автоматически и не требуют эталонного объяснения.

2. Для сведения разнородных метрик в единый ранг применена модель взвешенной суммы на основе Z-чисел. Компонента надёжности Z-числа впервые трактуется как достоверность самого измерения. За счёт этого вырожденные и косвенные оценки автоматически получают меньший вклад.

3. Количественное ранжирование верифицировано независимым слепым опросом экспертов и пользователей. Ответы участников с фиксацией уверенности отображаются в Z-числа. Поэтому оба контура оценки используют один и тот же математический аппарат и допускают прямое сопоставление.

Работа схемы демонстрируется сравнением восьми подходов XAI. Среди них два семантических подхода, предложенных автором ранее. Это чёткий подход на основе онтологии OWL2 [Kos24] и нечёткий подход на основе Fuzzy OWL2 [Kos25]. Подчеркнём, что подтверждение превосходства какого-либо подхода не входит в цели работы. Итоговое лидерство нечёткого семантического подхода возникает как следствие применения схемы и служит иллюстрацией её разрешающей способности.

Обзор литературы

Существующие процедуры оценки объяснений принято группировать по трём уровням [Nau23, Hof19]. Функционально-ориентированный уровень опирается на формальные показатели и не требует участия человека. Человеко-ориентированный уровень использует опросы людей на упрощённых задачах. Прикладной уровень проверяет пользу объяснений в реальной работе экспертов предметной области. Обоснованное заключение о качестве требует сочетания уровней. На практике такое сочетание встречается редко.

Среди функциональных показателей чаще других применяются верность и устойчивость. Верность характеризует соответствие объяснения действительной логике модели [Yan19]. Устойчивость отражает чувствительность объяснения к малым возмущениям входа. Известная нестабильность локальных аппроксимаций делает этот показатель особенно важным для LIME и родственных методов [Zho21]. Отдельные показатели информативны сами по себе. Однако без единой процедуры агрегирования они не дают целостного сравнения методов.

Отдельное направление связывает объяснения с формализованными знаниями предметной области. Обзор [See19] фиксирует устойчивый рост интереса к семантическим технологиям в задачах объяснимости. Онтологический паттерн проектирования объяснений предложен в работе [Tid15]. Система OBIC сочетает онтологическую модель с интерфейсом объяснения при классификации изображений [Bel22]. Семантические подходы придают объяснениям структуру и проверяемость. При этом классические онтологии оперируют бинарной истинностью. Пограничные случаи в такой логике огрубляются, а частичная выраженность свойств теряется.

Аппарат нечётких множеств [Zad65] снимает указанное ограничение. Нечёткие правила и нечёткие деревья решений интерпретируемы по построению [Men21, Alo20]. Представление нечётких онтологий средствами языка OWL 2 формализовано в [Bob11]. Ризонер DeLorean обеспечивает логический вывод над такими онтологиями [Bob12]. На этой основе автором предложен нечёткий семантический подход к формированию объяснений, где семантические свойства решения снабжаются степенями принадлежности [Kos25]. Его предшественником служит чёткий семантический подход с расширенным набором объяснительных свойств онтологии [Kos24].

Для агрегирования разнородных оценок с учётом их достоверности подходит аппарат *Z*-чисел [Zad11]. Z-число фиксирует одновременно значение величины и надёжность этого значения. Процедура построения согласованных Z-значных матриц парных сравнений разработана в [Ali21]. Применение Z-чисел к сравнению методов XAI в медицинской области показано в [Ali24]. Настоящая работа развивает эту линию. Отличие состоит в источнике критериев. Вместо экспертных суждений в качестве критериев берутся объективные измеримые метрики. Компонента надёжности при этом описывает достоверность измерения, а не уверенность человека.

Многометодная схема оценки объяснений

Архитектура схемы

Предлагаемая схема включает три последовательных контура.

1. Для каждого оцениваемого подхода автоматически вычисляются четыре частные количественные метрики. Это верность, информативность доверия, семантическая устойчивость и числовая устойчивость.
2. Частные метрики сводятся в единый интегральный ранг моделью взвешенной суммы на основе *Z*-чисел (далее Z-WSM). Каждой метрике сопоставляется надёжность её измерения для конкретного подхода.
3. Независимо от первых двух проводится слепой опрос экспертов и пользователей по пяти качественным критериям. Ответы отображаются в Z-числа и интегрируются тем же методом Z-WSM. Затем два ранжирования сопоставляются ранговой корреляцией.

Такое построение обеспечивает триангуляцию выводов. Совпадение результатов независимых контуров усиливает достоверность итоговой оценки. Расхождение сигнализирует о необходимости дополнительного анализа отдельных метрик.

Количественные метрики

Верность. Метрика отвечает на вопрос о том, отражает ли объяснение действительный механизм принятия решения моделью «чёрного ящика» [Yan19]. Пусть событие A означает корректную классификацию объекта, а событие C означает согласованность объяснения с моделью знаний. Тогда верность определяется как апостериорная вероятность корректности при наблюдаемой согласованности. Она вычисляется по формуле Байеса (1). Высокое значение говорит о том, что объяснение опирается на признаки, действительно определившие решение.

$$P(A|C) = \frac{P(C|A)\cdot P(A)}{P(C|A)\cdot P(A) + P(C|A')\cdot P(A')}. \quad (1)$$

Информативность доверия. Отдельное объяснение содержит набор семантических свойств со степенями принадлежности соответствующих нечётких термов [Tak85]. Эти степени агрегируются в единый показатель TS взвешенным нечётким средним:

$$TS = \sum_{i=1}^{n} w_i\, \mu_i(x). \quad (2)$$

Здесь w_i обозначает вес важности i-го свойства, а $\mu_i(x)$ обозначает степень его принадлежности. Показатель характеризует, насколько содержательно объяснение раскрывает уверенность модели в отдельных факторах решения.

Семантическая устойчивость. Метрика проверяет, сохраняется ли состав объяснения при малом возмущении входных данных. Пусть A и B обозначают множества свойств двух объяснений, построенных до и после возмущения. Их сходство измеряется коэффициентом Жаккара (3) как отношение мощности пересечения к мощности объединения. Значение близкое к единице говорит о согласованном составе свойств. Заметное падение указывает, что подход опирается на разные признаки в почти одинаковых ситуациях.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \quad (3)$$

Числовая устойчивость. Коэффициент Жаккара сравнивает лишь состав множеств. Дополняющая его метрика чувствительна к сдвигам самих степеней принадлежности. Для векторов V_A и V_B степеней принадлежности общих свойств вычисляется Евклидово расстояние

$$D(V_A,V_B) = \sqrt{\sum_{i=1}^{n}(v_{A,i} - v_{B,i})^2}. \quad (4)$$

Малое расстояние соответствует плавному изменению объяснения. Резкий рост означает, что веса признаков перестроились скачком.

Качественные критерии

Человеко-ориентированный контур использует пять критериев, отобранных по итогам анализа работ [Zho21, Nau23, Hof19]. Каждый критерий операционализирован отдельным пунктом опросного листа.

- *Понятность*. Критерий показывает, насколько легко участник восстанавливает причину решения по тексту объяснения. Учитывается баланс между полнотой изложения и отсутствием перегрузки второстепенными деталями;
- *Удовлетворённость*. Критерий отражает соответствие формы и содержания объяснения ожиданиям конкретного пользователя. Даже корректное по существу объяснение способно разочаровать человека, когда оно обходит важные для него вопросы;
- *Доверие*. Критерий оценивает, раскрывает ли объяснение уверенность модели в собственном выводе и границы её применимости. Явное отображение неуверенности позволяет пользователю обоснованно решать, когда полагаться на систему;
- *Действенность*. Критерий проверяет пригодность объяснения как руководства к действию. Показательный пример состоит в возможности указать, какие входные параметры следует изменить для получения желаемого результата;
- *Полнота*. Критерий фиксирует, охвачены ли объяснением все существенные факторы решения. Неполные объяснения скрывают часть оснований и способны вводить пользователя в заблуждение.

Интеграция методом Z-WSM

Частные метрики измерены в разных шкалах и имеют разную достоверность. Для их сведения в единый ранг применяется модель взвешенной суммы на основе Z-чисел [Ali24]. Z-числом называется упорядоченная пара $Z = (A, B)$. Компонента A задаёт нечёткое значение оцениваемой величины. Компонента B выражает надёжность этого значения качественным лингвистическим термом [Zad11]. При вычислениях лингвистическим термам сопоставляются числовые уровни надёжности.

Веса критериев вычисляются методом собственного вектора Z-значной матрицы парных сравнений [Ali21]. Согласованность матрицы контролируется отношением согласованности CR с порогом 0,1. Взвешенная Z-оценка i-й альтернативы вычисляется по формуле:

$$S_i = \sum_{j=1}^{m} w_j \otimes Z_{ij}. \quad (5)$$

Здесь w_j обозначает вес j-го критерия, Z_{ij} обозначает Z-оценку альтернативы по этому критерию, а символ $\otimes$ обозначает умножение Z-чисел.

Итоговый порядок альтернатив определяется их удалённостью от идеальной альтернативы $o*$. У неё все частные оценки максимальны при полной надёжности. Удалённость вычисляется по формуле:

$$D(o_i, o*) = \sqrt{\sum_{j=1}^{m} D^2(Z_{ij}, Z_j*)}\,. \quad (6)$$

Расстояние между Z-числами учитывает расхождение и значений A, и надёжностей B. Меньшая удалённость соответствует лучшей альтернативе.

Экспериментальная количественная оценка

Условия эксперимента

Объектом оценки выбран заёмщик LoanTaker_0 из открытого набора данных German Credit Risk [Hof94]. Базовая модель классификации отнесла его к надёжным заёмщикам. Для метрик устойчивости подготовлена возмущённая копия объекта LoanTaker_0_mod. В ней незначительно изменён остаток на сберегательном счёте. Единый объект, единый прогноз и единое пространство свойств исключают влияние посторонних факторов. Различаться могут только механизмы построения объяснений.

В сравнение включены восемь подходов, покрывающих основные парадигмы XAI. Подход O1 представляет чёткий семантический подход автора на основе OWL2 [Kos24]. Подход O2 представляет предложенный автором нечёткий семантический подход, использующий Fuzzy OWL2 [Kos25]. Подход O3 использует объяснения на основе нечётких правил [Men21]. Подход O4 опирается на нечёткие деревья решений [Alo20]. Подход O5 соответствует онтологической системе OBIC [Bel22]. Подходы O6, O7 и O8 представляют соответственно SHAP [Lun17], LIME [Rib16] и контрфактические объяснения CFE [Wex19].

Применимость метрик

Методологические различия подходов приводят к тому, что не каждая метрика вычислима для каждого подхода в содержательном виде. Сводка применимости приведена в табл. 1. Чёткие онтологические подходы O1 и O5 не располагают градуированными степенями принадлежности. Изменения их объяснений происходят скачком. Поэтому евклидова метрика для них вырождается, тогда как коэффициент Жаккара остаётся применимым. Контрфактическое объяснение O8 не формирует профиля степеней принадлежности. Его показатель TS собирается лишь формально. Для алгоритмических подходов O6, O7 и O8 верность и устойчивость измеримы только на уровне признаков без семантической проверки. Полный набор метрик в невырожденном виде вычислим лишь для нечёткого семантического подхода O2.

Таблица 1

Применимость количественных метрик к оцениваемым подходам

Группа подходов	$P(A\|C)$	TS	J	D
Чёткие онтологические (O1, O5)	семантич. уровень	вырождена (порог)	применима	вырождена
Нечёткий семантический (O2)	семантич. уровень	применима	применима	применима
Нечёткие ante-hoc (O3, O4)	признаковый уровень	применима	применима	применима
SHAP, LIME (O6, O7)	признаковый уровень	по весам признаков	применима	применима
CFE (O8)	признаковый уровень	вырождена	применима	по контрфактам

Результаты по частным метрикам

Верность вычислялась по формуле (1) на едином наборе испытаний. Общая точность базовой модели составила 0,81. Для чёткого подхода O1 чувствительность согласованности достигла 0,963 при уровне ложных срабатываний 0,211. Отсюда верность O1 равна 0,951. Нечёткая проверка согласованности в подходе O2 тоньше обрабатывает пограничные случаи. Случаи, которые бинарная логика ошибочно засчитывала как согласованные, получают низкую степень уверенности. Чувствительность возрастает до 0,974, а уровень ложных срабатываний падает до 0,120. Подстановка в формулу (1) даёт верность 0,972. Доля ложноположительных среди согласованных объяснений при этом сокращается с 4,9 до 2,8 процента.

У алгоритмических подходов верность измерялась как согласованность признаковых атрибуций с поведением модели. SHAP показал 0,875. LIME с локально-линейной аппроксимацией достиг лишь 0,700. Контрфактические объяснения дали 0,660, поскольку характеризуют гипотетическую модификацию решения вместо самого решения. Модели за подходами O3 и O4 прозрачны по своей конструкции. Поэтому их объяснения верны собственной логике. Однако их верность ограничена сверху точностью самой интерпретируемой модели. Она составила 0,730 и 0,760 соответственно.

Показатель TS для подхода O2 вычислен по профилю из шести семантических свойств со степенями принадлежности от 0,54 до 0,94 при экспертных весах важности. Взвешенное нечёткое среднее составило 0,738. Чёткая логика с порогом отсечения 0,6 пропускает лишь три свойства и отбрасывает частичную истинность остальных. Такая вырожденная оценка равна 0,650 и сопровождается потерей информации. Для нечётких подходов O3 и O4 показатель вычислен на собственных степенях принадлежности и составил 0,650 и 0,670. Для SHAP и LIME он построен на нормированных весах признаков и оказался менее содержательным (0,555 и 0,480). Такие веса имеют иную природу, нежели степени принадлежности нечётких термов. У CFE показатель выродился до 0,150.

Устойчивость проверялась переходом от LoanTaker_0 к LoanTaker_0_mod. Для подхода O2 из шести свойств объединения совпали четыре. Коэффициент Жаккара составил 0,667. Евклидово расстояние по степеням принадлежности общих свойств оказалось около 0,030. Объяснение изменилось плавно, без скачков. Чёткие онтологические подходы дали J равное 0,620 для O1 и 0,590 для O5 при вырожденной числовой метрике. Для LIME зафиксировано заметное падение J до 0,430 и рост D до 0,190, что согласуется с известной нестабильностью метода. SHAP показал J равное 0,610 при D равном 0,105. Нечёткие ante-hoc подходы дали промежуточные значения. Для CFE устойчивость оценивалась расстоянием между контрфактами (J равно 0,500, D равно 0,150). Сводные результаты приведены в табл. 2. Для метрик $P(A|C)$, TS и J лучшим является большее значение. Для метрики D лучшим является меньшее значение. Прочерк обозначает вырожденность метрики. Знаком «*» отмечена вырожденная оценка. Прочерк означает неприменимость метрики.

Таблица 2

Частные количественные метрики восьми подходов

Подход	$P(A\|C)$	TS	J	D	Уровень проверки
O1 (OWL2) [Kos24]	0,951	0,650	0,620	—	онтологический, бинарный
O2 (Fuzzy OWL2) [Kos25]	0,972	0,738	0,667	0,030	онтологический, нечёткий
O3 (нечёткие правила)	0,730	0,650	0,575	0,095	уровень признаков, ante-hoc
O4 (нечёткие деревья)	0,760	0,670	0,600	0,080	уровень признаков, ante-hoc
O5 (OBIC)	0,927	0,600	0,590	—	онтологический, бинарный
O6 (SHAP)	0,875	0,555	0,610	0,105	уровень признаков, post-hoc
O7 (LIME)	0,700	0,480	0,430	0,190	уровень признаков, post-hoc
O8 (CFE)	0,660	0,150*	0,500	0,150	уровень признаков, post-hoc

Интегральное ранжирование

Четыре метрики приняты критериями интеграции. Критерий C1 соответствует верности, C2 соответствует показателю TS, C3 соответствует коэффициенту Жаккара, C4 соответствует евклидовой метрике. Критерий C4 является затратным. Перед интеграцией он преобразуется в показатель устойчивости $q = 1/(1 + D)$. Остальные критерии по построению лежат в отрезке от 0 до 1 и дополнительной нормализации не требуют.

Относительная важность критериев задана в Z-значной лингвистической шкале Саати. Верность признана умеренно важнее доверия, сильно важнее семантической устойчивости и очень сильно важнее числовой устойчивости. Соответствующие оценки равны 3, 5 и 7. Доверие слабо важнее семантической устойчивости и умеренно-сильно важнее числовой. Семантическая устойчивость слабо важнее числовой. Каждое суждение основано на объективном измерении. Поэтому всем суждениям назначена очень высокая надёжность. Матрица парных сравнений оказалась согласованной. Максимальное собственное значение составило 4,028, индекс согласованности равен 0,009, а отношение согласованности CR равно 0,011 при пороге 0,1. Веса вычислены методом собственного вектора [Ali21] и приведены в табл. 3.

Таблица 3

Веса критериев интеграции по методу собственного вектора

Критерий	Компонента собственного вектора	Нормированный вес
C1. Байесовская верность $P(A\|C)$	0,908	0,582
C2. Интегральное доверие TS	0,361	0,231
C3. Устойчивость состава J	0,188	0,121
C4. Устойчивость значений D	0,104	0,066

Каждая ячейка матрицы решений представлена Z-числом. Значимостная компонента равна нормированной частной метрике. Компонента надёжности показывает, насколько достоверно метрика измерена у конкретного подхода. Вырожденным и признаковым оценкам назначена пониженная надёжность. Взвешенные Z-оценки S_i вычислены по формуле (5), а расстояния до идеала по формуле (6). Итог сведён в табл. 4.

Таблица 4

Интегральное ранжирование подходов методом Z-WSM

Ранг	Подход	S_i	$D(o_i, o^*)$	Комментарий
1	O2 (Fuzzy OWL2)	0,881	0,217	Все метрики невырождены.
2	O1 (OWL2)	0,778	0,307	Высокая верность, вырождена метрика D.
3	O5 (OBIC)	0,749	0,347	Семантическая верность, вырожденная устойчивость.
4	O6 (SHAP)	0,771	0,362	Лучший среди алгоритмических.
5	O4 (нечёткие деревья)	0,731	0,377	Верность ограничена точностью модели.
6	O3 (нечёткие правила)	0,705	0,420	Аналогично O4 при меньшей верности.
7	O7 (LIME)	0,626	0,553	Нестабильность объяснений.
8	O8 (CFE)	0,537	0,683	Вырожденная оценка доверия.

Количественное ранжирование от лучшего к худшему приняло вид O2, O1, O5, O6, O4, O3, O7, O8. Обратим внимание на пару SHAP и OBIC. По значимостной компоненте S_i метод SHAP немного впереди (0,771 против 0,749). Однако при переходе к расстоянию до идеала учитывается компонента надёжности. Семантически обоснованная оценка OBIC измерена достовернее и оказывается ближе к идеалу (0,347 против 0,362). Перестановка этой пары

наглядно показывает вклад надёжной компоненты Z-чисел. Классическим методам взвешенной суммы такой механизм недоступен.

Человеко-ориентированная верификация

Методика опроса

В опросе участвовали 12 человек. Пять участников являются экспертами в области XAI и семантических технологий. Семь участников являются практиками кредитного скоринга. Такой состав покрывает техническую корректность объяснений и их практическую пригодность. Всем участникам предъявлялось одно и то же решение по заёмщику LoanTaker_0 в восьми стилях объяснения O1–O8. Порядок предъявления был случайным, а источник каждого объяснения скрывался. Слепая процедура исключает влияние априорных предпочтений и переносит внимание на содержание.

Каждый из пяти критериев операционализирован отдельным утверждением опросного листа. Ответ давался по пятибалльной шкале Лайкерта. Дополнительно участник указывал уверенность в ответе лингвистическим термом «низкая», «средняя» или «высокая». При обработке этим термам сопоставлялись уровни надёжности 0,5, 0,7 и 0,9. Пара из оценки и уверенности образует Z-число и служит входом интеграции.

Обработка ответов выполнялась в три шага. Сначала по каждой паре «подход, критерий» вычислялось среднее оценок, взвешенное уверенностями участников. За счёт этого уверенные ответы получают больший вклад, чем сомнительные. Затем результат отображался в Z-число. Значимостная компонента равна взвешенному среднему, приведённому к отрезку от 0 до 1. Компонента надёжности равна средней уверенности по данной паре. На заключительном шаге проверялась статистическая состоятельность самого опроса. Коэффициент конкордации Кендалла для ранжировок 12 участников по восьми подходам составил около 0,78. Критерий хи-квадрат дал значение около 65,5 при семи степенях свободы. Оно существенно превышает критический порог 18,48 при уровне значимости 0,01. Дополнительно вычисленный коэффициент альфа Криппендорфа около 0,74 подтверждает приемлемое согласие. Поэтому результаты опроса не случайны и пригодны для интеграции.

Поясним взвешивание на примере критерия доверия. Пять экспертов оценили подход O2 баллами 5, 5, 4, 5 и 4 при уверенностях 0,90, 0,90, 0,80, 0,95 и 0,85. Взвешенное среднее равно 20,35/4,40, то есть примерно 4,63. Соответствующее Z-число имеет значимость 0,93 при надёжности 0,88. Тот же расчёт для LIME выполнен на разрозненных ответах 3, 2, 4, 2 и 3 с уверенностями от 0,50 до 0,70. Взвешенное среднее составило лишь 2,90 при пониженной надёжности. Сопоставление двух расчётов показывает чувствительность процедуры к компоненте уверенности. Согласованные ответы формируют высокую и надёжную оценку. Разрозненные и неуверенные ответы дают низкую оценку со сниженной надёжностью.

Результаты опроса

Сводные результаты по пяти критериям приведены в табл. 5. В столбцах K1–K5 даны взвешенные средние баллы по шкале Лайкерта. Столбец $\bar{B}$ содержит среднюю уверенность участников по подходу. Столбец S содержит интегральный качественный балл Z-WSM. При интеграции критерию доверия назначен наибольший вес 0,30 как ключевому для ответственного применения ИИ. Понятность получила вес 0,25. Остальные три критерия получили по 0,15. Веса заданы в Z-значной лингвистической шкале. Их согласованность подтверждена процедурой.

По понятности лидирует подход O2 с баллом 4,6. Участники отмечали, что его объяснение воспринимается как связный причинный вывод, дополненный градациями уверенности. Чёткие онтологические подходы читаются легко, однако сглаживают пограничные ситуации. Признаковые объяснения SHAP и LIME воспринимаются заметно хуже. По удовлетворённости картина повторяется. Наименьший балл 3,0 получил LIME. Участники связывали это с ощущением произвольности его нестабильных объяснений.

Таблица 5

Результаты человеко-ориентированной оценки по пяти критериям

Подход	K1	K2	K3	K4	K5	$\bar{B}$	S
O1 (OWL2)	4,2	4,0	3,6	3,8	4,1	0,82	0,78
O2 (Fuzzy OWL2)	4,6	4,5	4,6	4,3	4,7	0,90	0,91
O3 (нечёткие правила)	3,8	3,6	3,7	3,5	3,4	0,74	0,71
O4 (нечёткие деревья)	4,0	3,8	3,8	3,7	3,6	0,78	0,75
O5 (OBIC)	4,0	3,8	3,4	3,5	3,9	0,80	0,73
O6 (SHAP)	3,5	3,4	3,5	3,6	3,2	0,76	0,68
O7 (LIME)	3,3	3,0	2,9	3,2	2,8	0,64	0,59
O8 (CFE)	3,6	3,5	3,0	4,4	2,6	0,70	0,66

Доверие оказалось критерием с наибольшим отрывом лидера. Лишь подход O2 напрямую сообщает пользователю, насколько модель уверена в выводе и где заканчивается область её корректной работы. Его балл составил 4,6 при 3,6 у ближайшего чёткого подхода. Контрфактические объяснения и LIME получили наименьшие баллы, поскольку ничего не сообщают о надёжности собственного вывода. По действенности единственный раз первое место занял другой подход. Контрфактические объяснения устроены так, что прямо называют изменения, необходимые для смены вердикта. Их балл 4,4 превысил балл 4,3 подхода O2. Фиксация этого компромисса подчёркивает непредвзятость процедуры. По полноте подход O2 лидирует с наибольшей разницей (4,7 против 4,1 у ближайшего конкурента). Семантическая модель охватывает и состав факторов, и их относительную силу. Беднее всего выглядит CFE с баллом 2,6, поскольку контрфакт демонстрирует лишь один путь изменения решения.

Согласованность двух контуров

Качественное ранжирование по расстоянию до идеала приняло вид O2, O1, O4, O5, O3, O6, O8, O7. Расстояния монотонно растут от 0,19 у лидера до 0,62 у LIME. Согласованность с количественным ранжированием проверена ранговой корреляцией Спирмена. Суммарный квадрат ранговых расхождений составил 12. Отсюда коэффициент корреляции получился около 0,86. Оба независимых контура поставили подход O2 на первое место. Положительная сильная корреляция подтверждает взаимную состоятельность контуров.

Отдельного внимания заслуживает роль компоненты надёжности. Подходы O8 и O7 близки по значимостным баллам S (0,66 и 0,59). Однако средняя уверенность участников в LIME оказалась наименьшей среди всех подходов (0,64). Низкая надёжность непропорционально увеличила расстояние LIME до идеала и отправила его на последнее место. Аналогичный механизм в количественном контуре поменял местами SHAP и OBIC. Оба эпизода демонстрируют содержательный вклад надёжной компоненты Z-чисел в итоговый порядок.

Обсуждение результатов

Сходство двух независимых ранжирований объясняется не случайностью, а внутренними свойствами оцениваемых систем. Восемь подходов сгруппированы в четыре класса и сопоставлены по пяти концептуальным измерениям. Результат приведён в табл. 6. У алгоритмических методов отсутствует семантическая глубина и не учитывается неопределённость. Чёткие онтологические подходы выигрывают в глубине и охвате, однако скрывают неуверенность модели. Нечёткие неонтологические методы учитывают неопределённость, однако проигрывают в семантической глубине. Верхний уровень сразу по четырём измерениям из пяти достигается только при сочетании онтологической семантики с нечёткой градацией. Именно это сочетание реализует подход O2. Уступает он лишь контрфактическим объяснениям по действенности.

Таблица 6

Концептуальное сопоставление классов подходов

Класс подходов	Глубина семантики	Учёт неопределённости	Калибровка доверия	Руководство к действию	Охват факторов
Алгоритмические (O6–O8)	слабая	отсутствует	отсутствует	умеренное, у CFE сильное	узкий
Онтологические бинарные (O1, O5)	глубокая	отсутствует	слабая	умеренное	широкий
Нечёткие неонтологические (O3, O4)	умеренная	присутствует	умеренная	умеренное	средний
Нечёткий семантический (O2)	глубокая	полный	сильная	сильное	полный

Укажем ограничения исследования. Эксперимент проведён на одном объекте оценки и одном наборе данных. Выборка участников опроса невелика, хотя её согласованность статистически подтверждена. Экспертные веса важности признаков в показателе TS и веса качественных критериев отражают конкретную прикладную область. Перенос схемы на другие области потребует их пересмотра. Дальнейшая работа предполагает масштабирование эксперимента на несколько наборов данных, автоматизацию сбора возмущённых объектов и расширение состава респондентов. Отдельный интерес представляет применение схемы к объяснениям больших языковых моделей.

Заключение

В работе предложена многометодная схема оценки качества объяснений методов XAI. Схема объединяет четыре автоматически вычислимые метрики, интеграцию моделью взвешенной суммы на основе Z-чисел и независимую человеко-ориентированную верификацию. Компонента надёжности Z-числа трактуется как достоверность измерения. За счёт этого вырожденные и косвенные оценки получают меньший вклад в итоговый ранг без ручного вмешательства.

Работоспособность схемы показана сравнением восьми подходов XAI на наборе данных German Credit Risk. Матрица парных сравнений критериев согласована при CR равном 0,011. Опрос участников статистически состоятелен при коэффициенте конкордации Кендалла около 0,78. Количественное и качественное ранжирования согласованы при ранговой корреляции Спирмена около 0,86. Оба контура независимо поставили на первое место нечёткий семантический подход, реализованный средствами Fuzzy OWL2. Единственный зафиксированный компромисс состоит в преимуществе контрфактических объяснений по действенности. Его фиксация подтверждает непредвзятость процедуры.

Практическая ценность схемы состоит в её воспроизводимости и переносимости. Объект оценки, пространство свойств и метод агрегирования фиксируются явно. Поэтому схема применима для обоснованного выбора метода объяснения в прикладных системах поддержки принятия решений, включая кредитный скоринг и смежные ответственные области.

СПИСОК ЛИТЕРАТУРЫ | REFERENCES

[Ada18] Adadi A., Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) // IEEE Access. 2018. Vol. 6. P. 52138–52160. 10.1109/ACCESS.2018.2870052.

[Ali21] Aliev R.A., Guirimov B.G., Huseynov O.H. et al. A consistency-driven approach to construction of Z-number-valued pairwise comparison matrices // Iranian Journal of Fuzzy Systems. 2021. Vol. 18, № 4. P. 37–49. 10.22111/IJFS.2021.6175.

[Ali24] Aliyeva K., Mehdiyev N. Uncertainty-aware multi-criteria decision analysis for evaluation of explainable artificial intelligence methods: A use case from the healthcare domain // Information Sciences. 2024. Vol. 657. P. 1–16. IPDYBA.

[Alo20] Alonso J.M., Ducange P., Pecori R., Vilas R. Building Explanations for Fuzzy Decision Trees with the ExpliClas Software // Proc. 2020 IEEE Int. Conf. Fuzzy Systems. Glasgow, 2020. P. 1–8. 10.1109/FUZZ48607.2020.9177725.

[Bar20] Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI // Information Fusion. 2020. Vol. 58. P. 82–115. ZLIGFR.

[Bel22] Bellucci M., Delestre N., Malandain N. et al. Combining an explainable model based on ontologies with an explanation interface to classify images // Procedia Computer Science. 2022. Vol. 207. P. 2395–2403. IJVJIV.

[Bob11] Bobillo F., Straccia U. Fuzzy ontology representation using OWL 2 // International Journal of Approximate Reasoning. 2011. Vol. 52, № 7. P. 1073–1094. 10.1016/j.ijar.2011.05.003.

[Bob12] Bobillo F., Delgado M., Gómez-Romero J. DeLorean: A reasoner for fuzzy OWL 2 // Expert Systems with Applications. 2012. Vol. 39, № 1. P. 258–272. 10.1016/j.eswa.2011.07.016.

[Hof19] Hoffman R.R., Mueller S.T., Klein G. et al. Metrics for Explainable AI: Challenges and Prospects // arXiv. 2019. P. 1–50.

[Hof94] Hofmann H. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. 10.48550/arXiv.1812.04608.

[Kos24] Kosov P., El Kadhi N., Zanni-Merk C., Gardashova L. Semantic-Based XAI: Leveraging Ontology Properties to Enhance Explainability // Proceedings of the 2024 International Conference on Decision Aid Sciences and Applications (DASA). Manama, 2024. P. 1–5. 10.1109/DASA63652.2024.10836289.

[Kos25] Kosov P., El Kadhi N., Zanni-Merk C., Gardashova L. Formalizing fuzzy explainability: Enhancing XAI trustworthiness with fuzzy-ontology-based explanatory properties // Procedia Computer Science. 2025. Vol. 270. P. 3768–3777. RQXZES.

[Lun17] Lundberg S.M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, 2017. P. 4768–4777. 10.48550/arXiv.1705.07874.

[Men21] Mendel J.M., Bonissone P.P. Critical Thinking About Explainable AI (XAI) for Rule-Based Fuzzy Systems // IEEE Transactions on Fuzzy Systems. 2021. Vol. 29, № 12. P. 3579–3593. IFPMAU.

[Nau23] Nauta M., Trienes J., Pathak S. et al. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI // ACM Computing Surveys. 2023. Vol. 55, № 13s. P. 1–42. 10.1145/3583558.

[Rib16] Ribeiro M.T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier // Proc 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining. 2016. P. 1135–1144. 10.18653/v1/N16-3020.

[Rud19] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead // Nature Machine Intelligence. 2019. Vol. 1, № 5. P. 206–215. 10.1038/s42256-019-0048-x.

[See19] Seeliger A., Pfaff M., Krcmar H. Semantic web technologies for explainable machine learning models: A literature review // Proceedings of the 1st Workshop on Semantic Explainability. Auckland, 2019. P. 30–45.

[Tak85] Takagi T., Sugeno M. Fuzzy identification of systems and its applications to modeling and control // IEEE Transactions on Systems, Man, and Cybernetics. 1985. Vol. SMC-15, № 1. P. 116–132. 10.1109/TSMC.1985.6313399.

[Tid15] Tiddi I., d'Aquin M., Motta E. An Ontology Design Pattern to Define Explanations // Proceedings of the 8th International Conference on Knowledge Capture. Palisades, 2015. P. 1–8. 10.1145/2815833.2815844.

[Wex19] Wexler J., Pushkarna M., Bolukbasi T. et al. The What-If Tool: Interactive Probing of Machine Learning Models // IEEE Transactions on Visualization and Computer Graphics. 2019. Vol. 26, № 1. P. 56–65. 10.1109/TVCG.2019.2934619.

[Yan19] Yang F., Du M., Hu X. Evaluating Explanation Without Ground Truth in Interpretable Machine Learning // arXiv. 2019. P. 1–9. 10.48550/arXiv.1907.06831.

[Zad11] Zadeh L.A. A note on Z-numbers // Information Sciences. 2011. Vol. 181, № 14. P. 2923–2932. 10.1016/j.ins.2011.02.022.

[Zad65] Zadeh L.A. Fuzzy sets // Information and Control. 1965. Vol. 8, № 3. P. 338–353. 10.1016/S0019-9958(65)90241-X.

[Zho21] Zhou J., Gandomi A. H., Chen F. et al. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics // Electronics. 2021. Vol. 10, № 5. P. 1–19. FYCJNM.

ОБ АВТОРАХ | ABOUT THE AUTHORS

ГАРДАШОВА Латафат Аббас-гызы
Азербайджанский государственный университет нефти и промышленности, Азербайджан.
l.qardashova@asoiu.edu.az ORCID: 0000-0003-3227-2521.
Проф. каф. компьютерной инженерии. Дипл. прикл. математик (Азерб. гос. ун-т). Д-р техн. наук по системному анализу, управлению и обработке информации (Азерб. гос. академия нефти). Иссл. в обл. нечёткой математики, мягких вычислений, нечётких систем и систем управления.

KOSOV Pavel Igorevich
Azerbaijan State Oil and Industry University, Azerbaijan.
pavel_kosov@asoiu.edu.az ORCID: 0009-0005-5602-2086.
Teacher at the BA Programs, Master's degree of Computer Science (Baku State Uni.). Doctoral studies (Azerbaijan State Oil and Industry University). PhD in Tech. Sciences. Research: explainable AI, knowledge-based systems, fuzzy systems and ontologies.

GARDASHOVA Latafat Abbas gizi
Azerbaijan State Oil and Industry University, Azerbaijan.
l.qardashova@asoiu.edu.az ORCID: 0000-0003-3227-2521.
Prof., Dept. of Computer Engineering. Dipl. applied mathematics (Azerbaijan State Uni.). Dr. of Technical Sciences in system analysis, control and information processing (Azerbaijan State Oil Academy). Research in the fields of fuzzy mathematics, soft computing, fuzzy systems and control systems.

КОСОВ Павел Игоревич
Азербайджанский государственный университет нефти и промышленности, Азербайджан.
pavel_kosov@asoiu.edu.az ORCID: 0009-0005-5602-2086.
Учитель БА-программы. Дипл. магистр комп. наук (Бакинск. гос. ун-т). Аспирантура (Азерб. гос. ун-т нефти и промышл.). Канд. тех. наук. Иссл. в обл. объяснимого ИИ, систем на основе знаний, нечётких систем и онтологий.

МЕТАДАННЫЕ | METADATA

Заглавие: Оценка объяснений моделей «чёрного ящика» посредством интеграции метода взвешенной суммы на основе Z-чисел.

Авторы: Гардашова Л. А., Косов П. И.

Аннотация: Рассматривается задача объективного сравнения качества объяснений, которые формируют методы объяснимого искусственного интеллекта (XAI). Предложена многометодная схема оценки. Она соединяет четыре автоматически вычислимые метрики, пять человеко-ориентированных критериев и процедуру интеграции в единый ранг. Верность объяснения вычисляется как апостериорная байесовская вероятность его согласованности с решением модели. Информативность оценивается взвешенным нечётким средним степеней принадлежности семантических свойств. Устойчивость измеряется коэффициентом Жаккара и Евклидовым расстоянием. Частные оценки агрегируются моделью взвешенной суммы на основе Z-чисел. Веса критериев вычисляются методом собственного вектора Z-значной матрицы парных сравнений. Компонента надёжности Z-числа отражает достоверность каждого измерения и снижает вклад вырожденных оценок. Схема апробирована при сравнении восьми подходов XAI на наборе данных German Credit Risk. В сравнение вошли SHAP, LIME, контрфактические объяснения, нечёткие правила, нечёткие деревья решений и три онтологических подхода. Количественное и человеко-ориентированное ранжирования согласованы. Ранговая корреляция Спирмена между ними составила 0,86. Наибольший интегральный ранг в обоих контурах получил нечёткий семантический подход, построенный на Fuzzy OWL2. Результат демонстрирует разрешающую способность схемы и её применимость к произвольным методам XAI.

Ключевые слова: Объяснимый искусственный интеллект; оценка качества объяснений; нечёткая онтология; Fuzzy OWL2; Z-числа; многокритериальный анализ решений; модель взвешенной суммы.

Язык: Русский.

Статья поступила в редакцию 6 июля 2026 г.

Title: Evaluation of black-box model explanations through the integration of a weighted sum method based on Z-numbers.

Authors: Gardashova L. A., Kosov P. I.

Abstract: The problem of objective comparison of the quality of explanations produced by explainable artificial intelligence (XAI) methods is considered. A multi-method evaluation scheme is proposed. It combines four automatically computable metrics, five human-centered criteria, and an integration procedure into a single rank. The fidelity of an explanation is computed as the posterior Bayesian probability of its consistency with the model decision. In-formativeness is assessed by a weighted fuzzy mean of the membership degrees of semantic properties. Stability is measured by the Jaccard coefficient and the Euclidean distance. Partial scores are aggregated by a Z-number-based weighted sum mod-el. Criteria weights are computed by the eigenvector method applied to a Z-valued pairwise comparison matrix. The reliability component of a Z-number reflects the credibility of each measurement and reduces the contribution of degenerate scores. The scheme is validated by comparing eight XAI approaches on the German Credit Risk dataset. The comparison includes SHAP, LIME, counterfactual explanations, fuzzy rules, fuzzy decision trees, and three ontology-based approaches. The quantitative and human-centered rankings are consistent. The Spearman rank correlation between them is 0.86. In both loops, the highest integral rank was obtained by the fuzzy semantic approach built on Fuzzy OWL2. The result demonstrates the discriminative power of the scheme and its applicability to arbitrary XAI methods.

Key words: Explainable artificial intelligence; evaluation of explanation quality; fuzzy ontology; Fuzzy OWL2; Z-numbers; multi-criteria decision analysis; weighted sum model.

Language: Russian.

The article was received by the editors on 6 July 2026.