

УДК 519.841

doi: 10.15622/rcai.2025.102

ИНТЕРПРЕТИРУЕМЫЕ МОДЕЛИ-ЗАМЕСТИТЕЛИ С TREESHAP-АНАЛИЗОМ ДЛЯ СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ В ОБЛАСТИ ПОЖАРНОЙ БЕЗОПАСНОСТИ

Р.Ш. Хабибулин (*kh-r@yandex.ru*)

Академия ГПС МЧС России, Москва

В данной работе представлен формализованный подход к созданию и валидации моделей-заместителей машинного обучения для аппроксимации байесовских вычислений в задачах ранжирования комплексов противопожарных мероприятий. Предложенный подход основан на использовании древовидных алгоритмов (*Random Forest*, *XGBoost*, *LightGBM*) в сочетании с методом *TreeSHAP* для обеспечения интерпретируемости управленческих решений. На примере объектов хранения нефти и нефтепродуктов проведено экспериментальное исследование на синтетическом датасете из 2000 сценариев. Разработано программное обеспечение, включающее в себя комплексную систему валидации с метриками корреляции ранжирования и анализом влияния на принятие решений.

Ключевые слова: *TreeSHAP*, пожарная безопасность, системы поддержки принятия решений, интерпретируемое машинное обучение.

Введение

Принятие решений по выбору противопожарных мероприятий на различных объектах защиты представляет собой комплексную многокритериальную задачу, осложненную противоречиями между требованиями к точности, скорости, стоимости и объяснимости управленческих решений. Лица, принимающие решения (ЛПР), должны ранжировать альтернативные комплексы противопожарных мероприятий по многим критериям, включая такие важные как стоимость, время реализации и эффективность снижения пожарного риска в условиях возможной неопределенности исходных данных и субъективности экспертных предпочтений [Бурков, 2010]. Существующие широко известные методы многокритериального анализа (*TOPSIS* (ранжирование альтернатив на основе их близости к иде-

альному решению), *ELECTRE* (оценивает отношение превосходства одной альтернативы над другой), Метод анализа иерархий) в целом не предназначены для оценки интервальной неопределенности исходных данных и вариативности весов критериев, что может привести к принятию субоптимальных решений [Андрейчук, 2022]. Байесовские методы стохастического доминирования, хотя и обеспечивают математически строгий учет неопределенности через распределения Дирихле и итерации Монте-Карло, требуют тысяч вычислений для получения одного ранжирования комплекса противопожарных мероприятий, что делает их трудно применимыми в интерактивных системах поддержки принятия решений (СППР) [Агасиев и др., 2024]. Современные алгоритмы машинного обучения способны обеспечить требуемую скорость вычислений, однако их непрозрачность ("черный ящик") не удовлетворяет существующим требованиям к обоснованности принимаемых решений в области обеспечения пожарной безопасности [Кузнецова и др., 2018].

Таким образом, ключевая проблема заключается в отсутствии формализованного подхода, который одновременно обеспечивал бы математическую строгость байесовского подхода, вычислительную эффективность машинного обучения и полную интерпретируемость результатов для ЛППР. Таким образом, цель данной работы состоит в разработке и валидации моделей-заместителей (суррогатных моделей) машинного обучения с *SHAP*-интерпретацией (метод интерпретации моделей машинного обучения, основанный на теории значений Шепли из кооперативной теории игр) [Mitchell et al., 2022], способных аппроксимировать байесовские вычисления стохастического доминирования при сохранении точности ранжирования и обеспечении объяснений каждого управленческого решения. В работе применяется метод *TreeSHAP* (оптимизированный алгоритм вычисления значений *SHAP* для ансамблей деревьев) для объяснения решений в задачах управления противопожарными мероприятиями в рамках риск-ориентированного подхода и разработке комплексной системы валидации моделей-заместителей с учетом специфики ранжирования альтернатив.

1. Краткий обзор подходов и методов

Модели-заместители как метод аппроксимации вычислительно затратных процессов получили широкое распространение во многих предметных областях: от замены *CFD*-симуляций в аэродинамике до аппроксимации конечноэлементного анализа в конструкторском проектировании [Хведчук и др., 2018]. Однако применение таких моделей в задачах управления пожарной безопасностью остается малоизученным направлением, что объясняется специфическими требованиями к надежности и интерпретируемости решений в критически важных системах обеспечения безопасности. Параллельно развитие объяснимого искусственного

интеллекта привело к созданию универсальных методов интерпретации машинного обучения, таких как локальные линейные аппроксимации *LIME* (метод локальной интерпретации, не зависящий от типа модели) [Волков и др., 2023], теоретико-игровой подход *SHAP* на основе значений Шепли [Воробьев, 2021], и его реализация *TreeSHAP* для древовидных моделей [Deb et al., 2021]. *TreeSHAP* обеспечивает точные объяснения за полиномиальное время $O(TLD)$, где T - количество деревьев, L - листьев, D - глубина, что критично для интерактивных систем поддержки принятия решений.

Анализ классов алгоритмов машинного обучения для рассматриваемых задач моделирования в контексте поддержки управления пожарной безопасностью выявил различия в их применимости. Древовидные алгоритмы (*Random Forest*, *XGBoost*, *LightGBM*) демонстрируют оптимальное сочетание характеристик: нативную поддержку метода *TreeSHAP* без дополнительных вычислительных затрат, обработку нелинейных взаимодействий между интервальными признаками, достаточно высокую робастность к выбросам в данных экспертных оценок, быстрое действие как при обучении, так и при прогнозировании, и практическую интерпретируемость результатов. Нейронные сети, несмотря на высокую точность аппроксимации сложных нелинейных зависимостей, чувствительны к гиперпараметрам, имеют склонность к переобучению. Линейные модели, хотя и обладают высокой скоростью и естественной интерпретируемостью, не позволяют адекватно моделировать сложные взаимодействия между интервальными границами критериев и байесовскими весами, что приводит к существенной потере точности аппроксимации. Методы опорных векторов (*SVM*) демонстрируют хорошую обработку нелинейности и робастность, однако требуют применения модель-независимой версии *Kernel SHAP* для интерпретации, что увеличивает вычислительную сложность. Таким образом, древовидные алгоритмы представляют собой оптимальный выбор для моделирования в задачах ранжирования комплексов противопожарных мероприятий, обеспечивая необходимое сочетание точности, скорости и объяснимости решений.

2. Методология исследования

Исходная задача формализуется как байесовская модель стохастического доминирования для N альтернативных комплексов противопожарных мероприятий, где каждый комплекс i характеризуется интервальными оценками критериев для $k = 1, 2, 3$ (стоимость выполнения противопожарных мероприятий, время выполнения противопожарных мероприятий, дефицит пожарного риска на объекте защиты соответственно) и априорными параметрами распределения Дирихле $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ для весов критериев. Полная байесовская модель выполняет M итераций ме-

тодом Монте-Карло, генерируя на каждом шаге m случайные значения критериев \mathbf{c}_i и веса w_i , далее вычисляется нормированная взвешенная полезность каждой меры:

где $\frac{c_i}{\sum c_i}$ — нормированные значения критериев, $c_{i,m}$ — k -й компонент весового вектора в итерации m , U_i — полезность меры i в итерации m . Вероятность стохастического доминирования меры i над мерой j оценивается как:

где $I(\cdot)$ — индикаторная функция, M — количество итераций Монте-Карло, p_{ij} — вероятность доминирования. Интегральный балл меры для ранжирования вычисляется как $\sum U_i$, однако вычислительная сложность $O(M \cdot N^2)$ делает метод затруднительным для интерактивных систем поддержки принятия решений.

Модель-заместитель аппроксимирует байесовские вычисления через регрессионное отображение \hat{U}_i где вектор признаков \mathbf{x}_i формируется объединением интервальных границ и ожидаемых весов:

где \mathbf{c}_i^{\min} — нижняя и верхняя границы k -го критерия для комплекса противопожарных мероприятий i , w_k — ожидаемый вес k -го критерия при распределении Дирихле, $d = 9$ — размерность пространства признаков. Аппроксимация имеет вид:

где \hat{U}_i — предсказанный интегральный балл комплекса противопожарных мероприятий i , θ — параметры машинной модели, оптимизируемые минимизацией функции потерь $L(\hat{U}_i, U_i)$ между истинными и предсказанными баллами.

Процесс обучения включает генерацию сценариев с различными комбинациями интервалов и весов, запуск полной байесовской модели для получения истинных баллов, и обучение моделей-заместителей различных типов (*Random Forest*, *XGBoost*, *LightGBM*) с последующей валидацией. Система валидации учитывает специфику задач ранжирования через метрики корреляции рангов, где коэффициент корреляции Спирмена оценивает сохранение порядка альтернатив:

где r_i – разность рангов комплекса противопожарных мероприятий i в истинном и предсказанном ранжированиях, N – количество комплексов противопожарных мероприятий. Дополнительно вычисляется точность совпадения топ- K рекомендаций как $\frac{1}{K} \sum_{i=1}^K \mathbb{1}(i \in S)$, где S возвращает множество индексов K лучших комплексов противопожарных мероприятий.

Интерпретация решений модели-заместителя обеспечивается методом *TreeSHAP*, который для древовидных алгоритмов вычисляет значения Шепли для признака p комплекса противопожарных мероприятий i согласно формуле:

где F – множество всех признаков, S – подмножество признаков без p , ϕ_S – вектор признаков комплекса противопожарных мероприятий i с нулевыми значениями для признаков не из S , ϕ_p – вклад признака p в предсказание для комплекса противопожарных мероприятий i . Аддитивность *SHAP* гарантирует равенство $\phi_p = F - \phi_{S \cup \{p\}}$, где ϕ_0 – базовое значение модели, что обеспечивает полное объяснение каждого управленческого решения через распределение ответственности между влияющими факторами.

Основные этапы алгоритма для разработанной модели:

1. Ввод параметров объекта: тип объекта защиты

2. Установка ограничений (интервалы): максимальный бюджет, максимальное время реализации, минимальное требуемое снижение риска

3. Формирование альтернатив

Автоматическая генерация множества из 6 типовых комплексов противопожарных мероприятий

Загрузка характеристик из базы данных противопожарных мероприятий

Каждый комплекс мероприятий характеризуется интервальными оценками

4. Настройка критериев

Установка весов критериев (стоимость, время, риск)

Генерация параметров распределения Дирихле для учета неопределенности предпочтений

5. Выбор вычислительного метода

Выбор из 2 режимов:

Точный режим:

Байесовская модель стохастического доминирования

Время выполнения: ~52 мс

Максимальная точность

Быстрый режим:

Модель-заместитель XGBoost

Время выполнения: ~2.8 мс

Точность $R^2 = 0,9687$

6. Вычисление и ранжирование

Расчет интегральных баллов

Сортировка комплексов противопожарных мероприятий по убыванию баллов

Определение 3 лучших рекомендуемых решений

7. Интерпретация решений

TreeSHAP анализ для модели-заместителя

Вычисление SHAP значений для каждого фактора

Определение вклада каждого фактора в итоговое решение

8. Формирование результатов

Отчет включает:

Ранжированный список комплексов противопожарных мероприятий с баллами

3 лучших рекомендуемых решений с обоснованием

Визуализация (вортерфолл диаграммы, радарные графики)

Детальные объяснения через SHAP значения

9. Сохранение и итерация

Опциональное сохранение результатов в базу данных решений (база знаний)

Возможность проведения нового анализа с измененными параметрами

Накопление истории решений для дальнейшего анализа

3. Разработка программного обеспечения и проведение компьютерного моделирования

Экспериментальная валидация моделей-заместителей проводилась на синтетическом датасете из 2000 сценариев, где каждый сценарий включал $N = 15$ альтернативных комплексов противопожарных мероприятий с интервальными оценками критериев и случайными параметрами распределения Дирихле $\alpha \in [0,5, 5,0]$. Полная байесовская модель выполняла $M = 5000$ итераций Монте-Карло для каждого сценария.

Для проведения экспериментов было разработано специализированное программное обеспечение на языке *Python 3.9* с использованием библиотек: *NumPy* и *SciPy* для численных вычислений и статистических операций, *Pandas* для обработки табличных данных, *Scikit-learn* для базовых алгоритмов машинного обучения и метрик валидации, *XGBoost*, *LightGBM* для градиентного бустинга, *SHAP* для интерпретации моделей, *Matplotlib* и *Seaborn* для визуализации результатов (рис. 1).

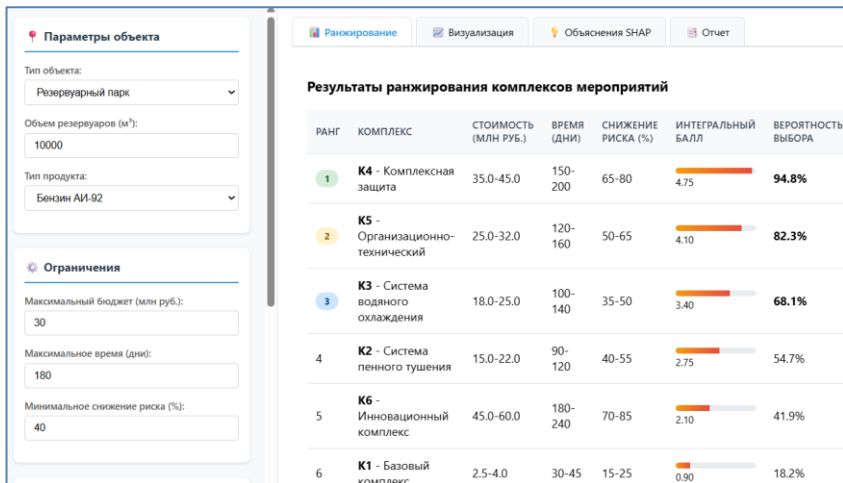


Рис. 1. Интерфейс программного обеспечения

Данные разделялись в пропорции 70%/15%/15% для обучения, валидации и тестирования соответственно. Обучались четыре типа моделей-заместителей: *XGBoost*, *Random Forest*, *LightGBM* и *Gradient Boosting* с последующей комплексной оценкой по метрикам регрессии, корреляции ранжирований и влияния на принятие решений.

XGBoost продемонстрировал превосходство по большинству метрик, достигнув коэффициента детерминации $R^2 = 0,967$ при минимальной частоте критических ошибок (0,3%), где худшая мера ошибочно выбиралась как лучшая. Корреляция Спирмена $\rho = 0,984$ указывает на высокое сохранение порядка ранжирования, что является ключевым требованием для задач принятия решений. Точность совпадения трех лучших рекомендаций составила 89%, что означает правильную идентификацию приоритетных мер в 9 случаях из 10.

TreeSHAP анализ 1000 объяснений (рис. 2) показал, что решения ЛПР в первую очередь определяются максимальным снижением риска (31,2% влияния) и минимальной стоимостью (24,7%), что соответствует принципу осторожного планирования в условиях неопределенности.

Ожидаемые веса критериев составляют 18,3% влияния, подтверждающая важность учета субъективных предпочтений экспертов в рассматриваемой модели.

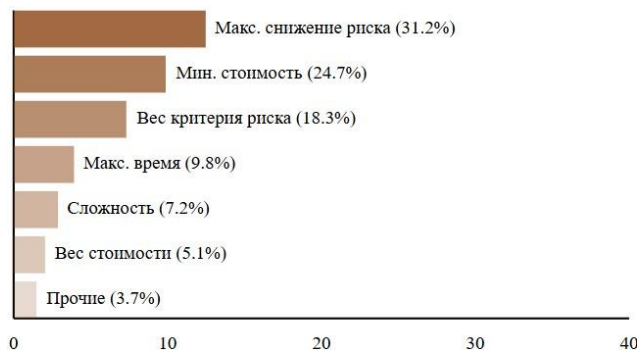


Рис. 2. SHAP-анализ ключевых факторов принятия решений

Локальные объяснения для конкретных комплексов противопожарных мероприятий (рис. 3) демонстрируют аддитивность *SHAP*-значений: например, комплекс противопожарных мероприятий №4 получает балл 2,0 как сумму базового значения (2,0) и вкладов отдельных факторов (-0,9 за экономичность, +1,5 за эффективность, -0,6 за длительность реализации), что обеспечивает полную прозрачность каждого решения для ЛПР.

Достигнутая точность ($R^2 = 0,9687$) превышает пороговое значение 0,95, рекомендуемое для критически важных приложений. Доверительный интервал [0,965, 0,972] указывает на статистическую устойчивость результата.

TreeSHAP обеспечивает полную декомпозицию каждого решения на вклады отдельных факторов. Доминирование фактора "максимальное снижение риска" (31,2%) соответствует принципу приоритета безопасности в управлении пожароопасными объектами.

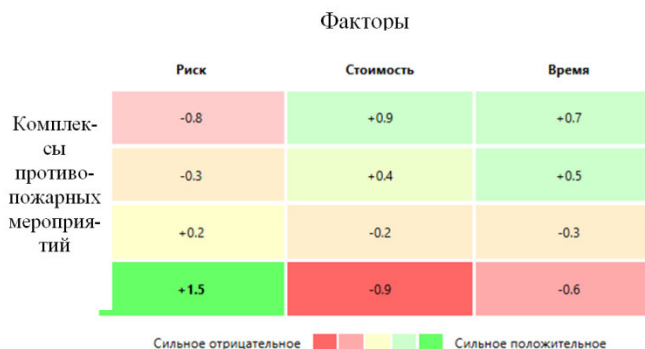


Рис. 3. Тепловая карта влияния факторов принятия решений

На основе модельных результатов разработаны практические рекомендации по выбору архитектуры моделей-заместителей: для критически важных решений в области пожарной безопасности рекомендуется модель *XGBoost* с параметрами и регуляризацией, обеспечивающими баланс точности и защиты от переобучения, тогда как для быстрого прототипирования подходит модель *Random Forest* со 100 деревьями. Установлены критерии приемлемости моделей-заместителей: коэффициент детерминации (R^2) $\geq 0,90$ для объяснения вариации, коэффициент корреляции Спирмена $\rho \geq 0,95$ для корреляции рангов, точность по тройке лучших (*Top-3 Accuracy*) $\geq 0,80$ для совпадения ключевых рекомендаций, частота критических ошибок (*Critical Error Rate*) $\leq 1\%$ для исключения катастрофических ошибок и среднюю абсолютную процентную ошибку (*MAPE*) $\leq 15\%$ для приемлемой точности.

Предложенный формализованный подход решает поставленную задачу создания быстрых, точных и интерпретируемых моделей для поддержки принятия решений в области пожарной безопасности. Достигнутое ускорение вычислений в 18–20 раз при сохранении точности выше 94% для практически значимых метрик делает подход применимым для создания интерактивных СППР, способных работать в режиме реального времени.

4. Заключение и выводы по результатам работы

Результаты проведенной работы направлены на решение противоречия между требованиями к математической строгости, вычислительной эффективности и интерпретируемости в задачах принятия решений по выбору комплекса мероприятий пожарной безопасности (на примере объектов хранения нефти и нефтепродуктов).

Экспериментальная валидация на 2000 сценариях подтверждает стабильность и робастность предложенного подхода с вариацией точности менее 1% при кросс-валидации, что обеспечивает надежность применения в реальных условиях рассматриваемых объектов защиты.

Практическая значимость исследования подтверждается созданием программного прототипа, демонстрирующего возможность интеграции практико-ориентированных технологий принятия решений в критически важные процессы управления пожарной безопасностью.

Перспективным направлением развития проводимого исследования являются методологические расширения: разработка мультифидельностных подходов с иерархией моделей различной точности (быстрый *Random Forest* для первоначальной оценки \rightarrow точный *XGBoost* для финального ранжирования \rightarrow полная байесовская модель для решений с наибольшей точностью). Дальнейшие исследования будут направлены на валидацию на реальных данных, рассмотрение комплексов противопожарных мероприятий на другие объекты защиты и интеграцию с системами автоматизированными СППР.

Список литературы

- [Агасиев и др., 2024] Агасиев Т.А., Карпенко А.П. Байесовская оптимизация с прогнозированием наилучших значений гиперпараметров суррогатной модели // Системы компьютерной математики и их приложения. – 2024. – № 25. – С. 62-67.
- [Андрейчук, 2022] Андрейчук А.А. Эффективный поиск ограниченно-субоптимальных решений задачи многоагентного планирования // Искусственный интеллект и принятие решений. – 2022. – № 1. – С. 57-70.
- [Бурков, 2010] Бурков Е.А. Определение субъективности и надежности экспертных оценок на основе анализа статистических данных // Известия СПбГЭТУ ЛЭТИ. – 2010. – № 9. – С. 33-38.
- [Волков и др., 2023] Волков Е.Н., Аверкин А.Н. Возможности применения объяснительного искусственного интеллекта для обнаружения глаукомы на примере метода LIME // Международная конференция по мягким вычислениям и измерениям. – 2023. – Т. 1. – С. 177-180.
- [Воробьев, 2021] Воробьев А.В. Метод выбора модели машинного обучения на основе устойчивости предикторов с применением значения Шепли // Экономика. Информатика. – 2021. – Т. 48, № 2. – С. 350-359.
- [Кузнецова и др., 2018] Кузнецова А.В., Сенько О.В., Кузнецова Ю.О. Преодоление проблемы "черного ящика" при использовании методов машинного обучения в медицине // Врач и информационные технологии. – 2018. – № S1. – С. 74-80.
- [Хведчук и др., 2018] Хведчук В.И., Антоник И.А. Базовые элементы FEM - анализа для электрических схем // Вестник Брестского государственного технического университета. Физика, математика, информатика. – 2018. – № 5(113). – С. 48-52.
- [Deb et. al., 2021] Deb D., Smith R.M. Application of Random Forest and SHAP Tree Explainer in Exploring Spatial (In)Justice to Aid Urban Planning // ISPRS International Journal of Geo-Information. – 2021. – Vol. 10, No. 9. – P. 629.
- [Mitchell et. al., 2022] Mitchell R., Frank E., Holmes G. GPUTreeShap: massively parallel exact calculation of SHAP scores for tree ensembles // PeerJ. Computer Science. – 2022. – Vol. 8. – P. e880. – DOI 10.7717/peerj-cs.880.