На правах рукописи

ПОЛИТЕХ
Санкт-Петербургский
политехнический университет
Петра Великого

Константинов Андрей Владимирович

Разработка и исследование новых интерпретируемых моделей машинного обучения на основе композиций слабых базовых моделей

1.2.2. Математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата физико-математических наук

Санкт-Петербург
2024

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования учреждении «Санкт-Петербургский политехнический университет Петра Великого».

Научный руководитель: **Уткин Лев Владимирович,**
доктор технических наук, профессор

Официальные оппоненты: **Бухановский Александр Валерьевич,**
доктор технических наук, директор мегафакультета трансляционных информационных технологий, федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО», Санкт-Петербург

Каплун Дмитрий Ильич,
кандидат технических наук, доцент, доцент кафедры автоматики и процессов управления, федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)»

Ведущая организация: Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук»

Защита состоится «11» декабря 2024 года в 16:00 на заседании диссертационного совета У.1.2.2.03 федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого» (195251, Санкт-Петербург, Политехническая ул., 29, 2 учебный корпус, аудитория 265).

С диссертацией можно ознакомиться в библиотеке и на сайте www.spbstu.ru федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого».

Автореферат разослан «___» _____ 2024 года.

Ученый секретарь
диссертационного совета
У.1.2.2.03,
к.т.н. Зайцева Надежда Игоревна

Общая характеристика работы

Актуальность темы. В настоящее время искусственный интеллект и машинное обучение широко применяются в самых различных прикладных областях, включая производственные системы, медицину, экономику. Эти области характеризуются табличным форматом данных, малыми размерами доступных выборок, а также сложностью формализации многих прикладных задач и необходимостью интерпретации и объяснения результатов моделирования. Учёт этих особенностей в совокупности не представлен в большинстве существующих на данный момент моделей машинного обучения, что не позволяет использовать их с максимальной эффективностью в этих областях. Отсутствие интерпретируемости также не позволяет в полной мере доверять результатам предсказаний, что является важнейшим элементом производственных и медицинских систем. Следует отметить, что существует ряд моделей машинного обучения, которые нашли применение в перечисленных областях. Однако эти модели обладают рядом существенных недостатков, включая низкую обобщающую способность при обучении на данных без предварительной обработки, необходимость ручной настройки большого числа параметров, сложность обработки наборов данных высокой размерности. Такие модели как, например, глубокие нейронные сети не обладают достаточной точностью при работе с малыми выборками табличных данных высокой размерности, что делает их малоэффективными при решении новых задач. Для построения более точных и эффективных моделей с учётом необходимости интерпретации их предсказаний требуются новые подходы, позволяющие обойти перечисленные выше недостатки.

Учитывая вышесказанное тема диссертационной работы, посвященная разработке и исследованию новых интерпретируемых моделей машинного обучения на основе композиций слабых базовых моделей, представляется черезвычайно актуальной. Модификация механизма внимания, широко применяемого в современных нейронных сетях для обработки последовательностей, и комбинирование его с ансамблевыми моделями, такими как случайный лес и градиентный бустинг, является ключевым элементом связи между современными моделями обработки табличных данных и нейронными сетями. Задача эффективной реализации разрабатываемых математических моделей в виде комплексов программ также является крайне актуальной с точки зрения их применения в реальных задачах. Поэтому в диссертационной работе представлены алгоритмы и программы, реализующие предлагаемые модели. Актуальной также является задача разработки новых методов интерпретации предложенных моделей и моделей вида «чёрный ящик», что обусловило решение данной задачи в диссертационной работе.

Степень разработанности темы. Построение композиций и ансамблей базовых моделей является одним из основных направлений для создания точных моделей по табличным данным. В частности, для задач восстановления регрессии и классификации был разработан ряд эффективных алгоритмов градиентного бустинга, включая XGBoost, LightGBM, CatBoost. Тем не менее, данные алгоритмы используют в качестве базовых моделей деревья решений, построение которых осуществляется «жадными» алгоритмами, что может рассматриваться как один из факторов снижения обобщающей способности. Другое направление составляет применение моделей на основе нейронных сетей. Среди таких моделей можно выделить TabPFN, TabR, SAINT, показывающие наилучшие результаты среди нейронных сетей. Все приведённые модели основаны на применении механизма внимания в различных формах. Однако подходы, совмещающие модели на основе ансамблей и механизма внимания ранее не рассматривались, что делает исследование в направлении данного сочетания перспективным. Основными подходами в интерпретации или объяснении моделей машинного обучения являются локальная и глобальная интерпретация с помощью обобщённых аддитивных моделей и оценка вкладов Шепли методом SHAP. Наиболее интенсивно исследуемый и применяемый подход к построению обобщённых аддитивных моделей, NAM, основан на применении нейронных сетей. Основным недостатком NAM является необходимость подбора параметров для построения модели с достаточной обобщающей способностью, что затрудняет быстрое получение локальной интерпретации модели. Метод SHAP позволяет оценивать вклад каждого признака в предсказание модели таким образом, что вклады удовлетворяют определённому набору свойств, обеспечивающему справедливость распределения вкладов по признакам. Из-за этого такой подход широко применяется для интерпретации моделей вида «чёрный ящик». Основным недостатком метода SHAP является экспоненциальная зависимость его вычислительной сложности от числа признаков, однако при рассмотрении объясняемой модели как «белого ящика» или как «серого ящика» возможно устранение данного недостатка. Поэтому создание методов оценки вкладов SHAP для разрабатываемых моделей является перспективным, так же как и разработка новых алгоритмов оценки вкладов, более эффективных с вычислительной точки зрения.

Целью данной работы является разработка и исследование новых интерпретируемых моделей машинного обучения на основе композиций слабых базовых моделей, позволяющих осуществлять моделирование зависимостей на основе данных и интерпретировать полученные модели и данные.

Для достижения поставленной цели было необходимо решить следующие **задачи**:

1. Разработать подход к моделированию зависимостей на основе ансамбля крайне слабых базовых моделей. В рамках данного подхода разработать модели, алгоритмы их обучения и методы интерпретации. Исследовать теоретическую сложность моделей.
2. Разработать подход к комбинированию ансамблей деревьев решений и механизма внимания. Реализовать его для моделей случайного леса и градиентного бустинга.
3. Разработать новые методы интерпретации для объяснения моделей вида «чёрный ящик» путём построения моделей частичной зависимости от признаков и путём аппроксимации чисел Шепли.
4. Разработать комплекс программ для реализации вычислительных экспериментов с реальными данными для оценки и использования предложенных методов и алгоритмов.

Научная новизна:

1. Разработан подход к моделированию зависимостей с помощью моделей градиентного бустинга на основе многомерных параллельных осям прямоугольников. Получены теоретические оценки сложности слабых базовых моделей в виде многомерных параллельных осям прямоугольников. В отличие существующих моделей, таких как случайный лес и градиентный бустинг, предложенные модели имеют минимальную вычислительную сложность и обладают более высокой обобщающей способностью.
2. Разработаны алгоритмы оценки вкладов Шепли для предложенной модели на основе многомерных параллельных осям прямоугольников для интерпретации модели. Доказана корректность разработанных алгоритмов. В отличие от существующих алгоритмов семейства SHAP, основанных на оценке вкладов Шепли, предложенные алгоритмы позволяют получить точные оценки вкладов для данных любой размерности. Один из алгоритмов имеет линейную сложность, в отличие от экспоненциальной сложности алгоритмов семейства SHAP.
3. Впервые разработан подход к комбинированию ансамблей деревьев решений и механизма внимания. В рамках подхода предложен класс моделей, где механизм внимания реализуется на различных уровнях деревьев решений случайных лесов и градиентного бустинга, с использованием интервальных моделей распределений вероятностей, а также нейронных сетей. В отличие от аналогичных моделей без механизма внимания предложенный подход позволил существенно повысить точность предсказаний.

4. Разработан новый метод интерпретации моделей вида «чёрный ящик» и экспериментальных данных на основе обобщённой аддитивной модели, реализуемой взвешенным ансамблем моделей градиентного бустинга. Метод отличается от существующего подхода на основе нейронных сетей возможностью получения точной интерпретации для малых обучающих выборок табличных данных.
5. Разработан новый метод оценки вкладов Шепли на основе механизма случайных подвыборок для интерпретации моделей вида «чёрный ящик», называемый Random SHAP, который в отличие от известных подходов на основе метода SHAP позволяет получить сколь угодно точные оценки для данных большой размерности.

Теоретическая значимость диссертационной работы определяется фундаментальностью и новизной полученных результатов, которые существенно расширяют область машинного обучения и искусственного интеллекта. Часть результатов является основой для новых классов моделей машинного обучения, связанных с объяснением и интерпретацией результатов моделирования. Часть результатов формирует новое направление в искусственном интеллекте, связанное с объединением современного аппарата механизма внимания с ансамблевыми моделями.

Практическая значимость. С практической точки зрения, разработанные новые методы и алгоритмы, реализованные в виде программных пакетов, могут быть использованы в таких прикладных областях как медицина, производственные системы и экономика. Они также могут использоваться для комплексов программ, ориентированных на другие области применения, в качестве подсистем.

Методология и методы исследования. Проведённые в работе исследования основываются на методах машинного обучения и анализа данных, теории статистического обучения, методах оптимизации, теории вероятностей и математического моделирования. Для количественного сравнения предложенных и существующих методов, алгоритмов и моделей проводились численные эксперименты на основе разработанных комплексов программ.

Основные положения, выносимые на защиту:

1. Подход к моделированию зависимостей с помощью моделей градиентного бустинга на основе многомерных параллельных осям прямоугольников. Получены теоретические оценки сложности слабых базовых моделей.
2. Алгоритмы оценки вкладов Шепли для предложенной модели на основе многомерных параллельных осям прямоугольников для интерпретации модели. Доказана корректность разработанных алгоритмов.

3. Подход к комбинированию ансамблей деревьев решений и механизма внимания. В рамках подхода предложен класс моделей, где механизм внимания реализуется на различных уровнях деревьев решений случайных лесов и градиентного бустинга, с использованием интервальных моделей распределений вероятностей, а также нейронных сетей.
4. Новые методы интерпретации моделей вида «чёрный ящик» и экспериментальных данных на основе обобщённой аддитивной модели, реализуемой взвешенным ансамблем моделей градиентного бустинга, и на основе оценки вкладов Шепли с помощью метода случайных подпространств.
5. Комплекс программ, реализующих разработанные методы и модели.
6. Вычислительные эксперименты с синтетическими и реальными данными, подтверждающие эффективность разработанных методов и алгоритмов.

Достоверность полученных результатов обеспечивается численными экспериментами, опубликованными в известных научных изданиях. Результаты находятся в соответствии с результатами, полученными другими авторами.

Апробация работы. Основные результаты работы, включая разработанные методы, алгоритмы и результаты численных экспериментов докладывались следующих международных научных конференциях:

- 28th Conference of Open Innovations Association (FRUCT), 25–29 января 2021 г., Москва;
- The 2nd International Conference on Cyber-Physical Systems and Control CPS&C 2021, 29 июня – 2 июля 2021 г., Санкт-Петербург;
- 31th Conference of Open Innovations Association (FRUCT), 27–29 апреля 2022 г., Хельсинки, Финляндия;
- XXXV Международная научная конференция «Математические Методы в Технике и Технологиях - ММТТ-35», 20 мая – 4 июня 2022 г., Ярославль;
- 11th International Young Scientist Conference on Computational Science, 12–17 сентября 2022 г., Санкт-Петербург;
- 8th International Conference on Interactive Collaborative Robotics, 25–29 октября 2023 г., Баку, Азербайджан;
- Международная научная мультиконференция: «Математические методы в технике и технологиях. ММТТ-36», 30 мая - 1 июня 2023 г., Нижний Новгород;
- XVI Всероссийская мультиконференция по проблемам управления (МКПУ-2023), 11–15 сентября 2023 г., Волгоград;

– 7th International Scientific Conference “Intelligent Information Technologies For Industry”, 25–30 сентября 2023 г., Санкт-Петербург;
– XIV Всероссийское совещание по проблемам управления (ВСПУ-2024), 17–20 июня 2024 г.

Результаты диссертации частично получены при выполнении следующих научно-исследовательских работ:

– Разработка новых моделей машинного обучения в задачах прогнозирования, интерпретации и объяснения результатов диагностики онкологических заболеваний (РНФ, 21-11-00116, 2020–2023 г.);
– Создание персонализированных методов оценки здоровья и риска онкологических заболеваний на основе интеллектуальной обработки больших массивов данных мультимодальной лучевой диагностики (РФФИ, 19-29-01004, 2019–2021 г.);
– Методы и алгоритмы интерпретации моделей машинного обучения и объяснительного интеллекта в анализе цензурированных данных и оценке эффекта воздействия (РФФИ, 20-01-00154, 2020–2021 г.);
– Фреймворк для разработки и применения алгоритмов машинного обучения на основе глубоких лесов (Deep Forest) (Код ИИ, 18ГУКодИИС12-D7/76723, 2022–2023г.);
– Разработка мультиагентного диспетчера управления ресурсами гетерогенной суперкомпьютерной платформой с использованием методов машинного обучения и искусственного интеллекта (Гос. задание FSEG-2022-0001, 2022–2023 г.).

Личный вклад автора заключается в непосредственном участии в постановке целей и задач исследований, разработке новых подходов, методов, алгоритмов, а также реализации комплексов программ. Все представленные экспериментальные и теоретические результаты получены автором. А.В. Константинов принимал непосредственное участие при написании научных статей и подготовке докладов, лично принимал участие в выступлениях на научных конференциях.

Публикации. Основные результаты по теме диссертации изложены в 29 печатных изданиях, 2 из которых изданы в журналах, рекомендованных ВАК, 16 — в периодических научных журналах, индексируемых Web of Science и Scopus, 11 — в тезисах докладов. Зарегистрированы 3 программы для ЭВМ.

Содержание работы

Во **введении** приведено обоснование актуальности темы исследований и степень её разработанности; формулируются цели и задачи исследования; описывается новизна, теоретическая и практическая значимость полученных результатов. В автореферате приведена только часть

всех лемм и теорем, представленных и доказанных в диссертационной работе.

В **первой главе**, «Задачи и основные элементы современных моделей машинного обучения», приведены основные понятия и элементы машинного обучения. Рассматривается задача восстановления зависимости на основе данных, функционал риска и эмпирический функционал риска. Даётся определение алгоритма построения модели машинного обучения, как алгоритма сопоставляющего обучающей выборке и набору гиперпараметров модель. Приводятся понятия ошибки обобщения, репрезентативности выборки и сложности Радемахера, теорема об ограничении ошибки обобщения сверху. Рассматриваются методы машинного обучения на основе ансамбирования, включая бэггинг, бустинг и стекинг; элементы современных моделей и ансамблей: деревья решений, экстремально рандомизированные деревья и механизм внимания (МВ). Описаны основные определения задачи интерпретации моделей машинного обучения, метод на основе линейного приближения функции LIME, методы на основе обобщённой аддитивной модели, метод оценки вкладов признаков на основе подхода из теории коалиционных игр SHAP.

Вторая глава, «Новые методы построения ансамблей на основе наиболее слабых базовых моделей», посвящена проблеме разработки наиболее слабых базовых моделей и адаптации к ним модели градиентного бустинга (МГБ) с целью получения модели, обладающей высокой обобщающей способностью. Классической базовой моделью в МГБ является дерево решений. Показано, что наиболее слабое дерево решений (имеющее глубину 1) не может быть использовано для построения моделей зависимостей от нескольких признаков, не представимых в виде суммы функций одной переменной. В качестве наиболее слабой базовой модели предлагается многомерный параллельный осям прямоугольник (МПП):

$$A_{r,v}(x) = \mathbb{I}\left[p_r(x)\right] v_{in} + \left(1 - \mathbb{I}\left[p_r(x)\right]\right) v_{out} = \begin{cases} v_{in}, & p_r(x) \\ v_{out}, & \neg p_r(x), \end{cases}$$

$$p_r(x) = \bigwedge_{k=1}^{d} \left(r_{left}^{(k)} \leq x^{(k)} \leq r_{right}^{(k)} \right),$$

где $\mathbb{I}$ – индикаторная функция, d – число признаков, $x^{(k)}$ – значение k-го признака, v_{in}, v_{out} – внутренее и внешнее значения соответственно, $r_{left}^{(k)} \in [-\infty; \infty)$, $r_{right}^{(k)} \in (\text{-}\infty, \infty]$ – левые и правые границы прямоугольника, соответственно, для которых верно $r_{left}^{(k)} \leq r_{right}^{(k)}$. Рассматриваются два частных случая МПП: d-мерный угол (У-МПП) – для каждого признака только одна из границ прямоугольника конечна; замкнутый МПП (З-МПП) – все границы конечны. Классы функций для данных моделей обозначаются $\mathcal{F}_C$ и $\mathcal{F}_R$, соответственно.

Для получения оценок сложности Радемахера полученных моделей определяются верхние границы мощности $\mathcal{B}(\mathcal{D})$ – множества векторов результатов применения всевозможных прямоугольников в каждой точке обучающей выборки $\mathcal{D}$:

$$\mathcal{B}(\mathcal{D}) = \left\{ \left\langle \mathbb{I}\left[p_r\left(x_i\right)\right] \right\rangle_{i=1}^{N} \mid r \in \mathcal{R} \right\}.$$

Лемма 1. Для произвольного обучающего множества $\mathcal{D}$ справедливы оценки сверху:

$$|\mathcal{B}_C(\mathcal{D})| \leq 2^d C_{N+d-1}^{N-1}, \qquad |\mathcal{B}_R(\mathcal{D})| \leq C_{N+2d-1}^{N-1},$$

где $\mathcal{B}_C$ и $\mathcal{B}_R$ соответствуют У-МПП и З-МПП. С помощью оценок из леммы 1 доказывается следующая теорема.

Теорема 1. Для $\delta > 0$, ограниченной 1-липшицевой функции потерь $|l(\hat{y}, y)| \leq c$, с вероятностью не менее $(1 - \delta)$, для любых β-ограниченных моделей $f \in \mathcal{F}_{C,\beta}$ переобученность У-МПП ограничена сверху:

$$\mathcal{O}_{l,\mathcal{P},\mathcal{D}}[f] \leq 2\beta\sqrt{\frac{2\ln\left(2^d C_{N+d-1}^{N-1}\right)}{d}} + 4c\sqrt{\frac{2\ln\left(4/\delta\right)}{N}}.$$

Аналогично определена оценка для З-МПП. Вводится ансамбль МПП и на основании теоремы определяется ограничение абсолютного значения β для обеспечения допустимой ошибки обобщения.

В работе предложены новые алгоритмы градиентного бустинга МПП. Пусть на очередной итерации алгоритма предсказание модели для x_i равно s_i, первая производная функции потерь $g_i = \frac{\partial l(y_i,z)}{\partial z}|_{z=s_i}$, вторая производная $h_i = \frac{\partial^2 l(y_i,z)}{\partial z^2}|_{z=s_i}$. Пусть также рассматривается некоторый прямоугольник r, для которого требуется определить оптимальный вектор значений $v = (v_{in}, v_{out})^T$. Для краткости обозначим $\mathbb{I}_{in}^i = \mathbb{I}[x_i \in r]$, $G_{in} = \sum_{i=1}^N \mathbb{I}_{in}^i \cdot g_i$, $H_{in} = \sum_{i=1}^N \mathbb{I}_{in}^i \cdot h_i$, и аналогично для $\mathbb{I}_{out}^i = \mathbb{I}[x_i \notin r]$, G_{out}, H_{out}. Первый алгоритм основан на использовании первой производной функции потерь и позволяет определить вектор v, обеспечивающий ограниченность абсолютного значения модели константой β. Получаемые оптимальные значения: $v_{in} = -\beta \cdot \mathrm{sign}(G_{in})$, $v_{out} = -\beta \cdot \mathrm{sign}(G_{out})$. В основе второго алгоритма лежит разложение в ряд эпирического функционала риска до второго порядка. Выведены оптимальные значения для случая отсутствия ограничения на абсолютное значение: $v_{in} = -\frac{G_{in}}{H_{in}}$, $v_{out} = -\frac{G_{out}}{H_{out}}$. Для обеспечения ограниченности, с целью повышения обобщающей способности алгоритма, к функционалу потерь добавляется штрафной компонент, осуществляющий регуляризацию вектора значений v: $\Omega(v) = \lambda_1 \|v\|_1 + \frac{\lambda_2}{2}\|v\|_2^2$. Для определения оптимальных значений используется следующая лемма.

Лемма 2. Оптимальное значение v_{in} с учётом регуляризации Ω равно:

$$v_{in} = \begin{cases} -\frac{G_{in}+N\cdot\lambda_1}{N\cdot\lambda_2+H_{in}}, & G_{in} < -N\cdot\lambda_1, \\ 0, & |G_{in}| \le N\cdot\lambda_1, \\ -\frac{G_{in}-N\cdot\lambda_1}{N\cdot\lambda_2+H_{in}}, & G_{in} > N\cdot\lambda_1. \end{cases}$$

Аналогично для v_{out}, заменой индексов in на out.

Наряду со стандартным типом регуляризации, приведённым выше, для повышения эффективности метода вводится новый тип регуляризации вида $\Omega_h = \frac{\eta}{2}\left\|v_{in} - v_{out}\right\|_2^2$, который до сих пор не использовался в моделях машинного обучения. Для определения оптимальных значений при использовании регуляризации Ω_h вводится следующая лемма.

Лемма 3. Оптимальное значение v_{in} с учётом регуляризации Ω_h равно:

$$v_{in}(\eta) = -\frac{G_{in} + N\cdot\eta\cdot(G_{in}+G_{out})/H_{out}}{H_{in} + N\cdot\eta\cdot(H_{in}+H_{out})/H_{out}},$$

Аналогично определяется оптимальное значение v_{out}. Отметим, что при использовании такого типа регуляризации v_{in} зависит в том числе от G_{out}, H_{out}, а также верно равенство $\lim_{\eta\to+\infty} v_{in}(\eta) = -\frac{G_{in}+G_{out}}{H_{in}+H_{out}}$, что существенно отличает данный тип от регуляризации Ω. Далее в работе предлагается способ определения минимальных значений параметров регуляризации при которых будет выполняться ограничение абсолютного значения МПП параметром β. Так, например, для l_2-регуляризации минимальное значение λ_2 равно $\frac{1}{N}\max(0,\ \frac{1}{\beta}|G_{in}| - H_{in},\ \frac{1}{\beta}|G_{out}| - H_{out})$.

Для построения отдельных МПП (параметров границ прямоугольника) приводится оптимальный, жадный, полуслучайный жадный и случайный алгоритмы. Предложен алгоритм обучения ансамбля МПП, включающий следующие этапы: вычисление производных функции потерь для каждой точки обучающей выборки; генерация и определение оптимальных параметров МПП; отбор МПП, не повышающих ошибку на отложенной на данной итерации выборке; перевыделение отложенной выборки. Приведён алгоритм предсказания для полученной модели. Описан способ преобразования уже построенного ансамбля в дерево решений или в ансамбль деревьев решений, а также преобразования дерева решений в ансамбль МПП. При этом предложенный алгоритм построения ансамбля МПП позволяет строить модели с большей обобщающей способностью, чем алгоритм градиентного бустинга деревьев решений, что подтверждено численными экспериментами.

Разработаны методы интерпретации ансамблей МПП, основанные на вычислении вклада Шепли каждого признака. Рассмотрены два подхода объяснения: объяснение модели без данных и объяснение модели с данными. В первом подходе модель рассматривается как функция, без учёта

распределения данных, что позволяет рассчитать точные значения вкладов за линейное время от числа признаков. Пусть ϕ_i – вклад i-го признака при объяснении отдельной МПП некоторого входного вектора x. В работе доказано следующее утверждение.

Теорема 2. Вклад i-го признака определяется единственным образом как:

$$\phi_i = \begin{cases} 0, & x^{(i)} \in r^{(i)} \\ \frac{v_{out} - v_{in}}{\sum_{j=1}^{d} \mathbb{I}\left[x^{(j)} \notin r^{(j)}\right]}, & x^{(i)} \notin r^{(i)}. \end{cases}$$

Во втором подходе кроме модели также учитывается распределение данных, за счёт использования обучающей выборки. Для этого аналогично классическому SHAP вводится функция зависимости от подмножества признаков Ψ, сопоставляющая заданному подмножеству индексов признаков S в качестве входа выходное значение модели. Используется определение этой функции в виде $\Psi(S) = \mathbb{E}[f(X) \mid X^{(S)} = x^{(S)}]$, где f – объясняемая модель, $x^{(S)}$ – вектор признаков с индексами из множества S. В качестве оценки $\widetilde{\Psi}$ функции Ψ используется среднее по обучающей выборке. В работе доказано следующее утверждение.

Теорема 3. Вклад i-го признака определяется единственным образом как:

$$\phi_i = \begin{cases} \frac{\widetilde{\Psi}(\{1,\ldots,d\}) - \widetilde{\Psi}(\emptyset) - \sum_{i=1}^{d} \phi_i \mathbb{I}[x_i \in r_i]}{\sum_{i=1}^{d} \mathbb{I}[x_i \notin r_i]}, & x^{(i)} \notin r^{(i)} \\ \sum_{S \subset d \setminus \{i\}} \frac{|S|!(d-|S|-1)!}{d!} \left(\widetilde{\Psi}\left(S \cup \{i\}\right) - \widetilde{\Psi}(S) \right), & x^{(i)} \in r^{(i)}. \end{cases}$$

На основе приведённых теорем строятся два соответствующих алгоритма вычисления вкладов признаков для ансамбля МПП, где агрегация вкладов разных МПП ансамбля осуществляется путём сложения, что корректно, поскольку вклады Шепли по определению аддитивны.

Для подтверждения полученных теоретических результатов разработан комплекс программ и проведено множество численных экспериментов с реальными данными. Выполнено сравнение разработанных моделей МГБ-МПП с существующими ансамблевыми методами, включая случайный лес, экстремально рандомизированные деревья и градиентный бустинг. На задачах регрессии МГБ-МПП превосходит все модели на 13 из 17 наборов данных; на задачах классификации – на 12 из 15 наборах данных. Исследованы зависимости точности МГБ-МПП от параметра регуляризации β, демонстрирующие сходные оптимальные значения для У-МПП и З-МПП. Кроме того, проведено сравнение типов МПП, по результатам которого У-МПП значительно превосходит З-МПП, что согласуется с полученными теоретическими оценками. Проведено сравнение времени работы классического алгоритма SHAP и двух разработанных алгоритмов на основе данных и без данных в зависимости от числа объясняемых примеров, по результатам которого предложенный алгоритм значительно превосходит классический SHAP. Оценки вкладов SHAP, полученные классическим алгоритмом и предложенным алгоритмом на основе данных совпадают.

Третья глава, «Методы усовершенствования ансамблей деревьев решений с помощью механизма внимания», посвящена построению более точных моделей на основе уже обученных ансамблей деревьев решений путём применения механизма внимания (МВ). В главе приводится описание основного предлагаемого подхода для введения весов ансамбля с помощью МВ. Для реализации подхода предлагаются постепенно усложняемые методы и модели с различными параметризациями и соответствующими задачами оптимизации, начиная от задач линейного и квадратичного программирования и заканчивая задачей обучения нейронной сети с МВ.

Основная идея модификации ансамбля с использованием МВ заключается в формировании пар векторов *ключ–значение* (в терминах МВ), соответствующих каждому элементу ансамбля (дереву решений) для нового входного вектора признаков x. Пусть дана обучающая выборка $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, где x_i – вектор признаков, y_i – целевое значение i-го примера. На выборке $\mathcal{D}$ построен ансамбль из T деревьев решений. Рассмотрим лист k-го дерева решений, в который попал *запрос* x, и обозначим множество индексов элементов обучающей выборки, попавших в данный лист, как $\mathcal{J}_k(x)$. В качестве *ключа* предлагается использовать среднее значение векторов признаков обучающей выборки, попавших в данный лист: $A_k(x) = \frac{1}{|\mathcal{J}_k(x)|} \sum_{i \in \mathcal{J}_k(x)} x_i$. Аналогичным образом определяется *значение*: $B_k(x) = \frac{1}{|\mathcal{J}_k(x)|} \sum_{i \in \mathcal{J}_k(x)} y_i$. Отметим, что в приведённых обозначениях предсказание классического ансамбля вычисляется как $\frac{1}{T} \sum_{k=1}^T B_k(x)$. В предлагаемом подходе вместо усреднения используется механизм внимания:

$$\hat{y}(x) = \sum_{k=1}^{T} \alpha_k(x; A_1(x), \ldots, A_T(x)) \cdot B_k(x),$$

где α_k – функция, определяющая неотрицательные веса внимания, удовлетворяющие условию: $\sum_{k=1}^T \alpha_k(x; A_1(x), \ldots, A_T(x)) = 1$. Иными словами, веса внимания формируют дискретное распределение вероятностей на элементах ансамбля (деревьях решений) для нового вектора признаков x, а предсказание вычисляется как ожидаемое значение $B(x)$. Обозначим для краткости $\alpha(x) = \langle \alpha_k(x; A_1(x), \ldots, A_T(x)) \rangle_{k=1}^T$.

В работе предлагаются различные варианты определения весов внимания. Первый вариант основан на статистической модели ϵ-засорения. Веса внимания определяются как выпуклая комбинация «динамических» весов, зависящих от расстояния от *запроса* x до *ключа* $A_k(x)$, и «статического» вектора весов w, состоящего из оптимизируемых (обучаемых) параметров:

$$\alpha(x) = (1 - \epsilon) \cdot \mathrm{softmax}\big(\langle -d(x, A_k(x)) \rangle_{k=1}^T\big) + \epsilon w,$$

где $d(x, A_k(x))$ – функция расстояния, которая также может иметь обучаемые параметры. В зависимости от вида функции $d(x, A_k(x))$ и наличия

в ней обучаемых параметров, формируются различные задачи обучения. Так, при задании функции $d(x, A_k(x)) = \frac{\|x - A_k(x)\|^2}{\tau}$ без оптимизируемых параметров (τ – некоторая константа), обучение предложенной модели сводится к задачам линейного или квадратичного программирования, в зависимости от выбранной функции потерь.

К другой задаче оптимизации приводит функция $d(x, A_k(x)) = \|(x - A_k(x)) \circ z\|^2 \cdot v_k$, где $\circ$ – покомпонентное произведение, z – вектор обучаемых параметров, соответствующих признакам, v_k для $k \in \overline{1,T}$ – соответствующие деревьям обучаемые параметры. Для поиска оптимальных параметров использовался алгоритм Бройдена-Флетчера-Гольдфарба-Шанно с ограниченным использованием памяти. Для учёта линейных ограничений на w, была произведена замена на функцию $w = \mathrm{softmax}(\omega)$, с новым оптимизируемым вектором ω, на который не накладываются ограничения. Также в качестве частного случая была рассмотрена задача с фиксированным $\epsilon = 0$.

На основе предложенного подхода разработана модификация механизма внимания для применения к ансамблям, построенным на основе алгоритма градиентного бустинга. Для этого, во-первых, пересчитываются значения остатков (в терминах градиентного бустинга) в листьях. Во-вторых, в функцию расстояния вводится необучаемый параметр дисконтирования $\delta \in (0, 1]$ следующим образом $d(x, A_k(x)) = \|x - A_k(x)\|^2 \delta^k$. Вследствие такого изменения понижается вклад деревьев, построенных на более поздних итерациях градиентного бустинга.

В работе формулируется общий подход к созданию новых МВ на основе крайних точек многогранника допустимых распределений вероятностей для реализации моделей, отличных от ϵ-засорения. Как и для модели ϵ-засорения, он позволяет вводить обучаемые параметры, чтобы функция весов внимания была по ним линейна. Идея подхода состоит в представлении весов в виде выпуклой комбинации M вершин некоторого выпуклого многогранника, форма которого определяется выбранной вероятностной моделью, а центр зависит от функции $d(x, A_k(x))$. В качестве обучаемых параметров выступают коэффициенты выпуклой комбинации. Итоговая модель ансамбля с МВ выглядит следующим образом:

$$\hat{y}(x) = \sum_{k=1}^{T} B_k(x) \sum_{j=1}^{M} v_j e_k^{(j)}(x),$$

где v_j для $j \in \overline{1,M}$ – обучаемые параметры, $e_k^{(j)}(x)$ – k-ая компонента j-ой вершины многогранника. В рамках предложенного подхода как частный случай рассматривается модель взаимного пари, задающая множество дискретных распределений вероятностей как:

$$\mathcal{M}(p) = \{\mu \in \Delta_T \mid \forall k \in \overline{1,T} \ \big(\mu_k \leq (1 + \theta) p_k\big)\},$$

где p – центр, $\theta > 0$ – необучаемый параметр, Δ_T – единичный симплекс. Такое множество представляет собой «перевёрнутый симплекс», и, как следствие, соответствующий многогранник может иметь переменное число крайних точек, в зависимости от положения центра p. Поскольку в рамках предложенного подхода число крайних точек должно быть постоянным, в работе предлагается ограничить центр многогранника следующим образом:

$$p_\lambda(x) = (1-\lambda)\frac{1}{T}\mathbf{1} + \lambda \cdot \operatorname{softmax}\big(\langle -d(x, A_k(x))\rangle_{k=1}^T\big),$$

где параметр λ рассчитывается с помощью следующей леммы.

Лемма 4. При фиксированном параметре модели взаимного пари $\theta > 0$ и заданном $\lambda = 1 - \frac{\theta T}{1+\theta}$, для любых x и $A_1(x), \ldots, A_T(x)$ число крайних точек $\mathcal{M}(p_\lambda(x))$ постоянно и равно T.

Другая предложенная модель на основе разработанного подхода введения весов ансамбля – многоуровневый лес с МВ, где веса присваиваются не только каждому дереву решений (листу), а каждому узлу дерева, через который проходил входной вектор x. Для этого вводятся векторы $A_{k,i}(x)$ и $B_{k,i}(x)$, соответствующие *ключам* и *значениям* для дерева k на уровне (глубине) i. Стоит отметить, что если k-е дерево решений не является полным, то для определённых векторов x узлы будут присутствовать не для всех уровней. Предложено два варианта решения данной проблемы: установить веса таких узлов равными 0, или повторять значения родительских узлов.

Следующая модель, предлагаемая в работе, рассматривает идею самовнимания для снижения влияния шума в обучающей выборке в применении к СЛ. Для этого в качестве *значений* предлагается сглаживающая оценка: $\hat{B}_k(x) = \sum_{j=1}^T \beta_{k,j}(x) \cdot B_j(x)$, где $\beta_{k,j}(x)$ называется весом самовнимания. Обучаемые параметры предложено вводить согласно общему подходу следущим образом:

$$\beta_{k,j}(x) = (1-\gamma)s_{k,j}(x) + \gamma \cdot u_j,$$

где функция $s_{k,j}(x)$ определяет близость листа дерева k к листу дерева j, $\gamma \in (0,1)$ – некоторая константа, u_j – «статический» вес дерева j. Пусть $s_k(x) = (s_{k,1}(x), \ldots, s_{k,T}(x))^T$. В работе предложено несколько вариантов определения «динамических» весов $s_k(x)$, включая:

1. Зависящие от *значений*: $s_k(x) = \operatorname{softmax}\big(\langle -\|B_k(x) - B_j(x)\|^2\rangle_{j=1}^T\big)$.
2. Зависящие от *ключей*: $s_k(x) = \operatorname{softmax}\big(\langle -\|A_k(x) - A_j(x)\|^2\rangle_{j=1}^T\big)$.

Задача определения параметров внимания w и параметров самовнимания u сводится к задаче квадратичного программирования за счёт следующей леммы.

Лемма 5. Функция предсказания $\hat{y}(x)$ линейно зависит от весов внимания и самовнимания.

Для уточнения оценок $A_k(x)$ и $B_k(x)$ в листьях вводится ещё одна новая модель случайного леса с двухступенчатым МВ, где сперва модель внимания применяется в пределах каждого листа, заменяя усреднение, использованное ранее. Пусть $\mathcal{J}_k(x) = (i_1, \ldots, i_m)$ – упорядоченный набор индексов векторов признаков обучающей выборки, попавших в тот же лист, что и x, где $m = |\mathcal{J}_k(x)|$. Уточнённое значение *ключа* определяется как: $A_k(x) = \sum_{j=1}^{m} \mu_j(x; x_{i_1}, \ldots, x_{i_m}) x_{i_j}$, где $\mu_j(x; x_{i_1}, \ldots, x_{i_m})$ – неотрицательные веса внимания, удовлетворяющие свойству: $\sum_{j=1}^{m} \mu_i(x; x_{i_1}, \ldots, x_{i_m}) = 1$. Аналогично определяется $B_k(x) = \sum_{j=1}^{m} \mu_j(x; x_{i_1}, \ldots, x_{i_m}) y_{i_j}$. Вторая модель внимания применяется на уровне леса в соответствии с общим подходом. Функция $\mu(x; x_{i_1}, \ldots, x_{i_m}) = \langle \mu_j(x; x_{i_1}, \ldots, x_{i_m}) \rangle_{j=1}^{m}$ определяется как $\mu(x; x_{i_1}, \ldots, x_{i_m}) = \mathrm{softmax}\big(\langle -\|x - x_{i_j}\|^2 \rangle_{j=1}^{m}\big)$.

В работе исследуется возможность обучения параметров ϵ и τ путём решения задачи квадратичной оптимизации. Для обучения ϵ предлагается изменить параметризацию таким образом, чтобы в качестве вектора оптимизируемых параметров использовался не w, а $\rho = \epsilon w$. Ограничения на новые переменные, в число которых входит ϵ, выглядят следующим образом: $\rho_k \geq 0$, $\sum_{k=1}^{T} \rho_k = \epsilon$, $0 \leq \epsilon \leq 1$. Вместо подбора оптимального параметра τ предложено использовать смесь моделей засорения следующего вида:

$$\alpha(x) = \sum_{i=1}^{M} (1 - \epsilon_i) \mathrm{softmax}\big(\langle -\|x - A_j(x)\|^2 / \tau_i \rangle_{j=1}^{T}\big) + \rho,$$

в которой содержится M различных значений параметра τ. Вместо τ осуществляется оптимизация параметров ϵ_i, удовлетворяющих следующим ограничениям:

$$\begin{cases} \epsilon_i \geq 0, \ i \in \overline{1, M} \\ \sum_{i=1}^{M} \epsilon_i < 1 \\ \rho_j \geq 0, \ j \in \overline{1, T} \\ \sum_{j=1}^{T} \rho_j = 1 - \sum_{i=1}^{M} \epsilon_i. \end{cases}$$

Таким образом, предложены подходы, позволяющие избежать ручного подбора параметров ϵ и τ.

Самой сложной моделью на основе предложенного подхода, с точки зрения числа параметров, является случайный лес с нейронным механизмом внимания (НМВ). Он имеет схему, аналогичную двухступенчатому МВ, и использует два различных МВ, реализуемых в виде нейронных сетей. Первый НМВ применяется в пределах каждого листа дерева решений для вычисления оценки целевого значения, а также реконструкции входного вектора. Второй НМВ используется для агрегирования промежуточных результатов, полученных на уровне каждого дерева, путём сопоставления входного вектора признаков (запроса в терминах МВ) с рассчитанными на

первой ступени реконструкциями (ключами в терминах МВ). Крайне интересно, что первый НМВ в таком случае выполняет роль «Трансформера» в том смысле, в котором данная модель применяется для моделирования естественного языка. Поскольку НМВ единый для каждого дерева леса, деревья можно рассматривать как генераторы «предложений», соответствующих новому вектору признаков x, где «словами» выступают объединённые векторы признаков и целевых значений. Целью НМВ является восстановление по входному вектору запроса x (не содержащему целевое значение) полного вектора «слова», содержащего как приближение x, так и оценку целевого значения y. Для получения финального предсказания полученные векторы «слов» агрегируются НМВ в единое предсказание. Такая трактовка позволяет использовать вместо одного слоя внимания произвольный «Трансформер», что может применяться для усложнения модели при обработке более крупных наборов данных. Тем не менее, более простой НМВ даёт возможность определять веса каждого элемента обучающей выборки, повлиявшего на предсказание, и осуществлять таким образом интерпретацию примером. Другим преимуществом разработанного подхода является отсутствие зависимости от конкретных деревьев. Так, дополнение ансамбля новыми деревьями может осуществляться без внесения изменений в обученные НМВ.

Для подтверждения результатов был разработан комплекс программ, реализующий обучение и применение предложенных моделей на основе ансамблей деревьев решений с механизмом внимания. Множество численных экспериментов продемонстрировало превосходство предложенных моделей по сравнению с классическими ансамблями деревьев решений.

Четвёртая глава, «Новые методы интерпретации на основе ансамблей», посвящена методам интерпретации моделей вида «чёрный ящик». Первый предлагаемый метод, интерпретируемый ансамбль моделей градиентного бустинга (ИА-МГБ), основан на обобщённой аддитивной модели. Новая модель, приближающая объясняемую модель вида «чёрный ящик» или зависимость по данным, реализуется в виде ансамбля взвешенных параллельных МГБ, где каждая МГБ является функцией от одного признака. Для обучения ИА-МГБ предложен итеративный алгоритм, состоящий из двух чередующихся фаз: подбора оптимальных весовых коэффициентов и уточнения всех МГБ ансамбля. Подбор оптимальных весов позволяет значительно сократить число итераций алгоритма. Для решения этой задачи используется метод Lasso, позволяющий найти разреженный вектор весов. На второй фазе одновременно уточняются все МГБ ансамбля. Для этого применяется один шаг градиентного бустинга. В работе показано, что остатки, которые необходимо приблизить при построении базовых моделей каждой МГБ ансамбля, линейно зависят от частной производной функции потерь и соответствующего веса. Таким образом, веса могут рассматриваться как адаптивные скорости обучения для каждой МГБ. Для численной

оценки «значимости» признаков производится корректировка весов, путём умножения на стандартное отклонение предсказаний МГБ. Такие оценки позволяют определять какие признаки не вносят вклад в предсказание объясняемой модели. Разработана программная реализация предлагаемого метода ИА-МГБ, что позволило провести численные эксперименты, которые продемонстрировали возможность интерпретации моделей вида «чёрный ящик», а также данных, где точность предложенной ИА-МГБ превосходит результаты существующего метода на основе нейронных сетей, а время обучения значительно меньше.

Второй метод, ансамбль случайных SHAP или Random SHAP, позволяет получить оценку вкладов Шепли для задач с большим числом признаков, когда применение классического метода SHAP затруднено ввиду его экспоненциальной вычислительной сложности. В основе предлагаемого алгоритма лежит метод случайных подпространств и предположение о том, что объсняняемая модель в виде «чёрного ящика» $f(x^{(1)}, \ldots, x^{(d)})$ реализует неаддитивные зависимости от подмножеств признаков лишь до определённого порядка. Для оценки вкладов признаков алгоритм SHAP применяется к Q отдельным вспомогательным моделям, построенным на основе исходной, и последующем агрегировании их оценок. Каждая вспомогательная модель использует случайное подмножество признаков размера S. Пусть I_j – множество индексов случайно выбранных признаков, используемых в j-ой вспомогательной модели, а исходное количество признаков равно d. Тогда j-ая вспомогательная модель является функцией от S выбранных признаков и для фиксированного вектора признаков x реализуется следующим образом: $\hat{f}_j(\langle \xi^{(i)} \rangle_{i \in I_j}) = f(\chi^{(1)}, \ldots, \chi^{(d)})$, где

$$\chi^{(i)} = \begin{cases} \xi^{(i)}, & i \in I_j \\ x^{(i)}, & i \notin I_j. \end{cases}$$

То есть если i-й признак входит в подмножество случайно выбранных признаков, будет использоваться значение соответствующего аргумента функции $\hat{f}_j$. В противном случае будет использовано значение признака фиксированного вектора x, для которого требуется оценить вклад Шепли. Для каждой вспомогательной модели предлагается рассчитать вклады Шепли с помощью классического алгоритма SHAP. Рассчитанные с помощью j-ой модели вклады обозначаются $\langle \phi_i^{(j)} \rangle_{i=1}^{d}$, где значения вкладов для признаков, от которых не зависела функция $\hat{f}_j$, то есть для всех $i \notin I_j$, далее не используются, и могут считаться равными нулю: $\phi_i^{(j)} = 0$. Для расчёта оценки вклада i-го признака исходной объясняемой модели f предложено использовать следующее выражение:

$$\phi_i = \frac{1}{n_i} \sum_{j : i \in I_j} \phi_i^{(j)},$$

где $n_i = \sum_{j:i\in I_j} 1$ – число вспомогательных моделей, зависящих от i-го признака. При $Q = 1$, $S = d$ реализуется классический алгоритм SHAP, требующий $O(2^d)$ вычислений модели f. Предложенный алгоритм требует $O(Q \cdot 2^S)$ вычислений модели f. Таким образом, при выборе Q, S, удовлетворяющих $\log(Q) + S < d$ предложенный подход более эффективен.

В целях повышения точности оценок вкладов Шепли предлагается использовать неравномерное дискретное распределение вероятностей при генерировании псевдослучайных подвыборок индексов признаков. Для этого строится дополнительная модель объяснения f, простая с вычислительной точки зрения, и позволяющая определять *значимости* признаков – случайный лес с глубокими деревьями решений. С помощью леса рассчитываются *значимости* (Feature Importance), обозначаемые $\nu_1, \ldots, \nu_d$, на основе которых рассчитываются вероятности распределния Больцмана: $p = \mathrm{softmax}(\langle -\nu_j/\tau \rangle_{k=1}^{d})$. Генерирование множеств индексов признаков $I_1, \ldots, I_Q$ осуществляется в соответствии с данным распределением вероятностей. Таким образом, в формируемые множества чаще попадают признаки с большей величиной *значимости*. В частности, удаётся избежать попадания в множества тех признаков, от которых модель не зависит.

В работе предлагается также другое обобщение предложенного метода на основе введения весов. Для этого оценка вкладов осуществляется не для одного вектора признаков, а для различных векторов в окрестности объясняемого вектора x: $z_1, \ldots, z_Q$. Для каждого z_j выбирается соответствующее множество индексов признаков I_j, как было описано ранее. Вклады предложено оценивать следующим образом:

$$\phi_i = \frac{1}{W_i} \sum_{j:i\in I_j} w_j \cdot \phi_i^{(j)},$$

где w_j – вес j-го примера, определяемый на основе гауссова ядра, $W_i = \sum_{j:i\in I_j} w_j$ – суммарный вес примеров, содержащих i-й признак. Данный метод сочетает преимущества двух подходов, ансамбля случайных SHAP и метода локальной линейной аппроксимации LIME.

Для реализации и исследования предложенных методов был разработан комплекс программ и проведены эксперименты с реальными наборами данных. Выполнено сравнение близости оценок вкладов Шепли и сравнение порядка (конкордации) значений вкладов. По результатам экспериментов предложенный подход позволяет получать за меньшее время оценки вкладов Шепли различной точности, в зависимости от выбора параметров Q, S. К наиболее точным оценкам с точки зрения индекса конкордации приводит алгоритм с предварительным заданием распределения вероятностей на признаках с помощью случайного леса.

В **заключении** приведены основные результаты работы, заключающиеся в следующем:

1. Предложен новый подход к моделированию зависимостей с помощью моделей градиентного бустинга на основе многомерных параллельных осям прямоугольников.
2. Получены теоретические оценки сложности слабых базовых моделей в виде многомерных параллельных осям прямоугольников, предложены границы для параметров регуляризации.
3. Разработаны алгоритмы оценки вкладов Шепли для предложенной модели на основе многомерных параллельных осям прямоугольников для интерпретации и доказана корректность разработанных алгоритмов.
4. Предложен подход к комбинированию ансамблей деревьев решений и механизма внимания. В рамках подхода предложен класс моделей, где механизм внимания реализуется на различных уровнях деревьев решений случайных лесов и градиентного бустинга, с использованием интервальных моделей распределений вероятностей, а также нейронных сетей.
5. Разработан новый метод интерпретации моделей вида «чёрный ящик» и экспериментальных данных на основе обобщённой аддитивной модели, реализуемой взвешенным ансамблем моделей градиентного бустинга,
6. Предложен новый метод оценки вкладов Шепли на основе механизма случайных подвыборок для интерпретации моделей вида «чёрный ящик», называемый Random SHAP.
7. Для выполнения поставленных задач разработаны соответствующие комплексы программ.
8. Численные эксперименты показали превосходство разработанных моделей и алгоритмов по сравнению с существующими.

Разработанные методы и подходы открывают множество направлений дальнейших исследований и разработки модификаций. Первое направление – разработка модификаций алгоритмов построения гиперпрямоугольников для применения к зачадам оценки эффекта лечения, в том числе в условиях цензурированности данных, а также к различным типам входных данных, таким как изображения, временные ряды и графы. Второе направление – дообучение моделей на основе предложенного подхода комбинирования ансамблей деревьев решений и механизма внимания с использованием новых поступающих обучающих данных для применения к данным с нестационарным распределением, за счёт обновления информации в листьях построенных деревьях решений, а также достраивания дополнительных деревьев решений, что становится возможным благодаря тому, что предложенный подход с использованием нейронных сетей позволяет использовать произвольное множество деревьев и примеров в листьях в качестве входа механизма внимания. Третьим направлением является разработка методов построения интервальных оценок вкладов Шепли и

функций формы на основе предложенных идей и методов, для интерпретации с учётом неточности аппроксимаций вкладов Шепли и функций формы.

Публикации автора по теме диссертации

В изданиях, входящих в международные базы цитирования Web of Science, Scopus и изданиях из списка ВАК РФ

1. *Konstantinov, A. V.* Interpretable ensembles of hyper-rectangles as base models / A. V. Konstantinov, L. V. Utkin // Neural Computing and Applications. — 2023. — Окт. — Т. 35, № 29. — С. 21771—21795.
2. *Konstantinov, A.* Multiple instance learning with trainable soft decision tree ensembles / A. Konstantinov, L. Utkin, V. Muliukha // Algorithms. — 2023. — Т. 16, № 8. — С. 358.
3. *Utkin, L. V.* Attention and self-attention in random forests / L. V. Utkin, A. V. Konstantinov, S. R. Kirpichenko // Progress in Artificial Intelligence. — 2023. — Т. 12, № 3. — С. 257—273.
4. *Konstantinov, A. V.* Interpretable machine learning with an ensemble of gradient boosting machines / A. V. Konstantinov, L. V. Utkin // Knowledge-Based Systems. — 2021. — Т. 222. — С. 106993.
5. A weighted random survival forest / L. V. Utkin [и др.] // Knowledge-Based Systems. — 2019. — Т. 177. — С. 136—144.
6. *Utkin, L. V.* Deep Forest as a framework for a new class of machine-learning models / L. V. Utkin, A. A. Meldo, A. V. Konstantinov // National Science Review. — 2019. — Т. 6, № 2. — С. 186—187.
7. *Utkin, L. V.* Attention-based random forest and contamination model / L. V. Utkin, A. V. Konstantinov // Neural Networks. — 2022. — Т. 154. — С. 346—359.
8. *Utkin, L.* Ensembles of random SHAPs / L. Utkin, A. Konstantinov // Algorithms. — 2022. — Т. 15, № 11. — С. 431.
9. A new adaptive weighted deep forest and its modifications / L. V. Utkin [и др.] // International Journal of Information Technology & Decision Making. — 2020. — Т. 19, № 04. — С. 963—986.
10. *Konstantinov, A. V.* Multi-attention multiple instance learning / A. V. Konstantinov, L. V. Utkin // Neural Computing and Applications. — 2022. — Т. 34, № 16. — С. 14029—14051.
11. *Konstantinov, A.* Heterogeneous treatment effect with trained kernels of the Nadaraya–Watson regression / A. Konstantinov, S. Kirpichenko, L. Utkin // Algorithms. — 2023. — Т. 16, № 5. — С. 226.

12. *Konstantinov, A. V.* Attention-like feature explanation for tabular data / A. V. Konstantinov, L. V. Utkin // International Journal of Data Science and Analytics. — 2023. — Т. 16, № 1. — С. 1—26.

13. *Konstantinov, A.* LARF: Two-level attention-based random forests with a mixture of contamination models / A. Konstantinov, L. Utkin, V. Muliukha // Informatics. — 2023. — Т. 10, № 2. — С. 40.

14. Improved anomaly detection by using the attention-based isolation forest / L. Utkin [и др.] // Algorithms. — 2022. — Т. 16, № 1. — С. 19.

15. *Utkin, L.* Random survival forests incorporated by the Nadaraya-Watson regression / L. Utkin, A. Konstantinov // Informatics and Automation. — 2022. — Т. 21, № 5. — С. 851—880.

16. BENK: The beran estimator with neural kernels for estimating the heterogeneous treatment effect / S. Kirpichenko [и др.] // Algorithms. — 2024. — Т. 17, № 1. — С. 40.

17. *Konstantinov, A. V.* Deep gradient boosting for regression problems / A. V. Konstantinov // Информатика, телекоммуникации и управление. — 2021. — Т. 14, № 3. — С. 7—19.

18. *Konstantinov, A.* Flexible deep forest classifier with multi-head attention / A. Konstantinov, L. Utkin, S. Kirpichenko // Информатика, телекоммуникации и управление. — 2023. — Т. 16, № 2. — С. 7—16.

Зарегистрированные программы для ЭВМ

19. *Свидетельство о гос. регистрации программы для ЭВМ.* Программа определения персонализированного эффекта лечения на основе обучаемых весов внимания в регрессии Надарая-Уотсона [Текст] / У. Л.В. [и др.] ; ФГАОУ ВО СПбПУ. — № 2022684638 ; заявл. 14.12.2022 ; опубл. 22.12.2022, 2022685218 (Рос. Федерация).

20. *Свидетельство о гос. регистрации программы для ЭВМ.* Программа автоматизированного анализа данных, предсказания и интерпретации на основе ансамблей гиперпрямоугольников [Текст] / К. А.В., У. Л.В., М. М.М. ; ФГАОУ ВО СПбПУ. — № 2023684700 ; заявл. 17.11.2022 ; опубл. 01.12.2023, 2023685938 (Рос. Федерация).

21. *Свидетельство о гос. регистрации программы для ЭВМ.* Программа классификации табличных данных на основе алгоритма глубокого леса с применением моделей внимания [Текст] / К. С.Р. [и др.] ; ФГАОУ ВО СПбПУ. — № 2023684740 ; заявл. 17.11.2022 ; опубл. 24.11.2023, 2023685294 (Рос. Федерация).

В сборниках трудов конференций

22. Neural attention forests: transformer-based forest improvement / A. V. Konstantinov [и др.] // Proceedings of the Seventh International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’23). — Cham : Springer Nature Switzerland, 2023. — С. 158—167.

23. GBMILs: Gradient boosting models for multiple instance learning / A. Konstantinov [и др.] // International Conference on Interactive Collaborative Robotics. — Springer. 2023. — С. 233—245.

24. *Utkin, L. V.* Attention-based random forests and the imprecise pari-mutual model / L. V. Utkin, A. V. Konstantinov, N. A. Politaeva // Cyber-Physical Systems Engineering and Control. — Springer, 2023. — С. 3—15.

25. A deep forest improvement by using weighted schemes / L. Utkin [и др.] // 2019 24th Conference of Open Innovations Association (FRUCT). — IEEE. 2019. — С. 451—456.

26. *Konstantinov, A.* AGBoost: Attention-based modification of gradient boosting machine / A. Konstantinov, L. Utkin, S. Kirpichenko // 2022 31st Conference of Open Innovations Association (FRUCT). — IEEE. 2022. — С. 96—101.

27. *Konstantinov, A.* Gradient boosting machine with partially randomized decision trees / A. Konstantinov, L. Utkin, V. Muliukha // 2021 28th Conference of Open Innovations Association (FRUCT). — IEEE. 2021. — С. 167—173.

28. *Konstantinov, A. V.* A generalized stacking for implementing ensembles of gradient boosting machines / A. V. Konstantinov, L. V. Utkin //. — Springer, 2021. — С. 3—16.

29. The Deep Survival Forest and Elastic-Net-Cox cascade models as extensions of the Deep Forest / L. Utkin [и др.] // Proceedings of International Scientific Conference on Telecommunications, Computing and Control: TELECCON 2019. — Springer. 2021. — С. 205—217.

30. *Utkin, L.* Modifications of SHAP for local explanation of function-valued predictions using the divergence measures / L. Utkin, A. Petrov, A. Konstantinov // Cyber-Physical Systems and Control II. — Springer, 2023. — С. 52—64.

31. *Konstantinov, A.* Multiple instance learning through explanation by using a histopathology example / A. Konstantinov, L. Utkin // 2022 31st Conference of Open Innovations Association (FRUCT). — IEEE. 2022. — С. 102—108.

32. Random forests with attentive nodes / A. V. Konstantinov [и др.] // Procedia Computer Science. Т. 212. — Elsevier, 2022. — С. 454—463.