**Министерство образования и науки Российской Федерации**
**Московский физико-технический институт (государственный университет)**

Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Выпускная квалификационная работа бакалавра

# Оценка взаимодействия признаков с помощью ансамблей консервативных деревьев

**Автор:**
Громаков Илья Алексеевич

**Научный руководитель:**
к.ф.-м.н. Ланге Андрей Михайлович


Москва 2025

**Аннотация**

Оценка взаимодействия признаков с помощью ансамблей консервативных деревьев

*Громаков Илья Алексеевич*

Оценка взаимодействия между признаками является одной из ключевых задач интерпретируемого машинного обучения. В данной работе предлагается оригинальный подход к выявлению взаимодействий между признаками у табличных данных с использованием ансамблей консервативных деревьев - модификации классических деревьев решений, обладающих специальным гиперпараметром консервативности $rit\_alpha$. Оценка взаимодействия основывается на анализе структуры решающих деревьев и построении статистик, определяющих его присутствие между признаками. Также приводятся результаты работы метода на синтетических данных и данных известных датасетов, показывающих пригодность его использования.

# Содержание

**1 Введение** **4**

**2 Постановка задачи** **6**

**3 Обзор существующих SOTA решений** **7**

**4 Теоретическая секция** **10**

**5 Исследование и построение решения задачи** **13**

**6 Вычислительные эксперименты** **22**

**7 Заключение** **29**

**Приложение** **32**

## 1 Введение

При построении моделей во многих научных и коммерческих исследованиях требуется определить, какие признаки наиболее важны и как именно они влияют на качество модели. Одним из важных аспектов такого построения является выделение пар (или групп) взаимодействующих между собой признаков. В ряде методов по определению важности признаков их рассматривают независимо друг от друга. Но для определения взаимодействия признаки следует рассматривать одновременно. Если их разделить, то эффект, который они дают вместе, невозможно будет определить или понять. Выявление взаимодействующих признаков даёт надежду на построение аддитивной структуры модели, что позволит упростить задачу, разложив её на компоненты меньшей размерности. Игнорирование влияния взаимодействия может приводить к ухудшению работы некоторых методов. Например, линейные модели или деревья малой глубины предполагают аддитивность признаков, и сильное взаимодействие каких-то из них может сильно влиять на качество таких моделей.

Кроме практической важности при построении моделей, знание о взаимодействии каких-либо признаков может быть критическим в научных областях, таких как биология, медицина, социология. Понимание того, что какие-то признаки взаимодействуют, раскрывает скрытые законы, которые впоследствии могут вести к новым научным гипотезам. Например, в биоинформатике взаимодействие генов (эпистазис) критически важно для понимания болезней. Дело в том, что для человеческой генетики важной задачей является выявление вариаций в последовательности ДНК, которые повышают или понижают устойчивость к заболеваниям. Восприимчивость к распространённым болезням зависит от нелинейных взаимодействий множества генов и факторов окружающей среды. Именно поэтому многие методы направлены на выявление именно генного взаимодействия [1][2][3][4].

Для решения задачи определения взаимодействия признаков было разработано множество методов. Прежде всего, так как взаимодействия являются важной частью статистического анализа, первые методы были параметрическими и требовали явного моделирования взаимодействий с применением мультипликативных слагаемых. Такие подходы давали возможность находить лишь неширокий круг различных взаимодействий [5][1][6].

Впоследствии начали применять более общий подход к решению подобной задачи. Эти методы основаны на построении модели машинного обучения и позволяют выявлять нелинейные зависимости в данных. Приведем здесь основные методы, которые применяются для поиска взаимодействий между признаками: нейронные сети, клеточные автоматы, ансамбли решающих деревьев и метод многомерного сокращения размерности.

Нейронные сети с оптимизацией архитектуры через эволюционные алгоритмы (GPNN) способны выявлять сложные взаимодействия между переменными. В [7] показали, что такие сети эффективно находят взаимодействия между признаками даже при низкой наследуемости ($<1\%$) и высокой

размерности задачи.

Клеточные автоматы моделируют взаимодействия между соседними клетками на решетке, где каждая клетка представляет признак (например, SNP). Автомат обновляет своё состояние по правилам, определяемым с помощью генетических алгоритмов. Но их использование ограничено применением преимущественно к дискретным признакам.

MDR группирует комбинации признаков в бинарные категории риска, упрощая задачу классификации. Он удобен в случае слабых основных эффектов и высокой степени взаимодействий (эпистазиса). Однако он страдает от комбинаторного взрыва при увеличении числа взаимодействующих признаков.

В методе ансамблирования строится множество деревьев решений, каждое из которых обучается на случайной подвыборке и случайном подмножестве признаков (RSM). Для выявления взаимодействий применяются различные методы, основанные как на структуре деревьев [8], так и на качестве предсказания самой модели [9][10].Ансамбли решающих деревьев устойчивы к переобучению и способны захватывать сложные взаимодействия без явного задания модели. В контексте аддитивных моделей, важно рассмотреть модели Random Forest [11] и Gradient Boosting [12] на CART [13].

В данной работе будет рассматриваться один из возможных методов определения взаимодействий с помощью анализа структуры консервативных решающих деревьев градиентного бустинга. Консервативные деревья - это модификация классических CART, которая вводится в данной работе. Они представляют собой решающие деревья с особым гиперпараметром $rit\_alpha$, регулирующим консервативность дерева. Под консервативностью понимается способность дерева повторно использовать для split-ов признаки, уже поучаствовавшие в разделении и обладающие наибольшей важностью. Такое свойство деревьев позволяет определять потенциально взаимодействующие признаки.

В работе будет представлен способ подбора гиперпараметра $rit\_alpha$ при помощи статистических гипотез. Также описан подход к определению взаимодействия признаков с помощью подсчета некоторой статистики, основанной на структуре консервативных решающих деревьев градиентного бустинга.

Эксперименты показывают, что метод способен определять значимые парные взаимодействия. Сильными сторонами метода являются его интерпретируемость и вычислительная доступность.

## 2 Постановка задачи

Формальная задача состоит в том, чтобы по табличным данным выяснить, существует ли взаимодействие между выбранной группой признаков, или же они не взаимодействуют:

Введем стандартные обозначения: $X$ - пространство объектов, $Y$ - множество ответов.

$f_j : X \to D_j, j = 1,...,p$ - признаки объектов (также будут обозначаться $x_j$).

Обозначим $F = \{f_j | j \in \{1, 2, ..., p\}\}$ - множество признаков объектов (считаем, что все объекты в табличных данных обладают одинаковым признаковым описанием).

Пусть $\mathcal{F} \subseteq \mathbf{F}$ - подмножество признаков, $|\mathcal{F}| \geq 2$.

Пусть $S(\mathcal{F})$ - какая-то мера взаимодействия в $\mathcal{F}$.

Требуется среди всех подмножеств $\mathcal{F}$ выделить такие, где взаимодействие между признаками значимо по сравнению с другими взаимодействиями этой размерности, а также отранжировать найденные множества.

В данной работе проверяется только сила взаимодействия между признаками относительно других взаимодействий в модели. Таким образом, требуется:

1. Для всех выбранных упорядоченных подмножеств $\mathcal{F}_m$ определить, есть у них значимое взаимодействие: $\forall k \in \{2, 3, ..., |F|\}$ найти $T_k = \{\mathcal{F} | S(\mathcal{F}) \geq S_0^k, |\mathcal{F}| = k\}$.

2. Ранжировать степень взаимодействия у различных подмножеств $\mathcal{F}$, то есть отранжировать найденные множества $T_k$.

Так как в работе используется метод, основанный на обучении некой модели, на основании которой уже и делается вывод о взаимодействии признаков, выделим некоторые критерии для этой модели и для метода в целом:

1. С помощью обученной модели можно определить, есть ли взаимодействие между какой-то парой признаков.

2. С помощью обученной модели можно определить, есть ли взаимодействие в какой-то группе признаков.

3. Модель должна уметь работать как с регрессией, так и с классификацией.

4. Метод должен быть устойчив к шуму и случайным взаимодействиям.

5. Метод должен уметь работать с числовыми и категориальными типами признаков.

6. Метод должен быть интерпретируемым.

## 3 Обзор существующих SOTA решений

Для решения задачи выявления взаимодействий между признаками по табличным данным существует множество методов, основанных как на статистических подходах к её решению, так и на построении некоторой модели, способной выявить взаимодействия.

Одним из наиболее активных направлений является использование методов, основанных на деревьях решений и их ансамблях, которые позволяют находить сложные нелинейные взаимодействия между признаками.

Приведем несколько уже существующих методов по обнаружению и оценке взаимодействий:

#### CART (Classification and Regression Trees) [13]

Это алгоритм построения решающих деревьев. Каждое дерево строится путём последовательного разбиения пространства признаков, выбирая оптимальный *сплит* по одному признаку в каждой вершине. Для определения взаимодействия используют то, как расположены признаки в разделяющих узлах этих деревьев. Знаком того, что между признаками есть взаимодействие, может являться поочередное включение рассматриваемых признаков в структуре дерева. Эта эвристика основывается на том, что при разделении данных в узловой вершине, оптимизируется специальная метрика (например Gini), направленная на уменьшение неоднородности данных. Предположим, что признаки $X_1$ и $X_2$ взаимодействуют, тогда для того, чтобы разделить данные максимально хорошо, требуются оба этих признака, что означает, что в дереве они будут, скорее всего, идти друг за другом. Но такие паттерны могут возникать и как следствие случайных вариаций (шум) или рекурсивной природы алгоритмов построения дерева. В данной работе будет предоставлено возможное решение этой проблемы. Недостатками этого метода является то, что с помощью дерева практически невозможно оценить многомерное взаимодействия, т.е. дерево подвержено переобучению при увеличении его глубины, а также деревья слабо устойчивы к шуму и шумовым признакам.

#### Random Forests (RF) и Gradient Boosting [11][12]

Эти методы строят модель - ансамбль CART-деревьев, обученных на различных подвыборках и с рандомизацией признаков при сплитах (если их много). Взаимодействие определяется с помощью деревьев, причем имеется способность захватывать многоразмерные взаимодействия за счёт ансамблирования. Имеют высокую предсказательную способность, но всё ещё подвержены тем же проблемам, что и обычные деревья, в виде появления паттернов, которые могут быть как следствием взаимодействия признаков, так и случайной природой деревьев.

**Node Harvest [14]**

Данная модификация ансамблей деревьев решений, использует отбор узлов (node selection), а не полных деревьев. Метод генерирует множество кандидатов-разбиений (узлов дерева), а затем с помощью регуляризации и отбора формирует линейную модель на их основе. Каждый узел интерпретируется как простое правило, и коэффициенты этих правил определяются через регуляризованную оптимизацию. Поскольку отбираются в основном "мелкие"и простые узлы, модель остаётся сколько нибудь интерпретируемой. Однако такая стратегия исключает возможность обнаружения взаимодействий между несколькими переменными, если они не встречаются в одном узле, то есть этот метод плохо находит многомерные взаимодействия. Более того, взаимодействия выше 2 порядка вообще практически не распознаются из-за того, что ограничена глубина деревьев и количество узлов в них.

**Forest Garrote [15]**

Данный метод является постобработкой ансамбля решающих деревьев, и он является некой аналогией методу LASSO, но применяемый к структуре RF. Сначала извлекаются признаки и взаимодействия в виде линейных комбинаций терминальных узлов (то есть путей дерева), после чего применяется L1-регуляризация (garrote), чтобы занулить неважные компоненты. Таким образом, метод позволяет не только уменьшить размер модели, но и улучшить её интерпретируемость. Однако этот подход зависит от начального ансамбля — если он не содержит интересующих взаимодействий, Forest Garrote их не обнаружит. Кроме того, метод плохо обрабатывает коррелированные взаимодействия.

**RuleFit [10]**

Этот метод представляет собой линейную модель, построенную на базе "правил извлечённых из деревьев решений. Каждое правило — это конъюнкция условий на признаки, соответствующая пути от корня к листу в дереве. После извлечения таких правил, они используются как бинарные признаки, и на них обучается Lasso-модель. Взаимодействие признаков в RuleFit формируется через правила, в которых участвуют более одного признака. Тем самым модель может явно отражать взаимодействия. Однако чувствителен к шуму: в правила могут попасть случайные корреляции. Кроме того, из-за ограниченного числа правил взаимодействия высокой степени могут быть проигнорированы.

**Iterative Random Forests (iRF) [16]**

Это метод, для устойчивого и воспроизводимого выявления взаимодействий признаков в табличных данных, направленный на выявление взаимодействий больших размерностей. Метод основан на классическом слу-

чайном лесе (Random Forest), но вводит итеративную процедуру обучения и специальную процедуру декодирования взаимодействий из путей дерева. Этот метод удовлетворяет всем критериям технического задания.

**Additive Groves [9]**

Это также метод ансамблевого обучения, специально разработанный для анализа взаимодействий между признаками на основе предсказательной силы модели. Метод основан на идее сравнения полной модели и ограниченной модели, в которой явно запрещено использовать определённые комбинации признаков. Основным компонентом Additive Groves является ансамбль решающих деревьев (grove), обучаемый по модифицированной схеме градиентного бустинга. Основным отличием от всех остальных методов является то, что здесь не используется непосредственная структура деревьев в ансамбле, а обучаются две полноценных модели, а затем оценивается разница в качестве их предсказаний.

**Специальные методы**

Также стоит упомянуть более специализированные методы, предназначенные для ген-ген взаимодействий с категориальными признаками: Logic Regression [17], Multifactor Dimensionality Reduction (MDR) [3], Bayesian Epistasis Mapping [2].

## 4 Теоретическая секция

**Определение взаимодействия**

Определение термина **взаимодействие признаков** может отличаться в различной литературе. Понятие статистического взаимодействия используется для описания нелинейных эффектов среди двух или более переменных (признаков) в функции (предсказательной модели). В этой работе будут использоваться следующие определения статистического взаимодействия:

пусть $a(x) = a(x_1, x_2, ..., x_p)$ - функция, аппроксимирующая целевую функцию, а $(x_1, x_2, ..., x_p)$ - $p$ признаков модели. Тогда, следуя нотации из [10]

**Определение:** вещественные признаки $x_i$ и $x_j$ будем называть **взаимодействующими**, если

$$E_{\mathbf{x}} \left[ \frac{\partial a(\mathbf{x})}{\partial x_i \partial x_j} \right]^2 > 0 \tag{1}$$

Для категориальных признаков (или других признаков, не являющихся числовыми), вводится эквивалентное определение:

**Определение:** признаки $x_i$ и $x_j$ будем называть **не взаимодействующими**, если $a(\mathbf{x})$ можно представить в виде суммы двух функций, одна из которых не зависит от $x_i$, а другая не зависит от $x_j$:

$$a(\mathbf{x}) = f_{\setminus x_i}(\setminus x_i) + f_{\setminus x_j}(\setminus x_j), \tag{2}$$

где $\setminus x$ означает все признаки кроме $x$.

Можно заметить, что такое определение взаимодействия эквивалентно первоначальному определению эпистазиса, данному Фишером в статье [6], а также классическому определению взаимодействия в регрессионных моделях через добавление коэффициента $\beta_{12}$ к линейной модели $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$ [16][1].

Данные определения легко обобщить на любое конечное множество взаимодействующих признаков, нужно лишь брать смешанную производную по каждому признаку из набора или же раскладывать функцию на большее количество слагаемых.

**Определение:** вещественные признаки $\mathcal{F}$ будем называть **взаимодействующими**, если

$$E_x \left[ \frac{\partial a(x)}{\prod_{x \in \mathcal{F}} \partial x} \right]^2 > 0 \tag{3}$$

**Определение:** признаки $\mathcal{F}$ будем называть **не взаимодействующими**, если $a(\mathbf{x})$ можно представить в виде суммы $N \leq |\mathcal{F}|$ функций, каждая из которых не зависит от хотя бы одной переменной $x \in \mathcal{F}$:

$$a(\mathbf{x}) = \sum_{x \in \mathcal{F}} f_{\setminus x}(\setminus x), \tag{4}$$

где $\backslash x$ означает все признаки кроме $x$ (заметим, что $f_{\backslash x}(\backslash x) = 0$ всегда является независимой от $x$).

Если существует взаимодействие между признаками $\mathcal{F}$, то будет называть это взаимодействие $K$-ого порядка, или $K$-мерное взаимодействие, где $K = |\mathcal{F}|$. Например, для $y = x_1 + \ln(x_2) + \exp(x_3)$ - нет никаких взаимодействий между признаками, а в $y = sh(x_1 + x_2) + cos(x_3) \cdot tg(x_1 - x_2)$ есть попарное взаимодействие между всеми признаками, а также их совокупное взаимодействие 3-его порядка.

Введение таких определений взаимодействия покрывает большую часть практических зависимостей, которые возникают в данных, но важно отметить некоторые моменты, когда взаимодействие не будет обнаружено с помощью таких определений. Следуя им, $K$-мерное взаимодействие невозможно, если не было хотя бы одного $K-1$-мерного взаимодействия между этими признаками. Рассмотрим такой пример, где признаки являются бинарными, то есть $x_i \in \{0, 1\}$, а целевая функция имеет вид $y = x_1 \cdot x_2 \cdot x_3$ (например, дверь, которая открывается только тремя ключами). Её нельзя представить в виде суммы 3-х функций из определения выше, что означает, что есть взаимодействие 3-его порядка между всеми признаками. Но в этой функции нет ни одного 2-х мерного взаимодействия!

Важно отметить различие между понятиями взаимодействия признаков и ассоциацией/корреляцией. Интуитивно, взаимодействие между признаками возникает тогда, когда эффект на целевое значение максимален, когда оба признака используются в модели, и сильно падает, если убрать какой-либо из них. Корреляция (корреляция Пирсона) же является лишь некой мерой линейного взаимодействия признаков, никак не отражающая его нелинейную часть. Ассоциация же между признаками возникает, когда один из признаков имеет сильный эффект на целевую переменную, если другой признак не берется во внимание, когда же они оба используются в модели, этот эффект от первого признака сильно снижается. Подробнее об отличии этих определений можно найти в [16].

**Решающие деревья**

Здесь опишем, что такое решающее дерево и введём основные обозначения. Решающее дерево - это логический алгоритм классификации, в основе которого лежит представление пространства признаков в виде последовательного деления по некоторым предикатам.

Формально, решающее дерево - это алгоритм классификации $a(x)$, задающийся деревом с корнем $v_0 \in V$ и множеством внутренних вершин $V = V_{\text{внутр}} \cup V_{\text{лист}}$, в котором каждой внутренней вершине $v \in V$ сопоставлен дискретный признак (предикат) $\beta_v : X \to D_v$, правило перехода в дочерние вершины $S_v : D_v \to V$, и каждой листовой вершине $\forall v \in V$ ставится в соответствие $y_v \in Y$.

На практике, в качестве предикатов используют признаки вида $\beta_v = [f_j(x) \geq t]$, где $f_j \in F$, $t$ - некий threshold, и тогда $D_v = \{0, 1\}$ и дерево становится бинарным. Далее, когда будет говориться о том, что в верши-

не/узле дерева находится признак, будет иметься в виду, что предикат в этой вершине использует в качестве разделителя этот признак. Для таких деревьев удобнее рассматривать вместо $S_v$ две другие функции $R_v$ и $L_v$ $\forall v \in V$, соответствующие правой и левой дочерней вершине вершины $v$. Таким образом, при классификации объект следует по предикатам от корня дерева $v_0$ вниз по дереву и спускается до какой-то листовой вершины, где ему присваивается метка $y_v$. Таким образом, для всех листовых вершин $v \in V$ существует булева конъюнкция $K_v : X \to \{0, 1\}$, которая определяет, что данный объект попадает в данную листовую вершину, а классы $\Omega_v = \{x \in X | K_v(x) = 1\}, \forall v \in V$ непересекающиеся и вместе составляют всё пространство объектов $X$, что гарантирует классификацию всех объектов.

Для того, чтобы выбрать, какой предикат $\beta_v$ использовать в вершине $v \in V$, используют специальные функции, показывающие меру неопределенности (impurity) распределения $p_y$, где $p_y = \frac{1}{|U|} \sum_{x_i \in U} [y_i = y]$, $U$ - рассматриваемая выборка объектов (изначально в корне дерева $U = X$). Такие функции обозначим $\Phi(U)$. Они удовлетворяют свойствам:

1. $\Phi(U)$ минимальна и равна 0, когда $p_y = \{0, 1\}$
2. Максимальна, когда $p_y = \frac{1}{|Y|}$
3. Симметрична, то есть не зависит от переименования классов

При выборе предиката неопределенность меняется, так как меняется распределение $p_y$:

$$\Phi(U|\beta) = \sum_k \frac{|U_k|}{|U|} \Phi(U_k), \tag{5}$$

где $U_k = \{x \in U | \beta(x) = k\}$.

В итоге считается выигрыш в уменьшении неопределённости после ветвления при использовании предиката $\beta$:

$$Gain(\beta, U) = \Phi(U) - \Phi(U|\beta) \tag{6}$$

Таким образом,

$$\beta_v = \arg\max_{\beta \in \mathscr{B}} Gain(\beta, U), \tag{7}$$

где $\mathscr{B}$ - множество предикатов, которое можно построить по выборке.

Для того, чтобы решить задачу регрессии с помощью деревьев (CART) в качестве меры неопределенности используют среднеквадратичную ошибку:

$$\Phi(U) = \min_{y \in Y} \frac{1}{|U|} \sum_{x_i \in U} (y - y_i)^2 \tag{8}$$

вместо же метки класса в $y_v$ используется МНК-решение:

$$y_v = \frac{1}{|U|} \sum_{x_i \in U} y_i. \tag{9}$$

Итак, классический рекурсивный алгоритм синтеза бинарного решающего дерева ID3:

**Алгоритм 1** ID3($U$)

**Вход:** $U$ — подмножество обучающей выборки $X$
**Выход:** $v$ — корень дерева

$\beta \leftarrow \arg\max_{\beta \in \mathcal{B}} \mathrm{Gain}(\beta, U)$
$U_0 \leftarrow \{x \in U \mid \beta(x) = 0\}$
$U_1 \leftarrow \{x \in U \mid \beta(x) = 1\}$
**если** $\mathrm{Gain}(\beta, U) < \mathrm{Gain}_0$ **то**
    создать новый лист $v$
    $y_v \leftarrow$ `majority_class`$(U)$ ▷ или МНК-оценка при регрессии
    **вернуть** $v$
**конец если**
создать новую внутреннюю вершину $v$ с $\beta_v := \beta$
$L_v \leftarrow$ `ID3`$(U_0)$
$R_v \leftarrow$ `ID3`$(U_1)$
**вернуть** $v$

Также решающие деревья могут подсчитывать важность признаков, собирая некоторую статистику по ним в деревьях, например, суммарный $Gain$, который дали предикаты, использующие этот признак. Особенно это полезно, когда рассматриваются ансамбли деревьев.

## 5 Исследование и построение решения задачи

Предлагаемый в этой работе метод, предполагает построение ансамблевой модели на основе консервативных решающих деревьев. Решение будет состоять из нескольких частей: сначала будет рассказано о консервативных деревьях и их свойствах. Затем будет показано, как эти свойства помогают решить проблему обычных деревьев решений. Далее будет обоснованно использование градиентного бустинга в качестве ансамблевой модели. Наконец, последует описание построения статистики, которая собирается на основе информации о структурах деревьев обученного ансамбля, и аргументы в пользу её релевантности.

#### Консервативные деревья решений

Консервативные решающие деревья - это модификация CART, в которой вводится новый гиперпараметр для дерева - $rit\_alpha$ (Redundancy Insensitive Trees). Этот параметр принимает значения от $-\infty$ до 1, $rit\_alpha \in (-\infty, 1)$. Использование rit_alpha дает признакам, которые уже встречались в дереве и были выбраны ранее, или тем, которые ещё не были выбраны, но участвовали в сплите, преимущество перед остальными признаками. Дело в том, что при использовании решающих деревьев в ансамблевых моделях частой практикой, улучшающей предсказательную способность и

декоррелирующей деревья, является RSM, когда при каждом сплите в дереве из всех признаков выбирается только $m$ на рассмотрение, поэтому консервативные деревья используются в контексте ансамблей деревьев.

Отличие консервативных деревьев от обычных решающих деревьев, как упоминалось выше, состоит в функции выбора признака во внутренней вершине. Дерево запоминает в каждом сплите, какой признак оно выбирало до этого в родительских вершинах, а также те признаки, которые были предложены для сплита (были кандидатами на сплит), но не были выбраны, давая им таким образом шанс. Следуя стандартному алгоритму, сначала выбирается признак, предикат по которому лучше всего максимизирует $Gain$, но уже среди признаков, которые были предложены для сплита, и тех, которые были выбраны для сплита до этого - $f_{best}$. Далее считается барьер - величина, равная $Gain$-у от предиката уже выбранного лучшего признака, умноженная на $(1 - rit\_alpha)$:

$$B = Gain(\beta_{f_{best}}, U)(1 - \text{rit\_alpha}) \tag{10}$$

Теперь рассматриваются признаки из тех, которые уже были выбраны в дереве в сплитах и тех, которые были кандидатами на это. И среди тех, что преодолеют порог $B$, выбирается наиболее важный. Как видно, $rit\_alpha = -\infty$, консервативное дерево превращается в обычное дерево решений, так как никакой признак уже не может переступить такой порог. При $rit\_alpha \approx 1$ все признаки из соответствующих множеств способны переступить порог и дерево становится супер-консервативным, то есть оно становится "одноцветным так как никакие другие признаки не пускаются в него. Псевдокод, описывающий этот алгоритм:

**Алгоритм 2** Обучение консервативного решающего дерева TreeLearn(...)

**Вход:** $U$ — подмножество обучающей выборки $X$
$F$ — множество признаков
$I$ — множество признаков, уже использованных в текущем дереве
$R$ — признаки из $F$, участвовавшие в отборе, но не выбранные
$rit\_alpha \in [0,1]$
**Выход:** $v$ — корень дерева

$M \leftarrow m$ признаков из $F$, выбранных для этого сплита
$f_{\text{best}} \leftarrow \arg\max_{f \in M \cup I} \text{Gain}(\beta_f, U)$
$B \leftarrow \text{Gain}(\beta_{f_{\text{best}}}, U) \cdot (1 - rit\_alpha)$
$C \leftarrow$ кандидаты из $I \cup R$ с Gain $\geq B$
**если** $C \neq \emptyset$ **то**
    $\beta \leftarrow$ элемент из $C$ с наибольшей важностью
**иначе**
    $\beta \leftarrow \beta_{\text{best}}$
**конец если**
$U_0 \leftarrow \{x \in U \mid \beta(x) = 0\}$
$U_1 \leftarrow \{x \in U \mid \beta(x) = 1\}$
**если** $\text{Gain}(\beta, U) < \text{Gain}_0$ **то**
    создать новый лист $v$
    $y_v \leftarrow$ `majority_class`$(U)$ ▷ или МНК-оценка при регрессии
    **вернуть** $v$
**конец если**
создать новую внутреннюю вершину $v$ с $\beta_v := \beta$
Обновить множества: $I, R$
$L_v \leftarrow$ `TreeLearn`$(U_0, F, I, R, rit\_alpha)$
$R_v \leftarrow$ `TreeLearn`$(U_1, F, I, R, rit\_alpha)$
    **вернуть** $v$

**Использование консервативных деревьев в задаче поиска взаимодействия**

Введение гиперпараметра rit_alpha в решающие деревья делает их более стабильными: позволяет признакам, имеющим большую важность, чаще появляться в деревьях. Принимая решение, метод отдает предпочтение ранее полезным признакам, даже если они проигрывают в $Gain$. Это увеличивает устойчивость деревьев к случайному шуму, позволяет фокусироваться на том, как признаки соотносятся друг с другом. Также это помогает решить проблему с занижением важности признака, когда слишком много других признаков с ним связано (correlation bias).

Более того, это позволяет делать анализ взаимодействий признаков, на чем и основана данная работа, так как, чтобы внешний признак был выбран в качестве сплитового, он должен преодолеть некоторый барьер, задаваемый параметром rit alpha. И если он это сделает, особенно при больших значениях rit alpha, то имеется большой шанс того, что признаки, чьи пороги он преодолел, могут взаимодействовать с ним. То есть, несмотря на барьер, этот признак более важен для разделения пространства признаков, чем остальные. Таким образом, появляется некоторая эвристика, состоящая в том, что в консервативном дереве признаки, идущие подряд

или находящиеся в каком-то пути из корня дерева в листовую вершину, взаимодействуют.

#### Ансамбли консервативных решающих деревьев

Одно консервативное дерево плохо справится с задачей выявления взаимодействия, потому что не сможет уловить всех взаимодействий между признаками, а при большой глубине оно к тому же и переобучится. Поэтому, по аналогии с уже существующими методами, в этой работе будет использоваться модель ансамбля консервативных деревьев.

**Random Forest (RF)**. Очевидным вариантом для решения поставленной задачи кажется использование ансамбля, который будет усреднять результат работы каждого отдельного дерева, то есть использование Random Forest с консервативными деревьями. Но при исследовании данного метода были обнаружены существенные недостатки, которые не позволяли в должной мере оценить влияние взаимодействия признаков. Разберём на простом примере: пусть есть данные $y = x_1 + x_2$, где $y$ - целевая переменная, $x_1, x_2$ - признаки. Модель должна уметь распознать по таким данным, что взаимодействие отсутствует, а также не должна терять качества при такой оценке.

К сожалению, эксперименты показывают, что качество такой аппроксимации падает 7. Дело в том, что при больших значениях $rit\_alpha$ деревья становятся "одноцветными" , то есть в их сплитах участвует один и тот же признак. Из-за этого каждое такое дерево пытается предсказать $\hat{y} = x_i + \bar{x_j} + \varepsilon$, где $\bar{x_j}$ - среднее значение признака, а $\varepsilon$ - случайный шум. Итого, RF вместо того, чтобы предсказать $y = x_1 + x_2$, аппроксимирует совершенно другую функцию $y = \frac{N_1 x_1 + N_2 x_2}{N} + \frac{N_1 \bar{x_2} + N_2 \bar{x1}}{N} + \varepsilon$, где $N_1, N_2$ - соответственно количество "одноцветных"деревьев с признаком $x_1, x_2$. При $N_1 = N_2$ получаем $y = \frac{x_1 + x_2}{2} + \frac{\bar{x_1} + \bar{x_2}}{2} + \varepsilon$. Именно это и является одной из основных причин падения качества. Поэтому для оценки взаимодействий в данной работе используется другой вид ансамблирования - Gradient Boosting (GB).

**Gradient Boosting (GB)**. Градиентный бустинг — это метод построения ансамбля слабых моделей (чаще всего деревьев решений), где каждая следующая модель обучается аппроксимировать антиградиент ошибки предыдущей. В данной работе используется стандартная модель градиентного бустинга, способная решать как задачу бинарной классификации, так и задачу регрессии, с тем лишь отличием, что вместо CART используются вышеупомянутые консервативные деревья поиска, к гиперпараметрам GB добавляется $rit\_alpha$, а также организуется процедура RSM. Будем называть такую модель ritGB. Приведем здесь краткое описание работы метода:

модель имеет вид:

$$F(x) = \sum_{t=1}^{T} \eta_t h_t(x), \tag{11}$$

где $h_t(x)$ — консервативные деревья решений, $\eta_t$ — скорость обучения (learning rate).

**Регрессия с MSE**

Функция потерь (MSE):

$$\mathcal{L}(y, F(x)) = \frac{1}{2}(y - F(x))^2 \tag{12}$$

**Алгоритм 3** ritGD regression

**Вход:** Обучающая выборка $\{(x_i, y_i)\}_{i=1}^n$, шаг обучения $\eta$, число итераций $T$
**Выход:** Финальная модель $F_T(x)$

1: Инициализировать: $F_0(x) := \frac{1}{n}\sum_{i=1}^{n} y_i$
2: **for** $t = 1$ **до** $T$ **выполнять**
3: Вычислить градиенты (residuals):

$$r_i^{(t)} := y_i - F_{t-1}(x_i), \quad \text{для всех } i = 1, \ldots, n$$

4: Обучить консервативное дерево $h_t(x)$ на выборке $\{(x_i, r_i^{(t)})\}$
5: Обновить модель:

$$F_t(x) := F_{t-1}(x) + \eta \cdot h_t(x)$$

6: **конец for**
7: **вернуть** $F_T(x)$

*Каждое дерево корректирует ошибку предыдущей модели, уменьшая MSE.*

**Классификация с log-loss**

Функция потерь (log-loss) для бинарной классификации:

$$\mathcal{L}(y, F(x)) = \sum_{i=1}^{n} \log\left(1 + e^{-y_i F(x_i)}\right) \tag{13}$$

**Алгоритм 4** ritGD classification

**Вход:** Обучающая выборка $\{(x_i, y_i)\}_{i=1}^n$, где $y_i \in \{0, 1\}$; шаг $\eta$, число итераций $T$
**Выход:** Финальная модель $F_T(x)$, предсказания $P(y = 1|x)$

1: Инициализация:

$$F_0(x) := \frac{1}{2} \log \left( \frac{p}{1-p} \right), \quad p = \frac{1}{n} \sum_{i=1}^{n} y_i$$

2: **for** $t = 1$ **до** $T$ **выполнять**
3: Вычислить псевдо-остатки (градиенты):

$$r_i^{(t)} := y_i - \sigma(F_{t-1}(x_i)), \quad \text{где } \sigma(z) = \frac{1}{1 + e^{-z}}$$

4: Обучить консервативное дерево $h_t(x)$ на выборке $\{(x_i, r_i^{(t)})\}$
5: Обновить модель:

$$F_t(x) := F_{t-1}(x) + \eta \cdot h_t(x)$$

6: **конец for**
7: Получить вероятности:

$$P(y = 1|x) := \frac{1}{1 + e^{-F_T(x)}}$$

8: **вернуть** $F_T(x)$, $P(y = 1|x)$

Эксперименты показали, что на данных, где, согласно определению, нет взаимодействия, качество ritGB не падает (значительно) с ростом $rit\_alpha$, а на данных, где это взаимодействие присутствует, качество заметно уменьшается при приближении $rit\_alpha$ к 1 см. примеры в 7. Такое поведение модели можно объяснить тем, что если целевая переменная имеет аддитивную природу по признакам, то ritGD, с помощью "одноцветного"дерева по какому-то признаку убирает зависимость от этого признака, добавляя какое-то смещение, но не может этого сделать, если признак взаимодействует с другими признаками. Аналогично, если присутствует какая-то аддитивная структура в данных, то деревья смогут убирать зависимость от слагаемых этой структуры. Разберем предыдущий простой пример с $y = x_1 + x_2$. Допустим, на первой итерации $ritGD$ построил "одноцветное"дерево по признаку $x_1$. Допустим также, что начальное приближение было нулевым (для наглядности). Тогда это дерево обучится на изначальной обучающей выборке и будет предсказывать $y = x_1 + \bar{x_2}$. Тогда на следующей итерации уже придется обучаться на residuals, то есть на антиградиенте по лосс функции, которая является разностью между целевым значением и предсказанием модели на данной итерации. В нашем примере $r = x_1 + x_2 - x_1 - \bar{x_2} = x_2 - \bar{x_2}$. Видно, что мы избавились от зависимости от $x_1$, что позволяет "одноцветному"дереву о $x_2$ качественно обучиться на $r$. Аналогичный эффект достигается и для классификации, так как градиент log-loss является аналогом градиента MSE, но для классификации.

**Статистика для определения взаимодействий и подбор гиперпараметров**

Для каждой модели ritGB подбор основных гиперпараметров проводится на обучающем датасете с помощью кросс-валидации по сетке гиперпараметров (например, с помощью GridSearchCV) с фиксированным $rit\_alpha$, так называемый "тюнинг"модели. Подбор гиперпараметра $rit\_alpha$ проводится с помощью статистических методов. Сначала обучается $N_{sample}$ оттюненных моделей с $rit\_alpha = -\infty$, что соответствует обычному бустингу на решающих деревьях, и берется их метрики качества. Так как нам хотелось бы, чтобы ritGB сохранял качество, мы перебираем значения параметра $rit\_alpha$ от 0 до 1, каждый раз обучая $N_{sample}$ оттюненных моделей с нынешним $rit\_alpha$ и беря метрики их качества, после чего с помощью критерия Манна-Уитни проверяем гипотезу о том, что качество не изменилось значимо, если гипотеза отклоняется, то берем последнюю $rit\_alpha$, когда отклонение ещё не было значимо. Формально:

Обозначим:

- $\mathcal{M}_\alpha = \{M_\alpha^{(1)}, \ldots, M_\alpha^{(N)}\}$ — множество $N$ моделей, обученных с параметром $rit\_alpha = \alpha$ на случайных подвыборках обучающей выборки одинакового размера;
- $\text{score}(M)$ — метрика качества модели $M$ (в работе использовалась $R^2$ и $roc - auc$).
- $S_\alpha$ - выборка с посчитанными метриками качества моделей.

Задача состоит в следующем:

$$\begin{cases} H_0: & \text{распределения } \text{score}(M_{\alpha_0}^{(i)}) \text{ и } \text{score}(M_{-\infty}^{(j)}) \text{ совпадают,} \\ H_1: & \text{качество при } \alpha_0 \text{ статистически ниже, чем при } \alpha = -\infty \end{cases}$$

Для проверки используется односторонний критерий Манна — Уитни (U-test) на уровне значимости $\delta = 0.05$:

$$p = \text{MannWhitneyU}\,(S_{\alpha_0}, S_{-\infty}, \text{alternative='less'})$$

Если $p > \delta$, то гипотеза $H_0$ **не отвергается**, и параметр $\alpha_0$ может быть принят как допустимый, не ухудшающий качество модели. Таким образом, выполняется следующий алгоритм подбора параметра $rit\_alpha$:

**Алгоритм 5** Подбор параметра `rit_alpha` с контролем качества

**Вход:** Данные $\mathcal{D}$, множество кандидатов $\{\alpha_k\}_{k=1}^{K}$, число повторов $N_{sample}$, уровень значимости $\delta = 0.05$

**Выход:** Оптимальное значение $rit\_alpha^*$

$S_{-\infty} \leftarrow$ массив из $N_{sample}$ метрик качества моделей с $rit\_alpha = -\infty$
**for** $k = 1$ **до** $K$ **выполнять**
    $\alpha_k \leftarrow$ текущий кандидат
    $S_{\alpha_k} \leftarrow$ массив из $N_{sample}$ метрик качества с $rit\_alpha = \alpha_k$
    Выполнить односторонний тест Манна — Уитни:

$$p \leftarrow \text{MannWhitneyU}(S_{\alpha_k}, S_{-\infty}, \text{alternative} = \text{'less'})$$

    **если** $p > \delta$ **то**
        **вернуть** $rit\_alpha^* := \alpha_k$ ▷ Качество не ухудшилось значимо
    **конец если**
**конец for**
**вернуть** $rit\_alpha^* := -\infty$ ▷ Не найдено подходящего значения

Ранее было объяснено, почему консервативные деревья в ritGD будут содержать в своих вершинах потенциально взаимодействующие признаки. Также было показано, как искать параметр $rit\_alpha$. Теперь нужно понять, как вытащить это знание из структуры деревьев. В данной работе предлагается считать статистики $S_i^{\mathcal{F}}$ для каждого дерева i из модели $M_\alpha$ (будем обозначать $i \in M_\alpha$), а затем усреднять их по всем деревьям, получая для каждого множества признаков $\mathcal{F}$ статистику $S(M_\alpha, \mathcal{F}) = \frac{1}{N_{M_\alpha}} \sum_{i \in M_\alpha} S_i^{\mathcal{F}}$, где $N$ — количество предикторов в $M_\alpha$.

Статистики $S_i^{\mathcal{F}}$ консервативного дерева $i$, являющегося предиктором в модели ritGD, предлагается считать следующим способом:

1. Для каждой листовой вершины $v \in V^i$ находим "путь"из корневой вершины к ней, который состоит из всех внутренних вершин, которые должен пройти объект, чтобы попасть в эту листовую вершину.

2. Для каждого такого "пути"строим список из признаков, которые лежат в вершинах этого "пути".

3. Находим длину списка $l_v^i$.

4. Для всех интересующих нас множеств признаков $\mathcal{F}$ по каждому списку считаем, количество раз, сколько раз множество $\mathcal{F}$ совпадало с множеством $|\mathcal{F}|$ подряд идущих признаков, обозначим это число $n_v^i(\mathcal{F})$. Например, если нас интересует взаимодействие между признаками $\mathcal{F} = \{x_1, x_2, x_3\}$, а список получился $[x_1, x_3, x_2, x_1, x_1, x_2, x_3, x_2]$, то множество совпадет 3 раза, $n_v^i(\{x_1, x_2, x_3\}) = 3$.

5. Теперь статистика $S_i^{\mathcal{F}}$ находится следующим способом:

$$S_i^{\mathcal{F}} = \frac{1}{|V^i_{\text{лист}}|} \sum_{v \in V_{\text{лист}}} \frac{N_v^i}{N^i} \frac{n_v^i(\mathcal{F})}{l_v^i}, \tag{14}$$

где $N_v^i$ - количество объектов из обучающей выборки, которые попали в листовую вершину $v$, $N^i$ - общее количество объектов в обучающей выборке.

Таким образом, итоговая статистика $S(M_\alpha, \mathcal{F})$ по модели ritGB $M_\alpha$ с $rit\_alpha = \alpha$ будет рассчитываться следующим образом:

$$S(M_\alpha, \mathcal{F}) = \frac{1}{N_{M_\alpha}} \sum_{i \in M_\alpha} \frac{1}{|V^i_{\text{лист}}|} \sum_{v \in V_{\text{лист}}} \frac{N_v^i}{N^i} \frac{n_v^i(\mathcal{F})}{l_v^i} \tag{15}$$

Теперь предлагается оттюнить $N_{st}$ моделей с подобранным $rit\_alpha$ и для каждой модели найти $S(M_\alpha, \mathcal{F})$, а затем усреднить результат.

Пусть $N_{st}$ моделей: $\{M_\alpha^1, M_\alpha^2, ..., M_\alpha^{N_{st}}\}$. Итоговая статистика $S(\mathcal{F})$ будет рассчитываться следующим образом:

$$S(\mathcal{F}) = \frac{1}{N_{st}} \sum_{j=1}^{N_{st}} S(M_\alpha^j, \mathcal{F}) = \frac{1}{N_{st}} \sum_{j=1}^{N_{st}} \frac{1}{N_{M_\alpha^j}} \sum_{i \in M_\alpha^j} \frac{1}{|V^i_{\text{лист}}|} \sum_{v \in V_{\text{лист}}} \frac{N_v^i}{N^i} \frac{n_v^i(\mathcal{F})}{l_v^i}. \tag{16}$$

Проанализируем эту формулу. Первое усреднение направлено на уменьшение дисперсии статистики. Второе усреднение направлено на уменьшение влияния шумовых статистик, которые могли бы случайно возникнуть в каком-то дереве, основные паттерны статистик будут наблюдаться в большом количестве деревьев. Третье усреднение по листьям в каждом дереве проводится на случай, если деревья будут иметь различное количество листьев, что может завысить или занизить определенные статистики. Усреднение по длине пути нужно, так как пути различной длины могут сильно разниться в показателях $n_v^i$. Вес $\frac{N_v^i}{N^i}$ вводится с целью устранить паттерны, которые маловероятно встречаются в деревьях. Если в один лист попало намного больше объектов при обучении, чем в другой, значит второй плохо отражает реальную ситуацию в данных (при условии неразреженности данных).

В техническом задании требовалось задать $\forall k \ S_0^k$. Исходя из экспериментов, можно утверждать что для построенной статистики проще всего определить $S_0^k \equiv 0 \ \forall k$.

## 6 Вычислительные эксперименты

Разработанный метод применяется для обнаружения взаимодействий на синтетических и реальных наборах данных. Эффективность метода можно оценить по результатам работы на синтетических данных, так как истинные взаимодействия известны. Для реальных же данных объяснение взаимодействий будет основываться на самих описаниях признаков. Во всех экспериментах проводится процедура подбора оптимального $rit\_alpha$ с помощью алгоритма, описанного выше. Расчет статистики происходит с параметрами $N_{st} = 20$, для более устойчивого её значения. Количество консервативных деревьев решений в модели ансамбля для поиска оптимального $rit\_alpha$ равно $N_{M_\alpha} = 100$, для подсчета статистики $N_{M_\alpha} = 1000$.

#### Эксперименты на синтетических данных

Генерация выборки для задачи регрессии задаётся с помощью функции

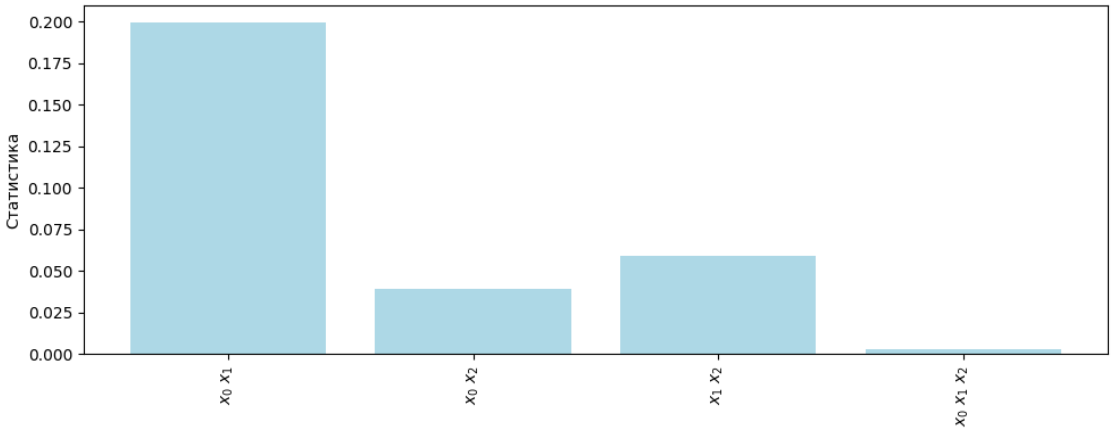
$$y = F(x_1, x_2, ..., x_p) + \varepsilon, \tag{17}$$

где $x_i$ берутся из равномерного распределения $U(0, 1)$, $\varepsilon \in \mathbf{N}(0, 0.1)$ — гауссовский шум.

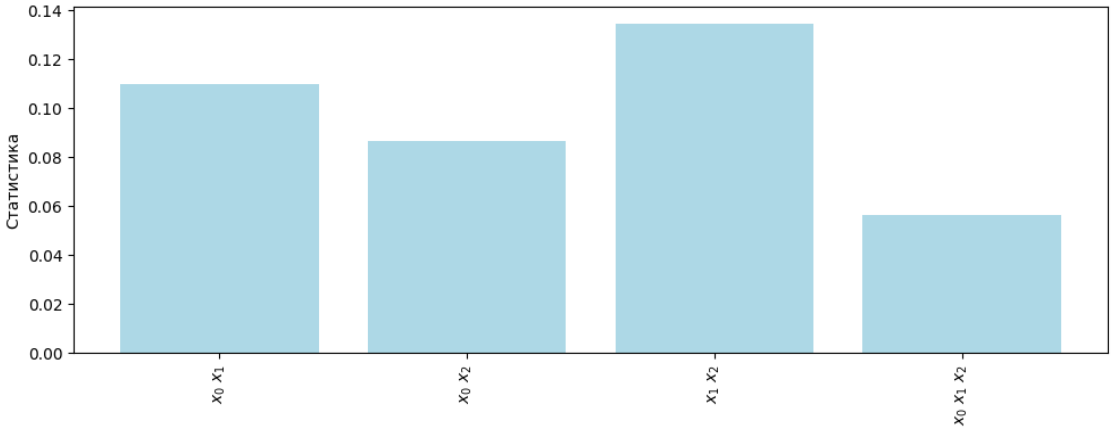
Для задачи классификации будут генерироваться выборки, делящие пространство $R^p$ на 2 класса. Будут рассмотрены только случаи $p = 2, p = 3$, так как они наиболее наглядные.

#### Задача регрессии

Сначала рассмотрим, как модель справляется на выборках малой размерности, результаты представлены на рис. 1




(a)



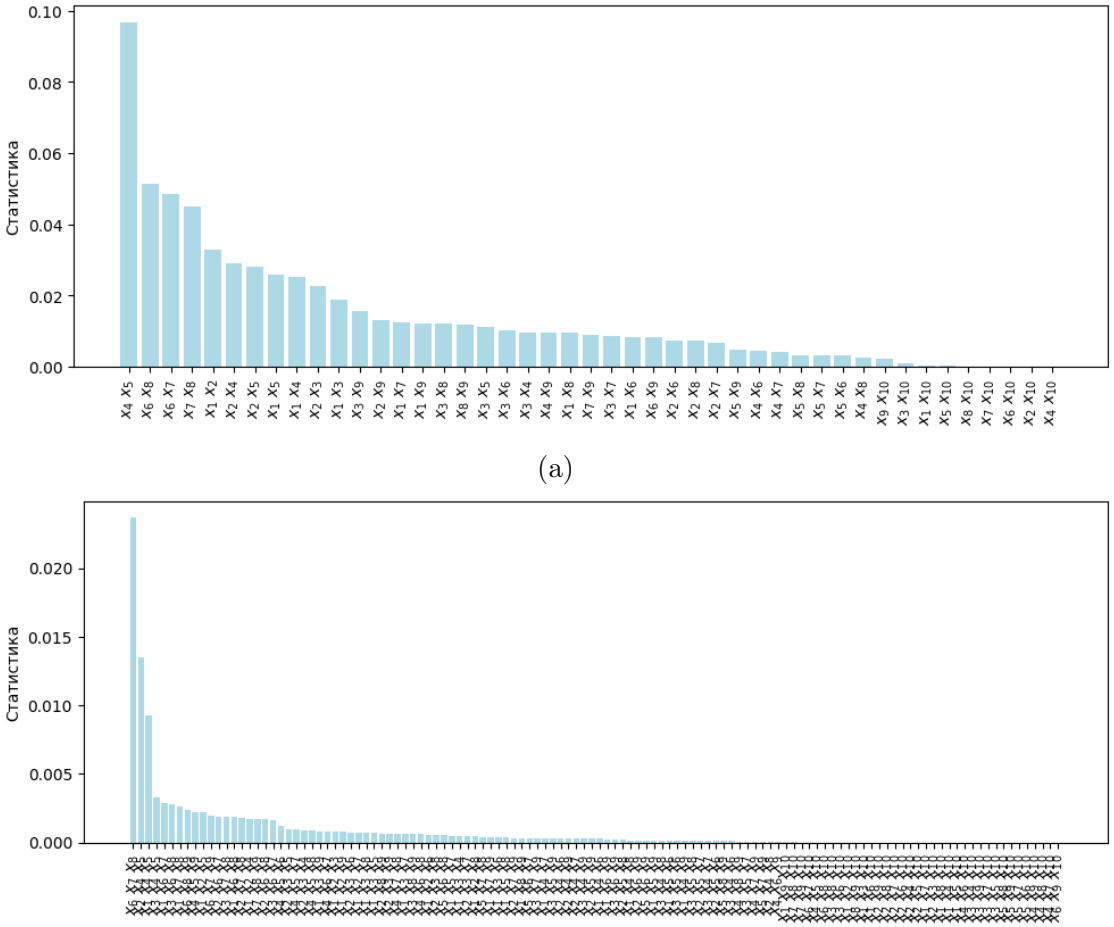

(b)

Рис. 1: (a) $y = x_1 \cdot x_2 + x_3$ (b) $y = x_1 \cdot x_2 \cdot x_3$

Как видим, на данных малых размерностей модель довольно хорошо определяет взаимодействие между признаками.

Рассмотрим более сложную модель с 10 признаками

$$F(x) = 2x_1 + 3x_2^2 + sin(x_3) + 5x_4x_5 + 4x_6x_7x_8 + 0.5x_9 + \varepsilon$$




(a)



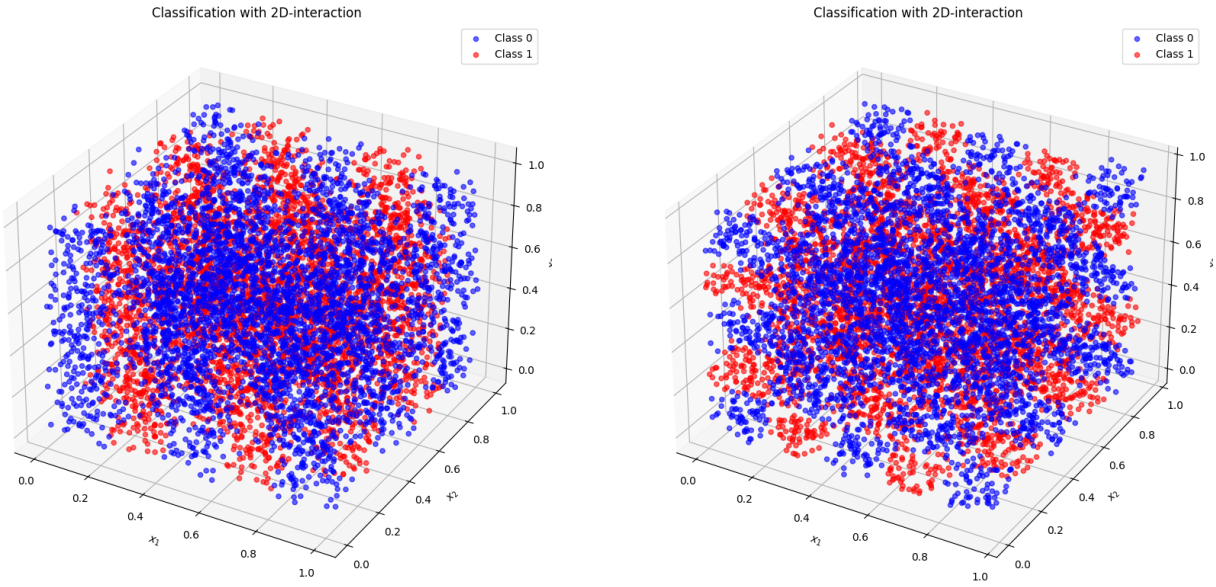

(b)

Рис. 2: (a) Парные взаимодействия (b) Тройные взаимодействия

Видно, что модель смогла уловить сильное взаимодействие $(x_4, x_5)$, а также попарные взаимодействия тройки $(x_6, x_7, x_8)$ и другие взаимодействия. Также были обнаружены 3-х мерные взаимодействия $(x_6, x_7, x_8)$ и $(x_2, x_4, x_5)$, второе из которых является следствием сильного взаимодействия $(x_4, x_5)$ и важного признака $x_2$.

**Задача классификации**

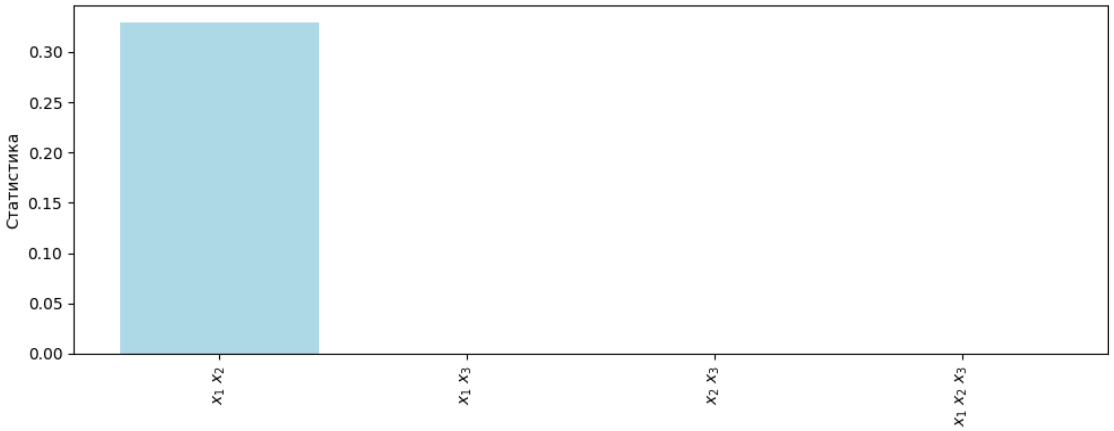
Проверяются датасеты, которые представлены на изображениях 4



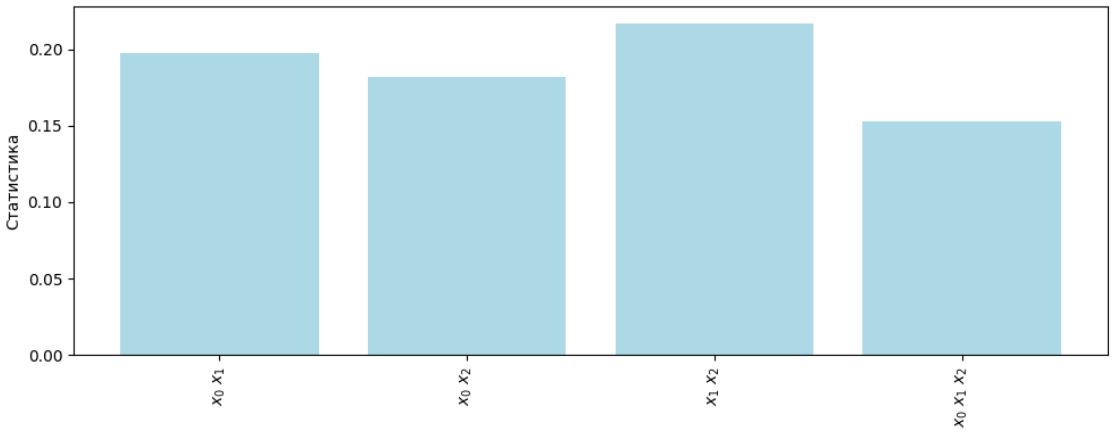

(a) (b)

Рис. 3: (a) нет взаимодействий с $x_3$ (b) присутствуют все взаимодействия

По изображению (a) мы можем наблюдать, что есть взаимодействие между признаками $x_1$ и $x_2$, а между $x_3$ и остальными взаимодействия нет. На изображении (b) видим, что есть взаимодействие между всеми признаками. Проверим, как справится с этим наш метод:




(a)




(b)

Рис. 4: (a) нет взаимодействий с $x_3$ (b) присутствуют все взаимодействия

Как видно, для датасета (a) было корректно найдено взаимодействие между $x_1$ и $x_2$, а для датасета (b) успешно были найдены попарные взаимодействия, а также 3-х мерное взаимодействие.

#### Эксперименты на реальных данных

Были проведены эксперименты на данных из 3-х датасетов регрессионных данных из коллекции Луиса Торго [18].

**California Housing.** Это набор регрессионных данных, представленный в (Pace & Barry, 1997). Он описывает, как цены на жильё зависят от различных переменных данных переписи населения. Признаки: MedInc - медианный доход жителей района, HouseAge - медианный возраст домов в районе, AveRooms - среднее количество комнат в домах, AveBedrms - среднее количество спальных комнат в домах, Population - численность населения в районе , AveOccup - среднее количество жителей в доме, Latitude - широта, Longitude - долгота.

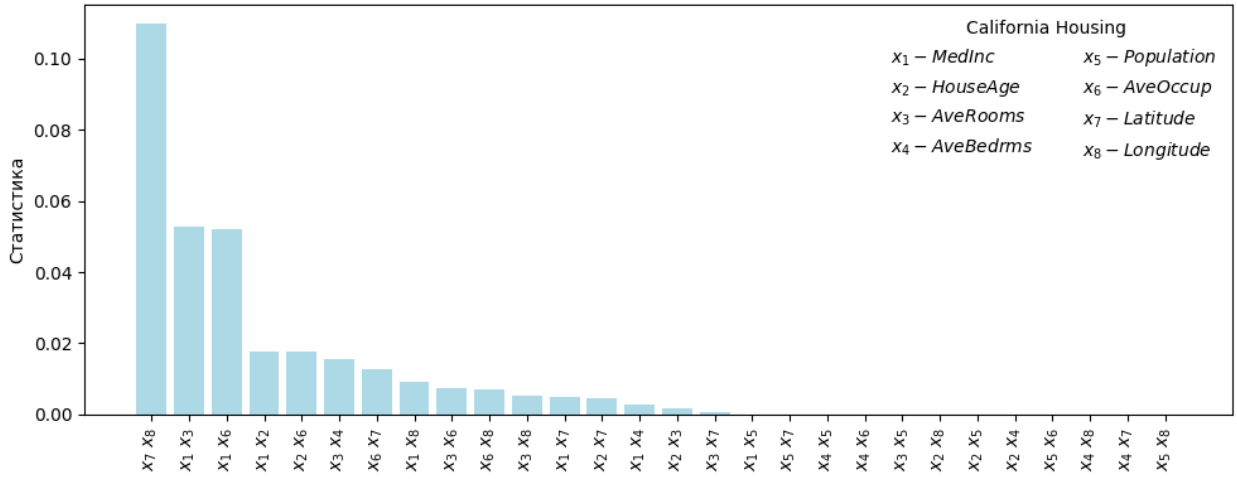



Рис. 5: Оценка попарных взаимодействий для California Housing

Видим, что модель смогла обнаружить сильное взаимодействие между обоими координатами района, что объясняется тем, что только вместе они правильно указывают местонахождение района. Также были обнаружены более слабые взаимодействия между медианным доходом и средним количеством комнат и жильцов в доме.

**Elevators.** Этот набор данных получен в результате выполнения задачи по управлению воздушным судном. Признаки: climbRate - скорость набора высоты, Altitude - текущая высота над уровнем моря, RollRate - скорость крена, curRoll - текущий угол крена, diffClb - производная от curRoll, diffDiffClb - вторая производная от curRoll.

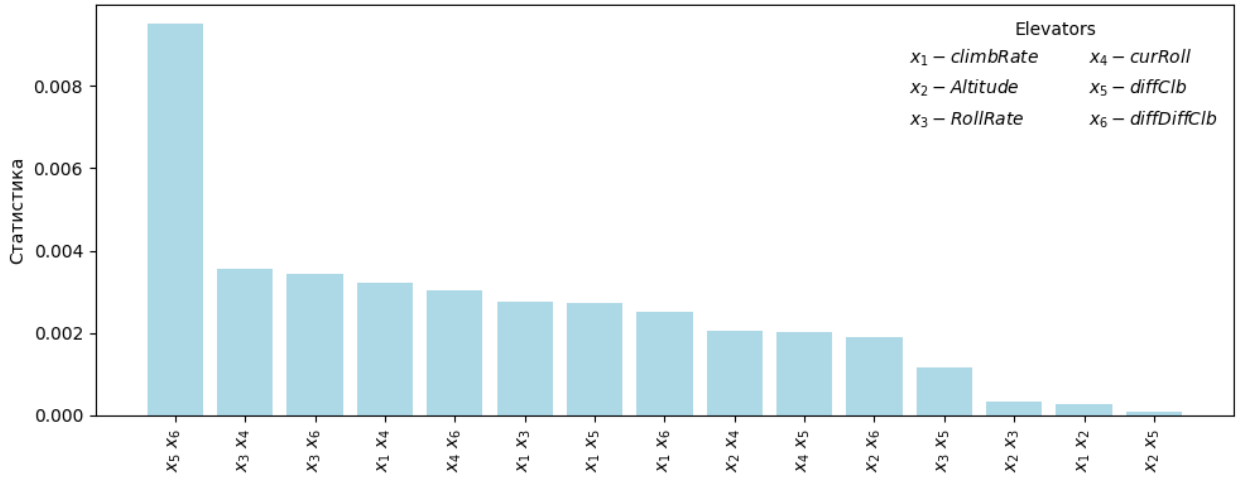



Рис. 6: Оценка попарных взаимодействий для Elevators

Метод обнаружил сильное взаимодействие между первой и второй производной *curRoll*, что объясняется тем, что они меняются одновременно при полете судна. Также было обнаружено 3-х мерное взаимодействие между Altitude, RollRate и diffDiffClb.

**Kinematics (kin8nm).** Набор данных kin8nm из репозитория Delve (Rasmussen et al., 2003) описывает моделирование движения 8-звенного манипулятора робота. Его входные переменные соответствуют угловым поло-

жениям суставов, и его создатели описывают его как сильно нелинейный. Признаки: $theta_{1-8}$.

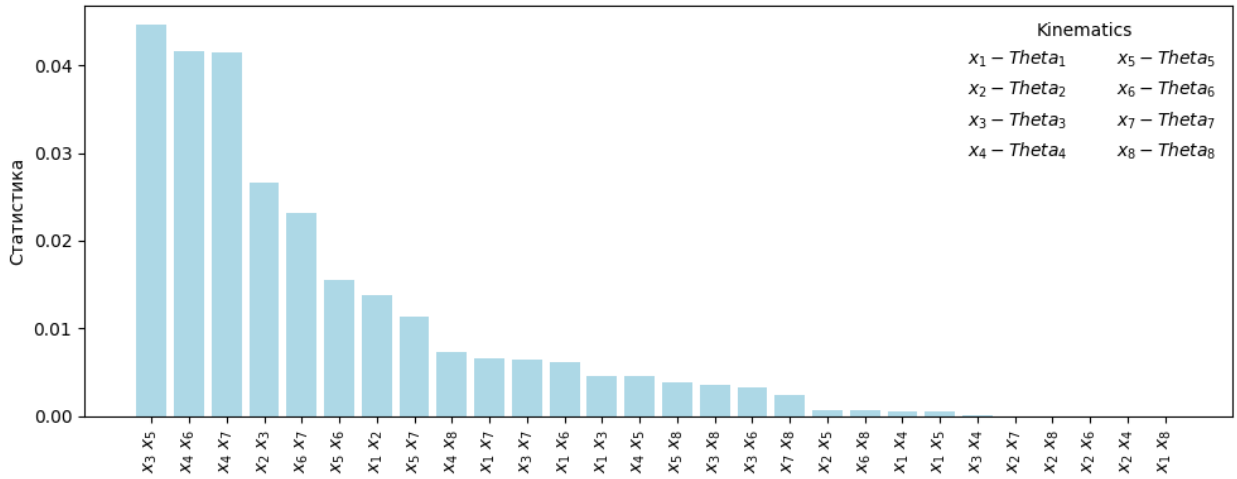



Рис. 7: Оценка попарных взаимодействий для Kinematics

Метод, действительно, обнаруживает множество парных взаимодействий между признаками.

## 7 Заключение

В работе было приведено описание метода по решению задачи определения взаимодействий признаков на табличных данных с помощью консервативного градиентного бустинга $ritGB$. Проверка корректности и качества метода была проведена на синтетических и реальных данных: метод показал свою способность улавливать значимые взаимодействия.

Приведенный в данной работе метод удовлетворяет требованиям тех. задания:

1. Для всех выбранных подмножеств признаков $\mathcal{F}$ определяется, есть ли у них взаимодействие.
2. Ранжирование происходит по подсчитанной статистике в каждой категории признаков.

Модель консервативного градиентного бустинга, использованная в работе, также удовлетворяет введенным критериям:

1. Модель позволяет определять взаимодействия за счёт консервативных решающих деревьев.
2. Ансамблевая структура модели позволяет её улавливать взаимодействия группы признаков.
3. Консервативные деревья решений - модификация CART, они способны работать как с регрессией, так и с классификацией.
4. Подобранный гиперпараметр $rit\_alpha$ в совокупности с ансамблевой структурой градиентного бустинга обеспечивает устойчивость к шумовым признакам и случайным взаимодействиям.
5. Консервативные деревья решений способны работать как с числовыми, так и с категориальными признаками.
6. Метод интерпретируемый, потому что используется анализ структуры решающих деревьев.

   Таким образом, поставленная задача решена.

## Список литературы


[1] *Cordell, Heather J.* Detecting gene–gene interactions that underlie human diseases. — 2009.

[2] *Phillips, Patrick C.* Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems / Patrick C. Phillips // *Nature Reviews Genetics*. — 2008. — Vol. 9, no. 11. — Pp. 855–867.

[3] Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer / Marylyn D. Ritchie, Lance W. Hahn, Nady Roodi et al. // *American Journal of Human Genetics*. — 2001. — Vol. 69, no. 1. — Pp. 138–147.

[4] *Zhang, Yu.* Bayesian inference of epistatic interactions in case–control studies / Yu Zhang, Jun S. Liu // *Nature Genetics*. — 2007. — Vol. 39, no. 9. — Pp. 1167–1173.

[5] *Cordell, Heather J.* Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans / Heather J. Cordell // *Human Molecular Genetics*. — 2002. — Vol. 11, no. 20. — Pp. 2463–2468.

[6] *Fisher, Ronald A.* XV.—The Correlation Between Relatives on the Supposition of Mendelian Inheritance / Ronald A. Fisher // *Transactions of the Royal Society of Edinburgh*. — 1918. — Vol. 52. — Pp. 399–433.

[7] *Koza, John R.* Genetic programming: On the programming of computers by means of natural selection (complex adaptive systems) / John R Koza // *A Bradford Book*. — 1993. — Vol. 1. — P. 18.

[8] *Lundberg, Scott M.* A Unified Approach to Interpreting Model Predictions / Scott M Lundberg, Su-In Lee // *Advances in Neural Information Processing Systems*. — 2017. — Vol. 30. — Pp. 4765–4774.

[9] Detecting Statistical Interactions With Additive Groves of Trees / Daria Sorokina, Rich Caruana, Mirek Riedewald, Dietrich Fink // *Proceedings of the 25th International Conference on Machine Learning (ICML)*. — 2008. — Pp. 1000–1007.

[10] *Friedman, Jerome H.* Predictive Learning via Rule Ensembles / Jerome H. Friedman, Bogdan E. Popescu // *The Annals of Applied Statistics*. — 2008. — Vol. 2, no. 3. — Pp. 916–954.

[11] *Breiman, Leo.* Random Forests / Leo Breiman // *Machine Learning*. — 2001. — Vol. 45, no. 1. — Pp. 5–32.

[12] *Friedman, Jerome H.* Greedy Function Approximation: A Gradient Boosting Machine / Jerome H. Friedman // *Annals of Statistics*. — 2001. — Vol. 29, no. 5. — Pp. 1189–1232.

[13] Classification and Regression Trees / Leo Breiman, Jerome H. Friedman, Charles J. Stone, Richard A. Olshen. — Belmont, CA: CRC Press, 1984.

[14] *Meinshausen, Nicolai.* Node Harvest / Nicolai Meinshausen // *The Annals of Applied Statistics.* — 2010. — Vol. 4, no. 4. — Pp. 2049–2072.

[15] *Meinshausen, Nicolai.* Forest Garrote / Nicolai Meinshausen // *Electronic Journal of Statistics.* — 2009. — Vol. 3. — Pp. 1288–1304.

[16] *Boulesteix, Anne-Laure.* Letter to the Editor: On the term 'interaction' and related phrases in the literature on Random Forests / Anne-Laure Boulesteix, Alexander Hapfelmeier // *Briefings in Bioinformatics.* — 2015. — Vol. 16, no. 2. — Pp. 338–341.

[17] *Ruczinski, Inke.* Sequence analysis using logic regression / Inke Ruczinski, Michael LeBlanc, Li Hsu // *Genetic Epidemiology.* — 2001. — Vol. 21. — Pp. S626–S631.

[18] *Torgo, Luís.* Data Mining with R: Learning with Case Studies / Luís Torgo. — Boca Raton, FL: Chapman and Hall/CRC, 2007.

[19] Iterative Random Forests to Discover Predictive and Stable High-Order Interactions / Sumanta Basu, K. Kumbier, J.B. Brown, B. Yu // *Proceedings of the National Academy of Sciences.* — 2018. — Vol. 115, no. 8. — Pp. 1943–1948.

# Приложение

**A1. Random Forest Score Experiments**

**A2. Gradient Boosting Score Experiments**