**Санкт–Петербургский государственный университет**

***Мусаева Аида Александровна***

**Выпускная квалификационная работа**

# *Применение вектора Шепли для интерпретации моделей машинного обучения*

Уровень образования: бакалавриат
Направление 01.03.02 «Прикладная математика и информатика»
Основная образовательная программа СВ.5005.2017 «Прикладная математика, фундаментальная информатика и программирование»
Профиль «Системный анализ, исследование операций и управление»

Научный руководитель:
профессор, кафедра математической теории игр
и статистических решений, д.ф. - м.н. Петросян Леон Аганесович

Рецензент:
доцент, кафедра математического моделирования
энергетических систем, к.ф. - м.н. Лежнина Елена Александровна

Санкт-Петербург
2021 г.

# Содержание

# Введение

Машинное обучение все чаще применяется в самых разных сферах жизни для решения задач прогнозирования, классификации и разработки рекомендательных систем. Одним из основных препятствий для широкого распространения машинного обучения является компромисс между интерпретируемостью и сложностью алгоритма. Чем сложнее внутренняя структура модели, тем более глубокие взаимосвязи между переменными она может находить, но и тем труднее она становится для понимания людьми.

Понимание того, какие переменные и каким образом влияют на работу модели, может помочь идентифицировать потенциальные проблемы в ней и дать информацию о том, какие еще переменные можно туда добавить, чтобы улучшить качество предсказания, а какие стоит из нее исключить, чтобы оптимизировать использование ресурсов.
Необходимость интерпретировать предсказания моделей возникла в силу многих причин.

Широкое использование моделей машинного обучения в таких отраслях, как медицина, финансы и политика породило требования к безопасности модели, обоснованию доверия ей, а также привело к существованию различных правил, регламентирующих работу автоматизированных систем принятия решений. Так с 2018 года в Европейском союзе начал действовать новый регламент о защите данных — General Data Protection Regulation (GDPR), одна из статей которого гласит, что каждый субъект данных имеет право на получение информации о том, почему автоматизированная система приняла то или иное решение.

Также интерпретация моделей тесно связана с классом задач состязательных аттак, целью которых является построение такого входа для алгоритма машинного обучения, на котором алгоритм бы сделал ошибку.

Глобальные методы интерпретации призваны показать, какие факторы в целом оказывают наибольшее влияние на структуру модели и на ее предсказания.

Локальные методы пытаются объяснить то, как было сделано данное конкретное предсказание (например, отказ в выдаче кредита клиенту). За-

частую локальные методы могут быть использованы как основа для более глобальной интерпретации, например, путем усреднения или визуализации.

Целью данной работы является рассмотрение локальных и глобальных методов интерпретации моделей машинного обучения, основанных на построении игровой модели, программная реализация выявленных подходов и иллюстрация применения на конкретной модели машинного обучения.

# Глава 1. Кооперативные ТП-игры

Кооперативная игра $n$ лиц с трансферабельными полезностями (полезности различных игроков могут быть оценены единой шкалой) описывается как отображение

$$v : S \to v(S) \in \mathbb{R}, S \subseteq I = \{1, 2, \ldots, n\}, S \neq \emptyset,$$

сопоставляющее всевозможным коалициям игроков $S$ их выигрыш.

Значение ТП кооперативных игр каждой игре $v : S \to v(S)$ ставит в соответствие сбалансированное распределение $x \in \mathbb{R}^n$ величины $v(I)$, т. е. $\sum_I x_i = v(I)$.

Пусть задана некоторая последовательность игроков $\pi = (i_1, i_2, \ldots, i_n)$. Этой последовательности ставится в соответствие вектор маргинальных вкладов $x(\pi) = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, заданный по формуле:

$$x_{i_1} = v(i_1), x_{i_2} = v(i_1, i_2) - v(i_1), \ldots, x_{i_n} = v(N) - v(i_1, i_2, \ldots, i_{n1})$$

Вектор маргинальных вкладов $x(\pi)$ определяет распределение общего выигрыша $v(I)$ в соответствии с заданным упорядочиванием игроков. Вектор Шепли это средний вектор, вычисленный по всевозможным упорядоченным последовательностям игроков.

Для получения соответствующей формулы, представим, что агенты из $I$ случайно упорядочены $(i_1, i_2, \ldots, i_n)$, причем вероятность каждого упорядочения одинакова. Вектор Шепли приписывает агенту $i$ среднее его маргинальной прибыли $v(S \cup \{i\}) - v(S)$, взятое по всем коалициям $S \cup I \setminus \{i\}$, включая пустое множество. Вес коалиции $S$ соответствует вероятности того, что в случайной очереди $(i_1, i_2, \ldots, i_n)$ перед агентом $i$ стоят в точности элементы из множества $S$. Непосредственное вычисление этой вероятности дает величину $\frac{s!(n-s-1)!}{n!}$, так как существует ровно $s!(n-s-1)!$ упорядочений $I$ таких, что первые $s$ элементов берутся из $S$, а последние $n - s - 1$ элементов берутся из $I \setminus (S \cup \{i\})$.

**Определение** Для ТП-игры $v : S \to v(S), S \subseteq I$ вектор Шепли $\phi$ рас-

пределяет выигрыш $v(I)$ максимальной коалиции следующим образом:

$$\phi_i(v) = \sum_{S \subseteq I \setminus \{i\}} \frac{|s|!(n - |S| - 1)!}{n!}(v(S \cup \{i\}) - v(S)), i \in I$$

Вектор Шепли удовлетворяет следующим свойствам:

1. **Линейность**. Для любой характеристической функции $v$, являющейся линейной комбинацией двух других характеристических функций $u$ и $w$ ($v(S) = \alpha u(S) + \beta w(S)$) значения Шепли $v$ равны соответствующей линейной комбинации значений Шепли $u$ и $w$ ($\phi_i(v) = \alpha\phi_i(u) + \beta\phi_i(w)$).

2. **Симметричность**. Если $v(S \cup \{i\}) = v(S \cup \{j\})$ для любого $S$, не содержащего $i$ или $j$, тогда $\phi_i(v) = \phi_j(v)$

3. **Аксиома болвана**. *Болваном* в теории кооперативных игр называется бесполезный игрок, не вносящий вклада ни в какую коалицию, то есть игрок $i$ такой, что для любого $S \subseteq I$ выполнено $v(S \cup \{i\}) = v(S)$. Аксиома болвана состоит в том, что если игрок $i$ — болван, то $\phi_i(v) = 0$.

4. **Эффективность**. Вектор Шепли позволяет полностью распределить имеющееся в распоряжении тотальной коалиции благосостояние, то есть $\sum_i \phi_i(v) = v(I)$.

**Теорема** (Шепли, 1953 [10]). Для любой кооперативной игры $v$ существует единственное распределение выигрыша, удовлетворяющее аксиомам 1 — 4. Этим распределением является вектор Шепли.

# Глава 2. Локальная интерпретация

Методы интерпретации моделей машинного обучения, не зависящие от вида модели, используют упрощенную объясняющую модель. Определим эту модель как любую модель, которая аппроксимирует первоначальную модель и может быть легко проинтерпретирована. Объясняющая модель зачастую использует упрощенный вход $x'$, который связан с первоначальными входными данными с помощью функции $x = h_x(x')$.

В работе [9] было показано, что многие локальные методы интерпретации используют одинаковую объясняющую модель и был введен класс методов, обобщающий эти подходы.

Пусть $f$ — модель, которую необходимо интерпретировать, а $g$ — объясняющая модель.

Методы с аддитивными значимостями признаков используют следующую объясняющую модель:

$$g(z') = \psi_0 + \sum_{i=1}^{N} \psi_i z'_i, \quad z' \in \{0,1\}^N,$$

$\psi_i \in \mathbb{R}$, $N$ — количество признаков модели.

Методы, использующие данную объясняющую модель, приписывают значимость $\psi_i$ каждому признаку, а суммирование по всем значимостям аппроксимирует выход исходной модели.

В работе [9] было доказано, что существует единственное решение в этом классе методов, удовлетворяющее трем желаемым свойствам, описанными ниже.

Желаемые свойства методов с аддитивными значимостями:

### 1. Локальная точность (Эффективность)

$$f(x) = g(x) = \psi_0 + \sum_{i=1}^{N} \psi_i x'_i,$$

Объясняющая модель $g(x')$ соответствует исходной модели $f(x)$ при $x = h_x(x')$

**2. Missingness**

$$x'_i = 0 \to \psi_i = 0,$$

Данное свойство подразумевает, что отсутствующий во входном векторе признак $x'_i$ получит нулевую оценку воздействия.

**3. Consistency**

Пусть $f_x(z') = f(h(z'))$ и $z'_{\setminus j}$ означает, что $z_j = 0$.

Обозначим за $f$ модель, аналогичную $f$ за исключением того, что входы $i$ и $j$ поменяны местами: для всех подмножеств $S$, не содержащих $i$ или $j$ выполняется $f'_x(S \cup \{i\}) = f_x(S \cup \{j\})$ и $f'_x(S) = f_x(S)$.

Если для любых двух моделей $f$ и $f'$ выполняется условие:

$$f'_x(z') - f'_x(z'_{\setminus j}) \geqslant f_x(z') - f_x(z'_{\setminus j})$$

для любых $z' \in \{0,1\}^N$, тогда

$$\psi_j(f', x) \geqslant \psi_j(f, x)$$

Свойство гласит, что если модель изменяется так, что вклад какого-либо признака увеличивается или остается неизменным независимо от других входных данных, то коэффициент воздействия признака не должен уменьшаться.

**Теорема** (Lundberg, 2017 [9]) Существует единственная объясняющая модель в классе методов с аддитивными значимостями признаков, которая удовлетворяет свойствам 1, 2, 3.

Коэффициенты объясняющей модели, являющиеся значимостями при-

знаков, определяются следующим образом:

$$\psi_i(f, x) = \sum_{z' \subseteq Z} \frac{|z'|!(N - |z'| - 1)!}{N!}(f_x(z') - f_x(z'_{\setminus i}), \tag{1}$$

$f_x(z') = f(h_x(z')) = E[f(z)|z_S]$,
$N$ — количество признаков,
$|z'|$ — количество ненулевых элементов в векторе $z'$,
$S$ — множество индексов ненулевых элементов в векторе $z'$

# Глава 3. Построение игровой модели локальной интерпретации

Далее будет изложено обобщение предыдущего подхода.

Пусть $f : \mathbb{X} \to \mathbb{R}$ — модель, отображающая $N$ — мерное пространство признаков $\mathbb{X}$ в вещественнозначные предсказания.

Аддитивные значимости признаков для $f(x)$ на определенном входе $\mathbf{x} = (x_1, \ldots, x_N) \in \mathbb{X}$ состоят из эталонной значимости $\phi_0$ и значимостей признаков $\phi = (\phi_1, \ldots, \phi_N)$, связанных с $N$ признаками следующим соотношением: $f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{N} \phi_i$.

На практике $\phi_0$ обычно представляет собой усредненный выход модели или выход модели для «базовых» входных данных (например, пустая строка для классификации тональности текста).

В настоящее время существует ряд методов для вычисления этих значимостей.

Чтобы объяснить предсказание модели, используя значения Шепли, необходимо сформулировать кооперативную игру с игроками, соответствующими признакам модели.

Пусть $D^{inp}$ — распределение, которое характеризует процесс, генерирующий входные данные модели, $\mathbf{x} = (x_1, \ldots, x_N)$ — вход, предсказание для которого необходимо интерпретировать, $\mathbf{x}_S = \{x_i : i \in S\}$ — вектор признаков, входящих в коалицию $S \subseteq N$, $\mathbf{r}$ — «эталонный» вход модели.

Комбинированным входом $z(\mathbf{x}, \mathbf{r}, S)$ будем называть следующий вектор:

$$z(\mathbf{x}, \mathbf{r}, S) = (z_1, \ldots, z_N), z_i = \begin{cases} x_i, i \in S \\ r_i, i \notin S \end{cases}$$

Заметим, что $z(\mathbf{x}, \mathbf{r}, \emptyset) = \mathbf{r}$, а $z(\mathbf{x}, \mathbf{r}, N) = \mathbf{x}$.

Функция выигрыша $v_{\mathbf{x}}$ в игре должна быть определена для каждого подмножества $S$ признаков так, чтобы $v_{\mathbf{x}}(S)$ отражала вклад $\mathbf{x}_S = \{x_i :$

$i \in S\}$ в прогноз модели. Это позволит нам вычислить возможные маргинальные вклады каждого признака в предсказание модели и получить значение Шепли.

Согласно определению аддитивных значимостей признаков и аксиоме эффективности вектора Шепли необходимо определить $v_{\mathbf{x}}(N) := f(\mathbf{x}) - \phi_0$ (выигрыш от полной коалиции должен быть разницей между интерпретируемым предсказанием модели и выходом модели для «базовых» или усредненных входных данных).

Определим характеристическую функцию игры следующим образом:

$$v_{\mathbf{x},D^{ref}}(S) = E_{\mathbf{R}\sim D^{ref}}[f(z(\mathbf{x},\mathbf{R},S))] - E_{\mathbf{R}\sim D^{ref}}[f(\mathbf{R})],$$

где $D^{ref}$ — распределение эталонных значений, $\mathbf{R}$ — случайная величина.

Имитация отсутствия признаков может быть реализована путем выборки отсутствующих признаков из условного распределения на основе значений существующих признаков. В таком случае характеристическая функция будет иметь вид:

$$v_x^{cond}(S) = E_{\mathbf{R}\sim D^{inp}}[f(z(\mathbf{x},\mathbf{R},S))|\mathbf{R}_S = \mathbf{x}_S] - E_{R\sim D^{inp}}[f(\mathbf{R})]$$

Значимости признаков, выраженные как значения Шепли для игры $(N, v_x^{cond}(S))$ соответствуют выражению 1.

В статье [8] было введено понятие *одноэталонной* игры — вспомогательной конструкцией для выражения вектора Шепли игры $v_{\mathbf{x},D^{ref}}$.

*Одноэталонная* игра $v_{\mathbf{x},\mathbf{r}}$ имитирует отсутствие признака, заменяя значение признака значением из определенного эталонного входа $\mathbf{r}$:

$$v_{\mathbf{x},\mathbf{r}}(S) = f(z(\mathbf{x},\mathbf{r},S)) - f(\mathbf{r})$$

Векторы Шепли игр $v_{\mathbf{x},\mathbf{r}}$ и $v_{\mathbf{x},D^{ref}}(S)$ связаны следующим соотношением:

$$\phi(v_{\mathbf{x},D^{ref}}) = E_{\mathbf{R}\sim D^{ref}}[\phi(v_{\mathbf{x},\mathbf{r}})] \tag{2}$$

Таким образом, значимости признаков можно посчитать, используя выражение 2.

Полученные значимости признаков можно суммировать по точечным прогнозам, чтобы получить глобальную значимость отдельной переменной, однако этот подход имеет серьезные недостатки: чрезмерное использованию отдельных наблюдений, отсутствие нормализованной меры для оценки относительной важности вклада каждой переменной, нестабильность при наличии аномалий данных, таких как поддельные данные, недостающие данные или выбросы.

# Глава 4. Модель глобальной интерпретации

Кривой Лоренца случайной величины $Y$ с математическим ожиданием $E(Y) = \mu$ называется графическое изображение функции

$$t \longrightarrow \mu^{-1} \int_0^t F_Y^{-1}(s)ds, 0 \leqslant t \leqslant 1,$$

где $F_Y^{-1} = \min\{y : F(y) \geqslant t\}$ — функция, обратная к функции распределения $F_Y$ случайной величины $Y$.

По данным, полученным в результате $n$ наблюдений, кривую Лоренца $L_Y$ случайной величины $Y$ можно определить как множество точек $\left(\frac{i}{n}, \sum_{j=1}^{i} \frac{y_{(i)}}{n\overline{y}}\right)$, $i = 1, \ldots, n$, где $y_{(i)}$ упорядоченные по неубыванию значения переменной $Y$, а $\overline{y}$ — выборочное среднее. Аналогично, упорядочивая значения переменной $Y$ в порядке невозрастания, получаем кривую $L'_Y$, двойственную к кривой Лоренца, которая определяется как множество точек $\left(\frac{i}{n}, \sum_{j=1}^{i} \frac{y_{(n+1-i)}}{n\overline{y}}\right)$.

Обобщением кривой Лоренца в $d$-мерном пространстве является так называемый зоноид Лоренца.

Рассмотрим множество $\mathbf{Y}^{d+}$ случайных величин в $\mathbb{R}^d$, имеющих конечные и положительные (в каждой компоненте) математические ожидания, а также подмножество $\mathbf{Y}_+^{d+} \subset \mathbf{Y}^{d+}$ векторов, содержащихся в $\mathbb{R}_+^d$.

Для $Y \in \mathbf{Y}^d$ введем обозначение

$$\tilde{\mathbf{Y}} = \left(\frac{Y_1}{E(Y_1)}, \ldots, \frac{Y_d}{E(Y_d)}\right)$$

Зоноид Лоренца случайного вектора $Y \in \mathbb{Y}^d$ — это выпуклый компакт в $\mathbb{R}^{d+1}$, определяемый следующим образом:

$$LZ(\tilde{\mathbf{Y}}) = \{E[g(\tilde{\mathbf{Y}}), g(\tilde{\mathbf{Y}})\tilde{\mathbf{Y}}], g : \mathbb{R}^d \to [0, 1]\},$$

где $g$ — измеримая функция.

Авторы статьи [6] показали, что зоноид Лоренца случайного вектора однозначно определяет его распределение, а также обладает свойством маргинальности и аддитивности.

Зоноид Лоренца можно использовать для оценки вклада независимых переменных в изменение переменной отклика.

Пусть $LZ_{d=1}(Y)$ — зоноид Лоренца переменной отклика $Y$, а $X_1$ — независимая переменная, такая, что $\hat{Y}_{X_1}$ — вектор оценочных значений, вычисленных с помощью модели линейной регрессии: $\hat{Y}_{X_1} = \hat{\alpha} + \hat{\beta}X_1$. Обозначим за $LZ_{d=1}(\hat{Y_{X_1}})$ зоноид Лоренца для $\hat{Y}_{X_1}$. Рассмотрим дополнительную независимую переменную $X_2$ и соответствующую модель линейной регрессии: $\hat{Y}_{X_2} = \hat{\alpha} + \hat{\beta}X_2$, обозначим за $LZ_{d=1}(\hat{Y}_{X_2})$ зоноид Лоренца для $\hat{Y_{X_2}}$.

В работе [4] было показано, что в одномерном случае зоноид Лоренца может быть выражен с помощью оператора ковариации:

$$LZ_{d=1}(Y) = \frac{2Cov(Y, F(Y))}{\mu},$$

где $\mu$ — математическое ожидание $Y$, а $F(Y)$ — функция распределения $Y$.

Таким образом, зоноиды Лоренца для $\hat{Y}_{X_1}$ и $\hat{Y}_{X_2}$ могут быть выражены следующим образом:

$$LZ_{d=1}(\hat{Y}_{X_1}) = \frac{2Cov(\hat{Y}_{X_1}, F(\hat{Y}_{X_1}))}{\mu},$$

$$LZ_{d=1}(\hat{Y}_{X_2}) = \frac{2Cov(\hat{Y}_{X_2}, F(\hat{Y}_{X_2}))}{\mu},$$

где $E(\hat{Y}_{X_1}) = E(E(Y|\hat{Y}_{X_1})) = \mu$, $E(\hat{Y}_{X_2}) = E(E(Y|\hat{Y}_{X_2})) = \mu$, $F(\hat{Y}_{X_1})$ и $F(\hat{Y}_{X_2})$ — функции распределения $\hat{Y}_{X_1}$ и $\hat{Y}_{X_2}$ соответсвенно.

Пусть $r(Y)$, $r(\hat{Y}_{X_1})$ и $r(\hat{Y}_{X_2})$ — ранговые оценки, соответствующие переменным $Y$, $\hat{Y}_{X_1}$ и $\hat{Y}_{X_2}$. Поскольку $r(\cdot)$ определяет эмпирическое пред-

ставление $F(\cdot) = \frac{r(\cdot)}{n}$, то:

$$LZ_{d=1}(Y) = \frac{2Cov(Y, r(Y))}{n\mu},$$

$$LZ_{d=1}(\hat{Y}_{X_1}) = \frac{2Cov(\hat{Y}_{X_1}, r(\hat{Y}_{X_1}))}{n\mu},$$

$$LZ_{d=1}(\hat{Y}_{X_2}) = \frac{2Cov(\hat{Y}_{X_2}, r(\hat{Y}_{X_2}))}{n\mu}.$$

Пусть имеется $n$ экземпляров данных, тогда:

$$LZ_{d=1}(y) = \frac{2Cov(y, r(y))}{n\overline{y}},$$

$$LZ_{d=1}(\hat{y}_{x_1}) = \frac{2Cov(\hat{y}_{x_1}, r(\hat{y}_{x_1}))}{n\overline{y}},$$

$$LZ_{d=1}(\hat{y}_{x_2}) = \frac{2Cov(\hat{y}_{x_2}, r(\hat{y}_{x_2}))}{n\overline{y}},$$

где $y$, $\hat{y}_{x_2}$ и $\hat{y}_{x_2}$ — векторы наблюдаемых и вычисленных значений, $\overline{y}$ — выборочное среднее.

Авторы работы [4] показали, что зоноид Лоренца является функцией суммы расстояний между значениями точек, лежащих на кривой Лоренца, и значениями точек, лежащих на биссектрисной кривой:

$$LZ_{d=1}(y) = \frac{2}{n\overline{y}} \left[ \frac{1}{n} \sum_{i=1}^{N} i y_{(i)} - \frac{n(n+1)}{2n} \overline{y} \right]$$

На основе зоноидов Лоренца были выведены следующие меры взаимосвязи переменных:

*Маргинальный вклад Джини* (MGC) — мера, позволяющая измерить абсолютную объясняющую мощность любой отдельной объясняющей переменной. Пусть $X_j$ — одна из $h$ объясняющих переменных ($j = 1, \ldots, h$).

Маргинальный вклад, вносимый объясняющей переменной $X_j$, определяется следующим образом:

$$MGC_{Y|X_j} = \frac{LZ_{d=1}(\hat{Y}_{X_j})}{LZ_{d=1}(Y)}$$

Чтобы понять, могут ли дополнительные переменные улучшить модель, необходимо определить меру частичного вклада объясняющей переменной.

Пусть $Y$ — переменная отклика, а $X_1, \ldots, X_h$ — набор объясняющих переменных. Чтобы оценить взаимосвязь между перменной отклика и независимыми переменными можно применить модель машинного обучения и получить соответствующие прогнозируемые значения $\hat{Y}_{X_1,\ldots,X_h}$.

Зоноиды Лоренца $Y$ и $\hat{Y}_{X_1,\ldots,X_h}$ будут представимы в следующем виде:

$$LZ_{d=1}(Y) = \frac{2Cov(Y, r(Y))}{n\mu},$$

$$LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_h}) = \frac{2Cov(\hat{Y}_{X_1,\ldots,X_h}, r(\hat{Y}_{X_1,\ldots,X_h}))}{n\mu}.$$

Отметим, что зоноид Лоренца обладает следующим свойством:

$$LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}}) \geqslant LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})$$

Пусть имеется $n$ экземпляров данных, тогда зоноиды Лоренца вектора наблюдаемых значений $y$ и предсказанных значений $\hat{y}_{x_1,\ldots,x_h}$ будут представимы в следующем виде:

$$LZ_{d=1}(y) = \frac{2Cov(y, r(y))}{n\overline{y}},$$

$$LZ_{d=1}(\hat{y}_{x_1,\ldots,x_h}) = \frac{2Cov(\hat{y}_{x_1,\ldots,x_h}, r(\hat{y}_{x_1,\ldots,x_h}))}{n\overline{y}}.$$

Пусть $\hat{Y}_{X_1,\ldots,X_h}$ — прогноз модели, включающей все объсняющие переменные, а $\hat{Y}_{X_1,\ldots,X_{h-1}}$ — прогноз модели, исключающей объясняющую переменную $X_h$.

*Частичный вклад Джини* (PGC) — вклад, связанный с включением объясняющей переменной, определяющийся следующим образом:

$$PGC_{Y,X_h|X_1,\ldots,X_{h-1}} = \frac{LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_h}) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{h-1}})}{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{h-1}})}$$

В статье [8] предложен следующий подход.

Зададим выигрыш числителем меры PGC:

$$p_{off} = LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_h}) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{h-1}})$$

Тогда вклад $X_k$-го признака в модель рассчитывается по формуле:

$$LZ^{X_k}_{d=1}(\hat{Y}) = \sum_{X'\setminus X_k} \frac{|X'|!(K-|X'|-1)!}{K!}(LZ_{d=1}(\hat{Y}_{X'\cup X_k}) - LZ_{d=1}(\hat{Y}_X)$$

Пусть имеется $n$ экземпляров данных, тогда:

$$LZ^{X_k}_{d=1}(y) = \sum_{X'\setminus X_k} \frac{|X'|!(K-|X'|-1)!}{K!}(LZ_{d=1}(\hat{y}_{X'\cup X_k}) - LZ_{d=1}(\hat{y}_{X'}) =$$

$$= \sum_{X'\setminus X_k} \frac{|X'|!(K-|X'|-1)!}{K!}\left(\frac{2}{n^2\overline{y}}\sum_{i=1}^{n} i(\hat{y}_{X'\cup X_k}(i) - \hat{y}_{X'}(i)\right)$$

В то время как стандартное разложение Шепли отражает вклады признаков на локальном уровне (для конкретного входа данных), разложение на основе зоноидов Лоренца описывает вклад признаков на глобальном уровне: дается описание модели с точки зрения объясняющих переменных, которые в основном определяют ее прогноз, кроме того данная мера выбора модели является нормированной.

# Глава 5. Детали реализации

## 5.1 Модель

Для иллюстрации методов интерпретации был использован набор данных Ирисов Фишера, состоящий из данных о 150 экземплярах ириса, по 50 экземпляров трех видов — Ирис щетинистый (англ. Iris setosa), Ирис виргинский (англ. Iris virginica) и Ирис разноцветный (англ. Iris versicolor). Для каждого экземпляра измерялись четыре характеристики (в сантиметрах):
$X_1$ — длина наружной доли околоцветника (чашелистика) (англ. sepal length);
$X_2$ — ширина наружной доли околоцветника (чашелистика) (англ. sepal width);
$X_3$ — длина внутренней доли околоцветника (лепестка) (англ. petal length);
$X_4$ — ширина внутренней доли околоцветника (лепестка) (англ. petal width).

На основании этого набора данных требуется построить правило классификации, определяющее вид растения по данным измерений. Это задача многоклассовой классификации.

**Таблица 1:** Пример экземпляров данных

| Длина чашелистика | Ширина чашелистика | Длина лепестка | Ширина лепестка | Вид ириса |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 5.6 | 2.5 | 3.9 | 1.1 | versicolor |
| 5.9 | 3.0 | 5.1 | 1.8 | virginica |

Выборка была разбита на обучающую (120 экземпляров данных) и тестовую (30 экземпляров данных).

В качестве алгоритма машинного обучения использовался основанный на методе опорных векторов SVM-классификатор, принимающий на

вход вектор признаков и возвращающий вектор вероятностей принадлежности к классам. Была достигнута точность предсказания в $95,833\%$ на тестовой выборке.

## 5.2 Интерпретация модели

Рассмотрим экземпляр данных $\hat{X}$ из тестовой выборки со следующими значениями признаков:

**Таблица 2:** Значения признаков экземпляра $\hat{X}$

| Длина чашелистика | Ширина чашелистика | Длина лепестка | Ширина лепестка |
|---|---|---|---|
| 5.8 | 2.8 | 5.1 | 2.4 |

Выход модели для $\hat{X}$:

**Таблица 3:** Спрогнозированные вероятности принадлежности к классам

| setosa | versicolor | virginica |
|---|---|---|
| 0.00927886 | 0.03090196 | 0.95981919 |

Усредненный выход модели, рассчитанный по обучающей выборке данных:

**Таблица 4:** Усредненные вероятности принадлежности к классам

| setosa | versicolor | virginica |
|---|---|---|
| 0.32137189 | 0.31069974 | 0.36792837 |

Значимости признаков, рассчитанные по выходу модели на входе $\hat{X} = \{5.8, 2.8, 5.1, 2.4\}$:

**Таблица 5:** Таблица значимостей признаков для входа $\hat{X}$ в соответствии с классами

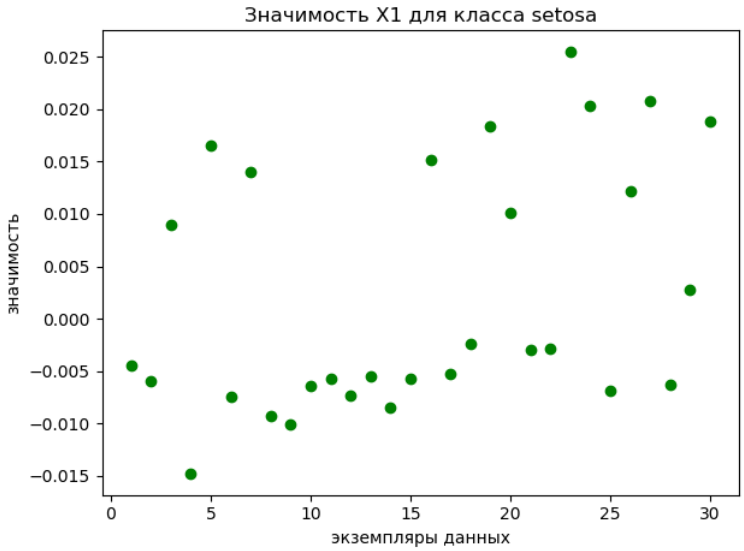
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| setosa | -4.41011371e-03 | -6.47742433e-03 | -2.60804036e-01 | -4.05337096e-02 |
| versicolor | 1.53848955e-03 | 1.13357938e-03 | -1.83453988e-02 | -2.61321665e-01 |
| virginica | 2.87162416e-03 | 5.34384494e-03 | 2.79149435e-01 | 3.01855375e-01 |

Чем больше по модулю значимость признаков, тем больший вклад признак вносит в прогнозирование наблюдаемого выхода модели. Отрицательность значимости говорит о том, что признак способствует отдалению прогнозируемой вероятности от усредненной вероятности, рассчитанной по обучающей выборке данных, к нулю.

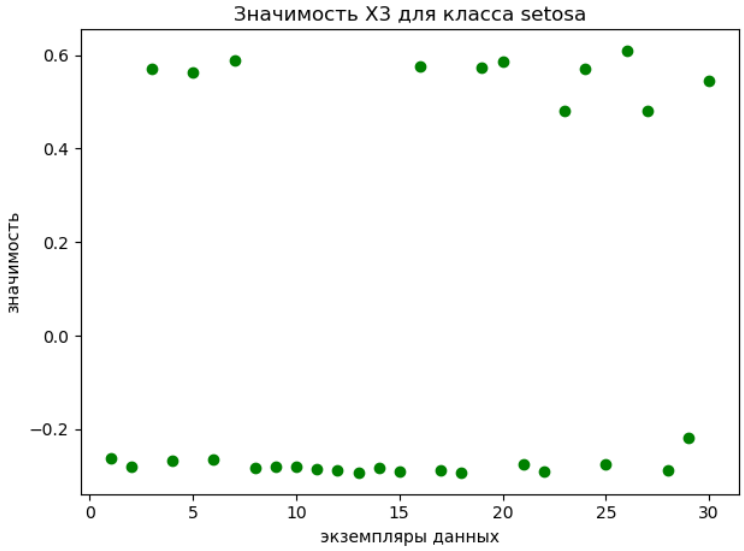
По получившимся данным, можно сделать вывод, что наибольший вклад в прогнозирование высокой вероятности ($\approx 95,98\%$) принадлежности к классу virginica для данного входа $\hat{X}$ внес признак $X_4$ (ширина лепестка).

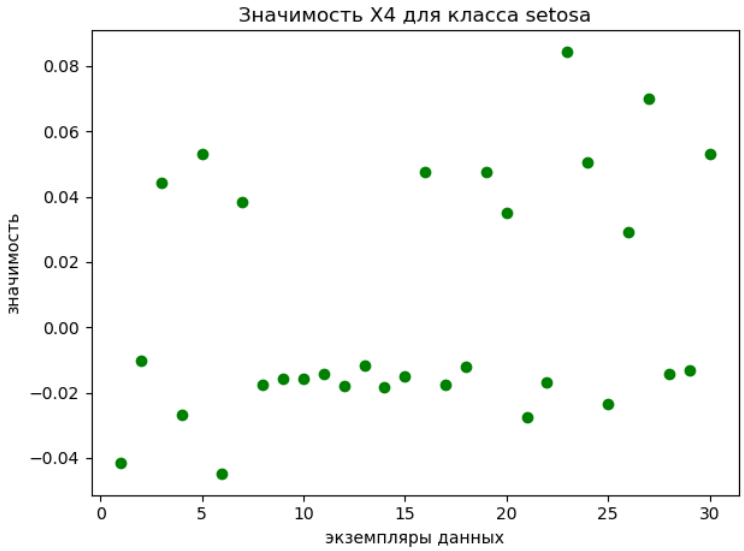
Наибольший вклад в прогнозировании более низкой вероятности ($\approx 30,9\%$) принадлежности к классу setosa для данного входа $\hat{X}$ внес признак $X_3$ (длина лепестка), а к классу versicolor — $X_4$ (ширина лепестка).

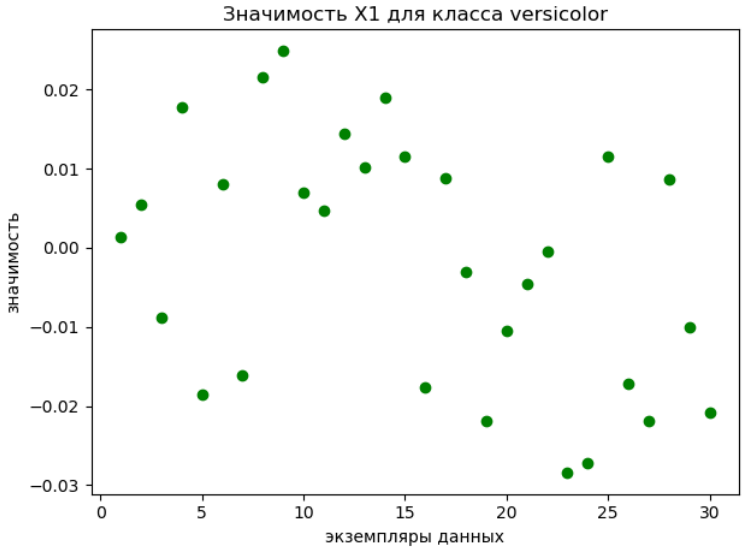
Значимости признаков для тестовых экземпляров данных в соответствии с прогнозами для классов представлены на графиках:

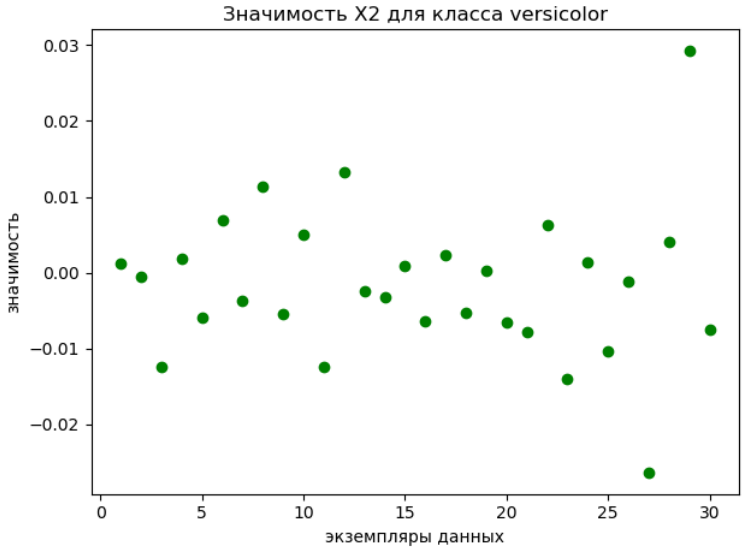

Значимость X1 для класса setosa
значимость
экземпляры данных

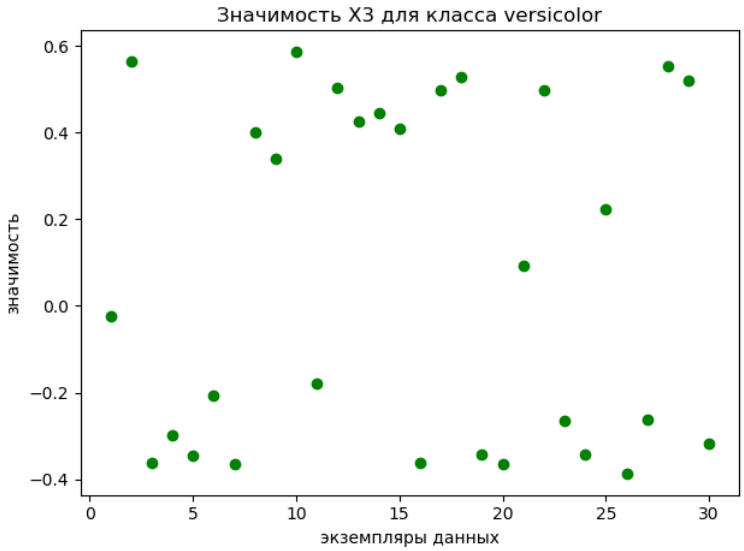

Значимость X2 для класса setosa
значимость
экземпляры данных

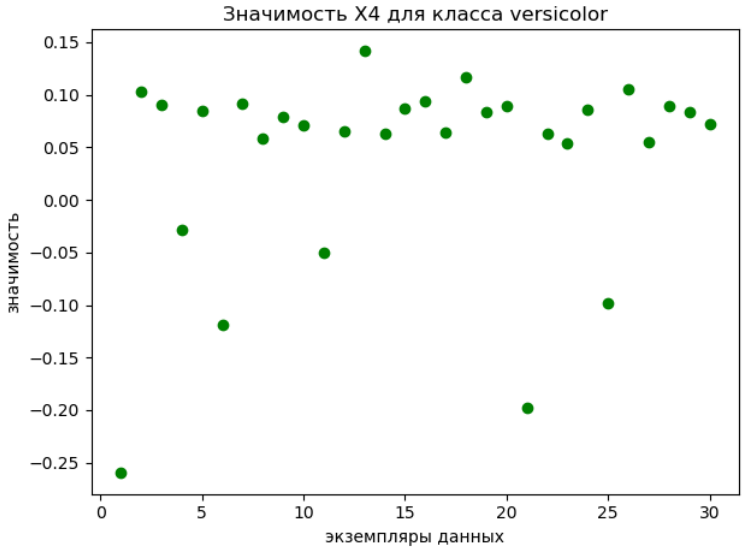

Значимость X3 для класса setosa
значимость
экземпляры данных

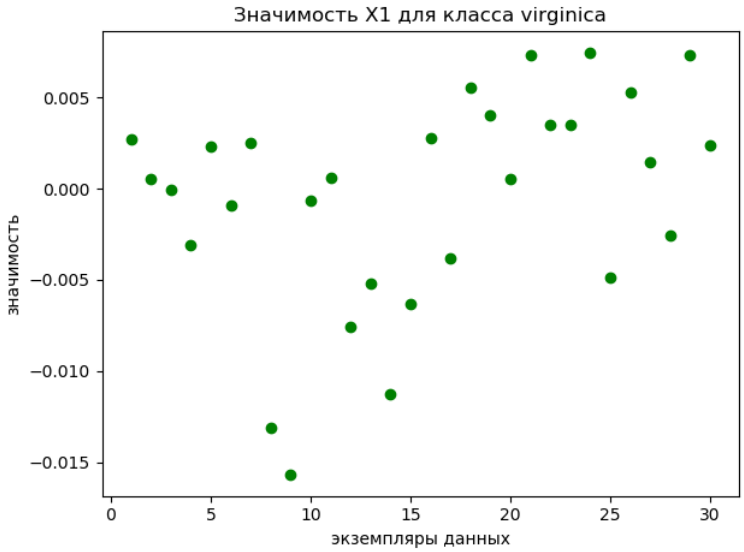

Значимость X4 для класса setosa
значимость
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
0
5
10
15
20
25
30
экземпляры данных

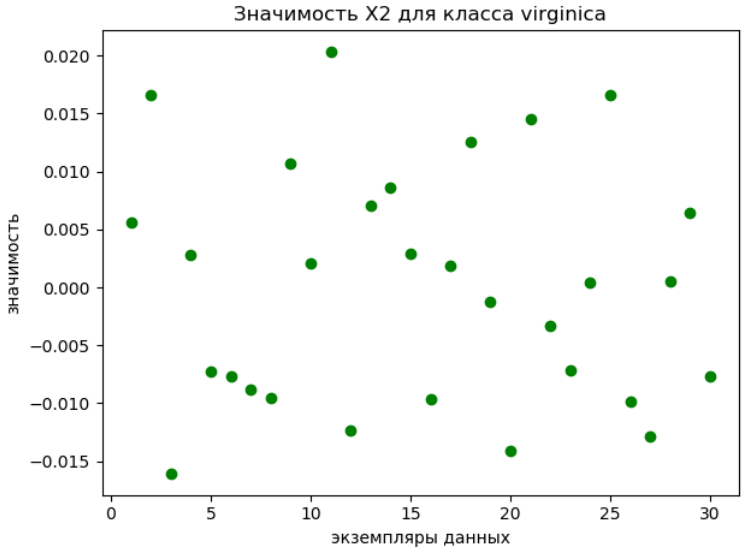

Значимость X1 для класса versicolor
значимость
0.02
0.01
0.00
−0.01
−0.02
−0.03
0
5
10
15
20
25
30
экземпляры данных

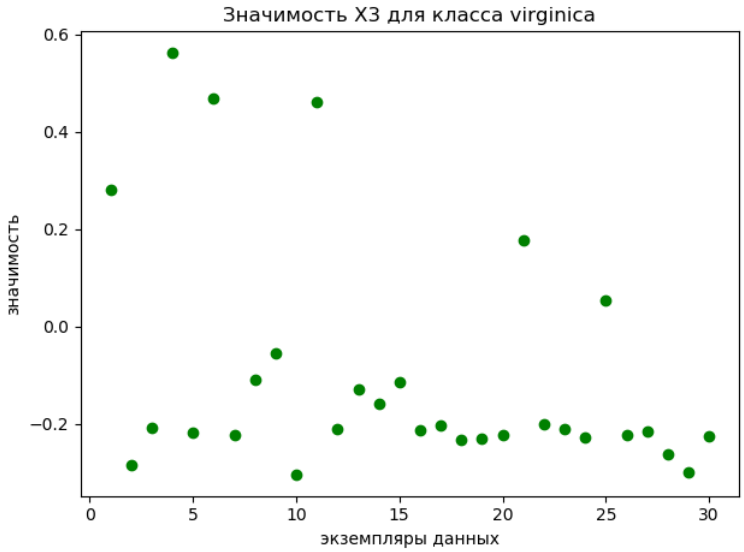

Значимость X2 для класса versicolor
значимость
0.03
0.02
0.01
0.00
−0.01
−0.02
0
5
10
15
20
25
30
экземпляры данных

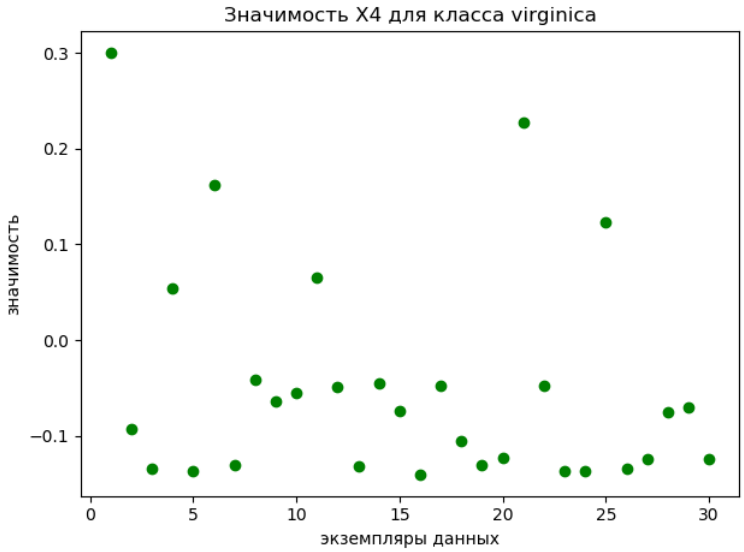

Значимость X3 для класса versicolor
значимость
экземпляры данных


Значимость X4 для класса versicolor
значимость
экземпляры данных


Значимость X1 для класса virginica
значимость
экземпляры данных




Рассчитав значимости признаков по прогнозам для всей тестовой выборки и проссумировав их абсолютные значения для каждого признака в

соответствии с классом, получаем:

**Таблица 6:** Таблица значимостей признаков в соответствии с классами

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| setosa | 0.29738504472012117 | 0.3077126579482167 | 11.439710277294981 | 0.9053557269061567 |
| versicolor | 0.4066588376028205 | 0.2156649317406732 8 | 11.059757495621232 | 2.8098600503706965 |
| virginica | 0.1351111885837297 | 0.2568043083774176 | 7.004344896839869 | 3.303427750600525 |

Далее представлены результаты реализации глобального метода интерпретации, основанного на зоноидах Лоренца:

**Таблица 7:** Таблица значимостей признаков в соответствии с классами

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| setosa | 0.01883331861407688 | 0.0118696898744596 98 | 0.5599857258942584 | 0.0562932218519 8023 |
| versicolor | 0.0073289070459745 | 0.0044674388265004 84 | 0.5131777943528874 | 0.1271046282408 2264 |
| virginica | 0.1351111885837297 | 0.2568043083774176 | 0.3043668968999869 | 0.1301063282408 2264 |

Отметим, что наибольшую значимость имеет признак $X_3$ (длина лепестка) для всех классов, вероятность принадлежности к которым прогнозирует модель, следующий по значимости — признак $X_4$ (ширина лепестка), признаки $X_1$ (длина чашелистика) и $X_2$ (ширина чашелистика) имеют малые

значимости и в зависимости от классов значимость одного становится больше другого.

# Заключение

В рамках данной работы были рассмотрены локальные и глобальные методы интерпретации моделей машинного обучения, универсальность которых заключается в том, что они не зависят от выбора модели.

Для класса методов с аддитивными значимостями признаков была формализована кооперативная игра и выведены значимости признаков как компоненты вектора Шепли.

Была произведена программная реализация выявленных подходов, основывающихся как на разложении локальных прогнозов модели, так и на разложении общей точности предсказания модели. Было проиллюстрировано на конкретной модели машинного обучения применение двух методов глобальной интерпретации, первый из которых основан на суммировании абсолютных значений локальных значимостей (описан в Главе 3), а второй основан на разложении с использованием зоноидов Лоренца (описан в Главе 4).

# Список литературы


[1] Акимов, W. Kerby. Значения для кооперативных игр. Обобщение теоремы единственности Шепли.// Искусственный интеллект и прянятие решений, №4. 2010. С. 77-80.

[2] Aas K., Jullum M.,// Anders Løland Explaining individual predictions when features are dependent: More accurate approximations to Shapley values.

[3] Giudici, P., Raffinetti, E.// Shapley-Lorenz Decompositions in eXplainable Artificial Intelligence (2020).

[4] Giudici, P., Raffinetti, E. //Lorenz Model Selection. J Classif 37, 754–768 (2020).

[5] Giudici, P., Raffinetti, E.// Shapley-Lorenz eXplainable Artificial Intelligence, (2021).

[6] Koshevoy, G., Mosler, K.// Multivariate Lorenz dominance based on zonoids . AStA 91, 57–76 (2007).

[7] Lundberg, S. M. and Lee, S.-I.// A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, pp. 4765–4774, 2017.

[8] Merrick L., Taly A. // The Explanation Game: Explaining Machine Learning Models Using Shapley Values.

[9] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, Su-In Lee// Explainable AI for Trees: From Local Explanations to Global

Understanding, c.64, (2020).

[10] Shapley LS // A value forn-person games. In: Kuhn HW, Tucker AW(eds) Contributions to the Theory of Games II, Annals of MathematicsStudies, vol 28, Princeton University Press, Princeton, pp 307–317, (1953).

[11] Shrikumar A., Greenside P., // Anshul Kundaje Learning Important Features Through Propagating Activation Differences, (2019).

[12] Shrikumar, A., Greenside, P., Shcherbina, A. & Kundaje, A.// Not just a black box: learning important features through propagating activation differences, (2016).

[13] Strumbelj E., Kononenko I.// An efficient explanation of individual classifications using game theory //The Journal of Machine Learning Research. – 2010. – T. 11. – C. 1-18.