Практический курс Machine Learning
Шкалирование, важность и интерпретируемость признаков

Игнатов Дмитрий Игоревич

[1] ML Instructor, BigData Team
[2] Национальный исследовательский университет Высшая школа экономики
Факультет компьютерных наук
Департамент анализа данных и искусственного интеллекта

2019

Обзор методов

На этой лекции:

1. Шкалирование признаков
2. Важность и интерпретируемость признаков

План лекции

Шкалирование признаков

Определение

Шкалирование признаков – разновидность преобразования признаков, заключающаяюся в приведении их значений к некоторой шкале, например, органиченному диапазону или множеству значений.

Примеры

- номинальное шкалирование (jargon: one-hot encoding)
- min-max нормализация (min-max normalization)
- стандартизованная оценка (z-score normalization)
- десятичное масштабирование (decimal scaling)
- ...

Номинальное шкалирование

jargon: one-hot encoding

$$f : Cat_A \to \{0,1\}^{|Cat_A|}, \text{ где}$$

Cat_A – множество значений шкалируемого признака A.

Пример

$Cat_A = \{red, blue, yellow\}$

- $f : red \mapsto (1,0,0)$
- $f : blue \mapsto (0,1,0)$
- $f : yellow \mapsto (0,0,1)$

Номинальное шкалирование

jargon: one-hot encoding

$$f : Cat_A \to \{0,1\}^{|Cat_A|}, \text{ где}$$

Cat_A – множество значений шкалируемого признака A.

Пример

$Cat_A = \{red, blue, yellow\}$

- $f : red \mapsto (1,0,0)$
- $f : blue \mapsto (0,1,0)$
- $f : yellow \mapsto (0,0,1)$

Q: Зачем?

Номинальное шкалирование

jargon: one-hot encoding

$$f : Cat_A \to \{0,1\}^{|Cat_A|}, \text{ где}$$

Cat_A – множество значений шкалируемого признака A.

Пример

$Cat_A = \{red, blue, yellow\}$

- $f : red \mapsto (1,0,0)$
- $f : blue \mapsto (0,1,0)$
- $f : yellow \mapsto (0,0,1)$

Q: Зачем?

A: Кодировать несравнимые значения из конечного множества.

Min-max нормализация

min-max normalization

Дан признак A, известны его минимальное и максимальное значения, min_A и max_A, тогда его значение v отображается в $v' \in [new\_min_A, new\_max_A]$.

$$v' = \frac{v - min_A}{max_A - min_A}(new\_min_A - new\_max_A) + new\_min_A.$$

Cat_A – множество значений шкалируемого признака A.

Пример

Минимальное и максимальное значение признака income в выборке 12000 и 98000, соответственно. Новый диапазон: [0,1]. Значение признака 73600 будет преобразовано как

$$\frac{73600 - 12000}{98000 - 12000}(1 - 0) + 0 = 0,716.$$

Стандартизованная оценка (z-score normalization)

zero-mean normalization

Дан признак A, известны его среднее значение, $\overline{A}$ и стандартное отклонение σ_A, тогда значение этого признака v отображается в v'

$$v' = \frac{v - \overline{A}}{\sigma_A}.$$

Пример

Среднее значение и стандартное отклонение признака income в выборке 54000 и 16000, соответственно. Значение признака 73600 будет преобразовано как

$$\frac{73600 - 54000}{16000} = 1,225.$$

Стандартизованная оценка (z-score normalization)

zero-mean normalization

Дан признак A, известны его среднее значение, $\overline{A}$ и стандартное отклонение σ_A, тогда значение этого признака v отображается в v'

$$v' = \frac{v - \overline{A}}{\sigma_A}.$$

Пример

Среднее значение и стандартное отклонение признака income в выборке 54000 и 16000, соответственно. Значение признака 73600 будет преобразовано как

$$\frac{73600 - 54000}{16000} = 1,225.$$

Q: Каковы среднее и стандартное отклонение отшкалированного признака по всей выборке?

Стандартизованная оценка (z-score normalization)

zero-mean normalization

Дан признак A, известны его среднее значение, $\overline{A}$ и стандартное отклонение σ_A, тогда значение этого признака v отображается в v'

$$v' = \frac{v - \overline{A}}{\sigma_A}.$$

Пример

Среднее значение и стандартное отклонение признака income в выборке 54000 и 16000, соответственно. Значение признака 73600 будет преобразовано как

$$\frac{73600 - 54000}{16000} = 1,225.$$

Q: Каковы среднее и стандартное отклонение отшкалированного признака по всей выборке?
A: 0 и 1, соответственно.

Десятичное масштабирование

decimal scaling

Дан признак A, тогда его значение v отображается в v'

$$v' = \frac{v}{10^j}, \text{ где}$$

j – наименьшее целое, такое что $Max(|v'|) < 1$.

Пример

Предположим, что $A \in [-986, 917]$, тогда его максимальное абсолютное значение равно 986. Для нормализации необходимо каждое значение разделить на 1000 ($j = 3$), таким образом $A_{new} \in [-0,986, 0,917]$.

Десятичное масштабирование

decimal scaling

Дан признак A, тогда его значение v отображается в v'

$$v' = \frac{v}{10^j}, \text{ где}$$

j – наименьшее целое, такое что $Max(|v'|) < 1$.

Пример

Предположим, что $A \in [-986, 917]$, тогда его максимальное абсолютное значение равно 986. Для нормализации необходимо каждое значение разделить на 1000 ($j = 3$), таким образом $A_{new} \in [-0,986, 0,917]$.

Q: Каковы среднее и стандартное отклонение отшкалированного признака по всей выборке?

Десятичное масштабирование

decimal scaling

Дан признак A, тогда его значение v отображается в v'

$$v' = \frac{v}{10^j}, \text{ где}$$

j – наименьшее целое, такое что $Max(|v'|) < 1$.

Пример

Предположим, что $A \in [-986, 917]$, тогда его максимальное абсолютное значение равно 986. Для нормализации необходимо каждое значение разделить на 1000 ($j = 3$), таким образом $A_{new} \in [-0,986, 0,917]$.

Q: Каковы среднее и стандартное отклонение отшкалированного признака по всей выборке?
A: 0 и 1, соответственно.

Резюме

Цели шкалирования

- Преобразование значения категориальных признаков в числовые.
- Предотвращение “перевешивания” значений одних признаков с большим диапазоном значений вклада других в метрических алгоритмах, например, доход в рублях и возраст в годах (или бинарный признак) для метода ближайших соседей в задаче классификации.
- Ускорение фазы обучения, например, при обучении нейронных сетей.
- ...

План лекции

1. Шкалирование признаков
2. Важность и интерпретируемость признаков

Важность признаков в линейных моделях

Пример задачи регрессии

Источник: C. Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2019

Признаки: park-nearby, cat-banned, area-50 и floor-2nd

Прогноз: 300000 евро

BIGDATA TEAM

Важность признаков в линейных моделях

Пример задачи регрессии

Известна средняя цена по выборке: 310000 евро.

Предсказание линейной модели:

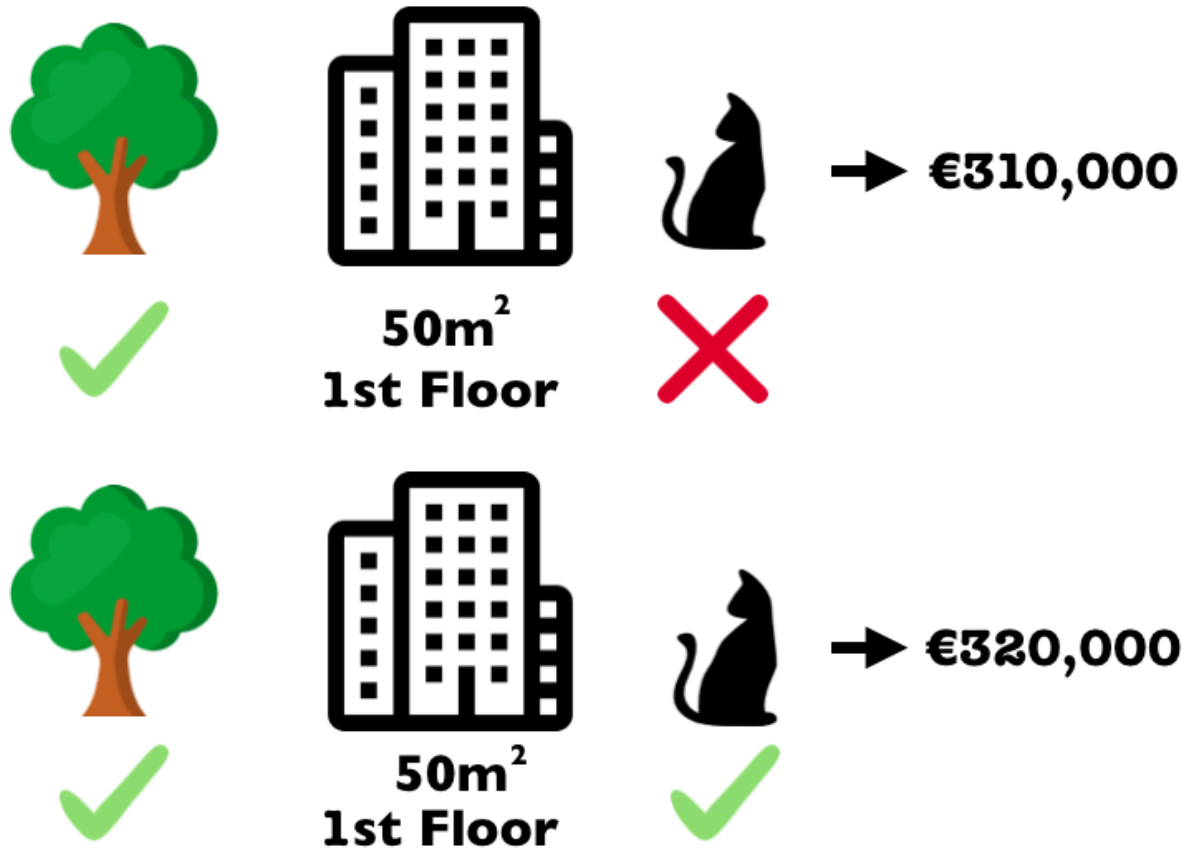
$$310000 + 30000 \cdot park\_nearby - 50000 \cdot cat\_banned + \\ + 10000 \cdot area\_50 + 0 \cdot floor\_2nd = 300000$$

Переменные бинарные, а их вклады заданы весами модели.

Значение Шепли: как оценить вклад одного признака?

Пример

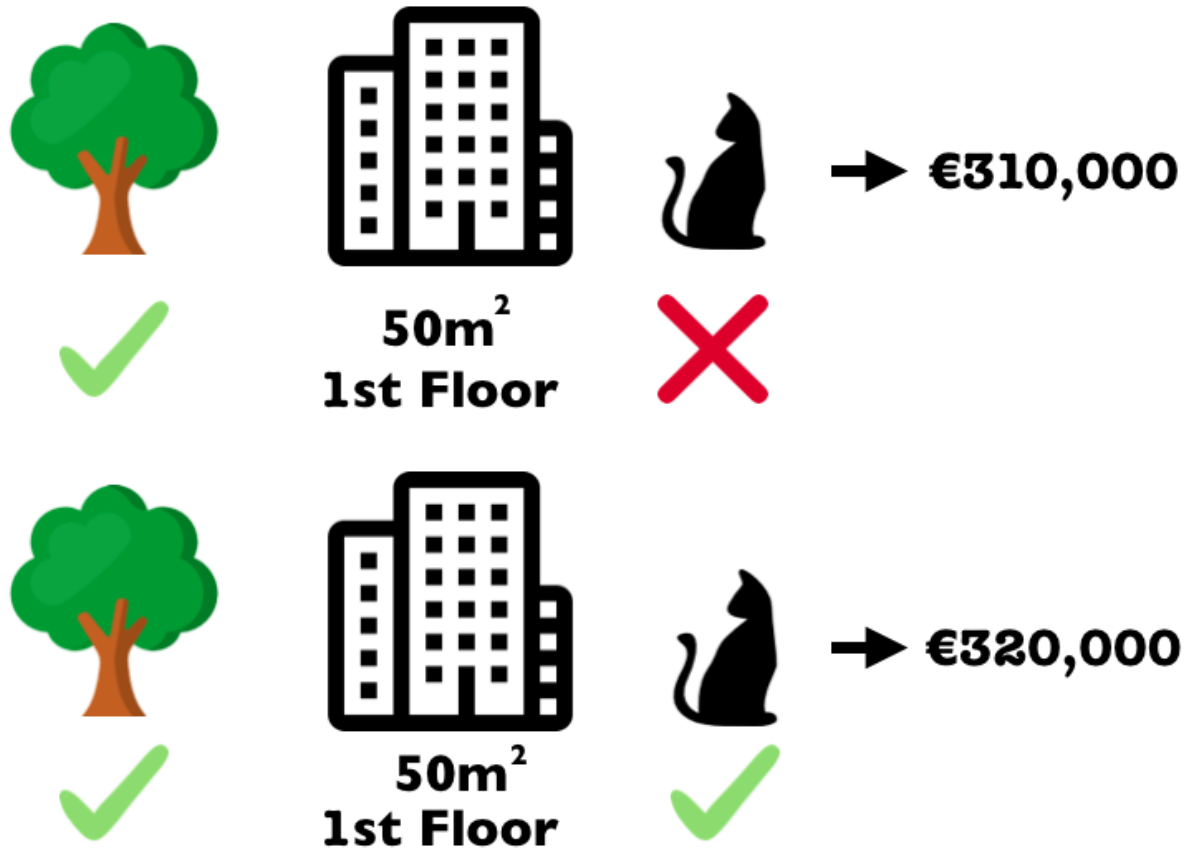
Игра: “коалиция” признаков – {park_nearby, area_50 }.
Q: Каков вклад *cat_banned* в выигрыш – прогноз?



Значение Шепли: как оценить вклад одного признака?

Пример

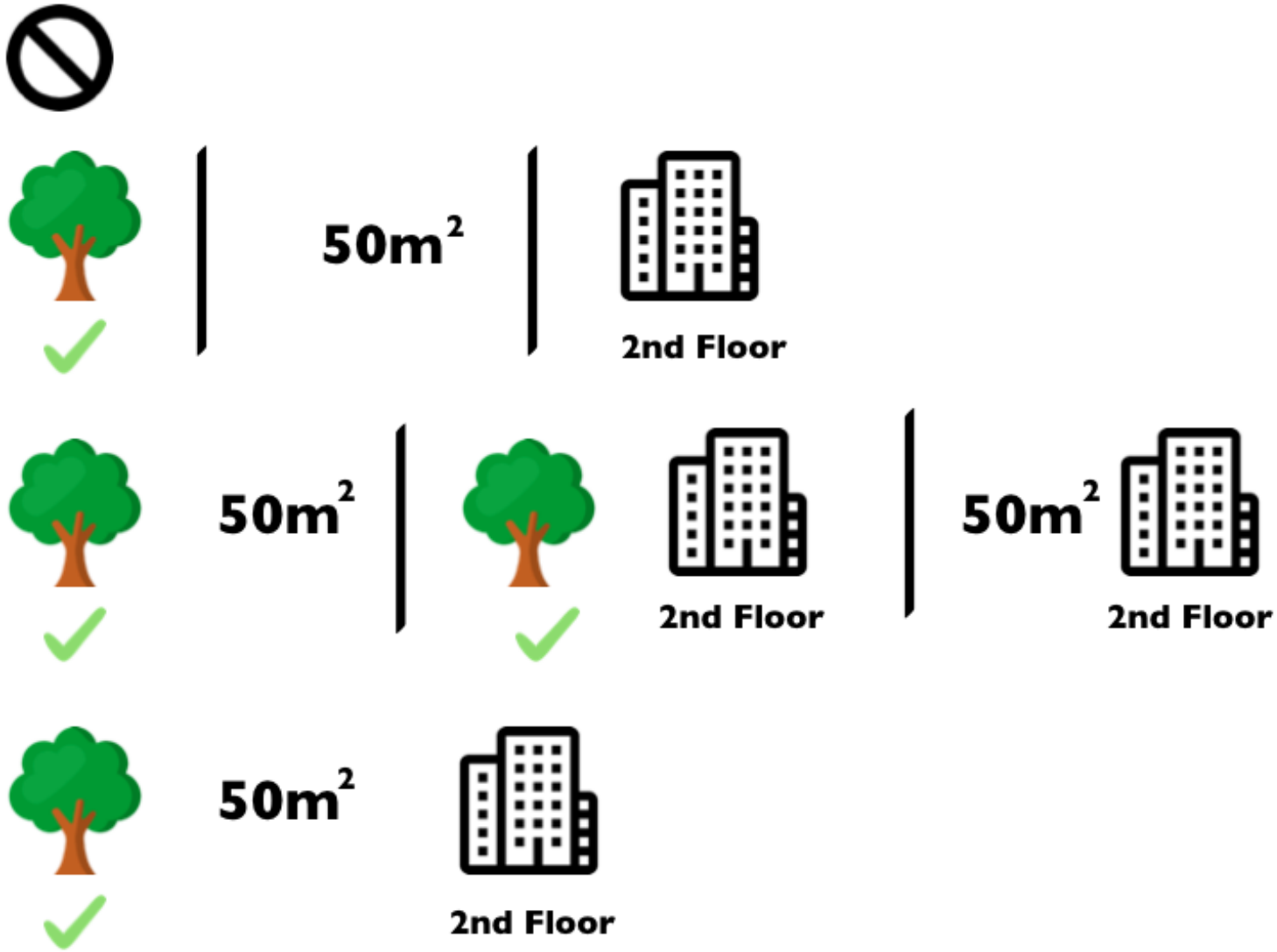
Игра: “коалиция” признаков – {park_nearby, area_50 }.
Q: Каков вклад *cat_banned* в выигрыш – прогноз?



A: 310000 - 320000 = – 10000 евро

Значение Шепли: коалиции

Пример



Значение Шепли: усреднение по коалициям

Пример

Вычисляем вклад как разность предсказаний с признаком cat-banned и без для каждой коалиции.

- No feature values
- park-nearby
- size-50
- floor-2nd
- park-nearby, size-50
- park-nearby, floor-2nd
- size-50, floor-2nd
- park-nearby, size-50, floor-2nd

Значение Шепли – среднее по всем таким вкладам.

Значение Шепли

Shapley value

Прогноз модели для объекта x:

$$\hat{f}(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Вклад признака j в прогноз:

$$\phi_j(\hat{f}) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j)$$

Вклад всех признаков:

$$\begin{aligned}\sum_{j=1}^{p} \phi_j(\hat{f}) &= \sum_{j=1}^{p} (\beta_j x_j - E(\beta_j X_j)) \\ &= (\beta_0 + \sum_{j=1}^{p} \beta_j x_j) - (\beta_0 + \sum_{j=1}^{p} E(\beta_j X_j)) \\ &= \hat{f}(x) - E(\hat{f}(X))\end{aligned}$$

Значение Шепли

Shapley value [Lloyd S. Shapley, 1953]

Q: Как посчитать вклад признака в прогноз для любой модели?

Значение Шепли

Shapley value [Lloyd S. Shapley, 1953]

Q: Как посчитать вклад признака в прогноз для любой модели?

A: Значение Шепли из кооперативной теории игр.

Значение Шепли

Shapley value [Lloyd S. Shapley, 1953]

Q: Как посчитать вклад признака в прогноз для любой модели?

A: Значение Шепли из кооперативной теории игр.

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{x_j\}\right) - val(S)\right)$$

Значение Шепли

Shapley value [Lloyd S. Shapley, 1953]

Q: Как посчитать вклад признака в прогноз для любой модели?

A: Значение Шепли из кооперативной теории игр.

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\dots,x_p\} \setminus \{x_j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{x_j\}\right) - val(S)\right)$$

Значение Шепли – это средний вклад признака в предсказание по различным коалициям. Важно: это не разность в прогнозах, когда мы удаляем этот признак из модели.

Значение Шепли

Shapley value [Lloyd S. Shapley, 1953]

Q: Как посчитать вклад признака в прогноз для любой модели?

A: Значение Шепли из кооперативной теории игр.

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\dots,x_p\} \setminus \{x_j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{x_j\}\right) - val(S)\right)$$

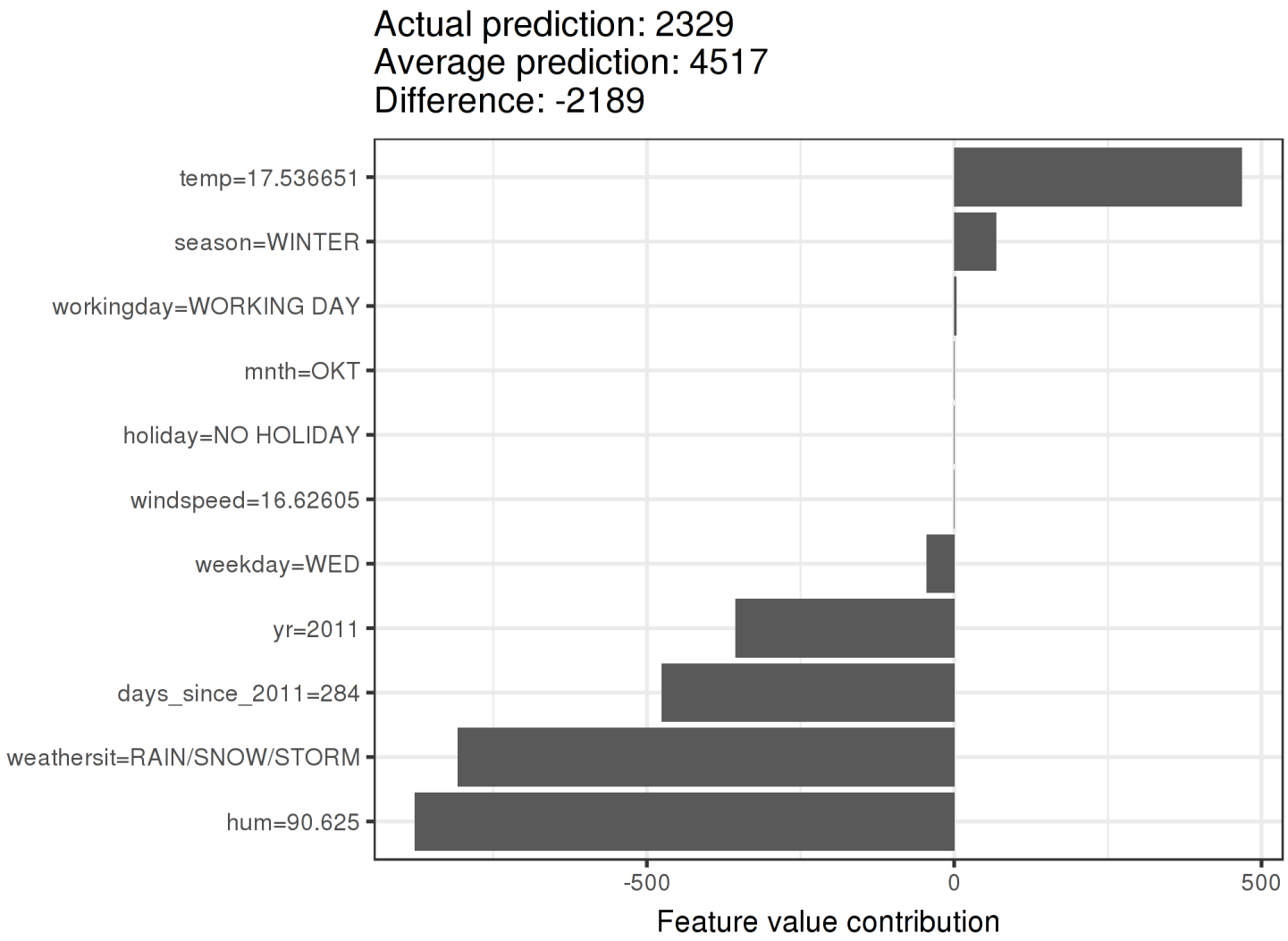
Значение Шепли – это средний вклад признака в предсказание по различным коалициям. Важно: это не разность в прогнозах, когда мы удаляем этот признак из модели.

Удовлетворяет аксиомам: эффективность, симметричность, нулевой вклад болвана, аддитивность.

Значение Шепли

Пример для случайных лесов

Данные: Bike rental system



Значение Шепли: приближенное вычисление

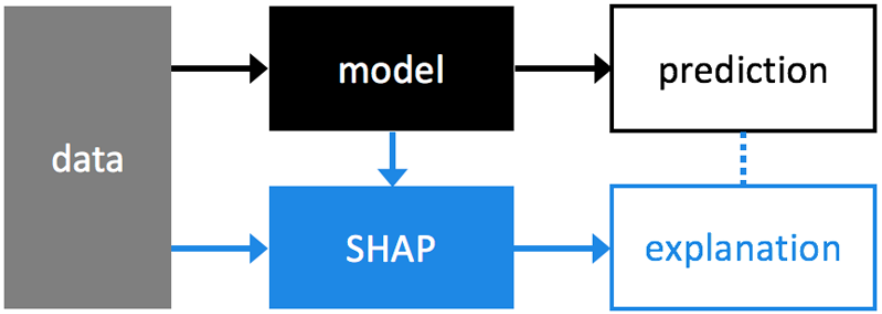
Монте-Карло алгоритм, [Strumbelj et al., 2014]

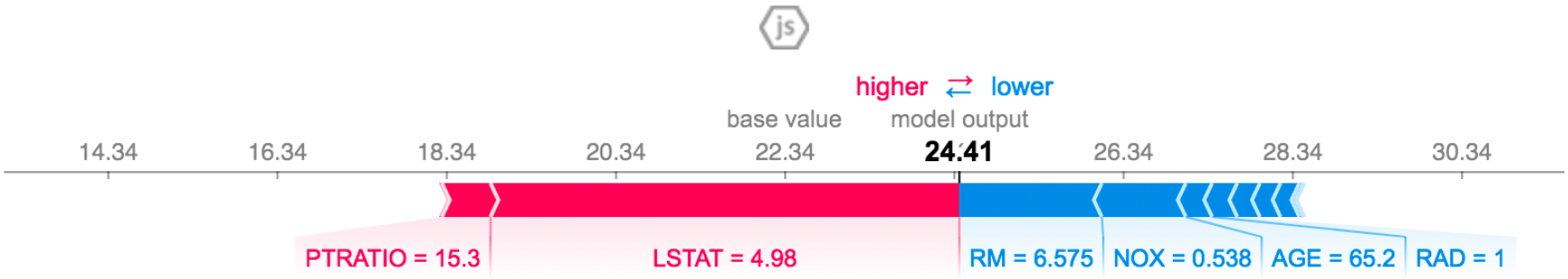
- Output: Shapley value for the value of the j-th feature
- Required: Number of iterations M, instance of interest x, feature index j, data matrix X, and machine learning model f
- For all m = 1,...,M:
  - Draw random instance z from the data matrix X
  - Choose a random permutation o of the feature values
  - Order instance x: $x_o = (x_{(1)}, \ldots, x_{(j)}, \ldots, x_{(p)})$
  - Order instance z: $z_o = (z_{(1)}, \ldots, z_{(j)}, \ldots, z_{(p)})$
  - Construct two new instances
  - $x_{+j} = (x_{(1)}, \ldots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \ldots, z_{(p)})$
  - $x_{-j} = (x_{(1)}, \ldots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \ldots, z_{(p)})$
  - $\phi_j^m = \hat{f}(x_{+j}) - \hat{f}(x_{-j})$
- Compute Shapley value as the average: $\phi_j(x) = \frac{1}{M} \sum_{m=1}^{M} \phi_j^m$

SHAP (SHapley Additive exPlanations)

Монте-Карло алгоритм, [Lundberg & Lee, 2017]

NIPS 2017 paper: A Unified Approach to Interpreting Model Predictions
Github: https://github.com/slundberg/shap

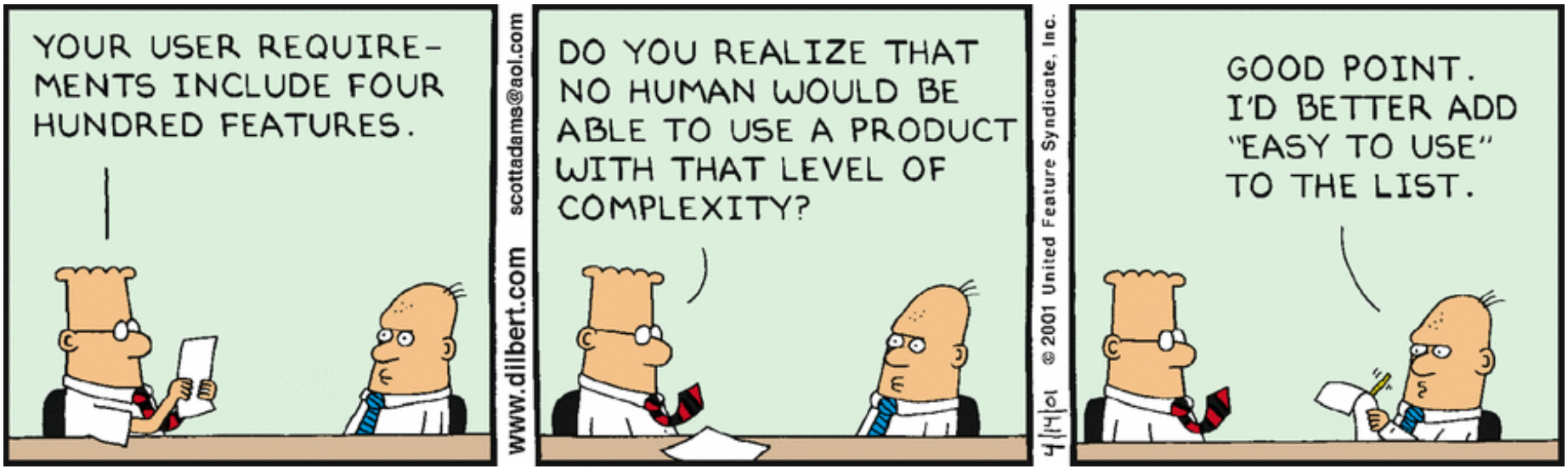




BIGDATA TEAM

Just for fun или шутки ради

dilbert.com



Вопросы и контакты

www.hse.ru/staff/dima

Спасибо!

dmitrii.ignatov@bigdatateam.org