Интерпретируем модели машинного обучения

Дмитрий Колодезев
ООО Промсофт, Новосибирск

План

- Что это и почему оно вас касается
- Интерпретируемые модели
- Основные подходы
- Немного о нейронках
- XGBoost Catboost
- Инструменты
- Литература

Зачем интерпретировать

- Отладка
- Очистка данных
- Извлечение знаний из модели
- Улучшение сбора данных
- Аудит — дискриминация, закладки и т.д.
- Требования регулятора (LR + бины?)
- Требования GDPR 12, 22, но это не точно
- Мы в любом случае строим модель модели

Объяснимость и предсказуемость

- Что она делает?
- Почему она так решила?
- Кто ее этому научил?
- Надежна ли она?
- На чем она сломается?
- Как повлиять на решение?
- Куда прятаться?
- Машина, робот-пылесос, светофор, модели ML

Когда не нужно

- Влияние модели мало
- Проблема хорошо разработана
- Класс моделей широко применяется
  - линейные модели
- Хотим скрыть алгоритм
  - скоринг
  - ранжирование
  - оценка качества

Свойства

- Понятность
- Контрастность contrastive
- Избирательность selectivity
- Стабильность robustness
- Локальность / глобальность
- Уместность (подходит этому пользователю)
- Последовательность (похожее объясняет похоже)
- Правдивость (нет контрпримера)
- В чем измерить интерпретируемость?

Что и как объясняем

- Входные данные
- Модель как белый ящик
- Модель как черный ящик
- Результаты работы модели

- Документируем внутренности
- Создаем сурогатную модель
- Объясняем на примерах

Интерпретируемые модели

- Линейные модели
- Деревья и списки правил
- kNN
- Примеры

Линейные модели

- Влияние на результат очевидно
- Масштаб. Сравнивать эффекты, не веса
- Значимость — p-value
- Иллюзия плотного распределения
- Проблемы с разреженными моделями
- LR — логарифм шанса
- GLM GAM — все сложно

Деревья

- Выучивают разбиение данных
- Очевидное представление
- Деревья небольшой глубины понятны

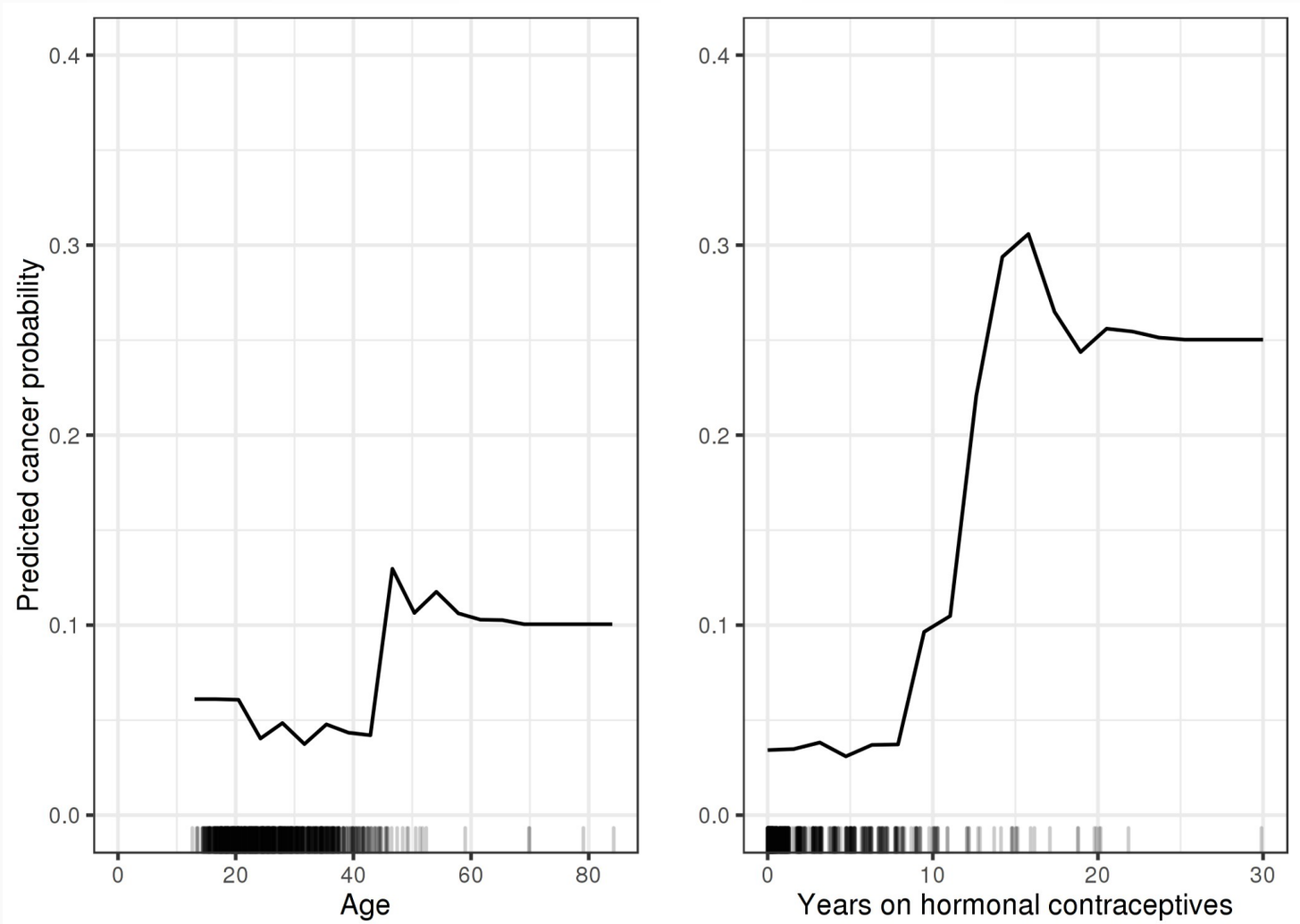
На примерах

- Если признаки интерпретируемы
- Картинки — да
- С табличными данными сложнее
- Понятный для человека подход
- Помогает построить ментальную метамодель
- kNN
- Деревья решений ищут похожие группы

Partial Dependence Plot

- Анализ частичной зависимости
- Как целевая метрика зависит от переменной
- Нескоррелированность (→ ALE)
- Глобальный, интуитивный
- Легко реализовать
- Не более двух переменных
- Разнонаправленные эффекты скрыты
- Skatter или самим написать

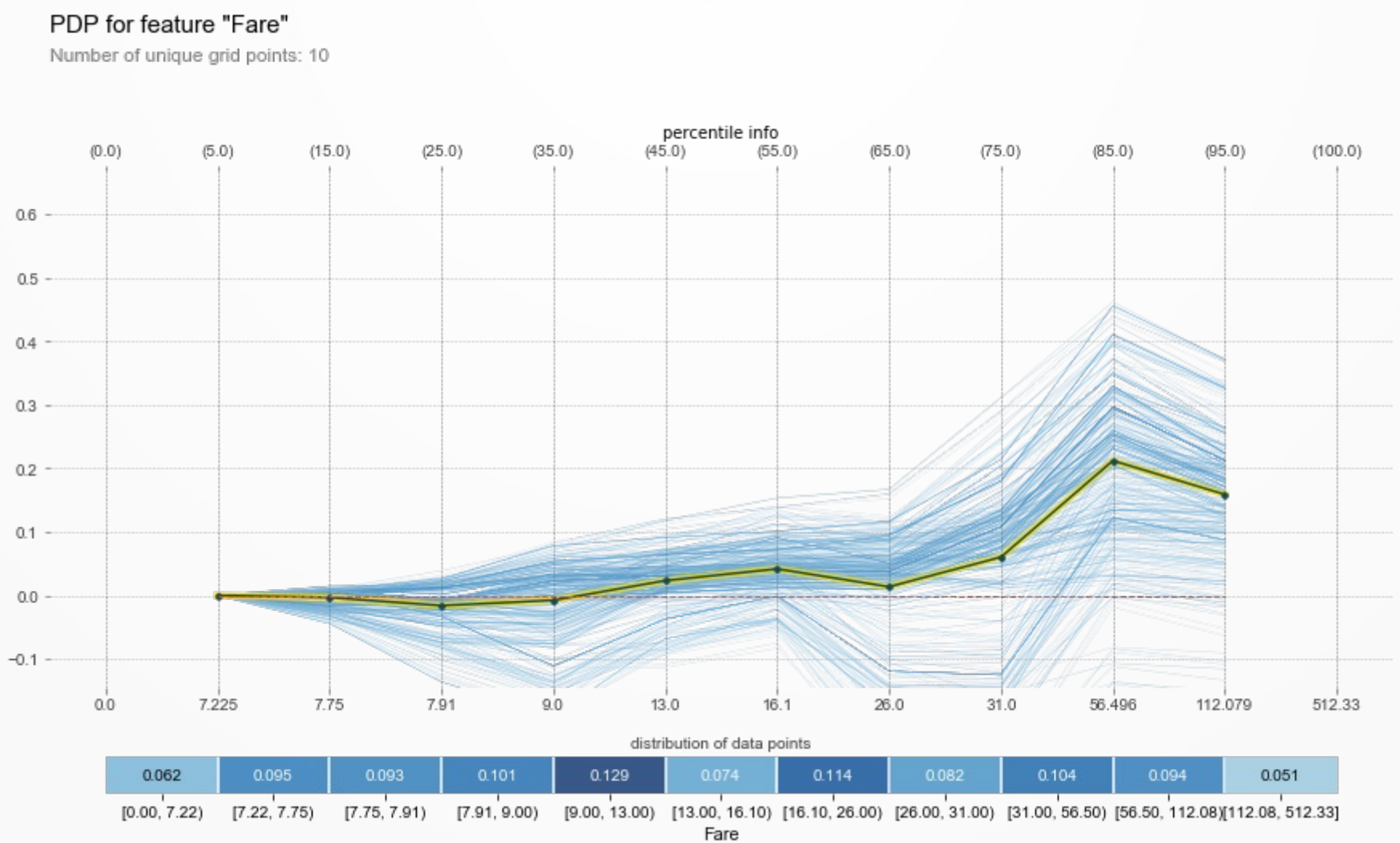
Partial Dependence Plot



Individual Conditional Expectation

- Индивидуальное условное ожидание
- Как целевая метрика изменялась бы для точки, если бы мы меняли один параметр
- Нескоррелированность, иллюзия плотности
- Самим написать, PyCEbox или PDPbox
- PD = среднее ICE по всем объектам

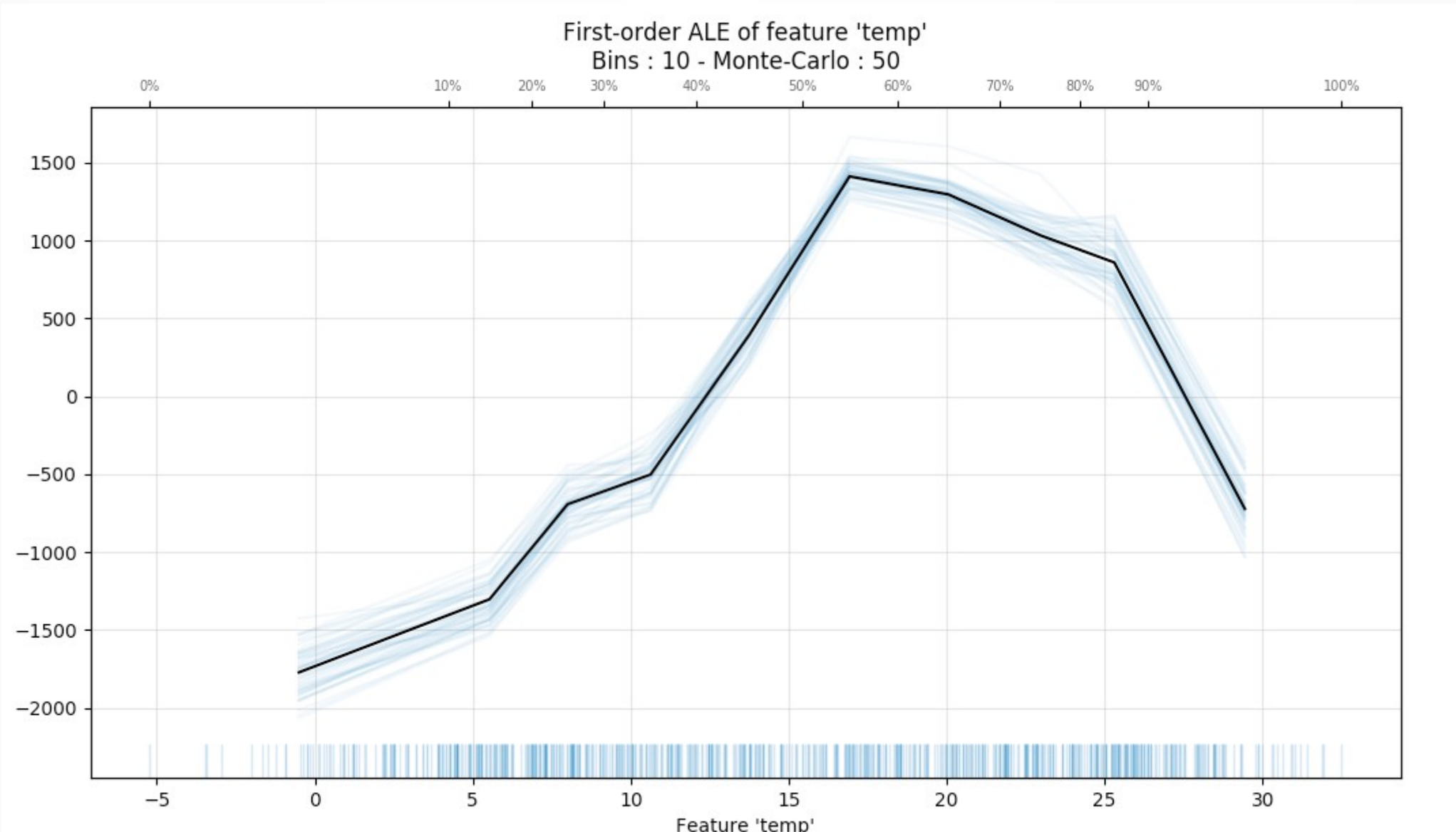
Individual Conditional Expectation



Accumulated Local Effects

- PDP в маленькой окрестности
- Уменьшает влияние корреляции признаков, но не убирает ее совсем
- Более корректное сэмплирование, т. е. не предполагает плотного распределения
- Быстрее считать, центрированы, понятны
- ALEPython

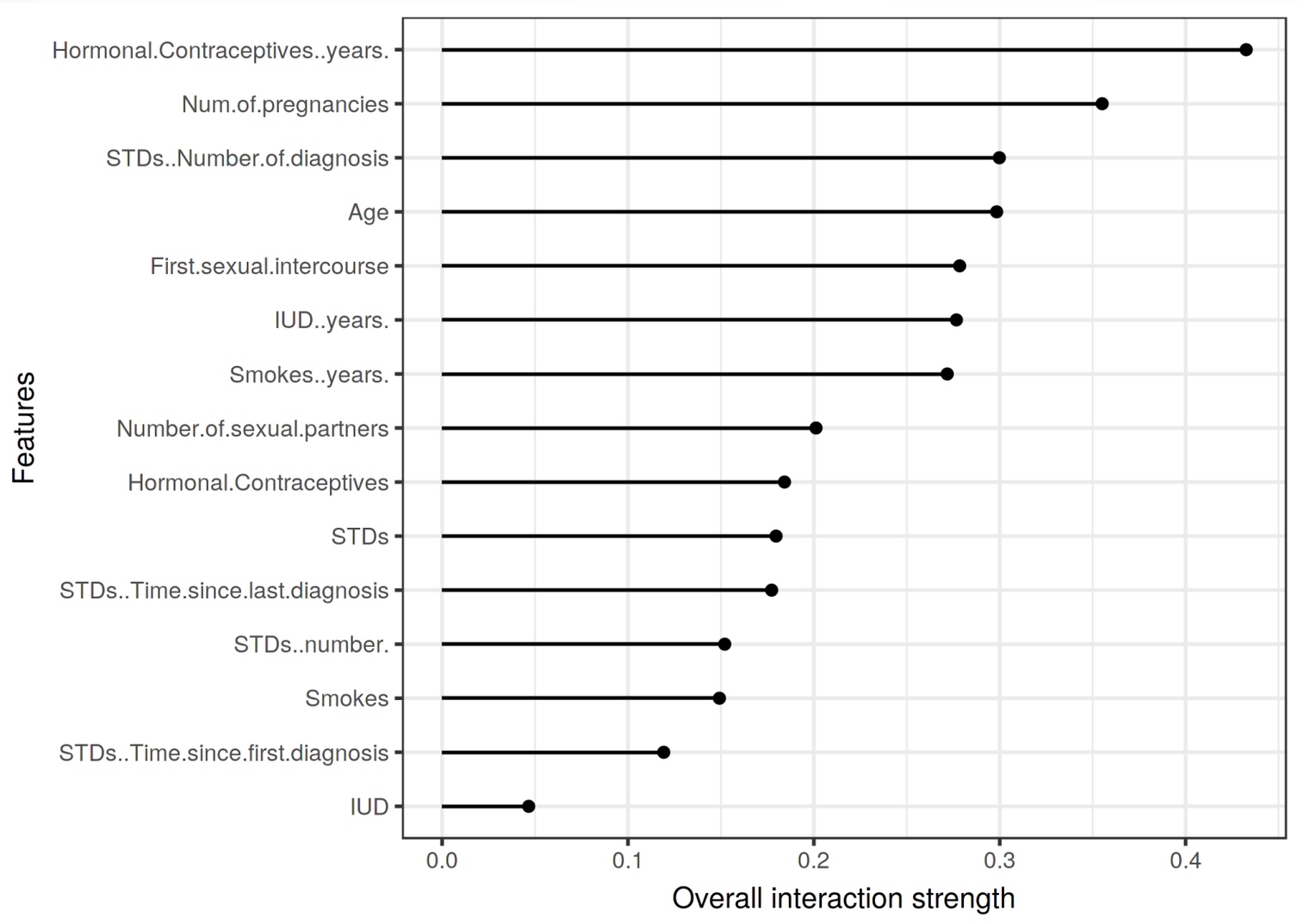
Accumulated Local Effects



Feature Interaction

- Мы можем посмотреть, насколько сумма PD по отдельным признакам отличается от заданной функции
- Если признаки нескоррелированы, целевая функция равна сумме PD с точностью до константы.
- Говорит о силе, не о форме взаимодействия
- H-статистика — объясненная дисперсия
- Есть в Catboost, можно считать руками

Feature Interaction



Feature Importance

- Насколько признак был важен для предикта
- Есть практически во всех пакетах
- Везде странно считается
- Правильно — насколько ухудшится при перемешивании, комбинаторно сложно
- Лучше отношение, а не разница
- Нестабильно
- При добавлении скоррелированных фич перераспределяется между ними.

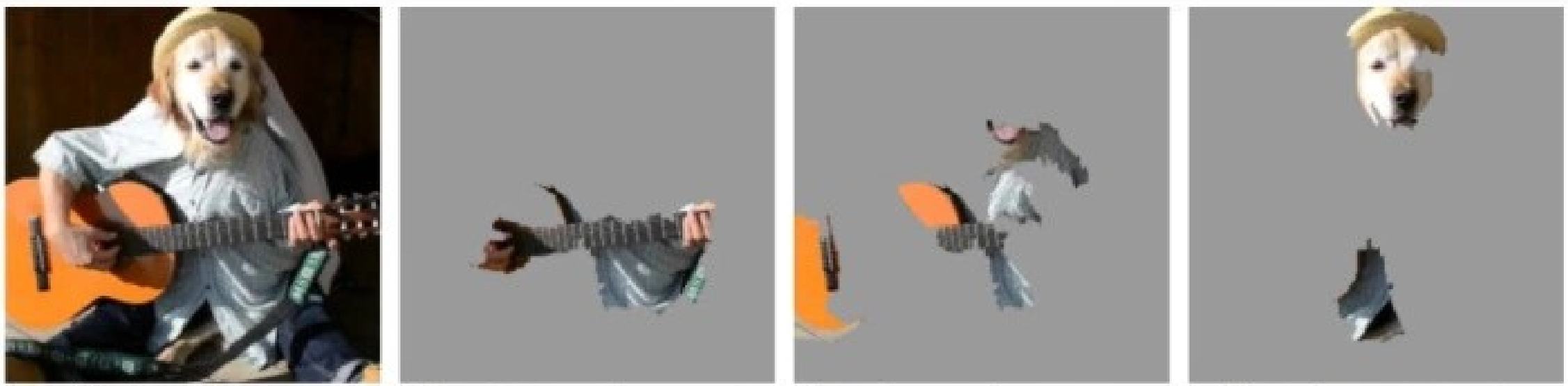
Глобальные суррогаты

- Обучаем простую модель предсказывать поведение сложной
- Предсказываем не мир, а модель!
- Можно измерить R^2. Сколько приемлимо?
- Гибкие
- Объяснимые (если суррогат объясним)
- Качество приближения может быть очень разным
- Можно использовать другой набор признаков!

LIME — локальные суррогаты

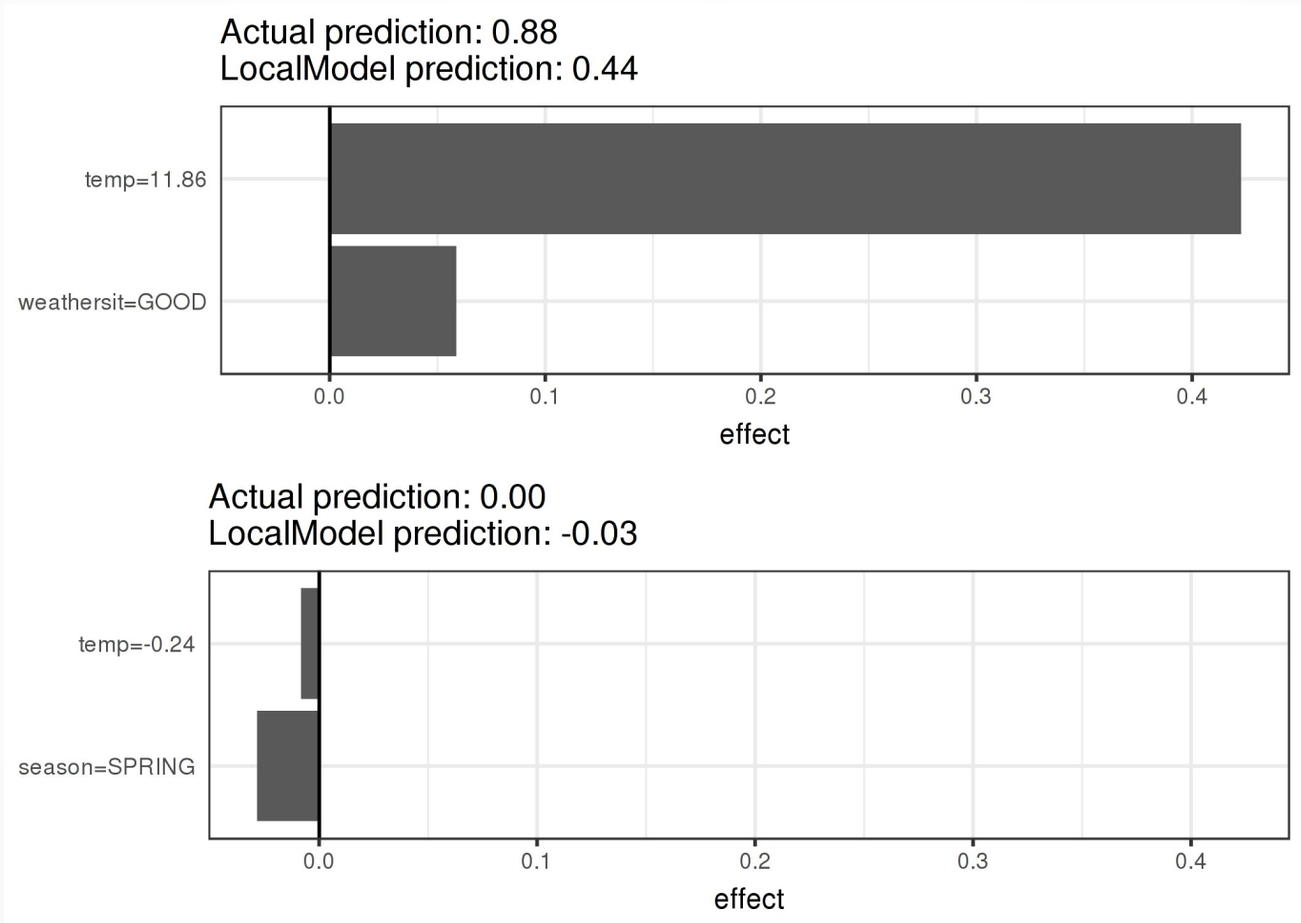
- Предсказание для конкретной точки
- Строим суррогатную модель для одной точки
- Интерпретируемую селективную модель — дерево или LASSO, взвешенное близостью
- Кого считаем соседями? Ширину ядра подбирать
- Количественные признаки сэмплируем из N
- Текст включаем-выключаем
- Картинки — суперпиксели — выключаем
- LIME, Eli5
- Может обучаться на других признаках!

LIME — локальные суррогаты



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

LIME — локальные суррогаты



LIME — локальные суррогаты

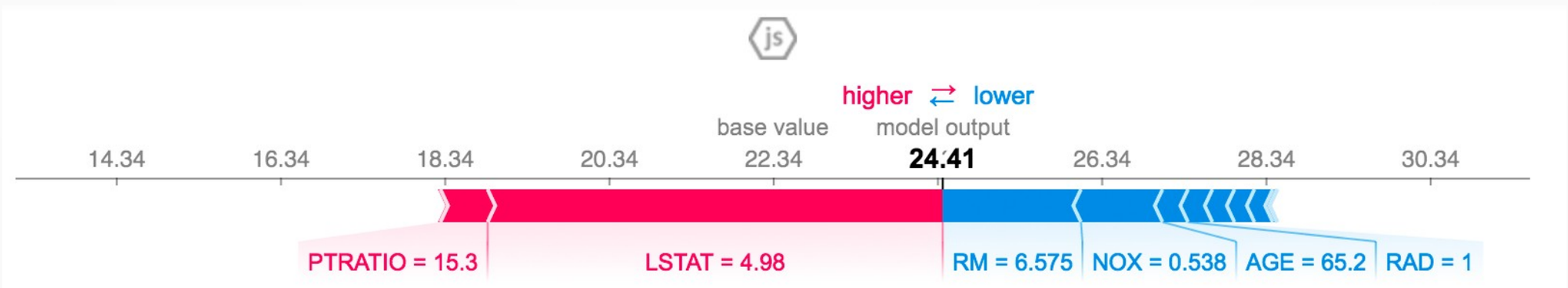
- For Christmas Song visit my channel! ;)

case	label_prob	feature	feature_weight
1	0.1701170	good	0.000000
1	0.1701170	PSY	0.000000
1	0.1701170	a	0.000000
2	0.9939024	channel!	6.180747
2	0.9939024	Song	0.000000
2	0.9939024	Christmas	0.000000

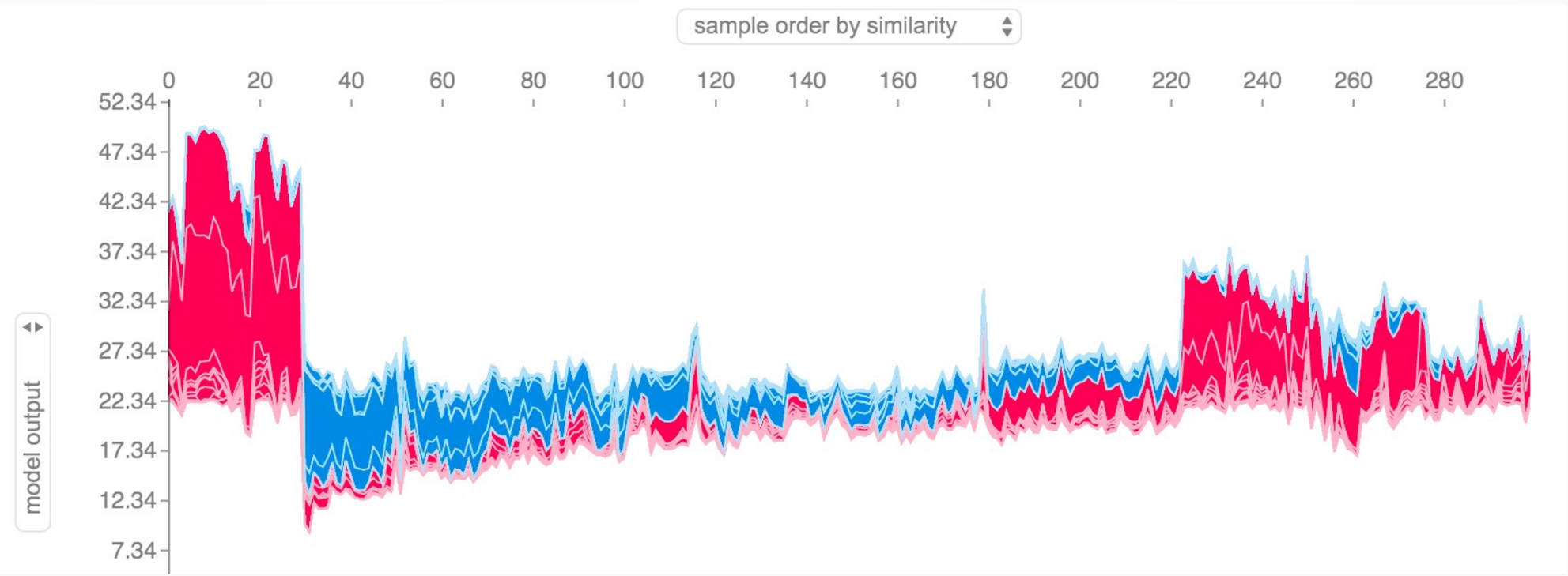
Shapley Values & SHAP

- Теория игр. Как делить приз между игроками
- Пробуем сыграть всеми возможными коалициями
- Средняя выгода от участия игрока — SV
- Делим добычу пропорционально SV
- Игроки — фичи.
- Считать напрямую комбинаторно сложно
- Аппроксимируют
- Для точки и среднее
- SHAP. Есть в Catboost

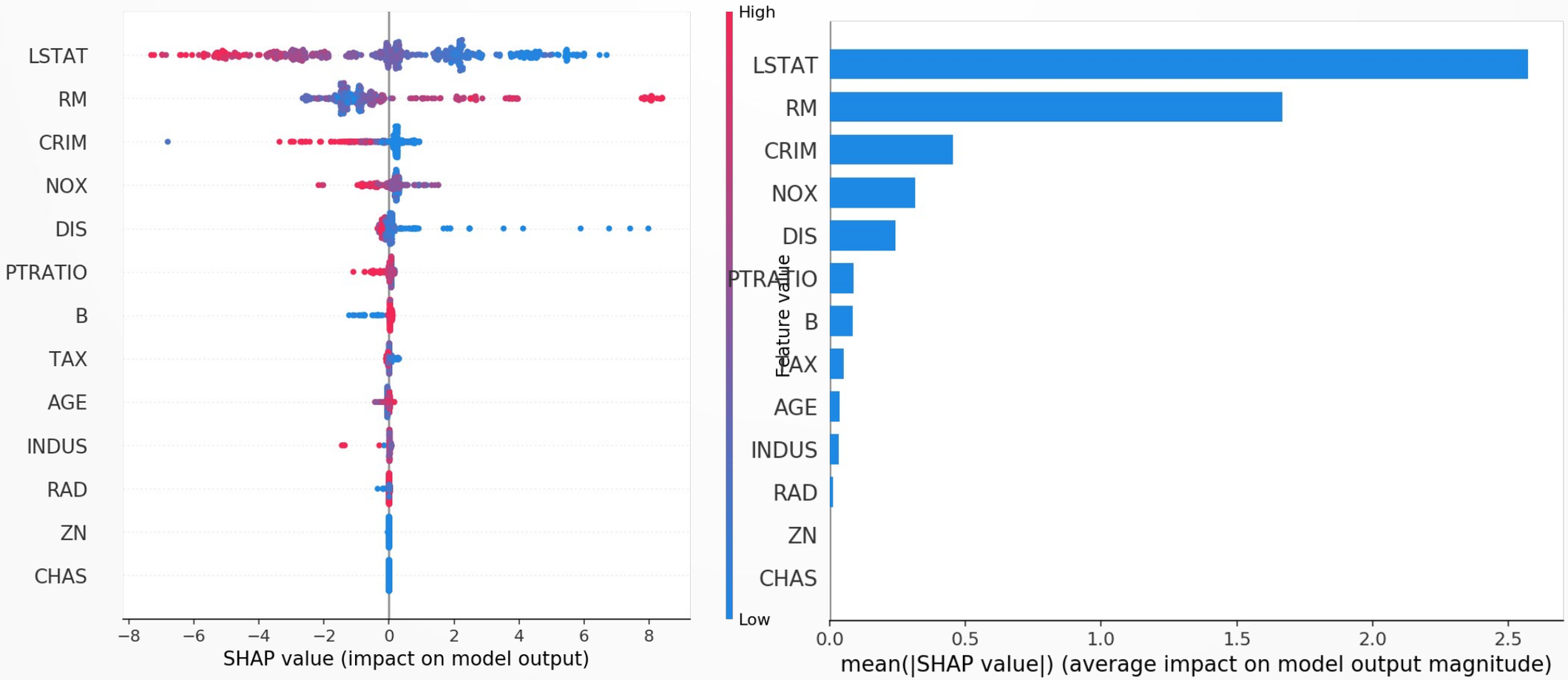
Shapley Values & SHAP



Shapley Values & SHAP



Shapley Values & SHAP



Shapley Values & SHAP

Shapley Values & SHAP

- Эффективно (сумма равна отклонению точки)
- Аддитивно
- Если SV = 0, влияния нет
- Имеет теоретическое обоснование
- Возможно, прокатит в продвинутом суде
- Не предиктивная модель!
- Для скоррелированных фич — маловероятные точки, как все перестановочные методы

Доказательство от противного

- Давайте найдем примеры, максимально похожие на нашу точку, но другого класса
- Разница между ними будет объяснением решения
- Не требует доступа к датасету
- Нет нормальных реализаций
- Пишется (расширяющаяся сфера)
- Может давать много разных объяснений

Anchors

- https://github.com/marcotcr/anchor
- Давайте найдем подмножество признаков, которые для заданной точки закрепляют предикт.
- Это и будут якоря — Anchors
- Хорошее объяснение
- Только для таблиц и текста
- Будет рассказ на датафесте

Adversarial Examples

- Дальнейшее развитие «от противного»
- Тысячи их, в основном для нейронок
- На полях презентации не хватило места для доказательства этой теоремы
- Некоторые могут быть использованы для аудита и отладки модели, но это неточно

Prototypes and Criticisms

- Давайте найдем типичные точки данных, и будем объяснять модель на примерах
- Рассмотрим внимательно примеры и поймем данные и модель
- Давайте найдем нетипичные точки данных, и будем использовать для отладки
- k-medoids возвращает прототипы. Находит наименее непохожие, попарно переставляя
- MMD-critic - не поддерживается, но какая статья!

Influential Instances

- Какие сэмплы сильнее всего повлияли на модель?
- Статистическая классика
  - Cook`s distance — если удаляем
  - DFBETA — как повлияли
- На предикт в точке или на всю модель
- Просто распечатать 10 самых влиятельных может быть недостаточно
- Можно построить линейную модель!
- Встроено в CatBoost
- Harmful Object Removal, Debugging Domain Mismatch

Интерпретация результатов

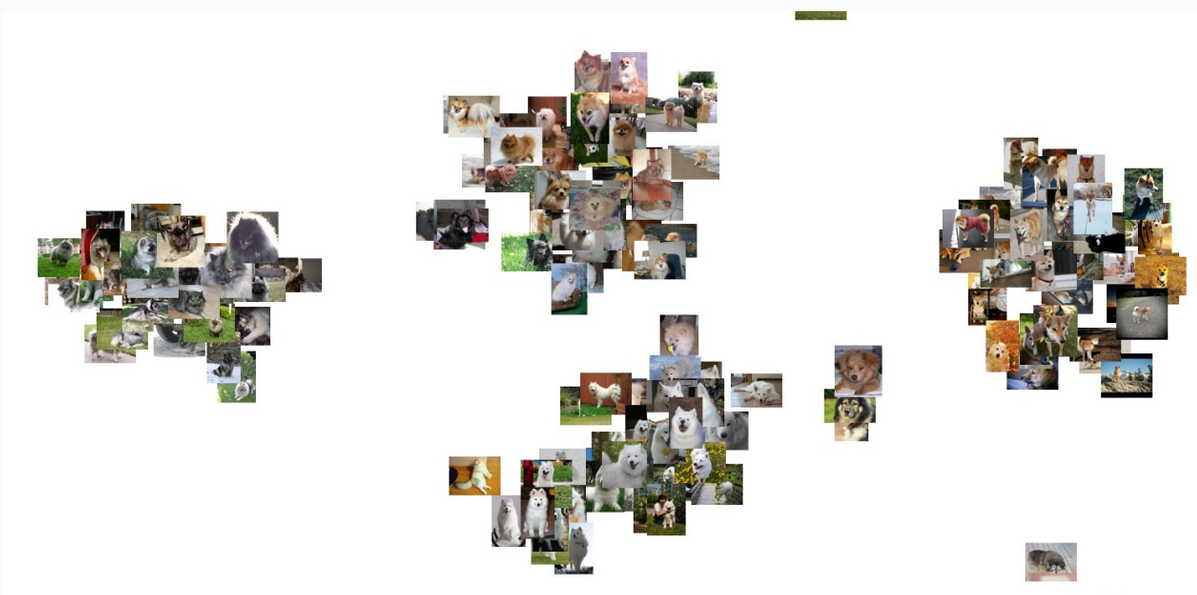
- 10 самых уверенных ошибок каждого класса
- 10 самых неуверенных предиктов
- Например в FastAI
- Легко делается везде

Про нейронки

- Градиент пиксела ~ необычность картинки
- Визуализация «внимания»
- Approximating CNNs with Bag-of-local-Features нарезали на патчи, Resnet фичи и LR

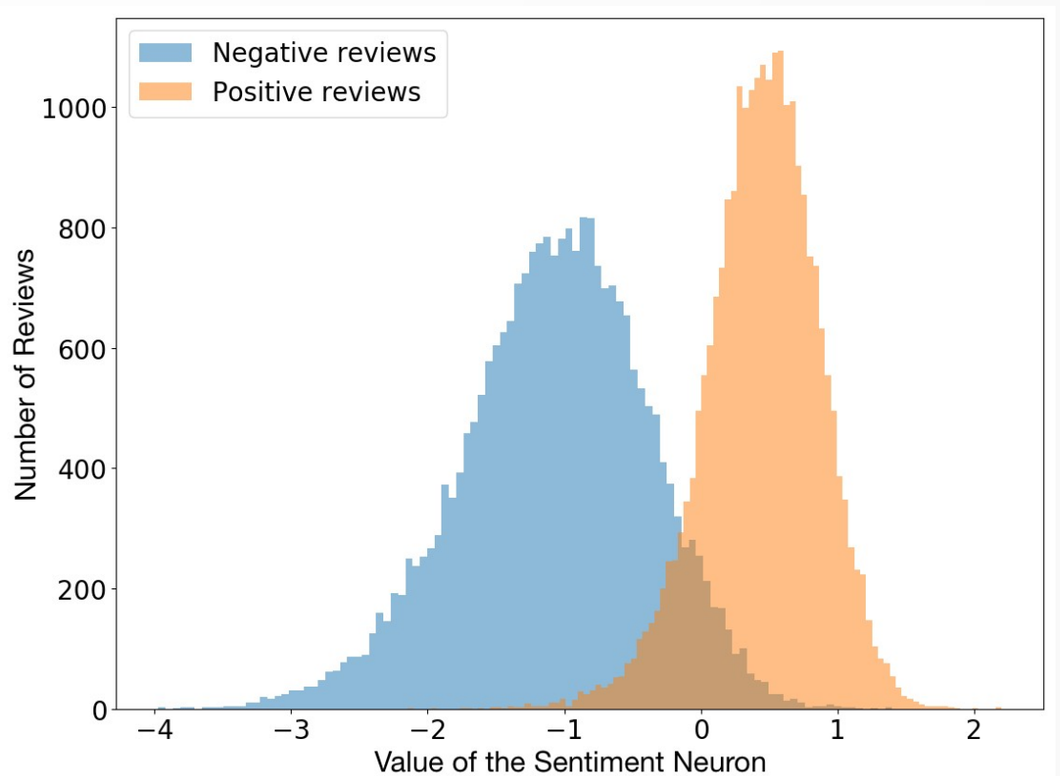
Нейронки — данные

- t-SNE
- CNN Visualizations with Dogs and Cats
- Do's and Don'ts of using t-SNE to Understand Visi on Models



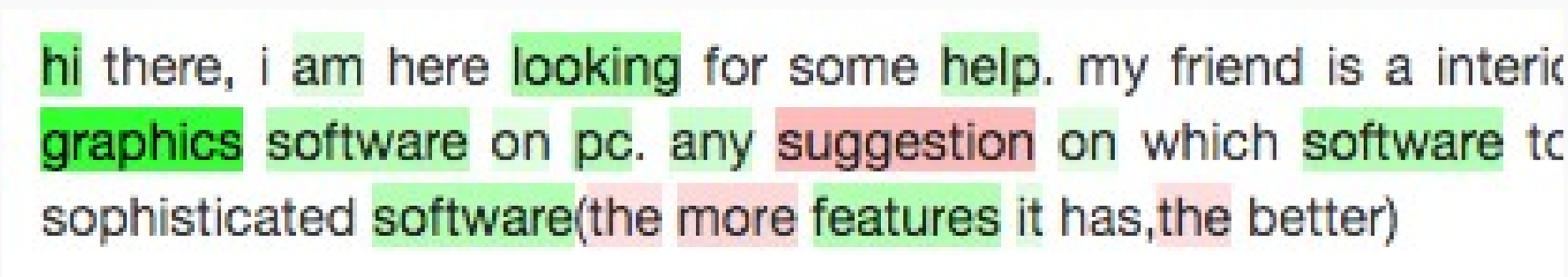
Нейронки — юниты

- Unsupervised Sentiment Neuron
- Похожие слои выучивают похожие концепты
- Выучат то, что в данных
- Не передать словами
- В списке литературы 2 мастеркласса



Eli5

- https://github.com/TeamHG-Memex/eli5
- https://www.youtube.com/watch?v=pqqcUzj3R90
- В основном про текст
- TextExplainer, crfsuite, XGBoost, LightGBM
- @kostia, @kmike



XGBoost

- Из коробки довольно грустно
  - Gain
  - Coverage
  - Weight
- LIME, SHAP, treeinterpreter, Eli5
- https://github.com/jphall663/interpretable_machine_learning_with_python

Catboost

- Feature importance
  - PredictionValuesChange
  - LossFunctionChange
  - ShapValues
  - Interaction
- Object importance
  - Average
  - PerObject

Почитать

- https://dyakonov.org/2018/08/28/интерпретации-чёрных-ящиков/
- https://www.kaggle.com/learn/machine-learning-explainability
- https://christophm.github.io/interpretable-ml-book/
- http://interpretable-ml.org/cvpr2018tutorial/
- https://interpretablevision.github.io/
- http://netdissect.csail.mit.edu/

Вопросы?

dkolodezev

promsoft

dkolodezev

d_key

dmitry_kolodezev

Пиши, если что...