Интерпретируем модели машинного обучения

Дмитрий Колодезев
ООО Промсофт, Новосибирск

DataConf Barnaul 25.05.2019

План

- Историческая справка
- Кому нужно и какое бывает
- Интерпретируемые модели
- Пристальное разглядывание данных
- Суррогатные модели
- Catboost
- Про нейронки
- Из личной практики
- Литература

История вопроса

— Вот, изволите видеть, так называемая эвристическая машина, — сказал старичок.

— Точный электронно-механический прибор для отвечания на любые вопросы, а именно на научные и хозяйственные.

А.Б. Стругацкие, «Сказка о Тройке»

Расёмон

- Лео Брейнман, RF Father - The Rashomon Effect:
  - Небольшие изменения в данных
  - Совершенно другие веса и правила
  - Разные важности признаков
  - Точно такие же предсказания
  - Множественность хороших моделей
  - Деревья, линейная регрессия, нейронки
  - Случайный лес должен был это решить.

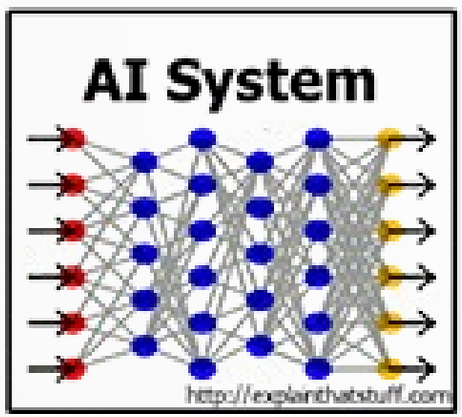
Statistical Modeling: The Two Cultures. Leo Breiman, 2001

FAT, ответственные алгоритмы

- Responsibility + Explainability + Accuracy
- Кому понадобятся объяснения?
- В каком виде?
- Объяснения — важная часть модели
- Данные, модели, процессы
- Решения модели
- Для коммерции

Principles for Accountable Algorithms
FAT* Conference 2018

DARPA, eXplainable AI



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

DoD and non-DoD Applications

Transportation

Security

Medicine

Finance

Legal

Military

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Explainable Artificial Intelligence (XAI)

Кому это нужно

- Инженерам,
  которые разрабатывают модели
- Бизнесу,
  которому мы их впариваем
- Конечным пользователям,
  которым с этим жить
- Регуляторам,
  которым не всё равно
- На соревнованиях (но это неточно)
- Мне (контроль качества, расширение команды)

Когда не нужно

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

Возможно ли это?

— «У мене внутре... гм... не... неонка»

— Внутре смотрите,
где у нее анализатор и думатель..

А.Б. Стругацкие, «Сказка о Тройке»

Можно попробовать

- Изучаем данные:
  на чем училась, что мешало
- Объясняем на примерах:
  показываем характерные точки
- Документируем внутренности:
  распечатываем веса, деревья
- Создаем суррогатную модель:
  делаем модель модели

Свойства

- Понятное и уместное — имеет смысл в мире пользователя
- Локальное или глобальное
- Избирательное — достаточный минимум информации
- Руководство к действию (actionable)
- Контрастное — показывает различие
- Стабильное
- Черный или белый ящик
- Последовательное (похожее объясняет похоже)
- Непротиворечивое (нет контрпримера)

Сложновато, да?

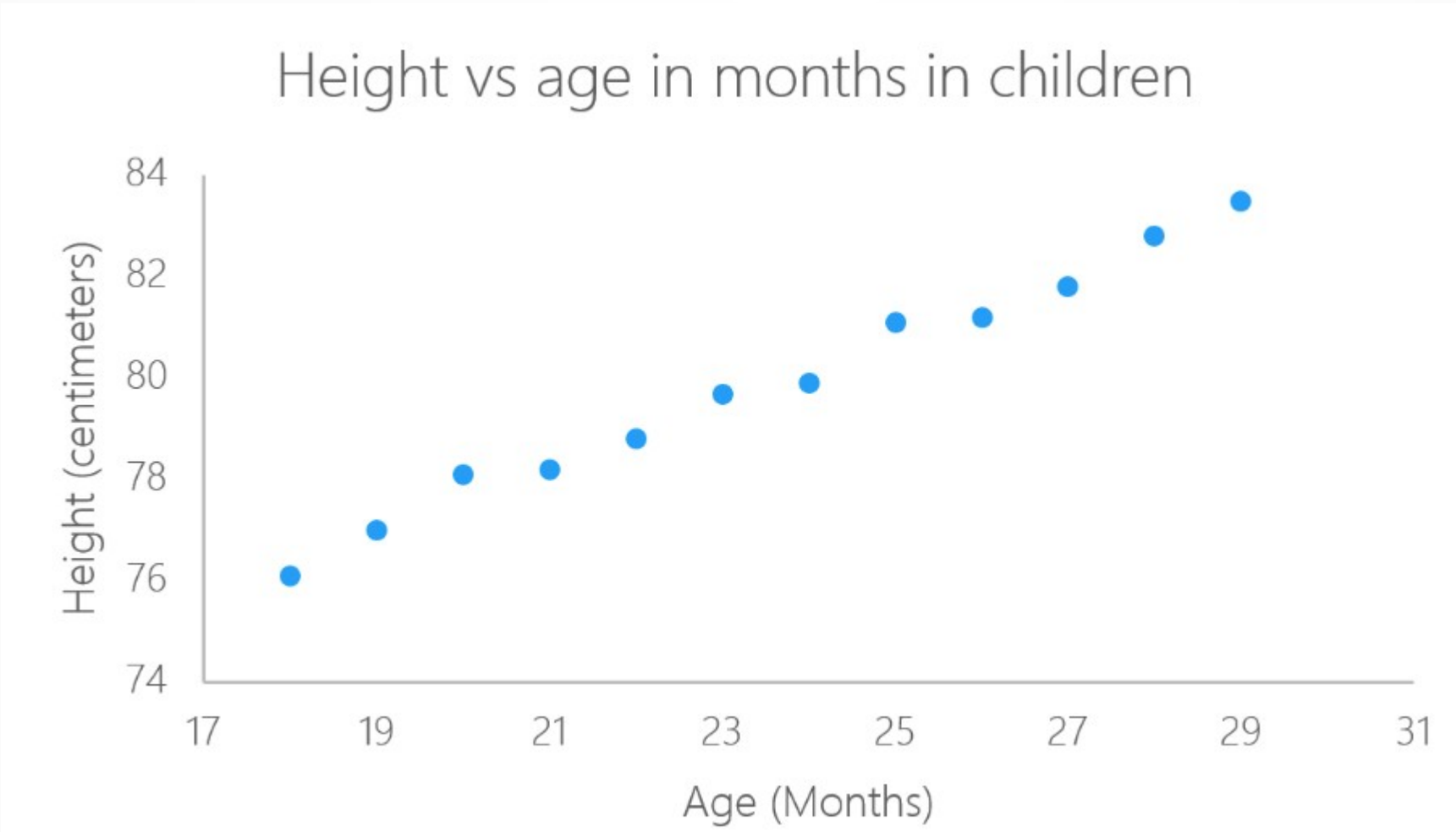
Нас спасут:

- Статистика
- Эвристика
- Здравый смысл

Интерпретируемые модели

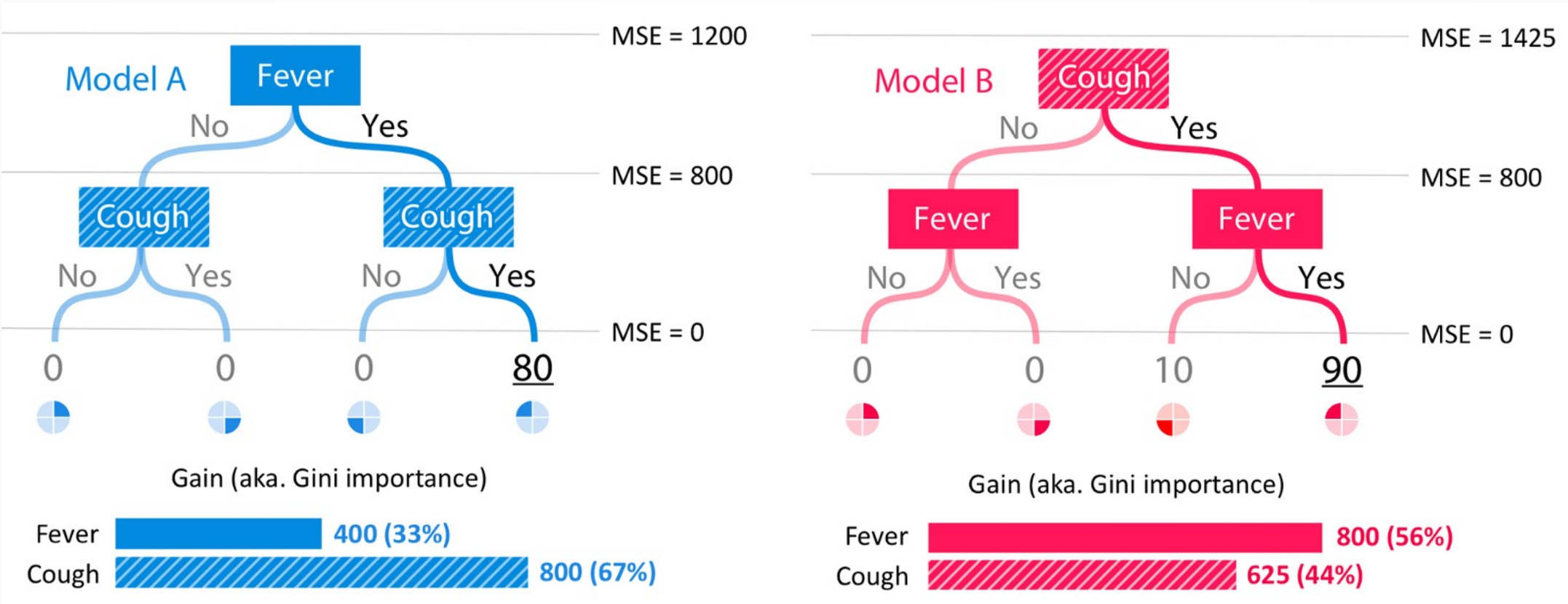
- Линейные модели
  - Сравнивать эффекты, не веса
  - Иллюзия плотности, разреженные данные
  - Взаимодействие признаков
- Деревья (неглубокие) и списки правил
  - Очевидны. Выучивают разбиение данных.
- kNN и на примерах
  - Когда точки интерпретируемы

Иллюзия плотности



https://www.datacamp.com/community/tutorials/linear-regression-R

Неглубокие деревья



https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27

Банальности

- EDA (разведочный анализ) — основа всего
- 10 самых уверенных неверных предсказаний
- 10 самых неуверенных предсказаний
- 10 самых уверенных правильных
- 10 самых «важных» признаков
- Найдите самое странное и там копайте

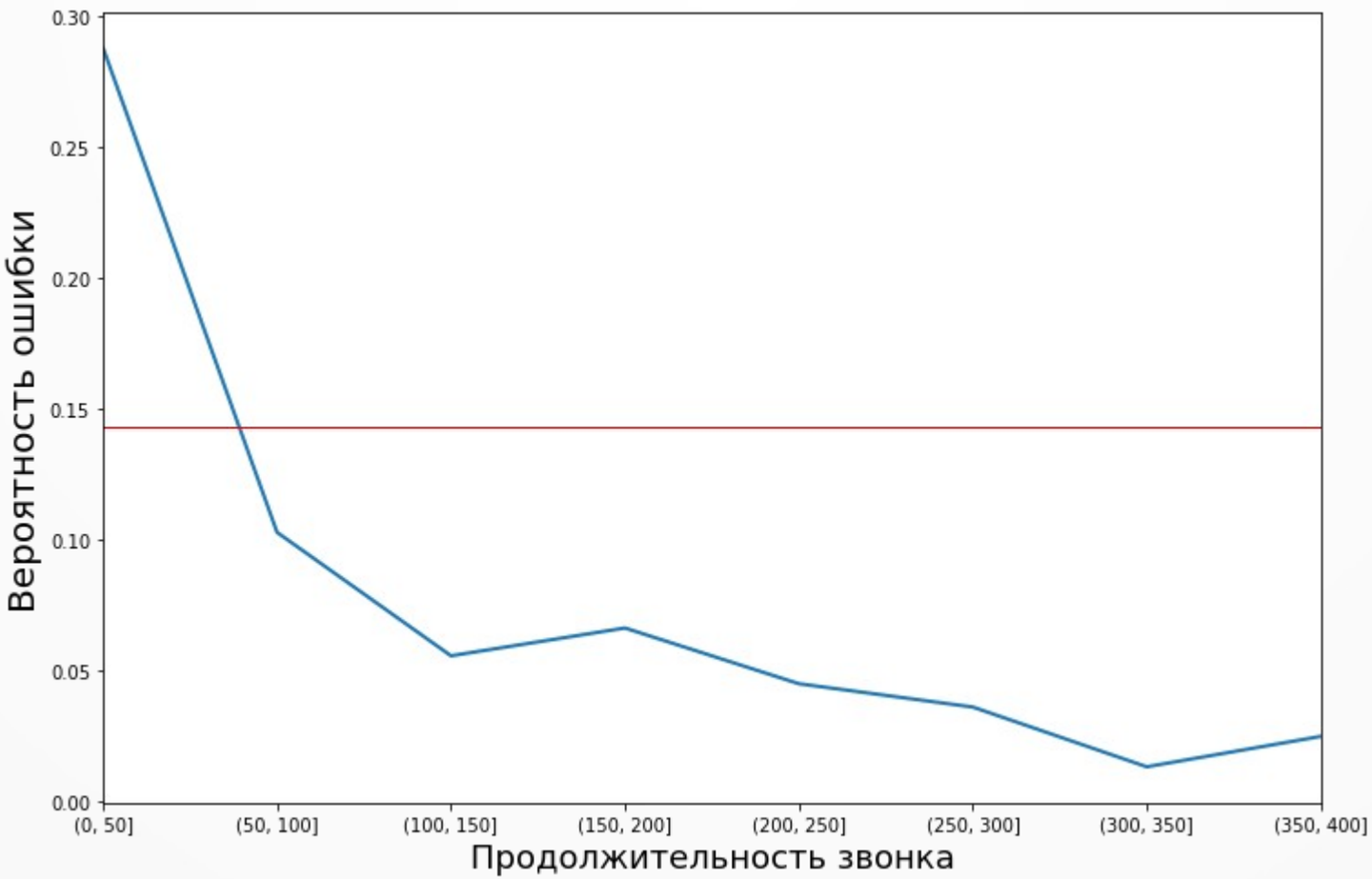
Важность признака

- Насколько признак был важен для предикта
- Есть практически во всех пакетах
- Иногда странно считается
- Правильно — насколько ухудшится при перемешивании, комбинаторно сложно
- Нестабильный показатель
- При добавлении скоррелированных фич перераспределяется между ними.
- CatBoost молодец

Влиятельные точки

- 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

Средний таргет по бинам



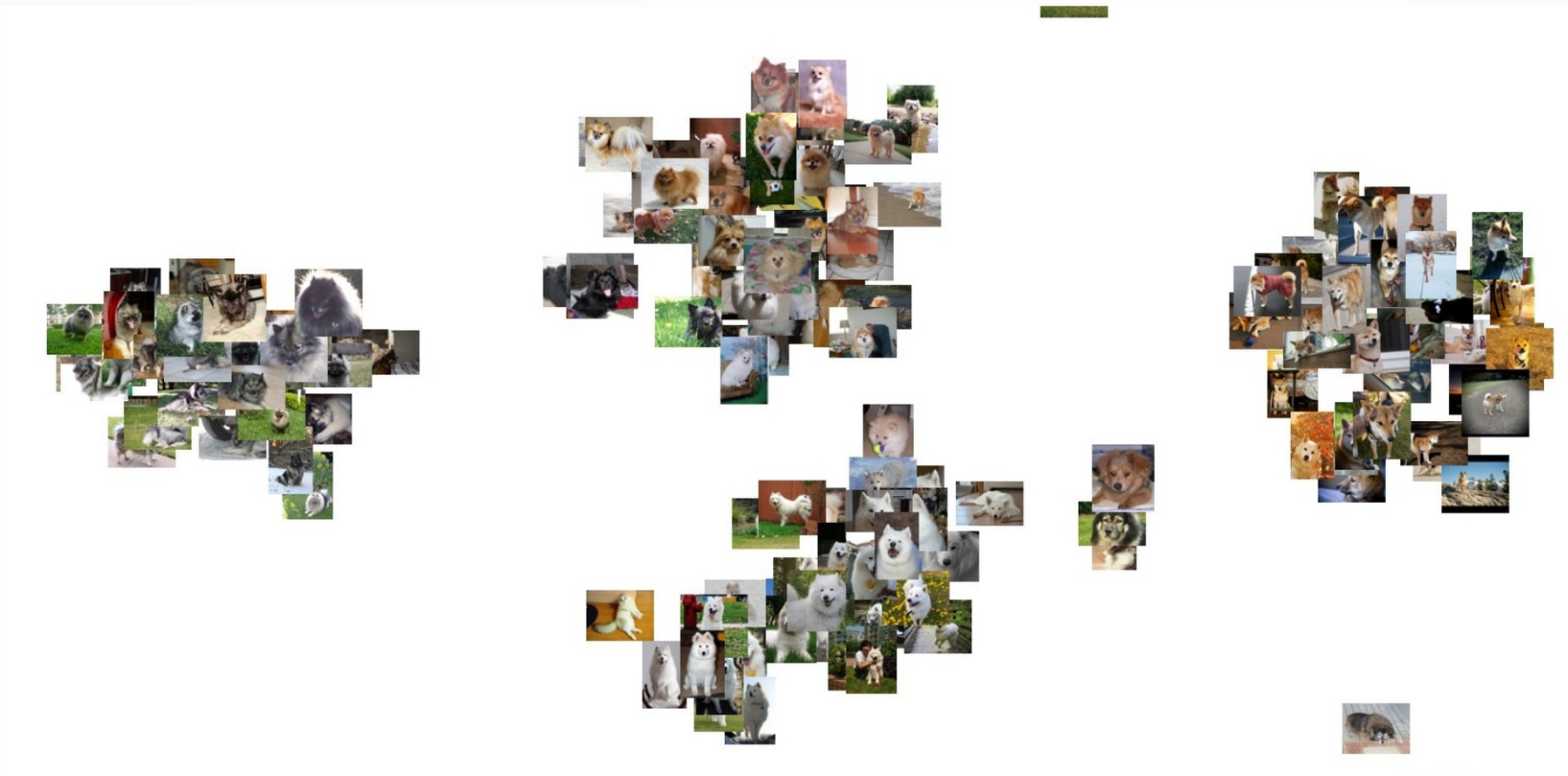
см. «Как не выстрелить себе в ногу», Данила Савенков

Прототипы и Критики

- Найдем типичные точки данных, объяснение
- Найдем нетипичные точки данных, отладка
- Поймем данные и модель
- MMD-critic + статья про это про это
- k-medoids возвращает прототипы
- Прототипы видно на t-SNE
- Критики — те, кого на CV болтает сильнее

t-SNE

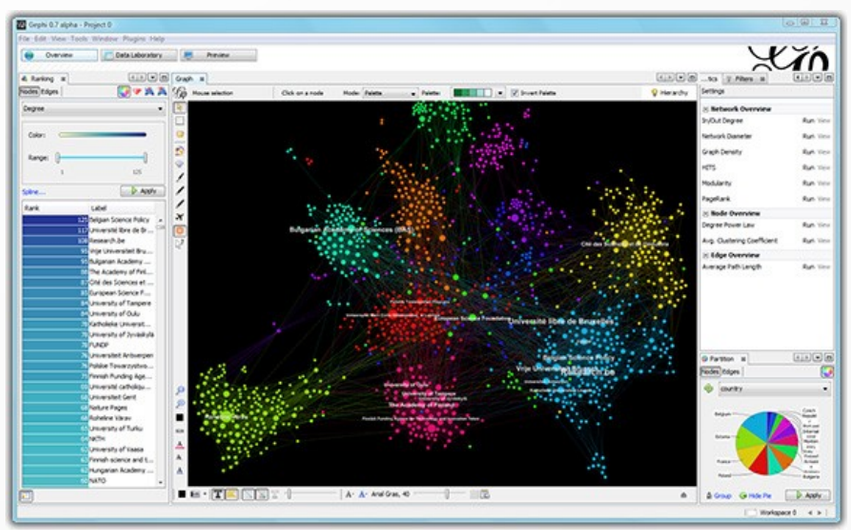
- Do's and Don'ts of using t-SNE
- CNN Visualizations with Dogs and Cats



Корреляция признаков

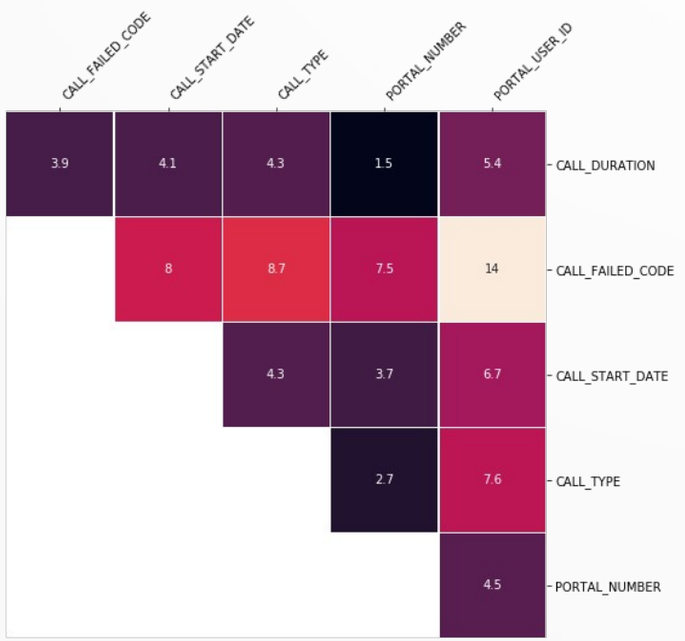
- Corrplot нечитаемый, если признаков много
- Генерируем граф, открываем в gephi

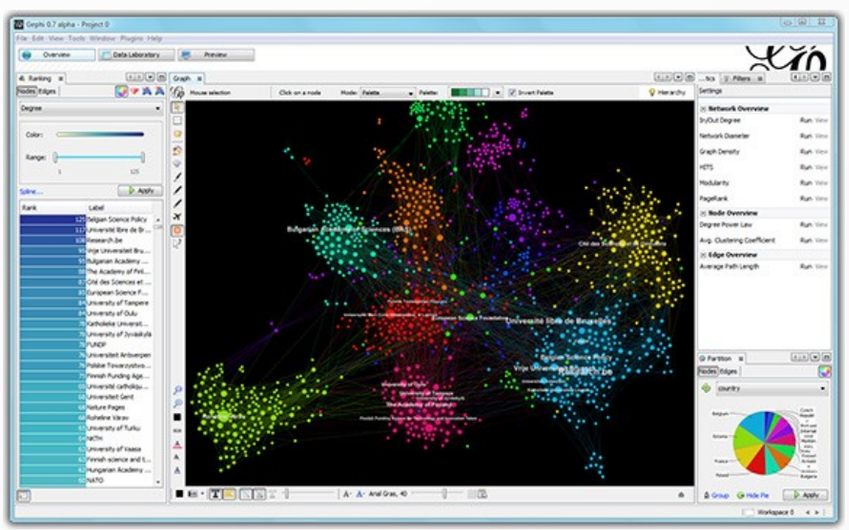
	0	1	2	3	4	5	6	7	8	9
0	1	0.35	0.4	0.46	0.073	-0.23	-0.73	0.48	-0.44	0.015
1	0.35	1	-0.28	0.57	-0.29	0.38	-0.36	0.64	0.25	0.19
2	0.4	-0.28	1	-0.52	0.15	-0.14	-0.093	0.016	-0.43	-0.38
3	0.46	0.57	-0.52	1	-0.23	-0.23	-0.48	0.47	0.28	0.45
4	0.073	-0.29	0.15	-0.23	1	-0.1	-0.15	-0.52	-0.61	-0.19
5	-0.23	0.38	-0.14	-0.23	-0.1	1	-0.03	0.42	0.21	0.095
6	-0.73	-0.36	-0.093	-0.48	-0.15	-0.03	1	-0.49	0.38	-0.35
7	0.48	0.64	0.016	0.47	-0.52	0.42	-0.49	1	0.38	0.42
8	-0.44	0.25	-0.43	0.28	-0.61	0.21	0.38	0.38	1	0.15
9	0.015	0.19	-0.38	0.45	-0.19	0.095	-0.35	0.42	0.15	1



Взаимодействие признаков

- Как признаки «сотрудничают» в модели?
- Для двух признаков $w_0x_0 + w_1x_1 + w_2x_0x_1 + C$
- Есть в Catboost, можно считать руками





Anchors

- https://github.com/marcotcr/anchor
- Давайте найдем подмножество признаков, которые для заданной точки почти закрепляют предсказание.
- Это и будут якоря — Anchors
- Локальное избирательное объяснение
- Таблицы, текст, картинки
- Рассказ на датафесте Юрия Гаврилина

ALIBI — якорь на стероидах

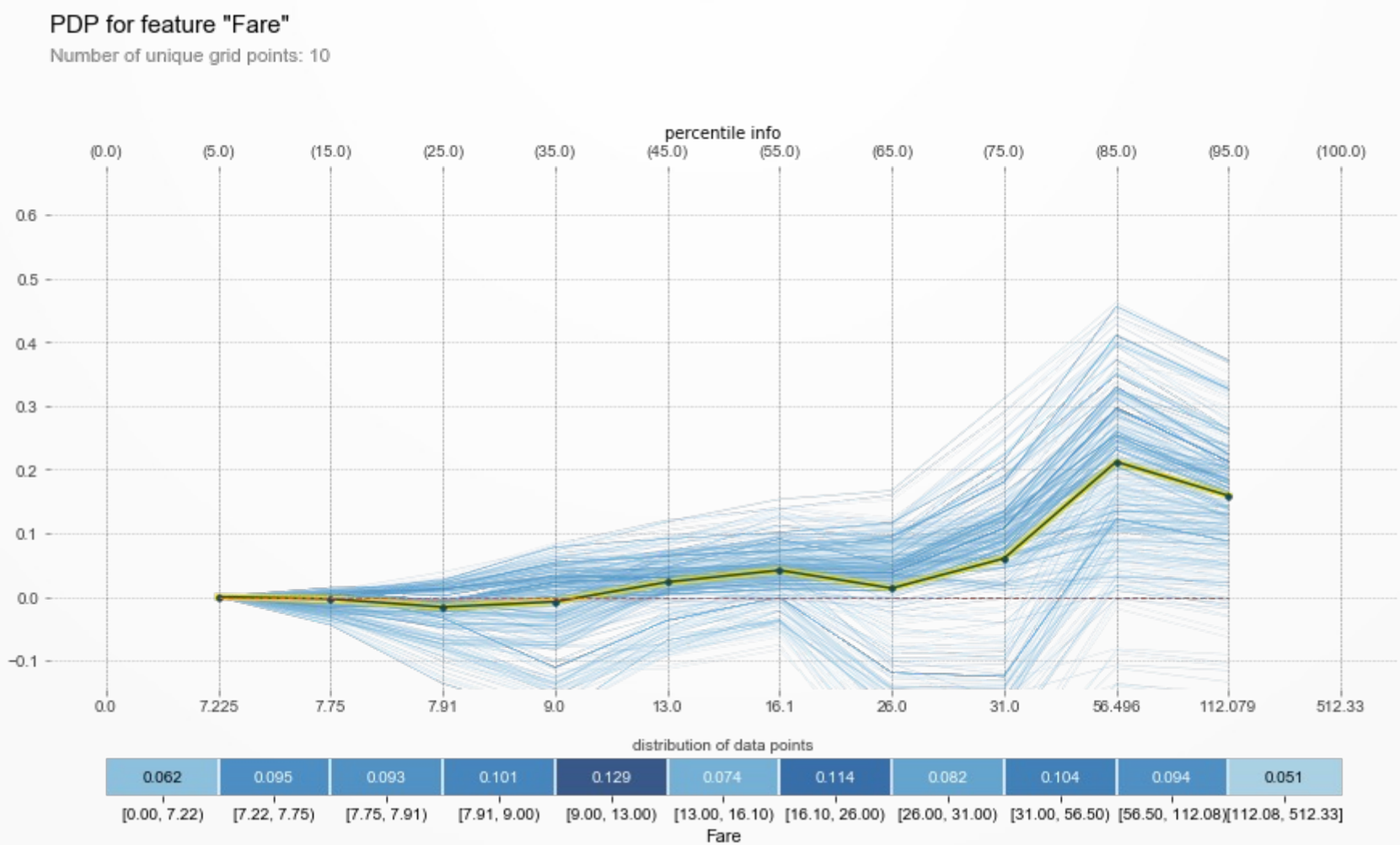
- https://docs.seldon.io/projects/alibi/
- Быстрый Anchors
- Contrastive explanation method:
  - Pertinent positive — что должно быть
  - Pertinent negative — чего не должно быть
- Trust Scores
  насколько сильно модель «выдумывала»

Adversarial validation

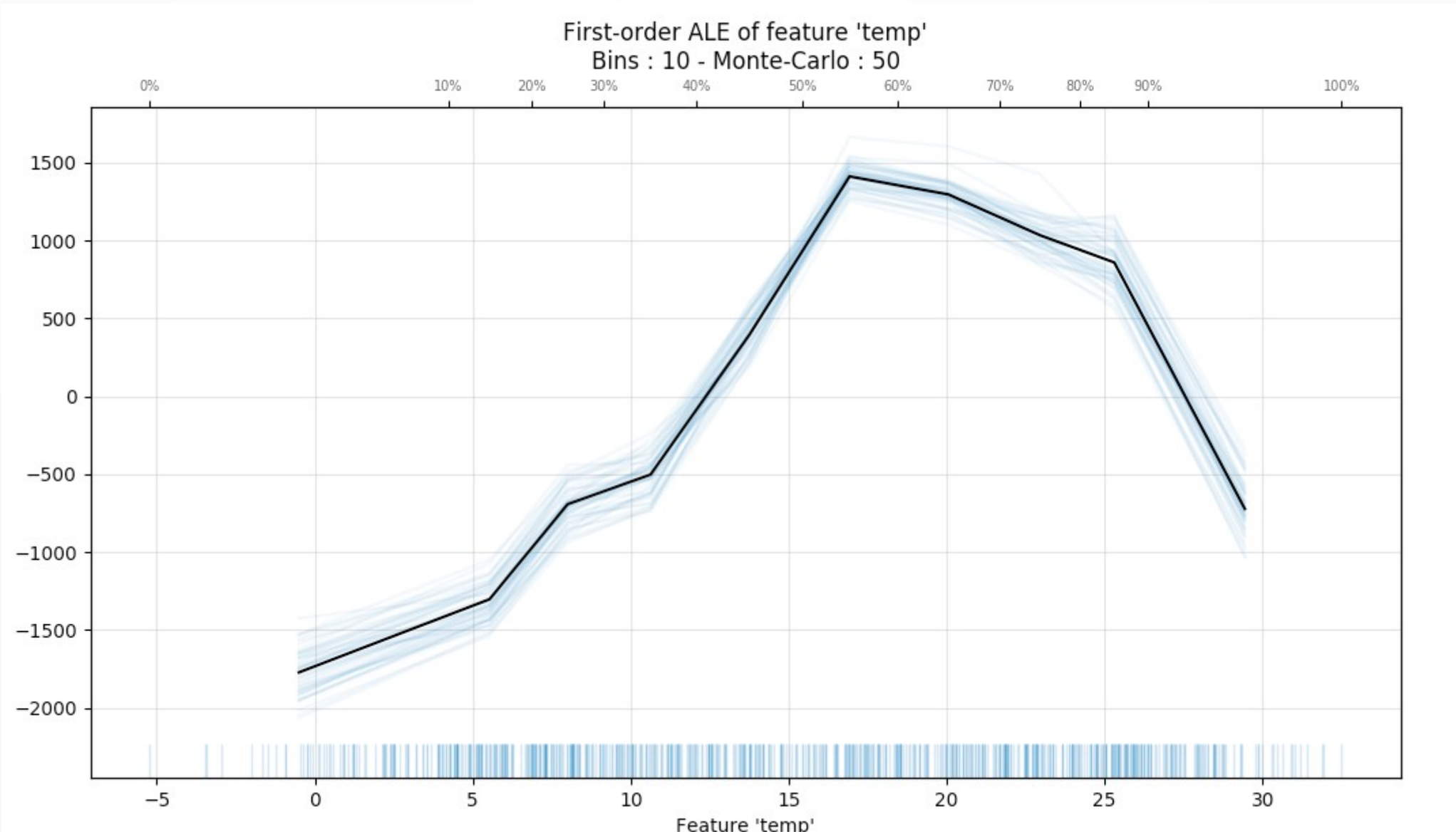
- Если модели на CV хорошо, а на тесте плохо
- Тест и трейн — из разных распределений?
- Классификатор тест-трейн
- Ищем «важные» признаки, выкидываем
- Выкидываем непохожие на тест точки
- Удаляем неправильно «влиятельные» точки
- Взвешиваем точки вероятностью теста

Partial Dependence Plot

Individual Conditional Expectation

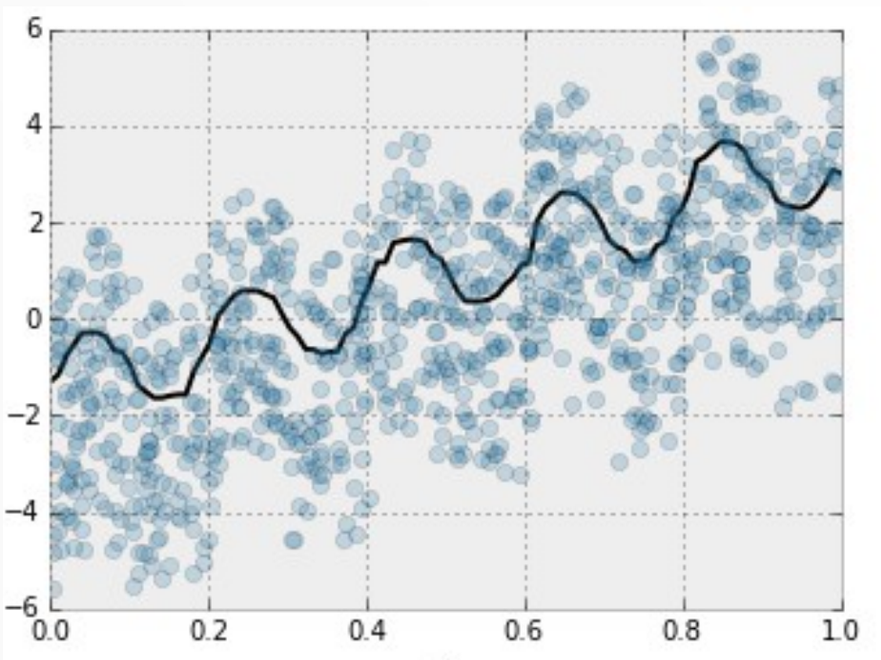


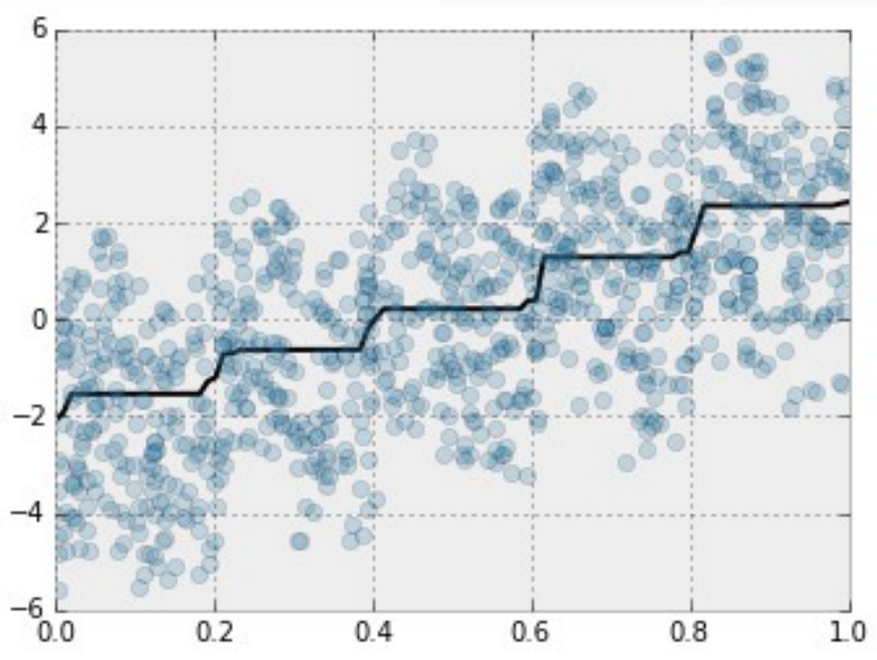
Accumulated Local Effects



Монотонные деревья

- Монотонные модели более интерпретируемы
- В XGBoost есть Monotonic Constraints



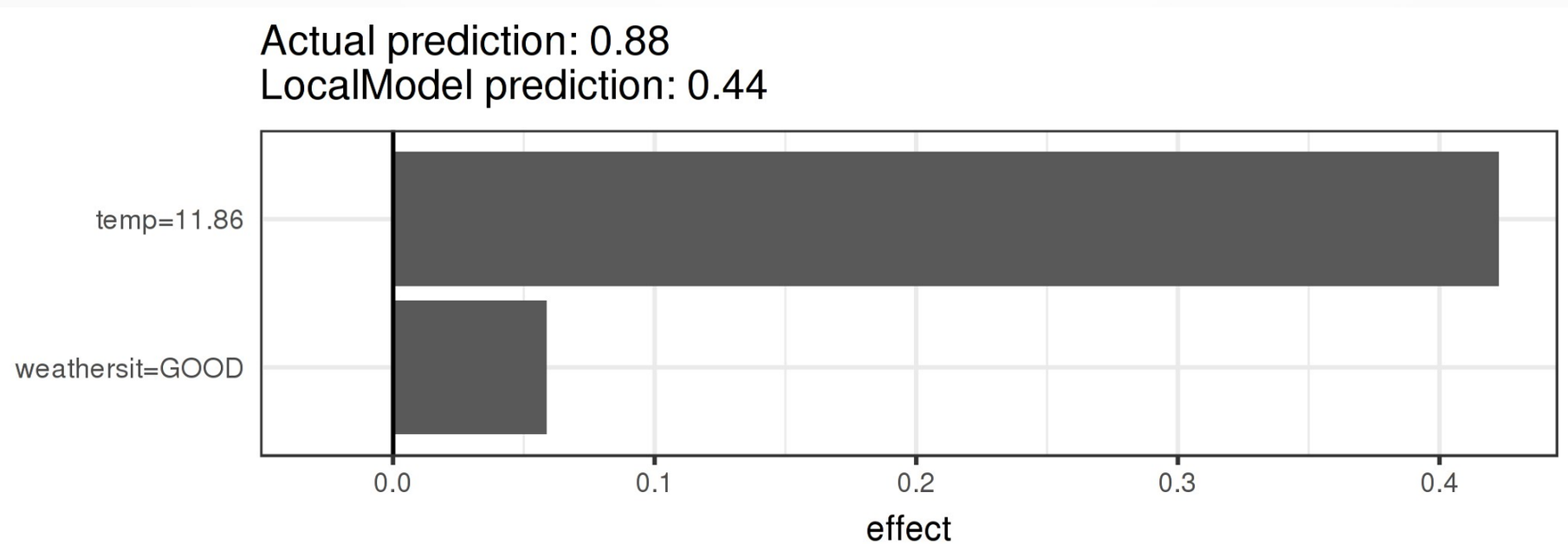


Глобальные суррогаты

- Обучаем простую модель предсказывать поведение сложной
- Предсказываем не мир, а модель
- Дерево
- Квантильная регрессия
- kNN
- Объяснимые (если суррогат объясним)
- Можно использовать другой набор признаков

LIME — локальные суррогаты

- Сурогатная модель для одной точки
- Интерпретируемуя селективная модель
- Может обучаться на других признаках

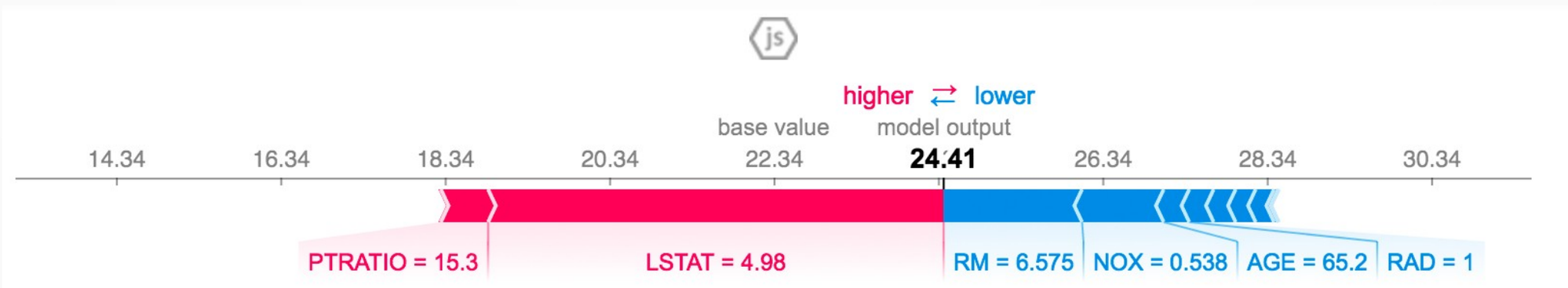


LIME — текст

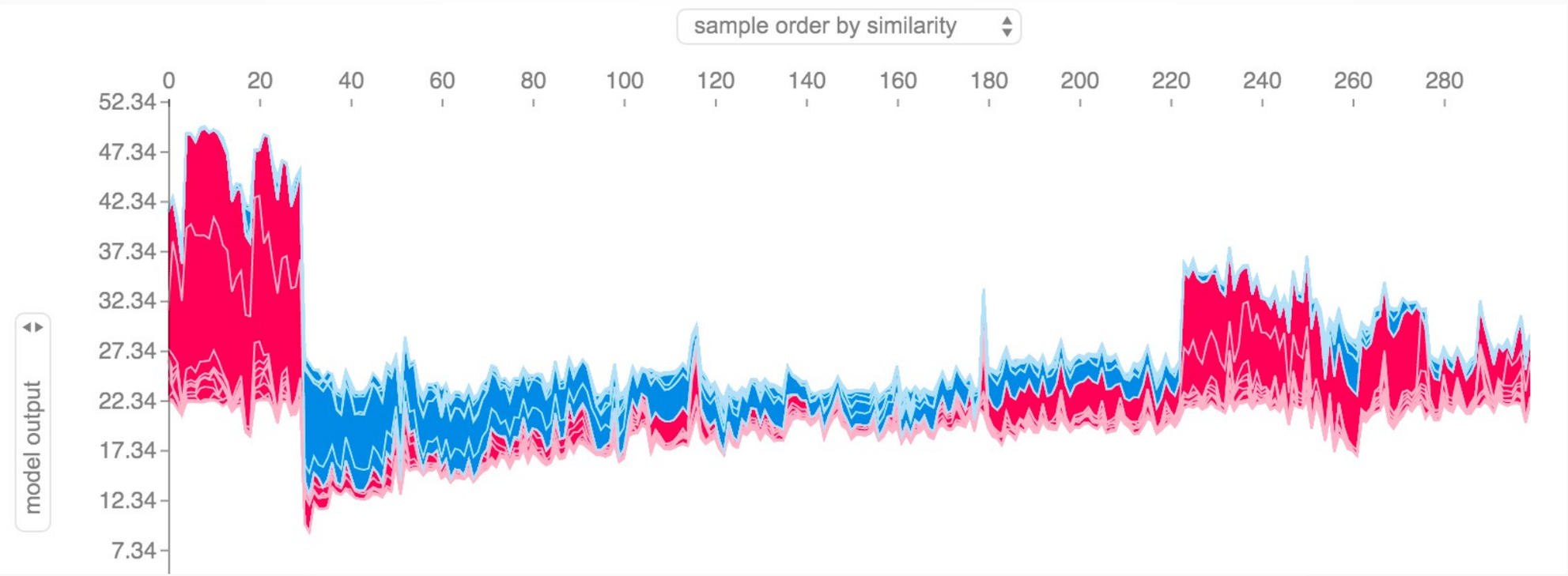
- For Christmas Song visit my channel! ;)

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

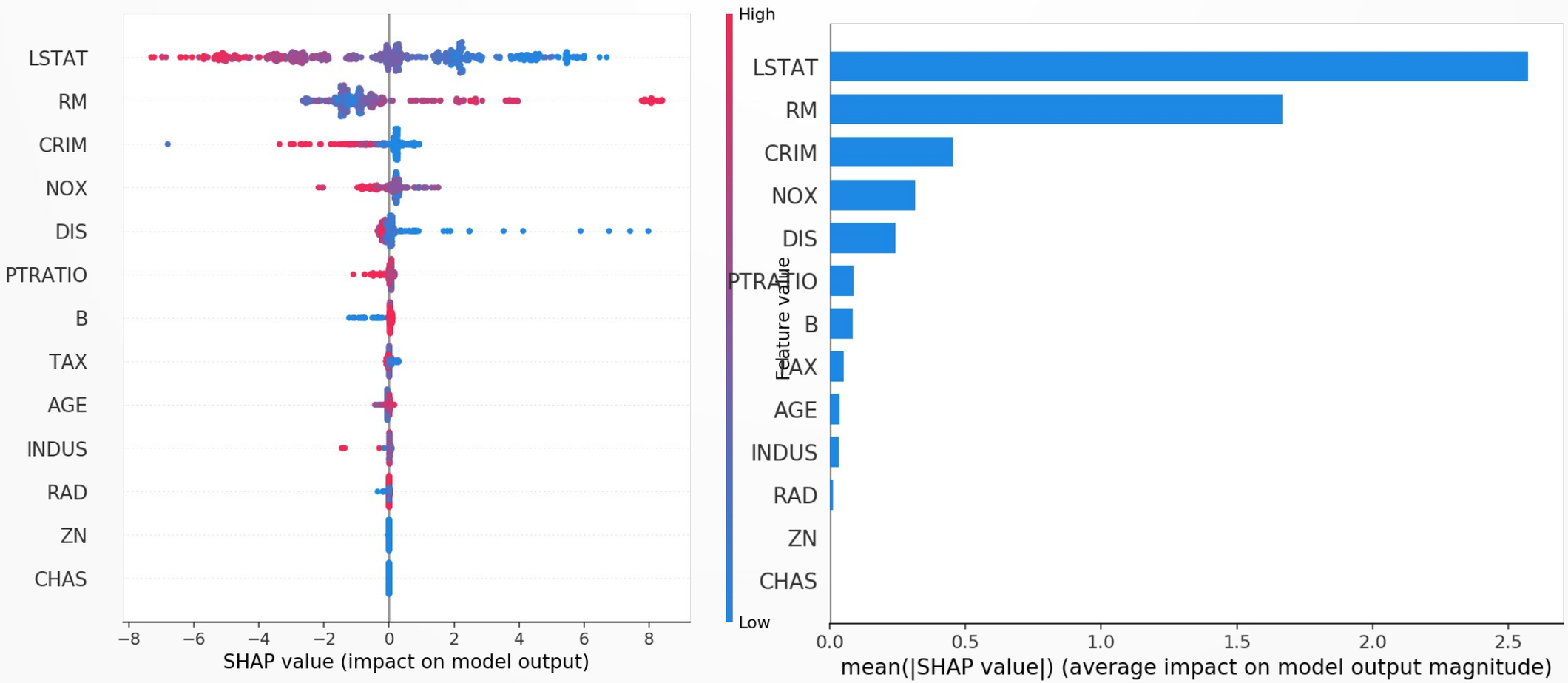
Shapley Values & SHAP



SHAP и структура данных

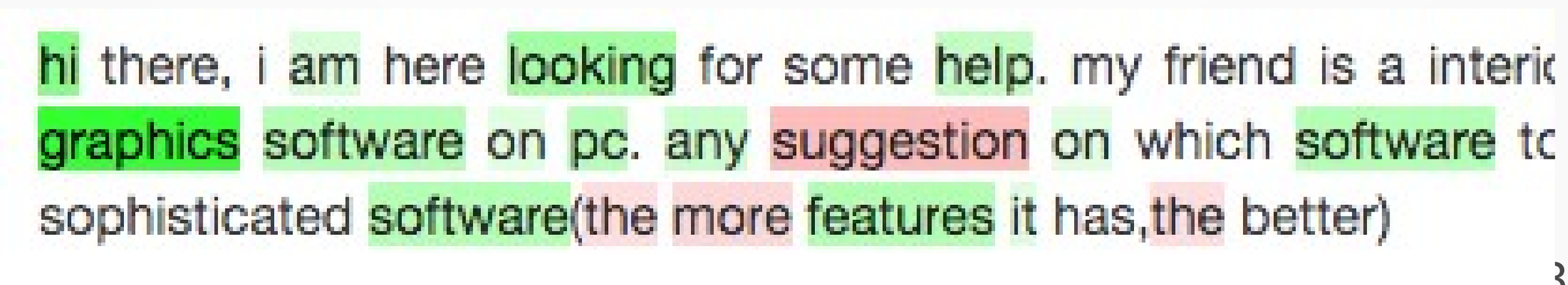


SHAP — глобальный



Eli5

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



Catboost

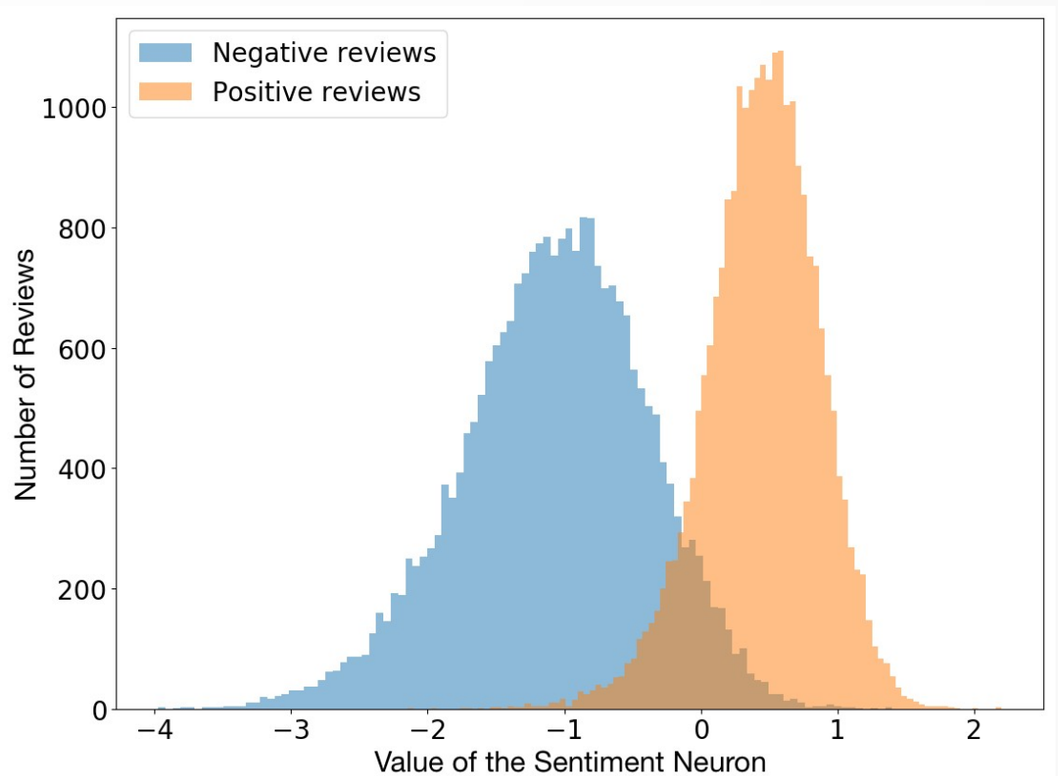
- Feature importance
  - PredictionValuesChange
  - LossFunctionChange
  - ShapValues
  - Interaction
- Object importance
  - Average
  - PerObject

Про нейронки

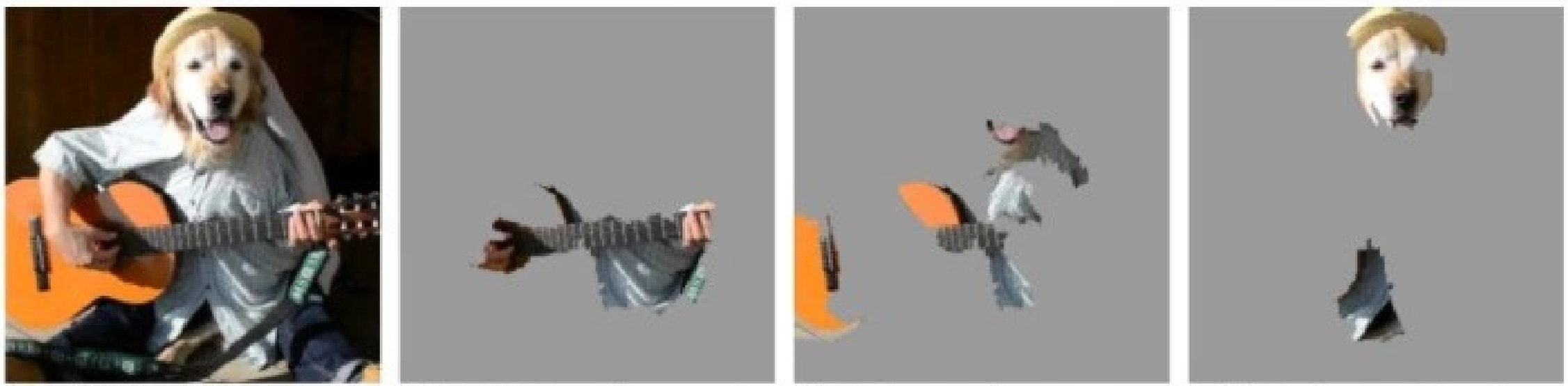
- Градиент пиксела ~ необычность картинки
- Grad-CAM и рассказ на DF6
- Approximating CNNs with Bag-of-local-Features
  BagNet: ансамбль из Resnet50 на патчах 33x33

Нейронки — юниты

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

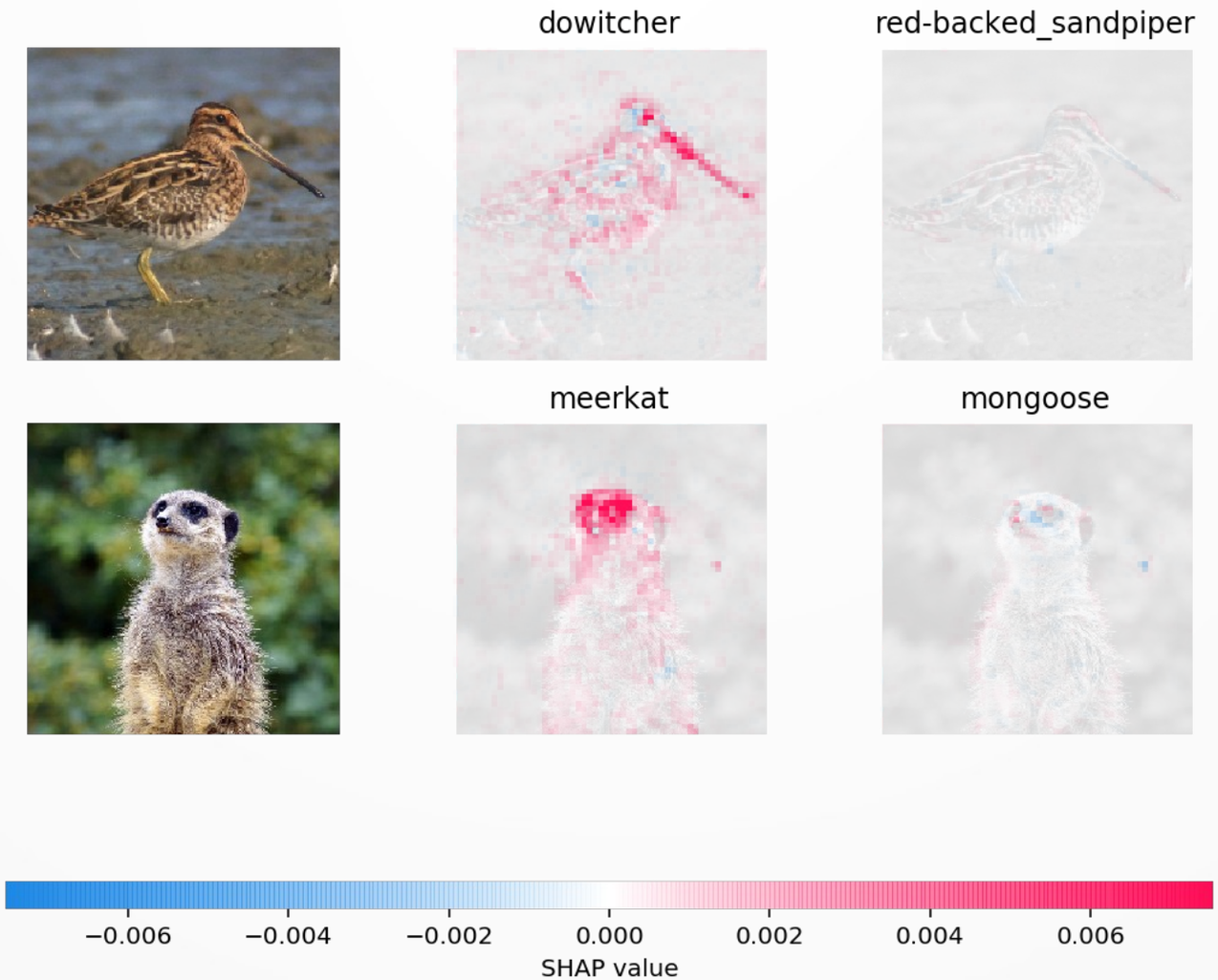


LIME — суперпиксели

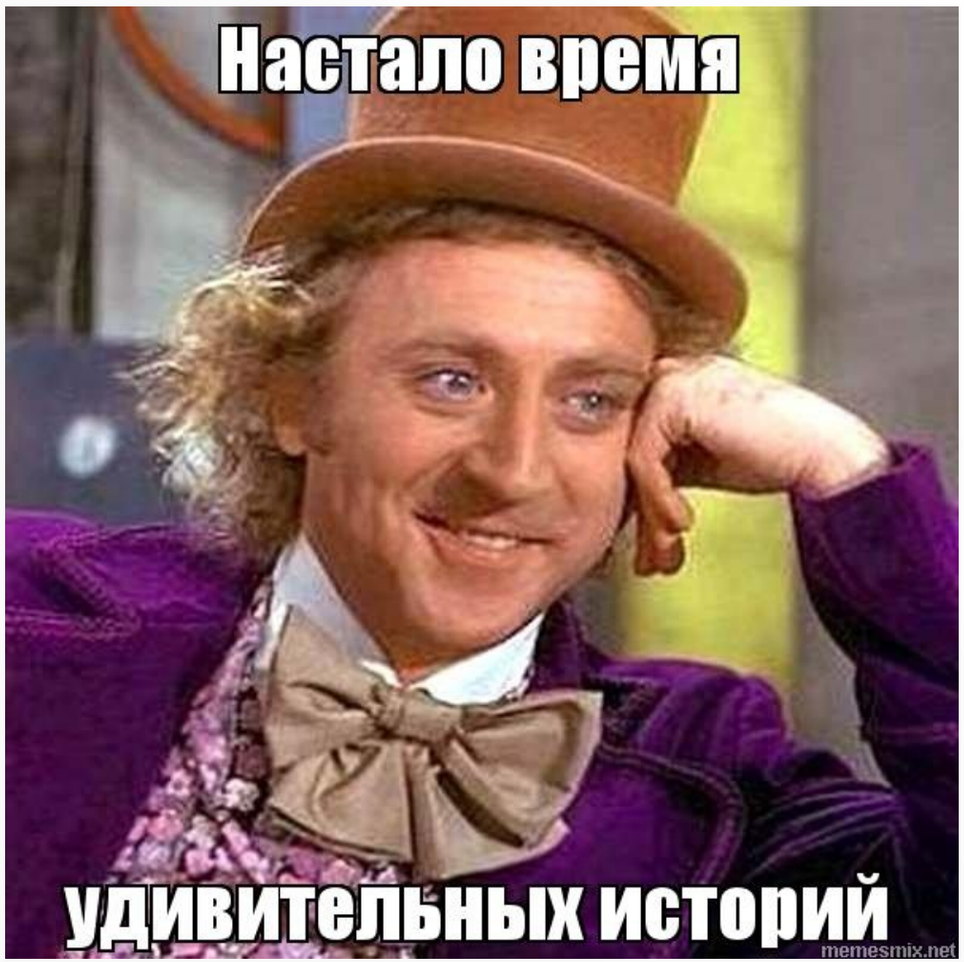


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

SHAP — влияющие точки



Три игрушечных примера



- Извлечение знаний.
  Что модель знает о мире?
- Бутстрап разметки.
  Когда данных мало,
  но нужно качество.
- Отладка.
  Что делать, когда
  ничего не получается?

Авария в колцентре

- Иногда возникает ошибка
- Шаблон не просматривается
- Три системы, у всех все работает
- Строим модель ошибки на логах
- Вызов: Извлечение знаний
- PDP по каждому признаку
  - пик после апдейта
- Взаимодействие признаков
  - обстоятельства звонка

Земельные участки

- Бутстрапим классификатор
- Разметки мало, дорогая, спорная, с заполнением
- CV не очень надежная
- Вызов: Повысить качество предразметки
- SHAPE — поняла ли модель задачу?
- Каких фич добавить, чтобы поняла?
- Влияющие точки: ошибки разметки, влияние заполнения.

Рецепт отладки

- Быстрая диагностика (ищем странное)
  - 10 самых уверенных ошибок, неуверенных и неустойчивых предиктов
  - Важность. На что модель обращает внимание?
- Что на входе: ошибки в данных, взаимодействие признаков
  - Влияющие точки для странных точек и в целом: убираем мусор
  - Адверсариал тест: трейн и тест вообще знакомы?
  - Граф корреляции и граф взаимодействия признаков: можно ли верить «важности» и PDP. Отбор и генерация признаков.
  - Таргет по бинам для наиболее «важных» признаков: важны ли они
- Что на выходе:
  - Локально SHAP / LIME. PDP / ICE для «важных» , SHAP, ALIBI
- Обсуждение модели с постановщиком задачи
  - Дерево, характерные и влияющие точки, таргет по бинам, SHAP

Почитать

- Дьяконов, Интерпретации чёрных-ящиков
- Becker, Machine Learning Explainability
- Molnar, Interpretable Machine Learning
- CVPR 2018 Tutorial
- ICCV 2019 Tutorial
- MIT Network Dissection

Вопросы?

Слайды тут

dkolodezev

promsoft

dkolodezev

d_key

dmitry_kolodezev

https://kolodezev.ru/download/slides-interpretation-v2.pdf