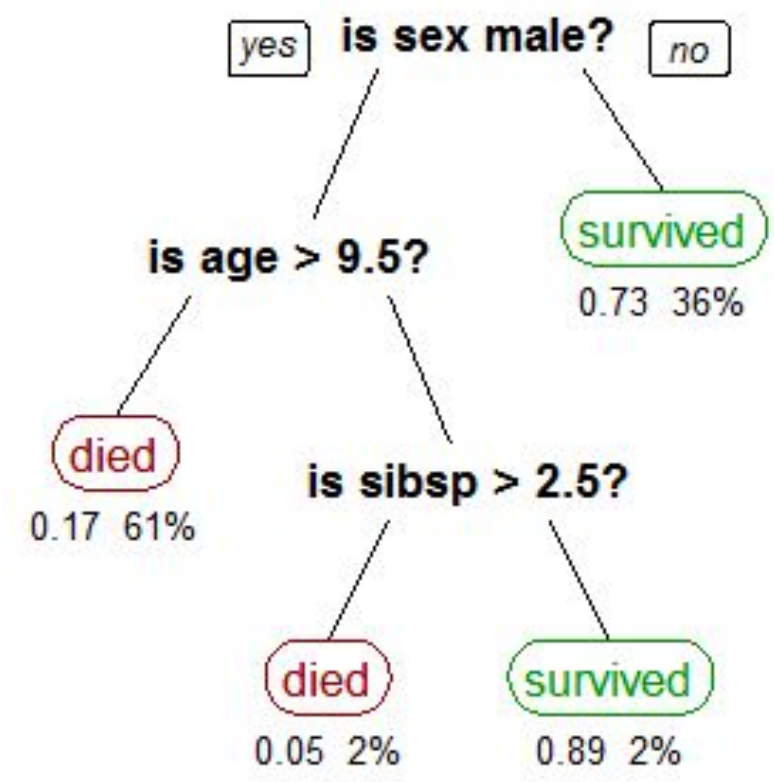
iML и XAI

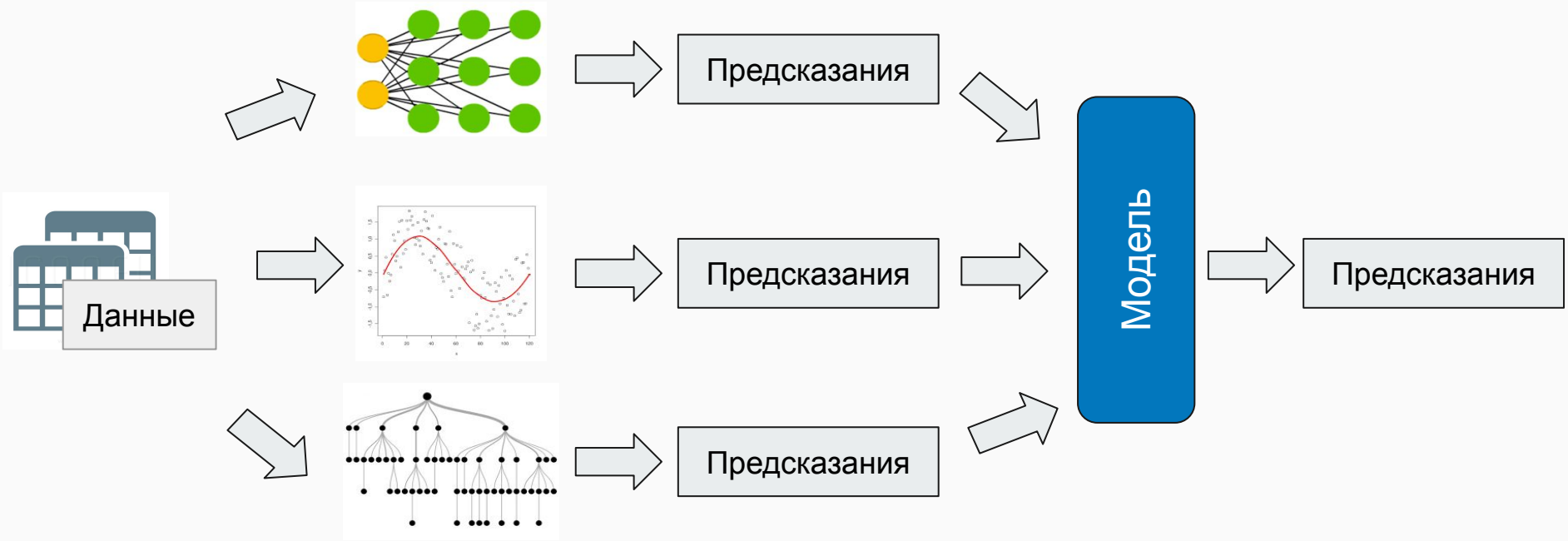
интерпретируемое машинное обучение

НУГ “Машинное обучение и социальный компьютинг” (апрель, 2020),
проект 20-04-024

Модели: дерево

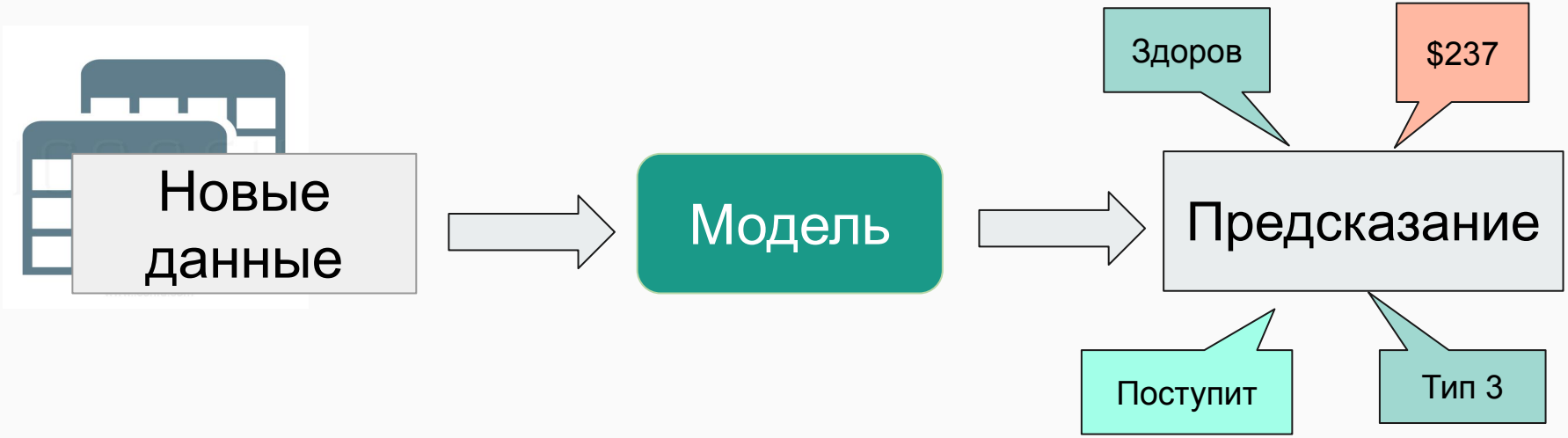
Классификационное дерево предсказания выживаемости на Титанике



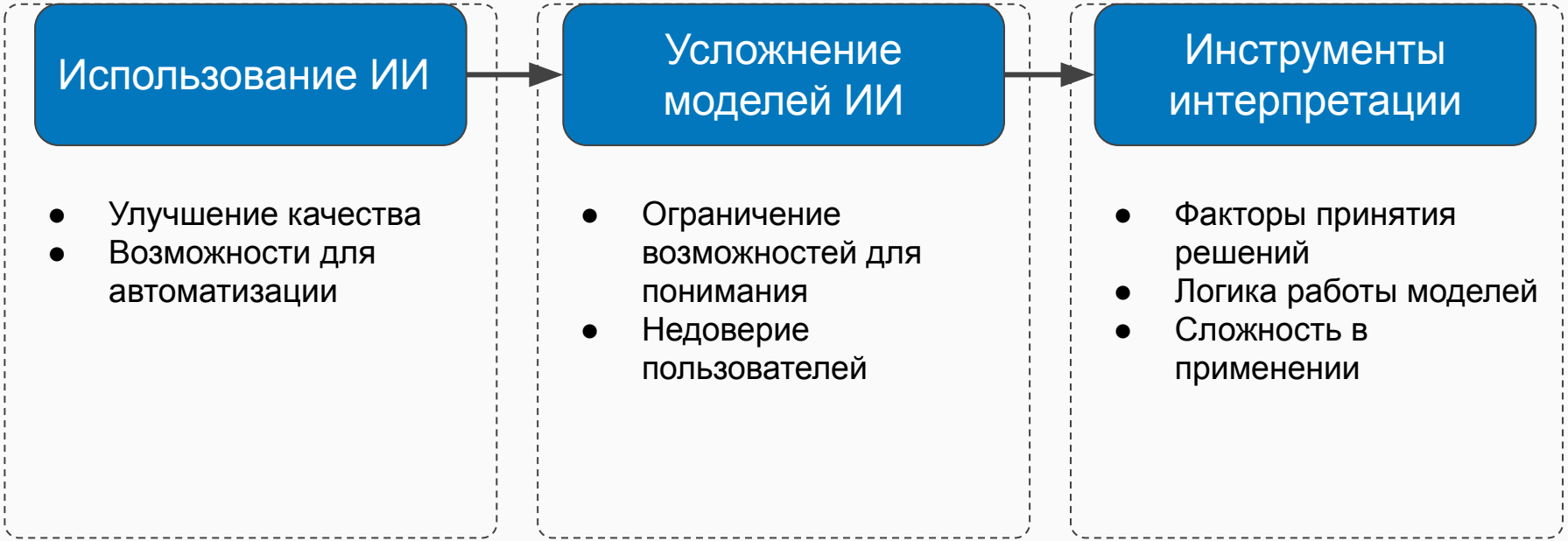


Пример: stacking

“Черный ящик”


Новые данные
Модель
Здоров
$237
Предсказание
Поступит
Тип 3

Зачем


Использование ИИ
Усложнение моделей ИИ
Инструменты интерпретации
Улучшение качества
Возможности для автоматизации
Ограничение возможностей для понимания
Недоверие пользователей
Факторы принятия решений
Логика работы моделей
Сложность в применении

VS

Хаски

Волк

Ribeiro M. T., Singh S., Guestrin C. Why should i trust you?: Explaining the predictions of any classifier (https://arxiv.org/pdf/1602.04938.pdf)

Ответ: волк

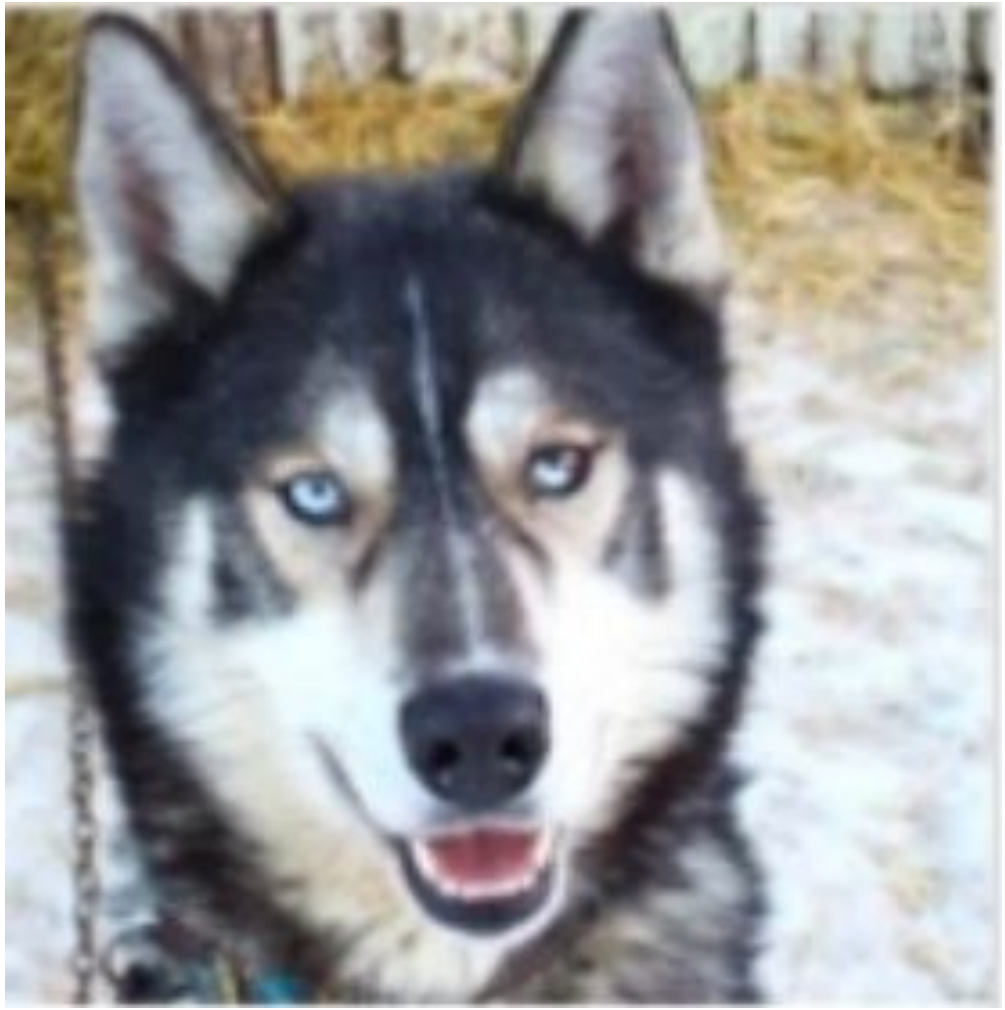
Алгоритм: волк

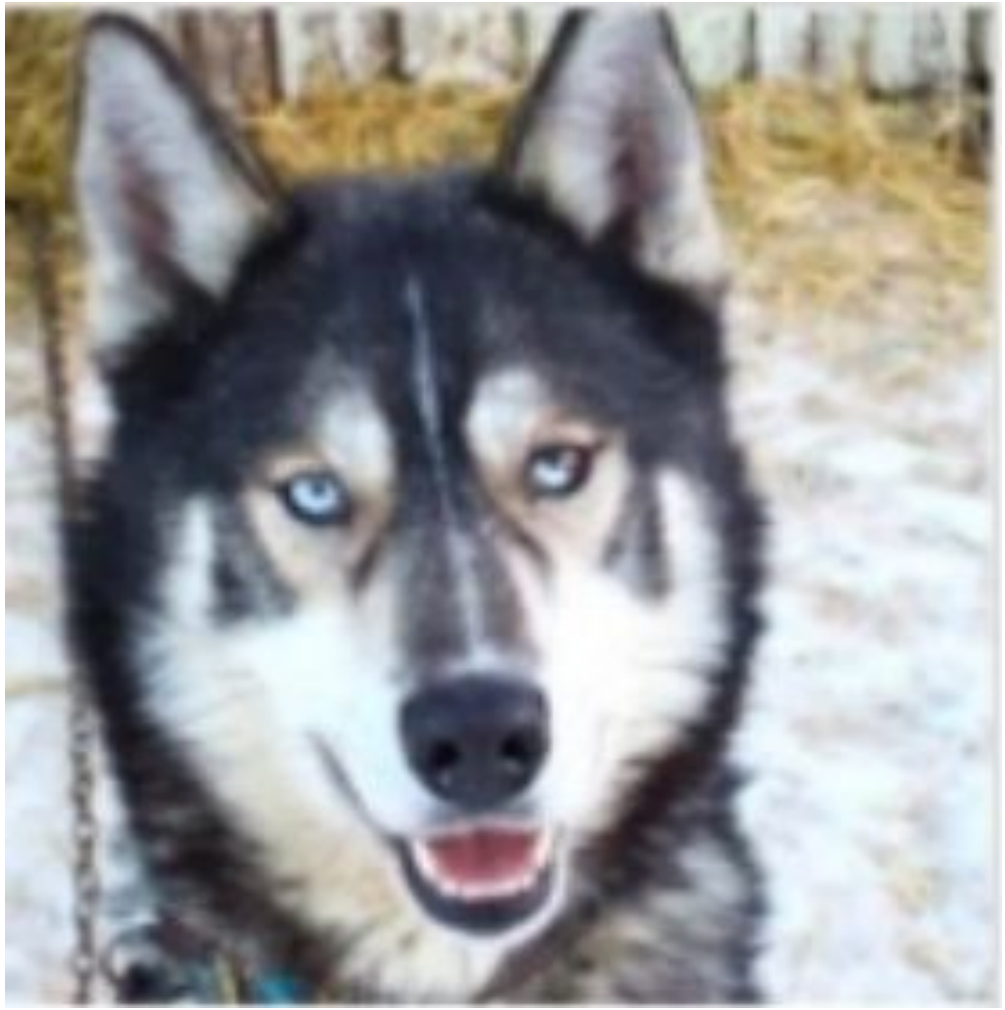
Ответ: хаски

Алгоритм: хаски

Ответ: волк

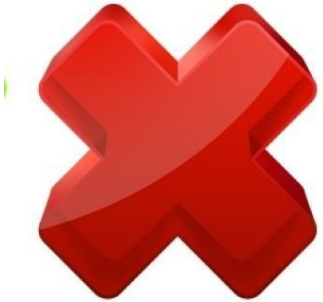
Алгоритм: волк

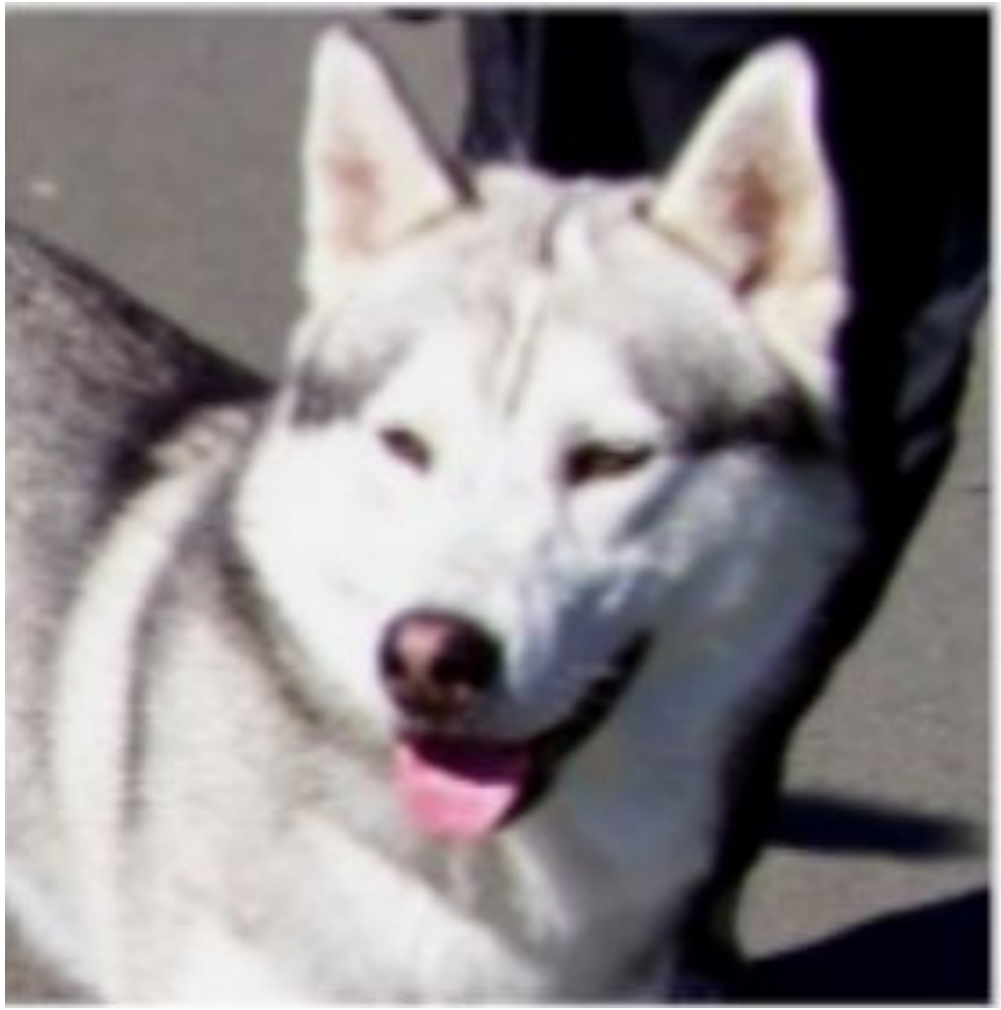




Ответ: хаски

Алгоритм: волк



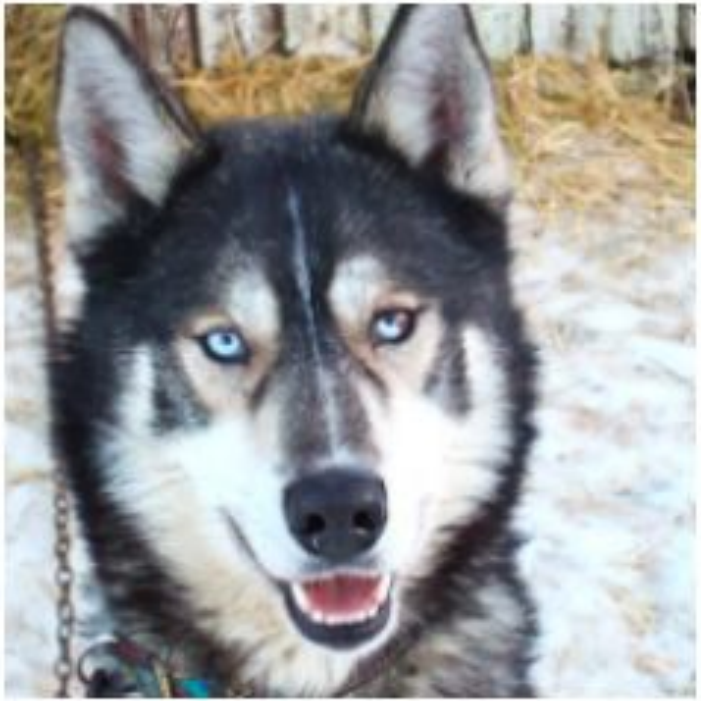


Ответ: хаски

Алгоритм: хаски

Ответ: волк

Алгоритм: волк



Интерпретировать неинтерпретируемое

Интерпретируемое машинное обучение

Что такое интерпретации?

Интерпретируемость — степень, до которой человек способен понять причины решения

Цель интерпретации — описание внутренней логики работы системы

Зачем нужны интерпретации?

- Обоснование принятия решений
- Выявление смещений в моделях
- Выполнение требований к “прозрачности” (GDPR)

Кому нужны интерпретации?

- Разработчики моделей машинного обучения
- Люди, принимающие решения (врачи, менеджеры)
- Потребители продуктов с ИИ

Новое направление

- Активно развивается: XAI (Explainable AI), Interpretable ML
- Секции на ведущих конференциях:
  - AI/ML & seeing through the black box (CHI 2020)
  - Coping with AI: not agAIn! (CHI 2020)
  - Explaining Machine Learning Predictions: State-of-the-art, Challenges, and Opportunities (NeurIPS 2020)
  - Algorithmic Fairness through the Lens of Causality and Interpretability (NeurIPS 2020)
  - Human-Centered Explainability for Healthcare (KDD 2020)
  - Interpretable Models (KDD 2020)
  - Explainable Models for Healthcare AI (KDD 2018),
  - Interpretable ML Symposium (NIPS 2017),
  - Explainable AI (IJCAI 2018),
  - Explainable artificial intelligence (XAI): Why, when, and how? (Strata Data Conference 2018)

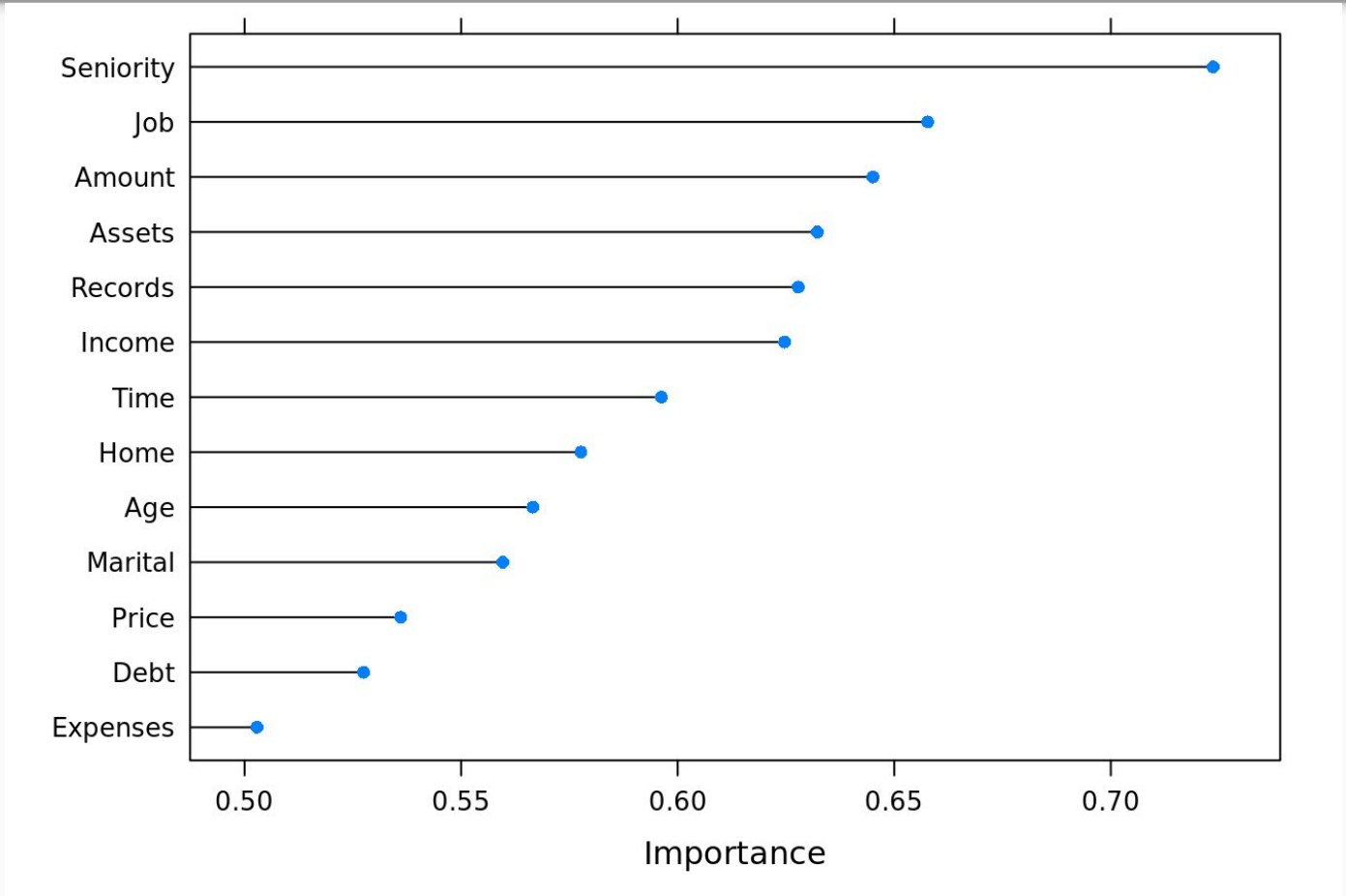
О чем

- Какие проблемы возникают (этика, общество)
- Подходы к решению: “подправленные алгоритмы”, использование только интерпретируемых моделей
- “Надстройка” интерпретации на черный ящик

Глобальная интерпретация

1. Оценка значимости признаков (importance), выделение значимых
    - зависящие от модели
    - не зависящие от модели (изменение качества предсказания)
2. Исследование изменения предсказания при изменении какой-то переменной (ICE графики)

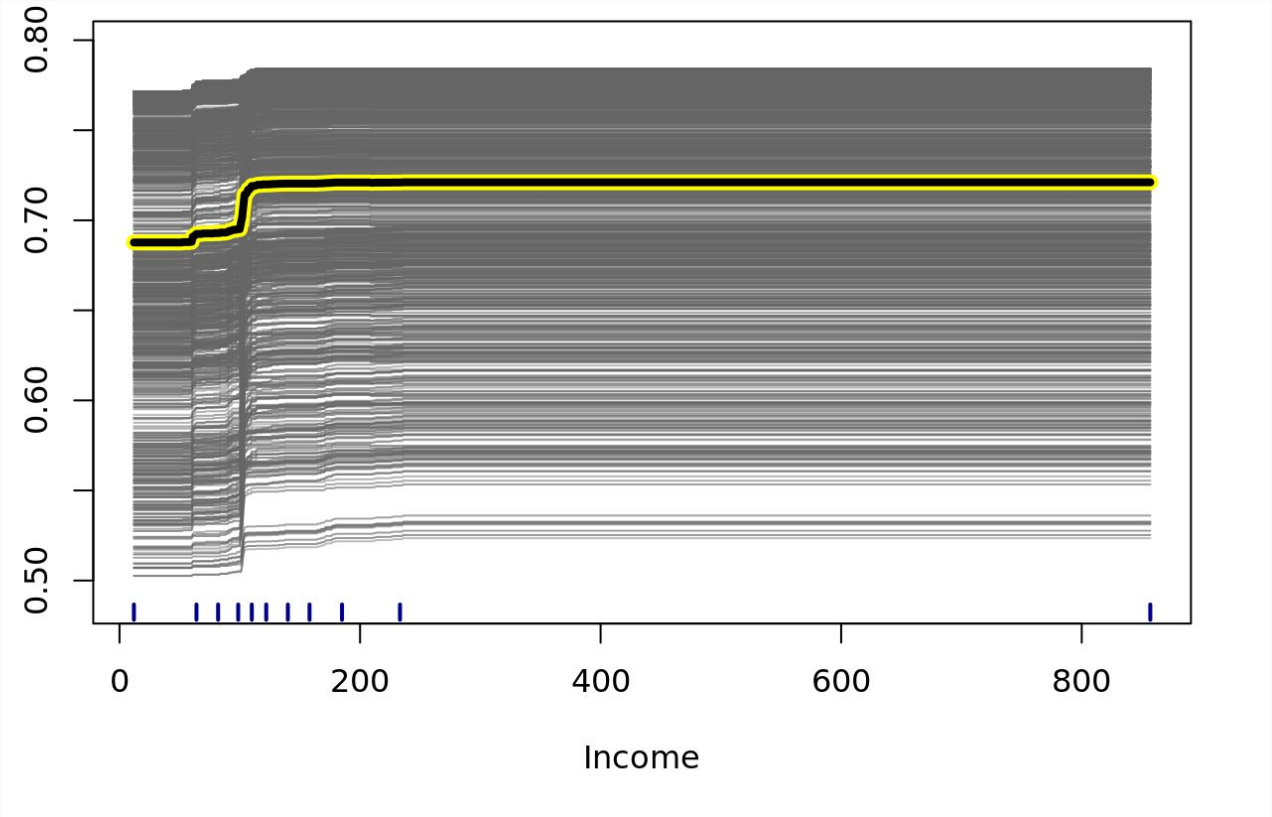
Значимость признаков



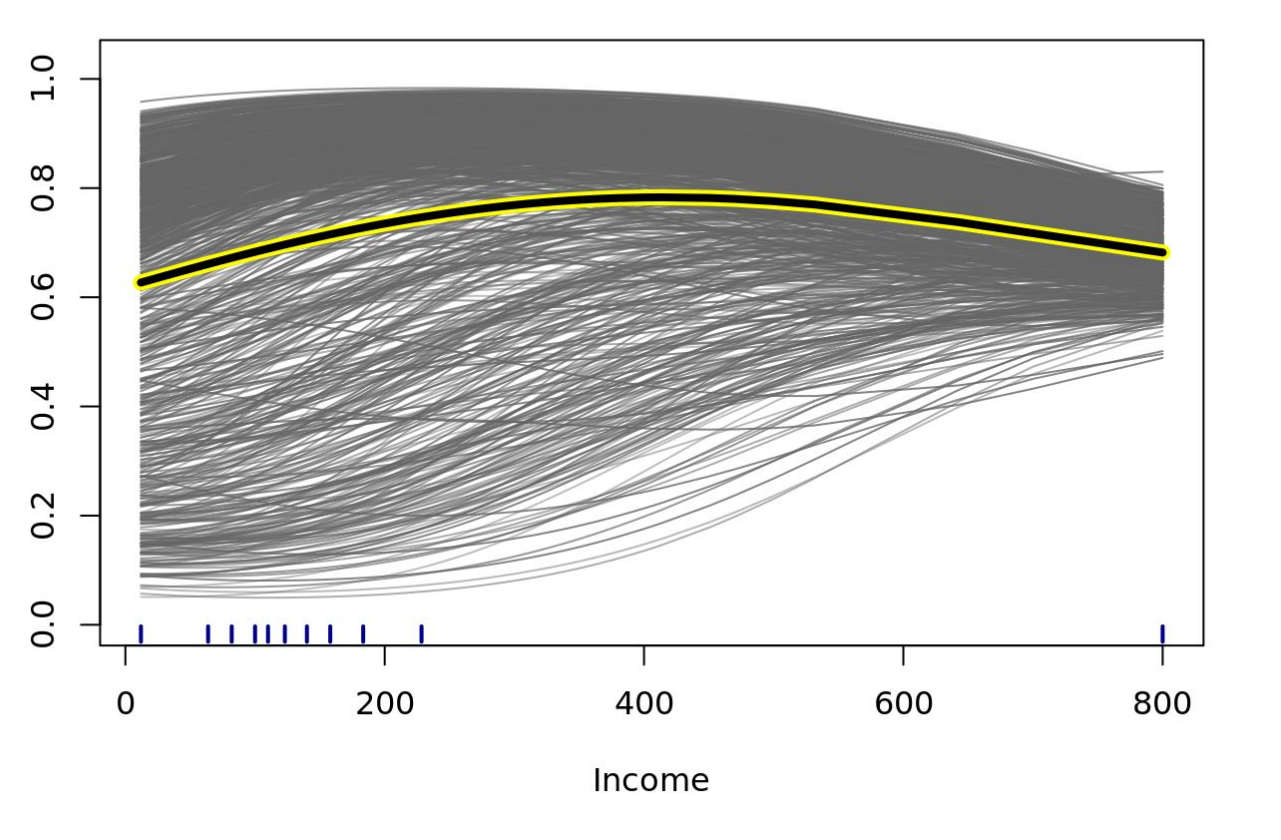
Визуальное исследование переменных

ICE (Individual Conditional Expectation) графики - показывают изменение предсказания при изменении одной из переменных

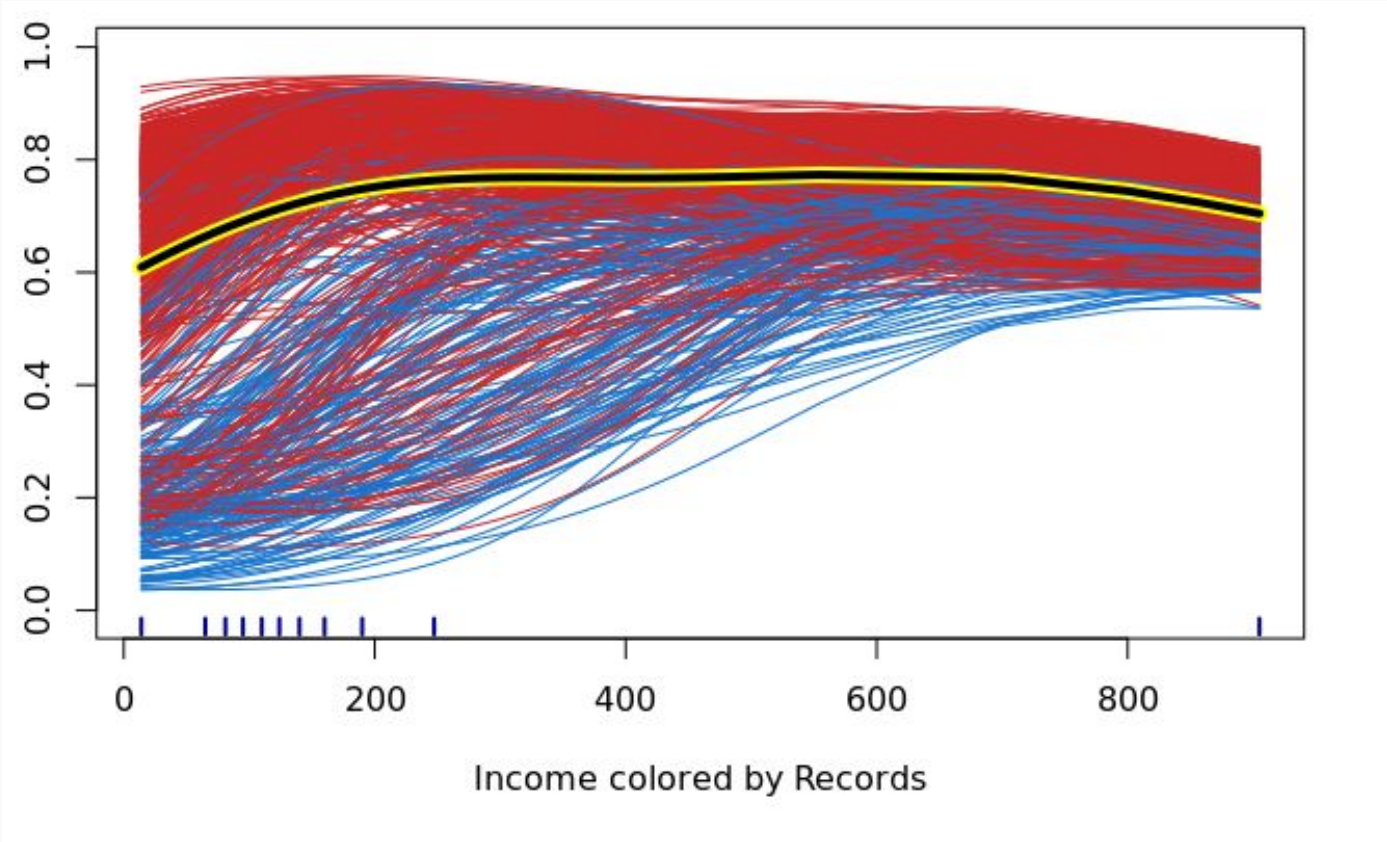
ICE график: доход (градиентный бустинг)



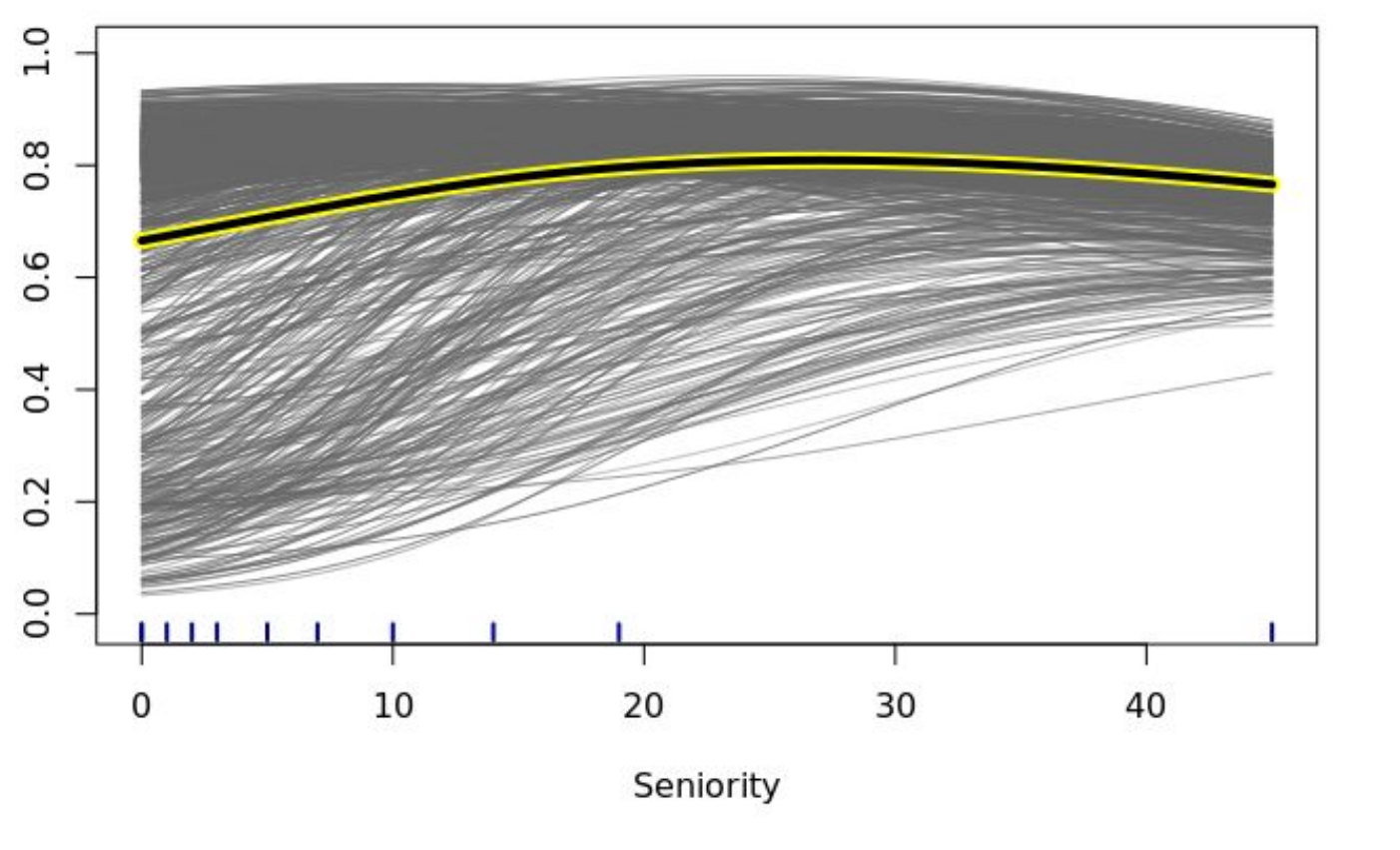
ICE график: доход (SVM)



ICE график: доход и задолженности (SVM)



ICE график: опыт работы (SVM)

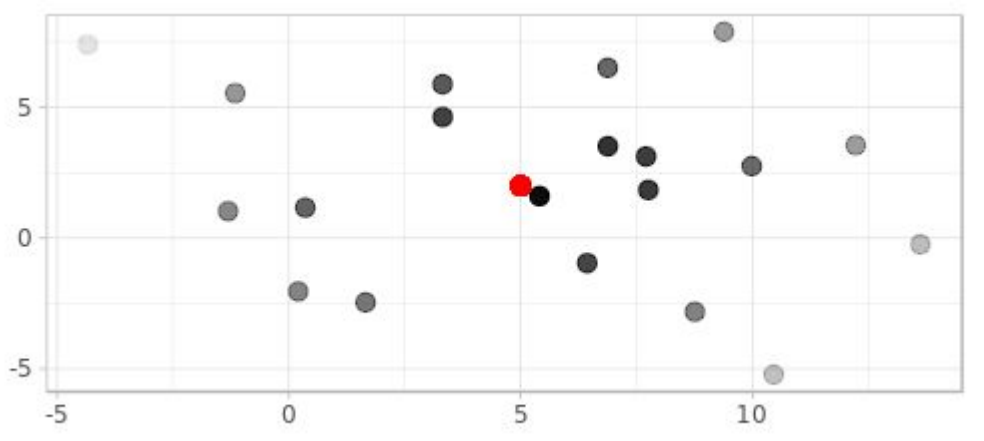


Локальная интерпретация

- Глобальная интерпретация получается достаточно обобщенной "усредненной" по всем данным
- Хотим исследовать конкретный пример, понять, какие факторы привели к тому, что у клиента плохой кредитный статус (и как можно его изменить)
- Для ответа на такие вопросы существуют алгоритмы локальной интерпретации

Локальная интерпретация: LIME

посмотрим на окрестности нашего примера. Т.е. будем изменять значения предикторов случайно и изучать, как это влияет на результат



LIME: приближение в окрестности

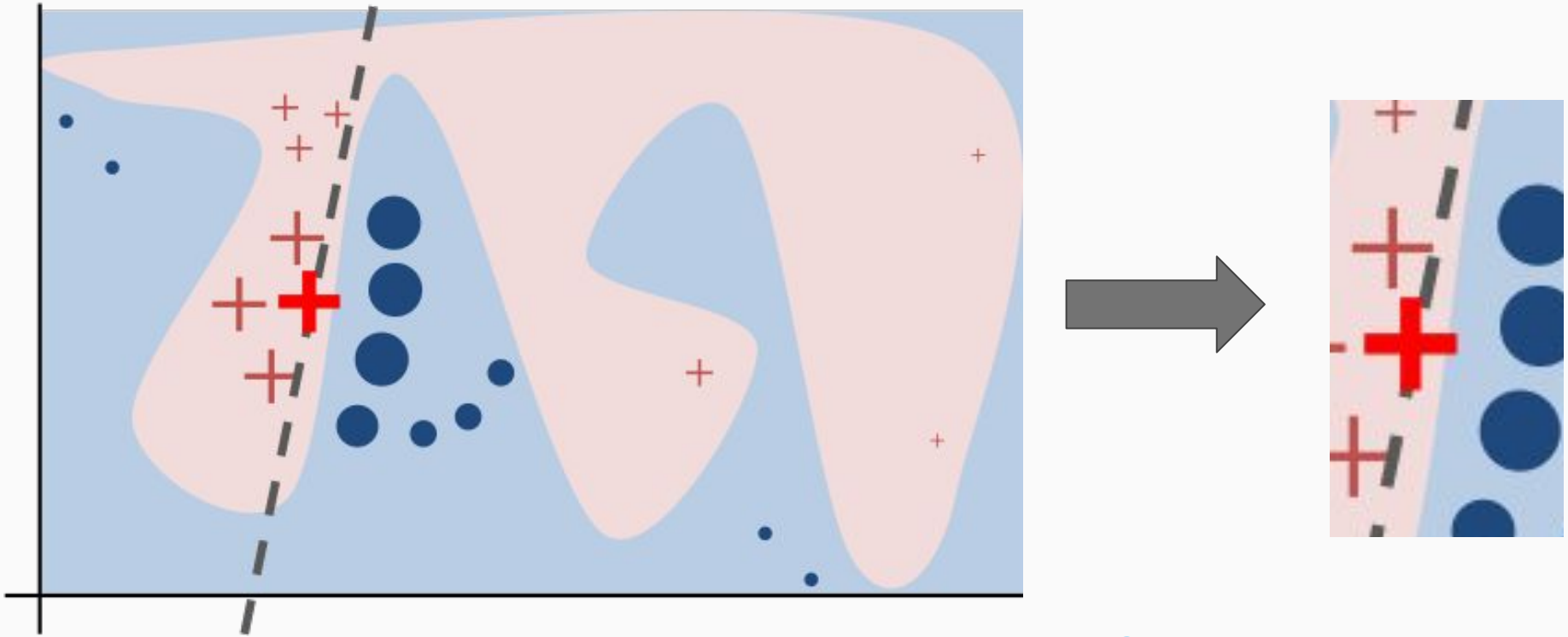
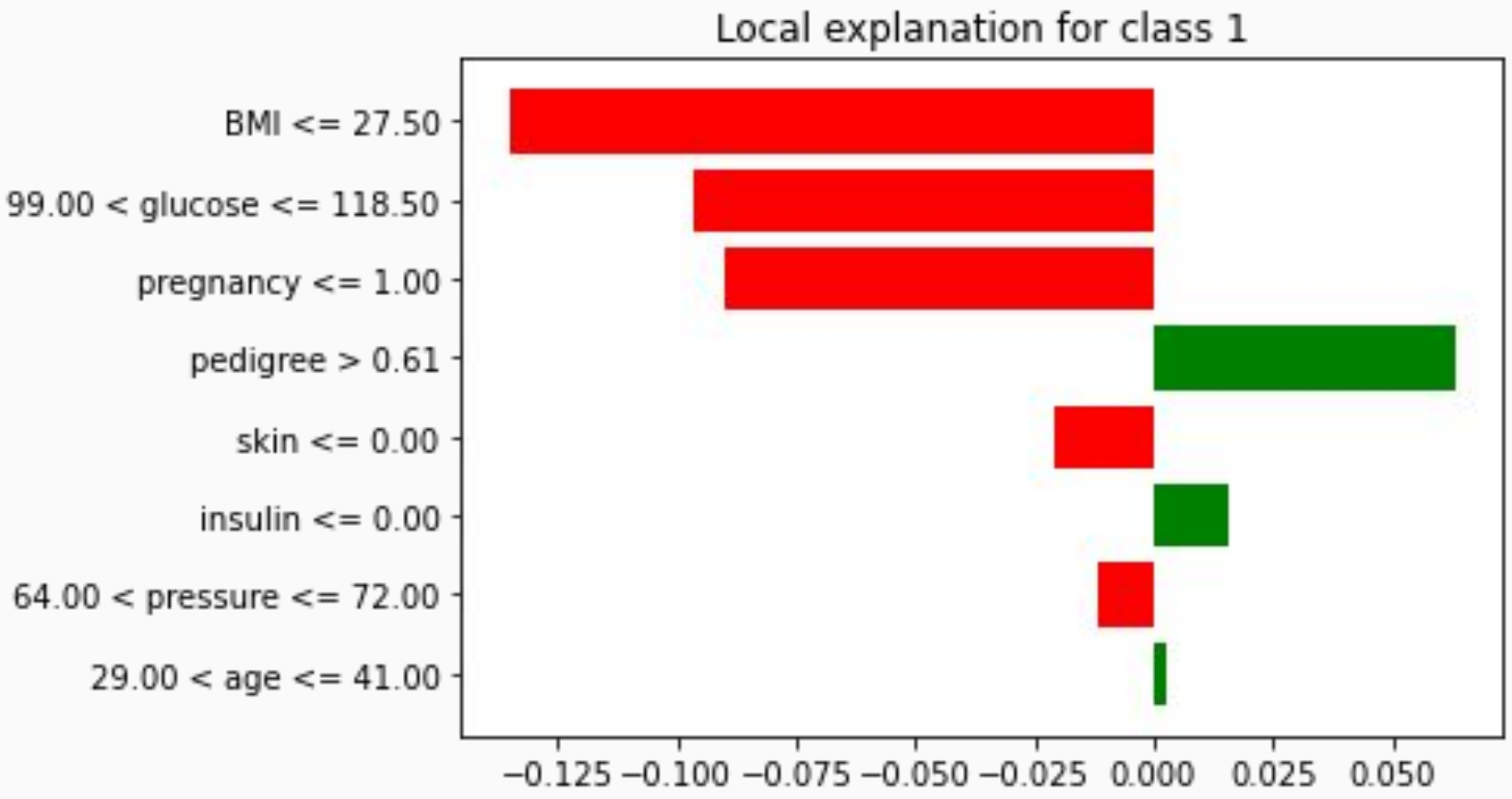


Рисунок из Github профиля автора метода

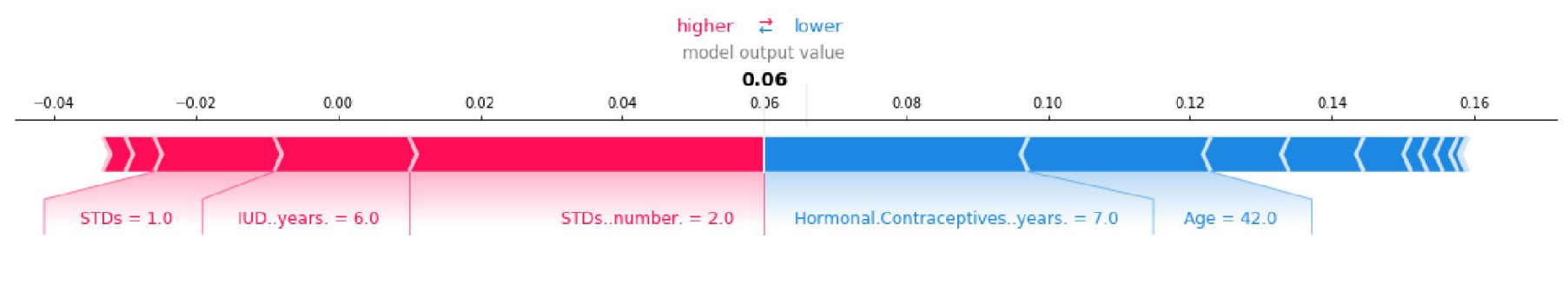
LIME: алгоритм

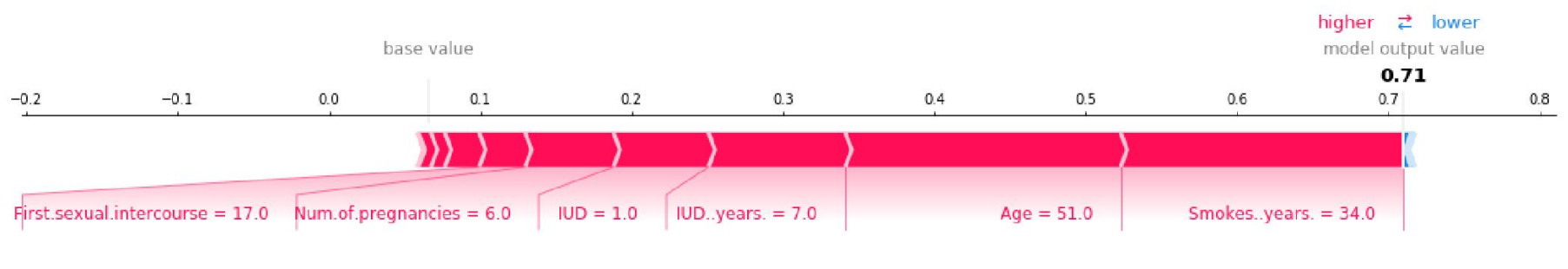
1) генерируем искусственные данные вокруг примера
2) получаем для них предсказание согласно нашей модели
3) используем какую-нибудь интерпретируемую модель (дерево/регрессию), чтобы связать 1 и 2. Важно: данные мы взвешиваем -- те, что ближе к исходному примеру (согласно какой-нибудь метрике близости), весят больше
4) интерпретируем результаты (справедливо **только** для окрестности примера)

LIME: пример, предсказание -- нет диабета



Не только LIME





Пример из книги [Interpretable Machine Learning]

Не только LIME

- Accumulated Local Effects (ALE) Plot
- SHAP (SHapley Additive exPlanations)
- Anchors (от авторов LIME, но результат в виде правил ЕСЛИ-ТО)
- Контр-примеры (“если я изменю признак X, то предсказание изменится на противоположное”)
- Похожие примеры
- Влиятельные наблюдения
- ...

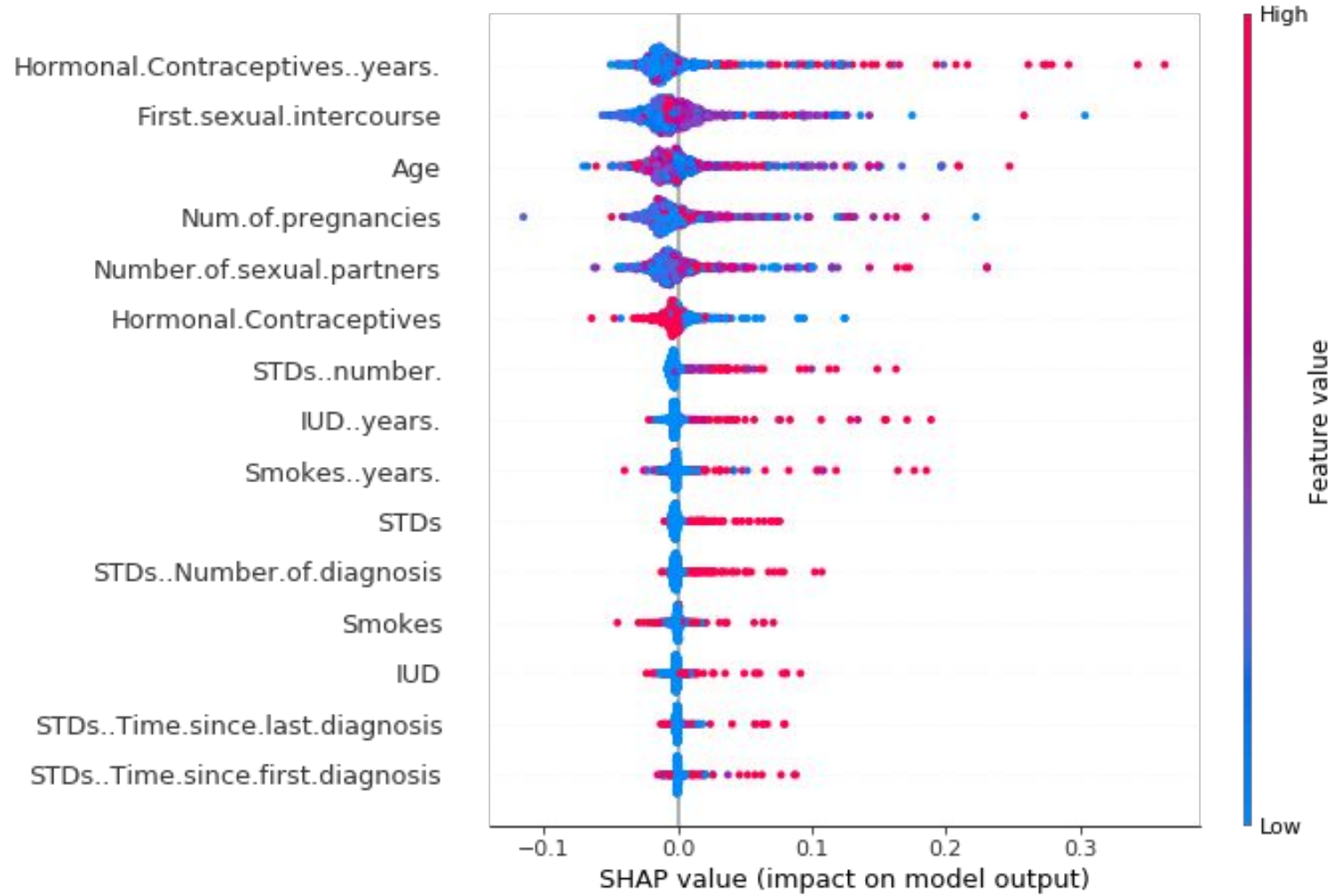
win-win: все отлично

- строим сложный многоуровневый ансамбль с высокой точностью (или другой целевой метрикой)
- поверх строим интерпретацию, чтобы понять, что на что влияет
- победа!

Зачем нам тогда простые модели?
И зачем нам всякие диагностики и т.д.?

Что может пойти не так

- эти методы интерпретируют **модель**, не реальность (другая модель -- другие выводы)
- приближение к приближению (“сломанный телефон”)
- модели могут быть плохими (низкое качество) и модели интерпретации моделей тоже могут быть плохими
- неправильные выводы из модели интерпретации
  - локальные методы дают локальную интерпретацию, нельзя делать выводы в общем
  - нет понимания, что именно показывает тот или метод / визуализация



Пример из книги [Interpretable Machine Learning]

Полезные ссылки

- Interpretable Machine Learning by Christoph Molnar
- Hands-on Machine Learning Model Interpretation (примеры в python)
- библиотека lime
- Local Interpretable Model-Agnostic Explanations (LIME): An Introduction
- (рус) Интерпретируемая модель машинного обучения (очень короткое введение)