

# Машинное обучение: работа над проектом

MADE academy  
Эмели Драль

# Прикладное машинное обучение

1. Предпроектное исследование: от постановки задачи до оценки потенциального эффекта
2. Проектная работа: оптимизация и валидация модели, демо-стенд, разработка сервиса
3. Поддержка и сопровождение сервиса. Чек-лист data-саентиста.

# Этапы работы над проектом

- Постановка задачи
- Определение метрик и критериев успеха
- Оценка доступных данных
- Обучение моделей
- Тестирование моделей (эксперимент)
- Разработка сервиса
- Тестирование качества работы сервиса
- Мониторинг и поддержка качества сервиса, регулярное дообучение модели

# Работа над проектом

1. Оптимизация модели
2. Валидация модели
3. Тестирование в production

# Оптимизация модели

# Оптимизация модели

## Quality vs Complexity

Quality: чем меньше ошибка модели, тем лучше

Complexity: чем проще модель, тем стабильнее она работает

# Оптимизация модели

## Quality vs Complexity

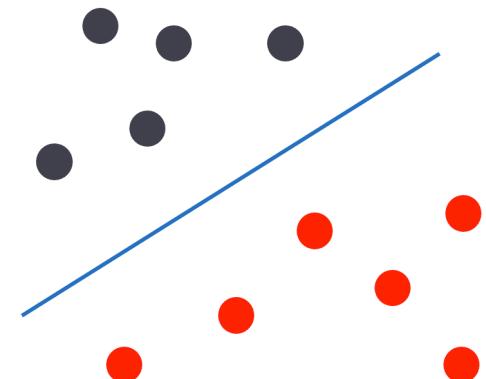
Quality: чем меньше ошибка модели, тем лучше

Complexity: чем проще модель, тем стабильнее она работает

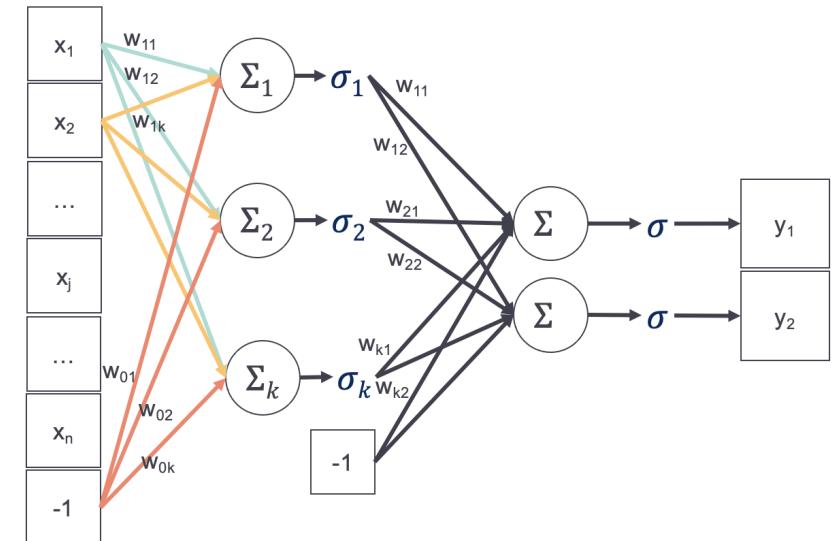
Часто, более сложные модели (или комбинации моделей) дают меньшую ошибку, но для использования в сервисе выбирают ближайший по качеству более простой аналог

# Оптимизация модели

# Quality vs Complexity

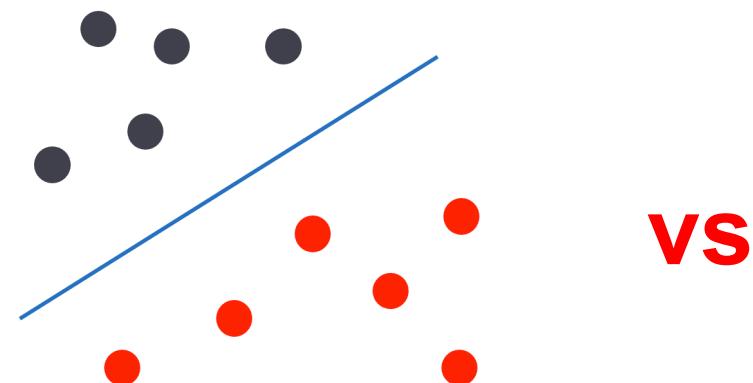


vs



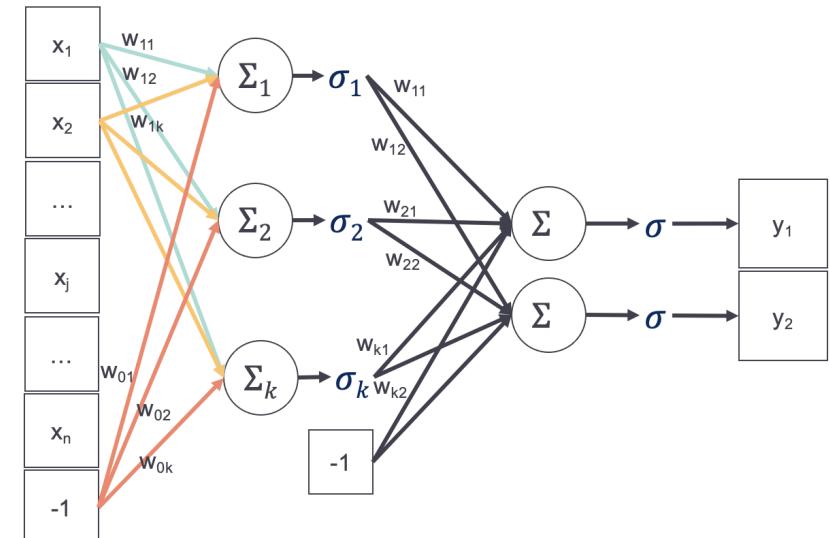
# Оптимизация модели

## Quality vs Complexity



vs

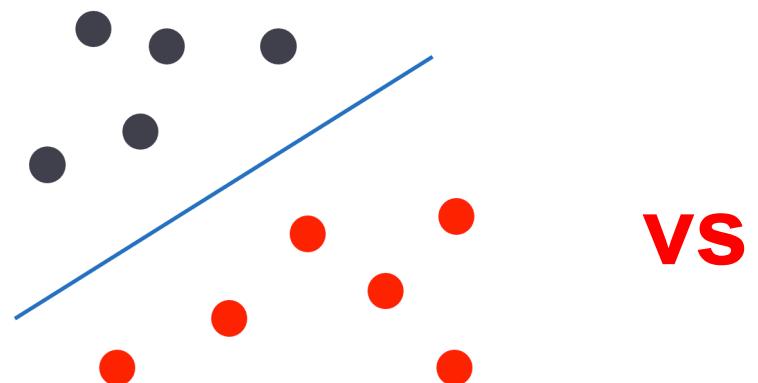
ROC AUC = 0,74



ROC AUC = 0,79

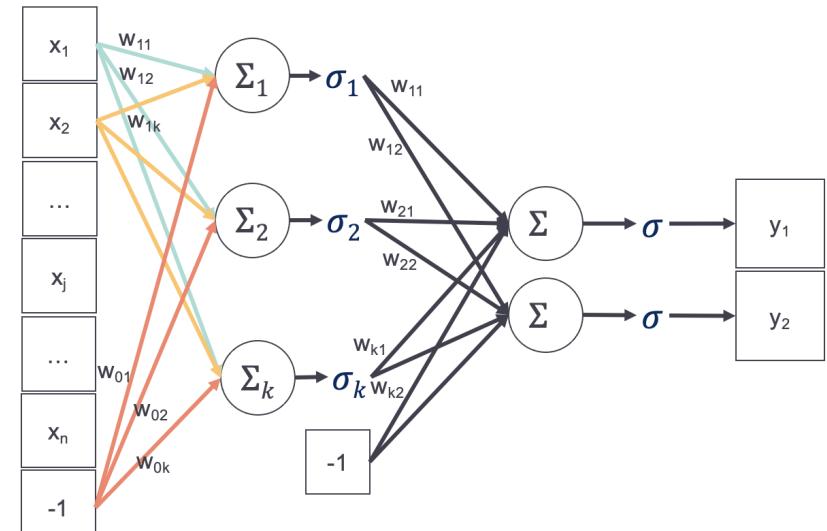
# Оптимизация модели

# Quality vs Complexity



vs

ROC AUC = 0,74



ROC AUC = 0,79

- Связь качества модели и экономического эффекта:  
сколько нам стоит 0.05 ROC AUC?
- Готовы ли мы ради этого эффекта усложнить  
архитектуру для поддержки нейронных сетей?

# Quality vs Complexity

Важно иметь несколько моделей-кандидатов разной сложности и понимать, какой прирост в качестве и эффекте дает усложнение модели

Оптимизация  
модели

# Оптимизация модели

## Модели-кандидаты

Важно иметь несколько моделей-кандидатов разной сложности и понимать, какой прирост в качестве и эффекте дает усложнение модели

Пример:

1. Constant model
2. Simple model with numeric features only
3. Complex model with numeric features only
4. Simple model with some feature engineering
5. Complex model with some feature engineering
6. Hybrid model

# Оптимизация модели

## Constant model

1. Самый популярный класс в задаче классификации
2. Среднее или медиана (посчитанные по обучающей выборке!) в задаче регрессии
3. Last value (можно с учетом сезонности) в задаче прогнозирования
4. Most popular items для рекомендательной системы

# Оптимизация модели

## Constant model

1. Самый популярный класс в задаче классификации
2. Среднее или медиана (посчитанные по обучающей выборке!) в задаче регрессии
3. Last value (можно с учетом сезонности) в задаче прогнозирования
4. Most popular items для рекомендательной системы
  - для каждой задачи можно подобрать условно оптимальную константу
  - это важный benchmark, позволяющий понять ценность решения

# Оптимизация модели

## Constant model

В некоторых индустриях метрики качества даже устроены таким образом, чтобы оценивать относительный прирост качества модели.

Пример: задача прогнозирования оттока в телеком.

Метрика  $lift@k$  - во сколько раз ранжирование среди top  $k\%$  абонентов согласно модели лучше случайного ранжирования?

$$lift@k = \frac{precision@k}{precision@all} = \frac{precision@k}{churn rate}$$

# Оптимизация модели

## Simple model

1. Регрессия по одному или нескольким признакам
2. Дерево решений небольшой глубины
3. Метод ближайших соседей по нескольким признакам
4. Rule-based (часто, это текущее production решение)

# Оптимизация модели

## Модель другого типа

Часто, текущее production решение не является моделью машинного обучения

1. Rule-based system
2. Математическая модель (аналитическая формула)
3. Физическая модель

Их не вполне справедливо считать простыми, но это также хороший benchmark

# Оптимизация модели

## Complex model

Дальнейшее снижение ошибки модели возможно за счет:

- feature engineering
- более сложный алгоритм с большим количеством параметров
- комбинации более сложного алгоритма и feature engineering

# Оптимизация модели

## Complex model

Дальнейшее снижение ошибки модели возможно за счет:

- feature engineering
- более сложный алгоритм с большим количеством параметров
- комбинации более сложного алгоритма и feature engineering

Полезно проанализировать **остатки модели**, чтобы оценить наличие оставшегося сигнала в данных;

Имеем смысл смотреть на **feature importance** добавленных признаков, особенно если их сложно рассчитывать

# Оптимизация модели

## Hybrid model

Альтернативный способ снижения ошибки – использование комбинации из нескольких подходов к решению задачи.

Подходов очень много, например:

- Стандартный stacking
  - Content based + collaborative filtering recommender system
  - Бинарная классификация + регрессия для одного из классов
  - Физико-химическая модель + ml модель
  - Термодинамическая модель + ml модель
- и пр.

# Оптимизация модели

## Quality vs Complexity

Модель	Precision@10% (cv mean)
Constant model	0.08
Physical model	0.71
Linear model (num features)	0.61
GB (feature engineering)	0.76
Physical model + GB on residual	0.82
Ideal model	0.9

Полезно дать рекомендацию о том, какую модель вы считаете оптимальной.

# Оптимизация модели

## Quality vs Complexity

Модель	Precision@10% (cv mean)
Constant model	0.08
<b>Physical model</b>	<b>0.71</b>
Linear model (num features)	0.61
GB (feature engineering)	0.76
<b>Physical model + GB on residual</b>	<b>0.82</b>
Ideal model	0.9

Полезно дать рекомендацию о том, какую модель вы считаете оптимальной.

# Валидация модели по историческим данным

# Валидация модели

## Что нужно оценить?

- Качество модели
- Экономический эффект
- Дополнительные свойства:
- Скорость устаревания модели
- Bias & fairness
- Интерпретация

# Валидация модели

## Качество модели

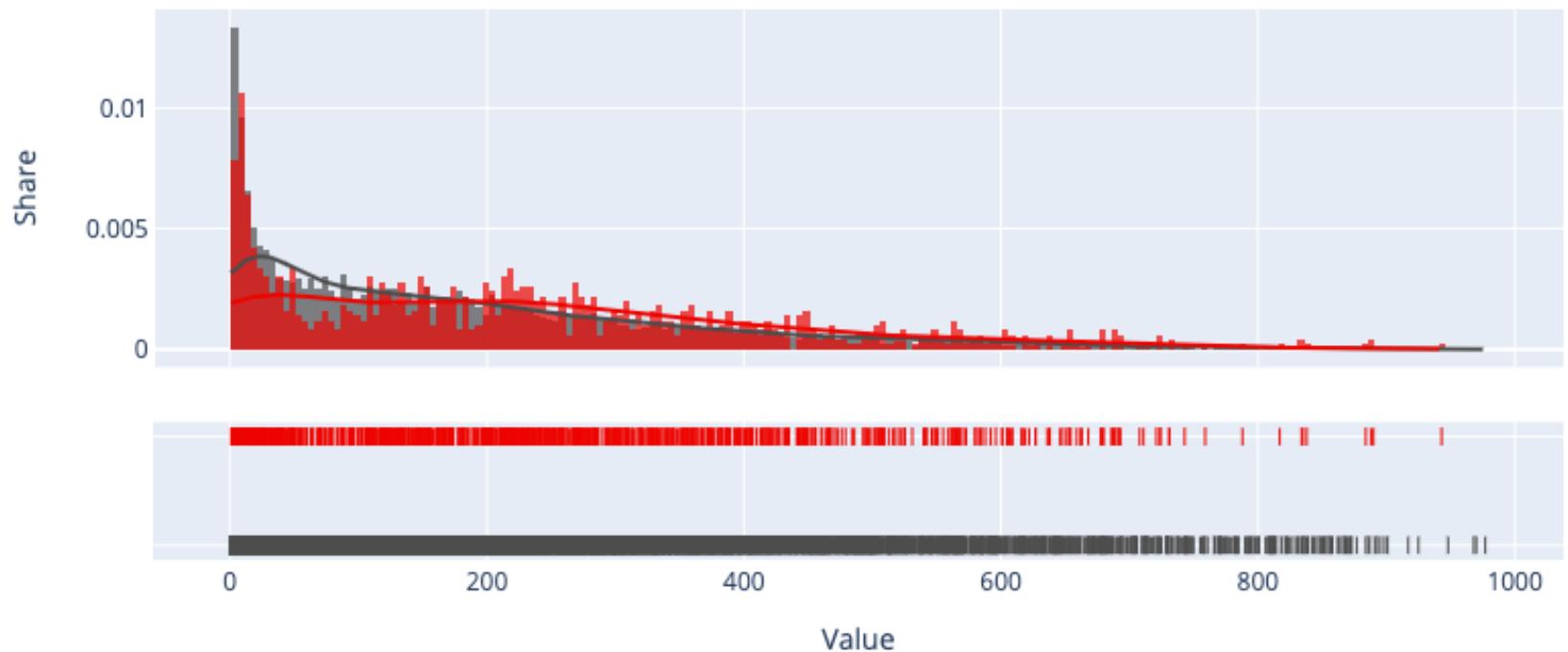
- Можно смотреть на несколько метрик, хотя оптимизируем всегда одну
- Интервальные оценки лучше точечных
- Cross-validation + hold-out test

Также, с помощью cross-validation можно оценить стабильность модели:

- меняется ли качество от фолда к фолду?
- меняется ли feature importance от фолда к фолду?

## Валидация модели

# Качество модели

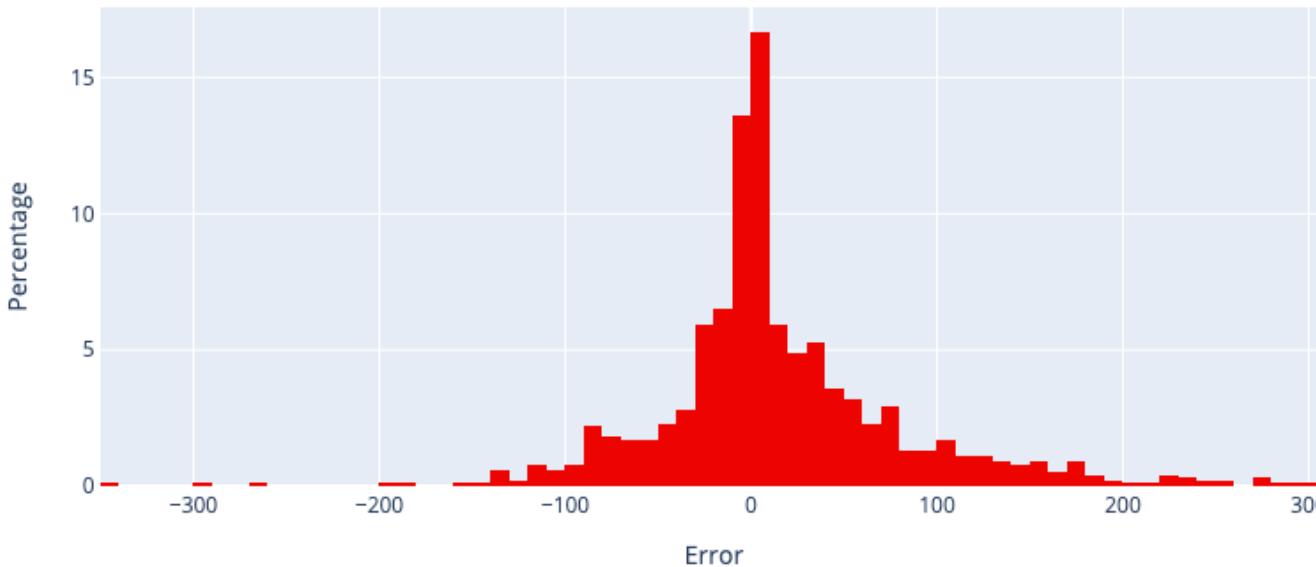


Дополнительно, имеет смысл сравнить:

- распределение target на обучении и отложенной выборке
- распределение model output на обучении и отложенной выборке

# Валидация модели

## Качество модели

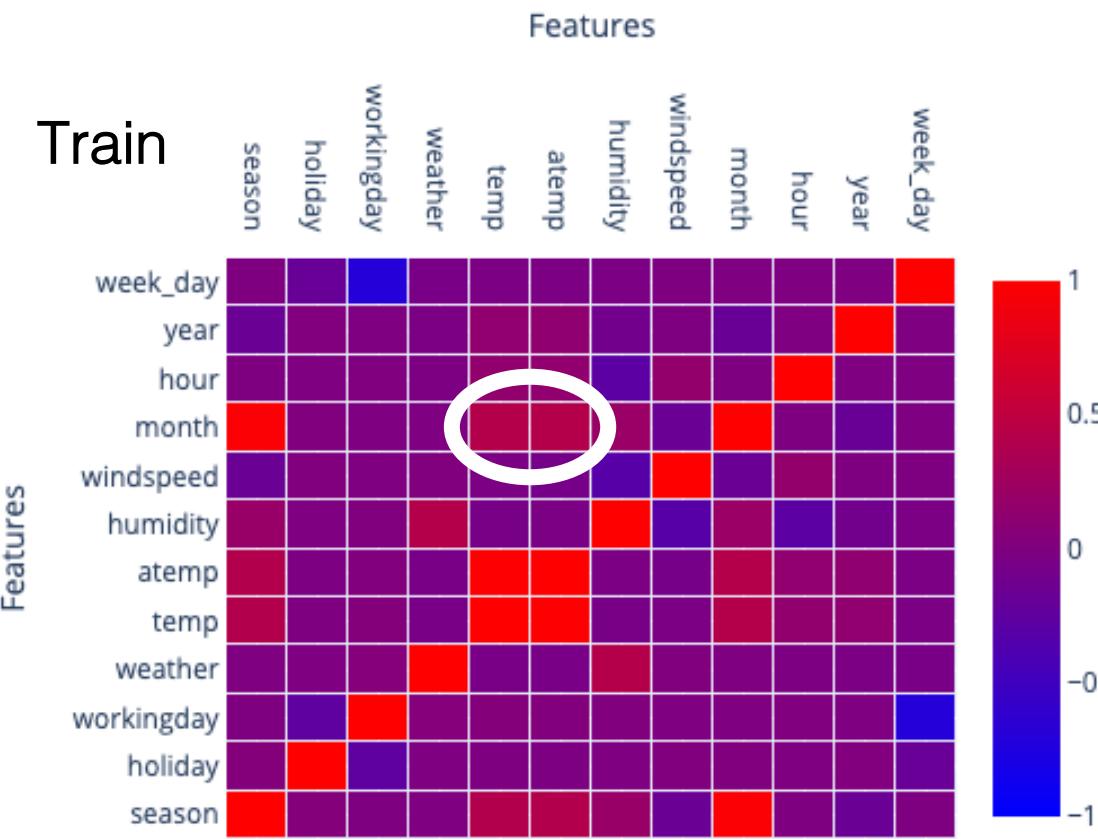


Распределение ошибок поможет понять:

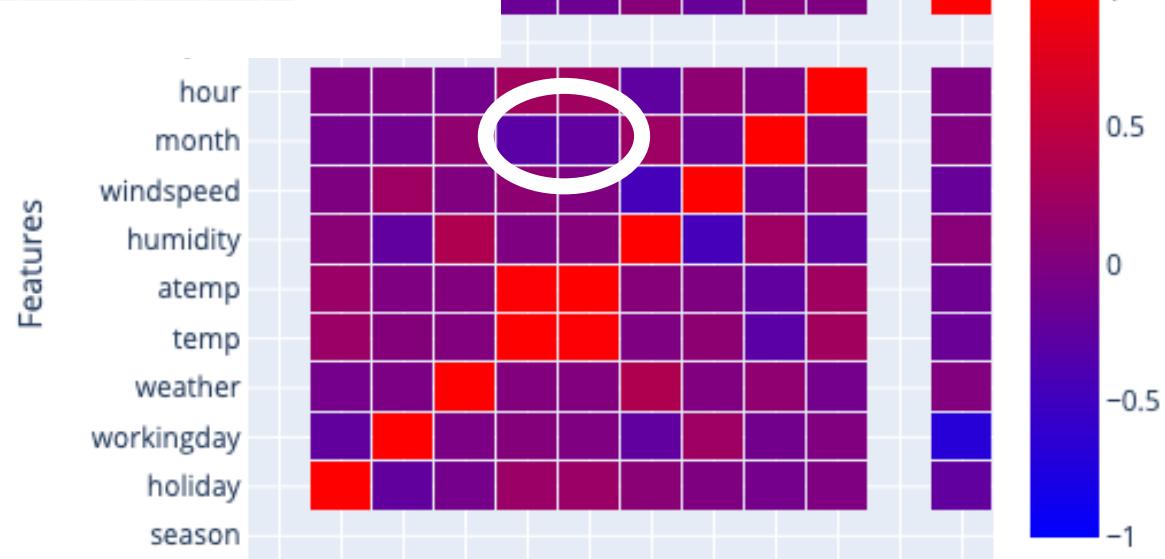
- склонна ли модель к недо/переоценке целевой функции
- остался ли сигнал в данных
- есть ли выбросы или сегменты с большей ошибкой

# Валидация модели

## Train



# Hold-out test



# Экономический эффект

## Валидация модели

Экономический  
эффект

Можете ли вы оценить  
возможный экономический  
эффект за месяц?

Можете ли вы оценить  
возможный эффект,  
например, при улучшении  
прогноза на 1%?

Для выбора оптимальной модели из имеющихся кандидатов полезно понимать связь эффекта и качества.

# Валидация модели

## Дополнительные свойства

Такие характеристики модели, как:

- калибровка
  - качество в топе прогнозов
  - ошибка в разрезе выбранных сегментов
  - линейность по выбранным признакам
- и пр.

# Валидация модели

## Скорость устаревания

Важная характеристика, на основе которой можно сделать вывод о необходимой частоте переобучения модели

Подход к оценке: (**обучение**, ошибка внутри ожидаемого интервала, ошибка за пределами интервала)

Быстрое устаревание:



Модель не устаревает:



Среднее устаревание:



# Валидация модели

# Валидация модели и предотвращение ошибок



# Избежание предвзятости

## Валидация модели



DHH ✅ @dhh · 7 нояб. 2019 г.

The [@AppleCard](#) is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

## Amazon scraps secret AI recruiting tool that showed bias against women

In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

# Валидация модели

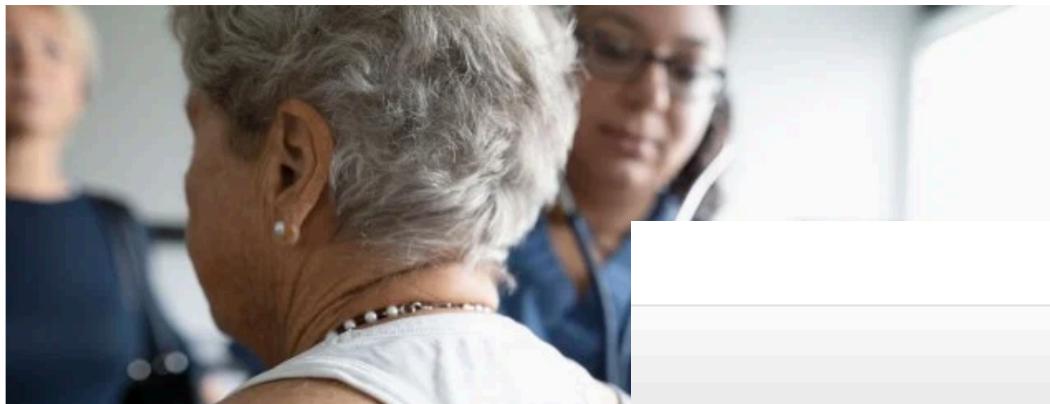
# Избежание предвзятости

MIT  
Technology  
Review

Artificial intelligence Oct 25

...

**A biased medical algorithm favored white people for health-care programs**



The New York Times

*Facial Recognition Is Accurate, if You're a White Guy*

# Доверие

More CEOs (84%) ‘agree’ that AI-based decisions need to be explainable than that AI is good for society (79%).

## Валидация модели

Mark J. Girouard, an employment attorney at Nilan Johnson Lewis, says one of his clients was vetting a company selling a resume screening tool, but didn’t want to make the decision until they knew what the algorithm was prioritizing in a person’s CV.

After an audit of the algorithm, the resume screening company found that the algorithm found two factors to be most indicative of job performance: their name was Jared, and whether they played high school lacrosse. Girouard’s client did not use the tool.

<https://az.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased>

<https://hackernoon.com/dogs-wolves-data-science-and-why-machines-must-learn-like-humans-do-41c43bc7f982>

<https://www.pwc.com/mu/pwc-22nd-annual-global-ceo-survey-mu.pdf>

# Валидация модели

## Регуляторные требования

### Европа – GDPR

The right to access meaningful information about the logic involved, as well as the significance and the envisaged consequences of automated decision-making”

### США - Equal Credit Opportunity Act

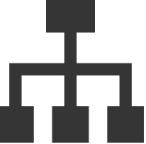
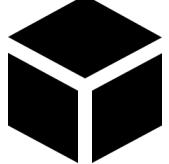
Statement of reasons for adverse action, must be specific and indicate the principal reason(s) for the adverse action

# Интерпретация модели

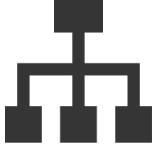
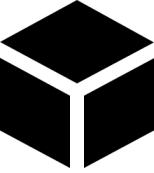
Валидация  
модели



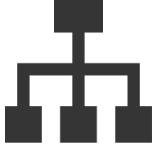
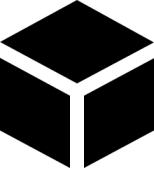
# Подходы к оценке вклада признаков в модель

	 MODEL SPECIFIC	 MODEL AGNOSTIC
GLOBAL	<ul style="list-style-type: none"><li>• Зависят от модели</li><li>• «Объясняют» модель целиком</li><li>• LM, DT, Shap.TreeExplainer, Shap.DeepExplainer...</li></ul>	
LOCAL	<ul style="list-style-type: none"><li>• Зависят от модели</li><li>• «Объясняют» отдельные прогнозы</li><li>• Shap.TreeExplainer, Shap.DeepExplainer...</li></ul>	

# Подходы к оценке вклада признаков в модель

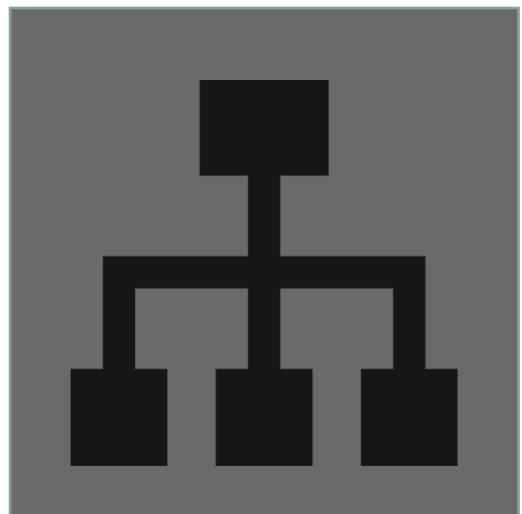
	 MODEL SPECIFIC	 MODEL AGNOSTIC
GLOBAL	<ul style="list-style-type: none"><li>• Зависят от модели</li><li>• «Объясняют» модель целиком</li><li>• LM, DT, Shap.TreeExplainer, Shap.DeepExplainer...</li></ul>	<ul style="list-style-type: none"><li>• <b>НЕ</b> Зависят от модели</li><li>• «Объясняют» модель целиком</li><li>• DPD, ALE, SHAP.KernelExplainer...</li></ul>
LOCAL	<ul style="list-style-type: none"><li>• Зависят от модели</li><li>• «Объясняют» отдельные прогнозы</li><li>• Shap.TreeExplainer, Shap.DeepExplainer...</li></ul>	<ul style="list-style-type: none"><li>• <b>НЕ</b> Зависят от модели</li><li>• «Объясняют» отдельные прогнозы</li><li>• ICE, LIME. SHAP.KernelExplainer</li></ul>

# Подходы к оценке вклада признаков в модель

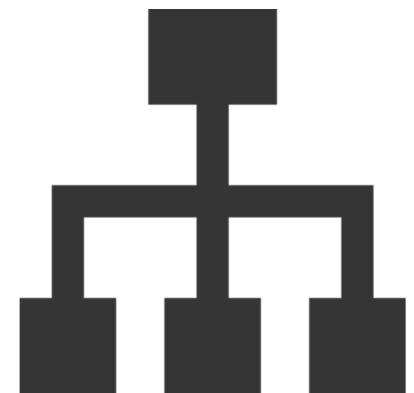
	 MODEL SPECIFIC	 MODEL AGNOSTIC
GLOBAL	<ul style="list-style-type: none"><li>• Зависят от модели</li><li>• «Объясняют» модель целиком</li><li>• LM, DT, Shap.TreeExplainer, Shap.DeepExplainer...</li></ul>	<ul style="list-style-type: none"><li>• <b>НЕ</b> Зависят от модели</li><li>• «Объясняют» модель целиком</li><li>• DPD, ALE, SHAP.KernelExplainer...</li></ul>
LOCAL	<ul style="list-style-type: none"><li>• Зависят от модели</li><li>• «Объясняют» отдельные прогнозы</li><li>• Shap.TreeExplainer, Shap.DeepExplainer...</li></ul>	<ul style="list-style-type: none"><li>• <b>НЕ</b> Зависят от модели</li><li>• «Объясняют» отдельные прогнозы</li><li>• ICE, LIME. SHAP.KernelExplainer</li></ul>

## Валидация модели

Лучшее объяснение модели –  
она сама



=



# Валидация модели

## Интерпретируемые модели



- Объяснение точное
- Для построения объяснения не требуется данные, только модель
- Позволяет в какой-то степени объяснять и индивидуальные прогнозы

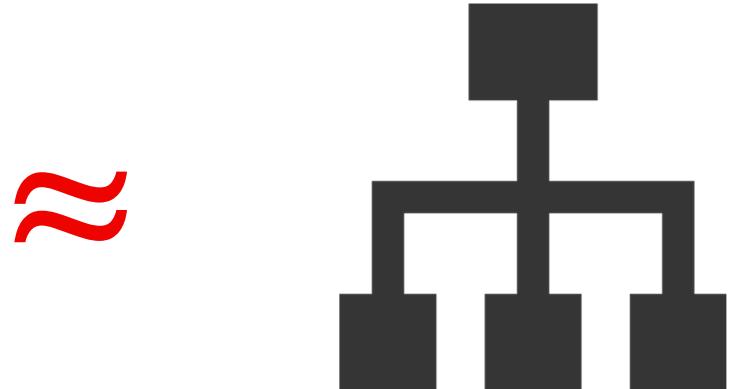
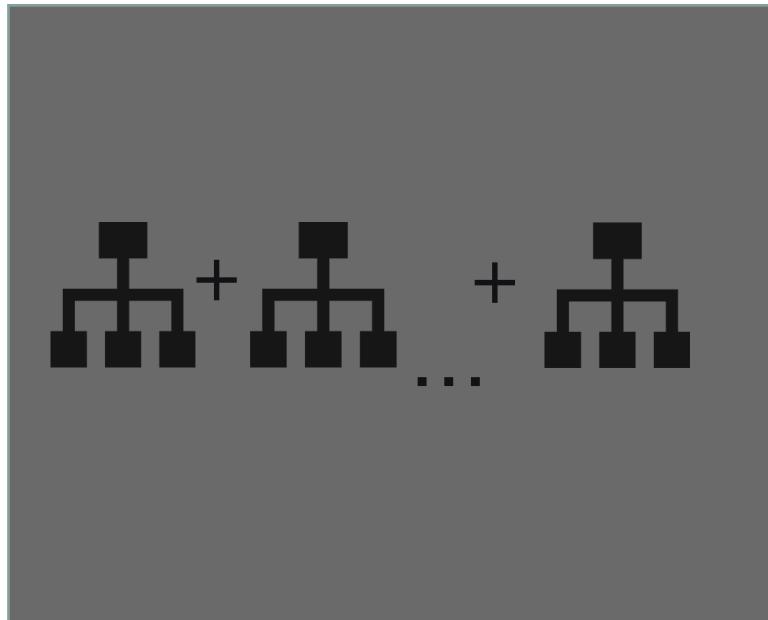


- Подходит только для простых моделей
- Иногда всё равно слишком сложно, требует некоторых представлений о принципах работы модели

# Global Surrogate

Для аппроксимации сложной модели выбирается более простая, но интерпретируемая модель

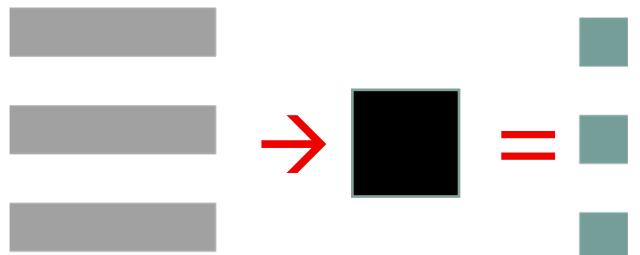
Валидация  
моделей<sup>+</sup>



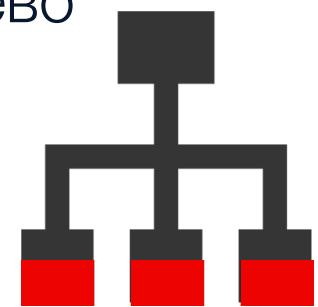
## Валидация модели

# Пример суррогатной модели

Генерируются  
прогнозы бустинга



Интерпретируется  
новое решающее  
дерево



Обучается дерево,  
которое вместо  
таргета приближает  
прогнозы бустинга

Приближение бустинга над решающими деревьями  
одним деревом

## Валидация моделей

# Суррогатные модели



- Можно применить для абсолютно любой модели
- Можно интерпретировать любой из интерпретируемых моделей
- Можно выбрать достаточно простую модель и интерпретация не потребует дополнительных знаний

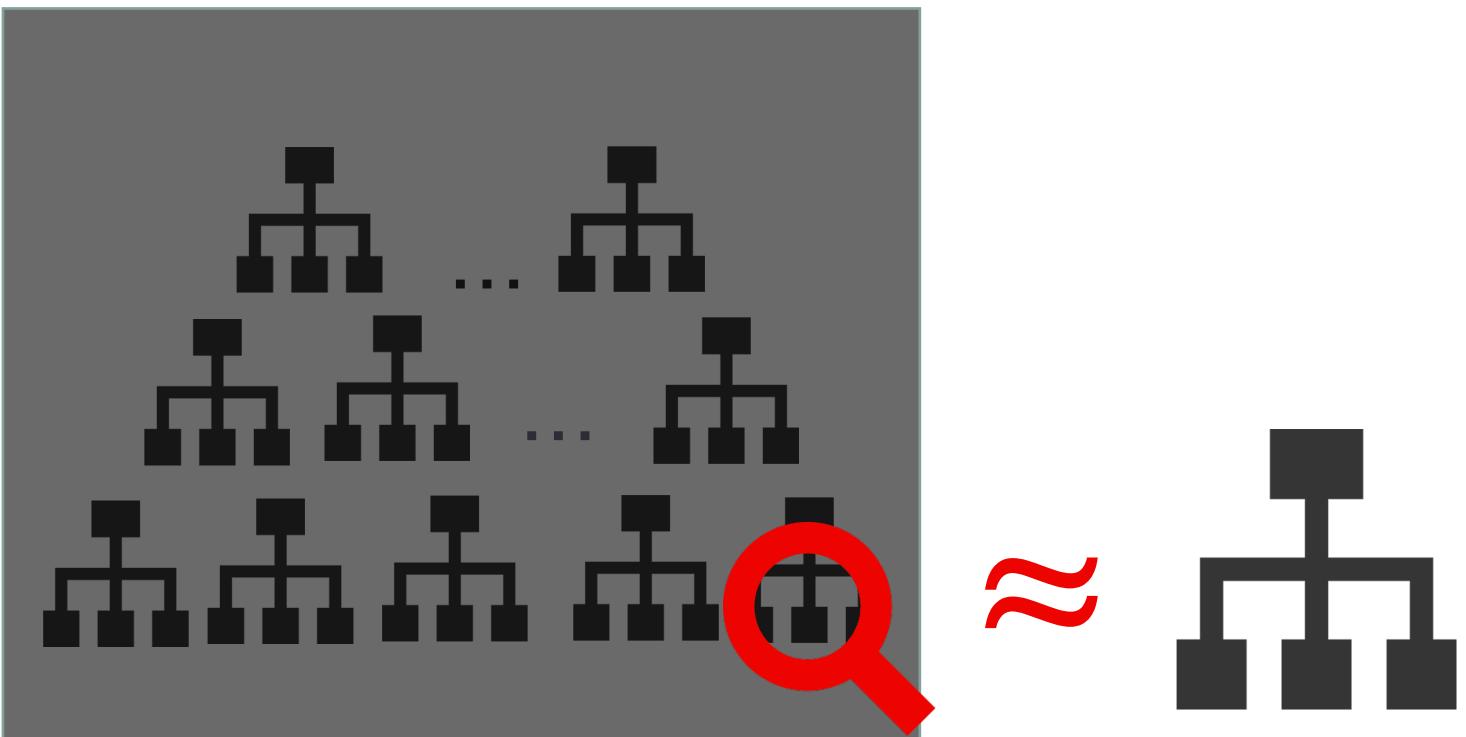


- Результат не консистентный, слишком много степеней свободы
- Интерпретация не использует настоящую модель, поэтому она приблизительная

# Local Surrogate

Строится интерпретируемая модель, которая хорошо аппроксимирует сложную модель в **окрестностях конкретной точки**

Валидация  
моделей



# Local Surrogate: пример

Пример: выбираем объект и приближаем бустинг линейной моделью.

$$\text{explanation}(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Здесь:

- $x$  - выбранный объект
- $f$  – сложная модель, например, градиентный бустинг
- $g$  – интерпретируемая модель, например, линейная ( $G$  – семейство таких моделей)
- $L$  – функция потерь, например, MSE
- $\Omega(g)$  – сложность объясняющей модели
- $\pi_x$  - мера близости между объектом  $x$  и другим объектом

# Валидация модели

## Local Surrogate: пример

Другими словами, делаем следующие шаги:

- Выбираем объект
- Генерируются (иногда выбираются, но редко) новые объекты на основе целевого
- На основе близости к целевому объекту новым объектам назначаются веса
- Считается прогноз сложной модели на новых объектах
- По полученному данных обучается интерпретируемая модель

# Валидация моделей

## Local Surrogates



- Можно применить для абсолютно любой модели
- Можно интерпретировать любой из интерпретируемых моделей
- Один и тот же датасет может быть использован для разных моделей



- Результаты не стабильны, зависят от результатов семплирования
- Веса тоже не стабильны, зависят от того, как взвешивать
- В целом, многовато степеней свободы (также модель), отчего иногда похожие объекты получают очень разные объяснения

# SHapley Additive exPlanations

Общий принцип:

- Обучается набор линейных моделей с конкретным признаком и без
- Оценивается вклад признака в смещение прогноза модели
- Каждый признак оценивается по набору подмножеств, а не одной паре

... поэтому работает долго

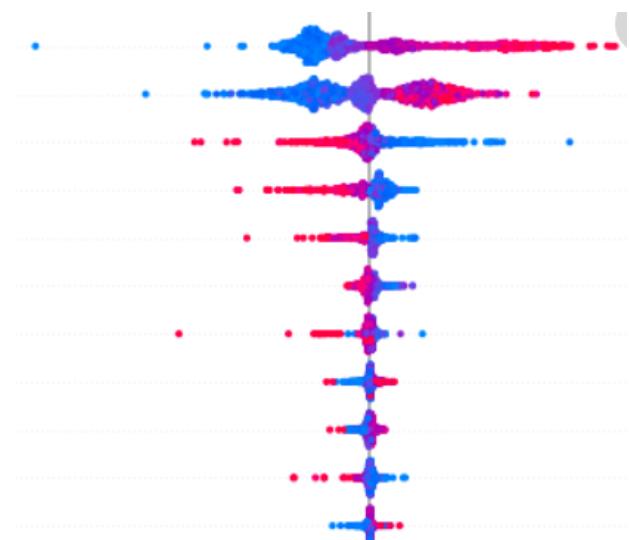
Валидация  
моделей

# Валидация моделей

## SHapley Additive exPlanations

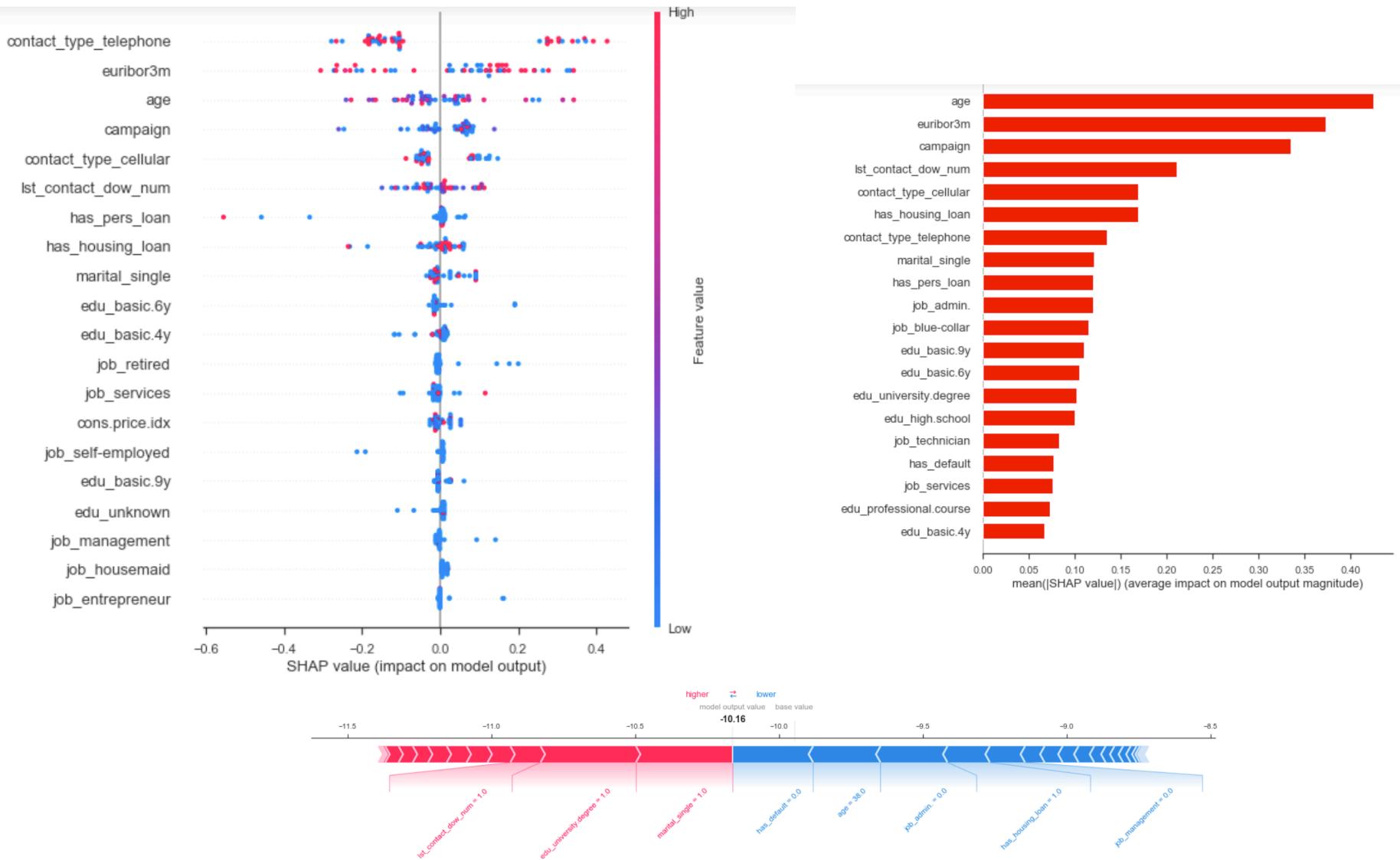
Объяснение бустинга с помощью TreeExplainer или KernelExplainer:

- выбираем часть датасета для оптимизации расчета
- рассчитываем shap values построением моделей по подмножествам признаков
- получаем оценку смещения прогноза от добавления признака



# SHapley Additive exPlanations

## Валидация моделей



# Валидация моделей

## SHapley Additive exPlanations



- Есть универсальная версия (KernelExplainer) и версии, специфичные для алгоритмов (TreeExplainer, DeepExplainer, ...)
- Более стабилен, чем другие методы на основе суррогатных моделей, меньше степеней свободы
- Для похожих объектов выдает похожие объяснения



- Из-за перестановок вычислительно сложный метод, в процессе обучается много моделей
- Из-за выбора набора данных для расчета shap values результаты могут меняться

# Валидация моделей

## SHapley Additive exPlanations

Библиотека SHAP и подходящий Explainer:

- TreeExplainer
- DeepExplainer
- GradientExplainer
- KernelExplainer

Тестирование в production

# Тестирование в production

## Тестирование в production

Варианты тестирования в боевых условиях:

- Пилотный тест
- АБ-тестирование

# Тестирование в production

## Пилотный тест

- Изменение применяется в течение ограниченного промежутка времени
- Изменение может быть применено на небольшой группе пользователей/объектов
- Сравниваются значения метрик до и во время пилота
- Оценка проверяется на практическую и статистическую значимость



# Тестирование в production

## Пилотный тест



### Пилот:

- Сравнивают значения метрик до и во время пилота
- Часто для большей достоверности сравнивают также значения метрик во время и после пилота

Тестирование  
в production

# Пилотный тест



В чем ограничения данного метода тестирования?

# Тестирование в production

## Пилотный тест

В чем ограничения данного метода тестирования?

- Сложно изолировать влияние изменения от влияния внешних факторов
- В частности влияние сезонных факторов, трендов
- Требуется длительное время для проведения серии экспериментов

# АБ-тестирование

Что если усложнить пилотный тест?

Тестирование  
в production

# Тестирование в production

## АБ-тестирование

Что если усложнить пилотный тест?

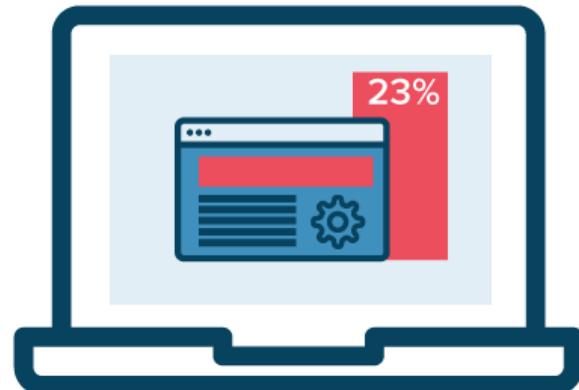
АБ-тестирование:

- Пользователи/объекты делятся на контрольную и тестовую группы (сегменты, “сплиты”) – А и Б
- Изменение производится только в одной из групп (группа Б)
- Едновременно тестируется только одно изменение
- Оцениваются отличия между контрольной и тестовой группой
- Оценка проверяется на практическую и статистическую значимость

Тестирование  
в production

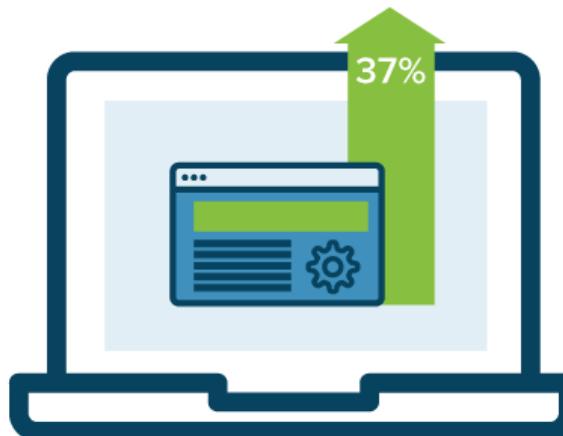
# АБ-тестирование

A



CONTROL

B



VARIATION

В чем преимущества и ограничения АБ-тестирования  
по сравнению с пилотным тестом?

# Тестирование в production

## АБ-тестирование

### Преимущества АБ-тестирования

- Легче изолировать влияние от внешних факторов
- Возможно отделить влияние сезонных факторов, трендов
- Возможно проведение серии экспериментов одновременно (если данных достаточно)

### Ограничения АБ-тестирования

- Требуется разбиение на А и Б группы, пригодное для сравнения
- Может потребоваться длительное время для достижения значимого результата
- Риск совершения ошибок I и II рода: ложная детекция результата или пропуск значимого результата при неудачном дизайне эксперимента

# Машинное обучение: предпроектное исследование

Спасибо!  
Эмели Драль