Лекция 1. Принципы обоснованного Machine Learning. Проверка адекватности моделей

Практикум по программированию, 5 семестр

Иван Евгеньевич Бугаенко,
ассистент каф. ПМиФИ

Цель курса

Основная цель – научиться работать с ML-проектами и строить обоснованные модели на основе аналитики данных

- принятие решений на основе статистического и графического анализа данных и аргументация своих действий
- обоснование выбора методов предобработки данных и моделей для прогнозирования
- интерпретирование работы модели
- применение современных методов анализа данных и лучших практик

Формат работы

Курс состоит из:

- 8 лекционных занятий (16 баллов)
- 7 лабораторных работ (35 баллов + 9 доп. баллов)
- домашнего задания (20 баллов)
- демоэкзамена (20 баллов)

Итого: 100 баллов

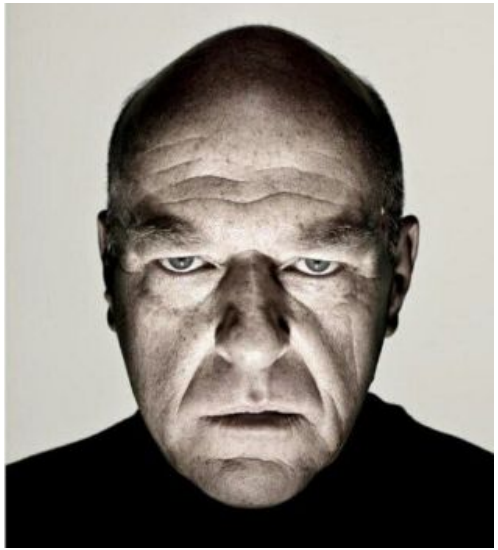
Все материалы выкладываются на wiki.pmifi.ru

Рейтинг ведется в таблице

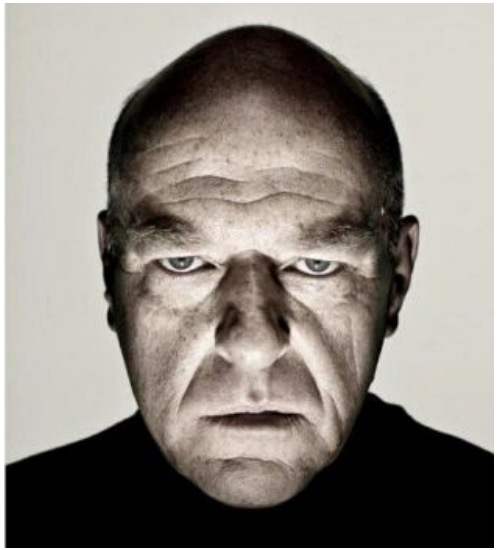
Зачем нужно обосновывать свои решения в ML?

Итоговая цель каждого ML-проекта – разработать жизнеспособную модель

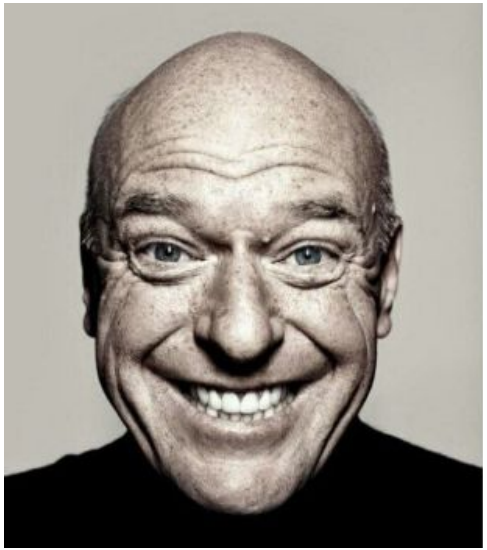
Как вы себе представляете работу над ML-проектом?



натыкать наугад



перебрать все варианты



обоснованный пайплайн работы с данными

Проблемы других подходов

- невозможно перебрать ВСЕ варианты
- шаги по работе с данными могут совершаться интуитивно и наугад, что может испортить качество модели
- нет обоснования, почему были совершены те или иные действия, что затрудняет читаемость работы и ее воспроизводимость

Проблемы других подходов

- невозможно перебрать ВСЕ варианты
- шаги по работе с данными могут совершаться интуитивно и наугад, что может испортить качество модели
- нет обоснования, почему были совершены те или иные действия, что затрудняет читаемость работы и ее воспроизводимость

Работа становится одним большим черным ящиком!
Итоговой модели нельзя доверять

Что такое аналитика данных?

Аналитика данных – это набор подходов и инструментов по извлечению информации из данных, которая используется для обоснования шагов, которые совершаются в ходе решения ML-задачи

Обосновывая свои шаги, исследователь получает жизнеспособную, надежную и понимаемую модель

Как понять, что модель ML корректна?

Что такое корректная ML-модель?

Корректная модель = метрика + адекватные предсказания +
+ грамотная работа с данными

Инструменты анализа корректности модели

1. сложность алгоритма ML
2. подбор гиперпараметров модели (профилактика переобучения + достигаются лучшие метрики)
3. метрики качества (проверка обобщающей способности модели)
4. графики остатков (Predicted vs Actual Plot и Residuals vs Predicted) – для задачи регрессии
5. ROC/AUC кривые и PR-кривые – для задачи классификации

Какие проблемы могут быть с метриками?

Иногда возникает такая ситуация, когда при разделении выборки на обучение и тест данные о тестовой выборке попадают в обучающую. Такая проблема называется **утечкой данных**

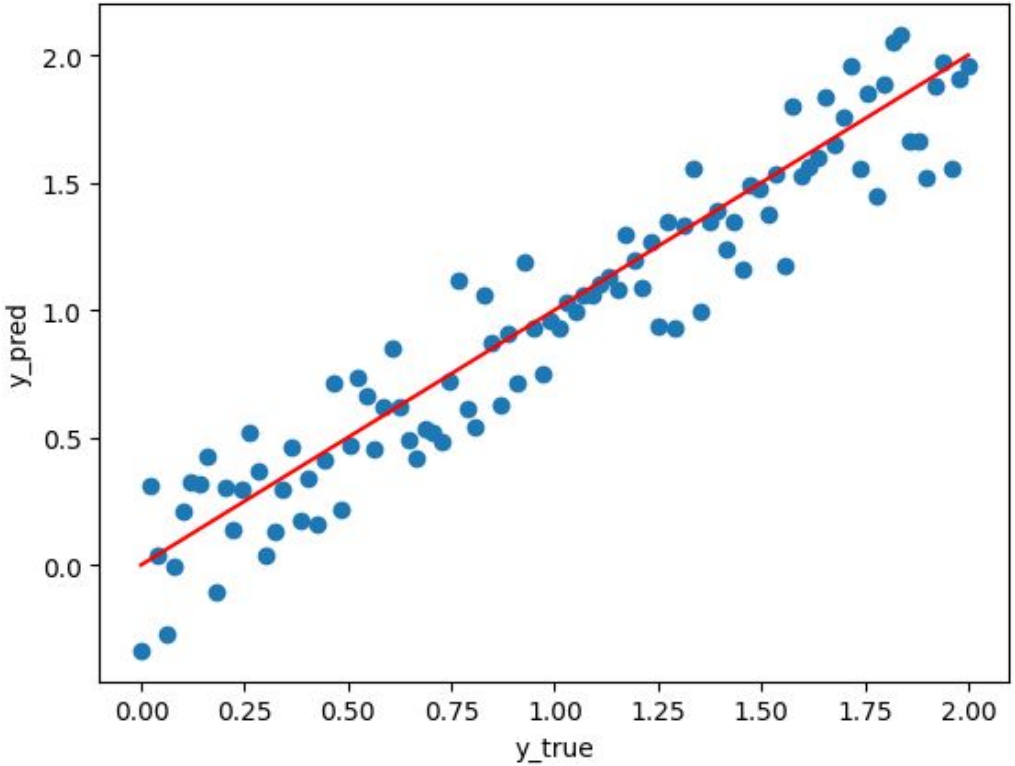
Это очень опасно, поскольку модель знает фрагменты информации о тестовой выборке, следовательно, метрики становятся автоматически выше. И они получаются некорректными

Что такое график остатков?

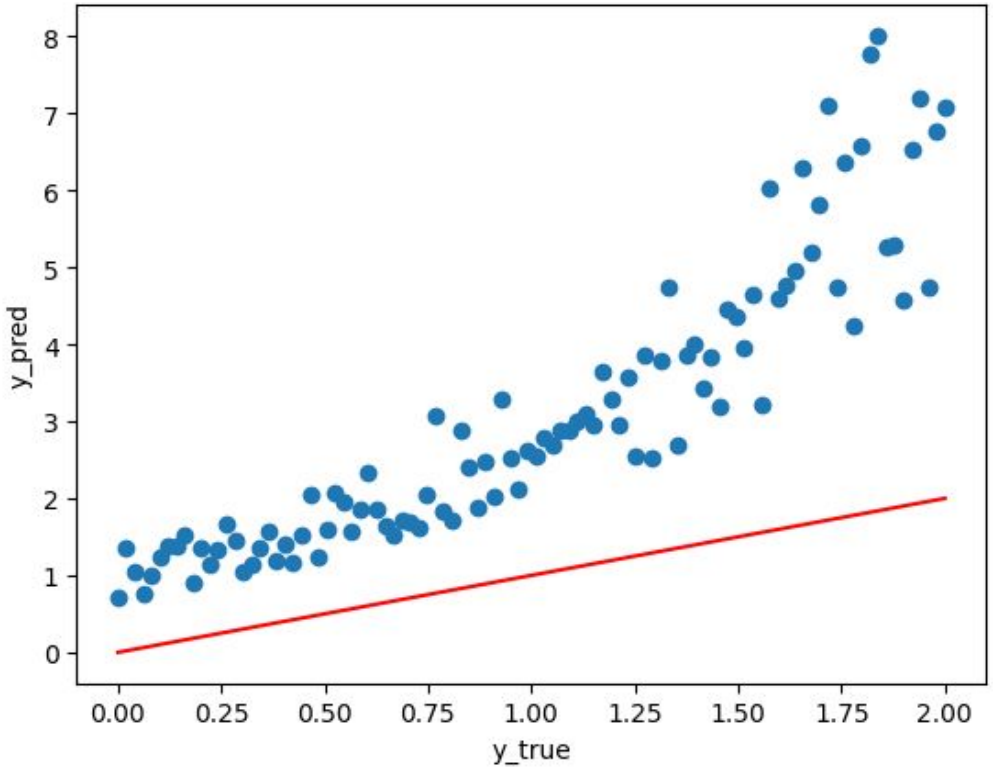
График остатков – это точечный график, предназначенный для оценки предсказанных значений

Для графика **Predicted vs Actual Plot** на оси X откладываются истинные значения, а на оси Y – предсказанные

Для графика **Residuals vs Predicted Plot** на оси X откладываются предсказанные значения, а на оси Y – разница между истинным значением и предсказанным



отличная модель



недообучение/слабая модель