Семинар 1

ИИ и БА, УБ 3 курс

Группа в ТГ

Пару правил:

1. Общаемся культурно
2. Бесконтрольный флуд переводим в ЛС
3. Не спамим рекламой «1ХБет»

Группа для вас, чтобы вы могли быстро получить нужную информацию, спросить, уточнить

О себе

Петяева Елизавета (Лиза)

- ВШЭ БИ (бака)
  Лаборатория алгебры Клиффорда ВШЭ
  Лаборатория сложных систем ВШЭ
- ML Engineer Kuper (Сбермаркет)
  Ex Tripple Whale

Никита Лебедев (Никита)

- ВШЭ Экономика (бака)
  ФКН ФТИАД (мага)
- Руководитель стратегии ИИ в X5
  Ex Ozon Bank, Яндекс Реклама, Сбер, Oliver Wyman Consulting Group

О курсе (1/3)

Инструменты

Python

«У нас в семье нет программистов»

Закодированный дед

Orange

Yandex DataLens

https://yandex.cloud/ru/training/datalens

О курсе (2/3)

Текущий контроль			Экзамен
Активность (А)	**Отчет по проекту** (О)	**Итоговый тест** (Т)	**Защита проекта** (З)
Домашние и др. практические задания *(команда 3-5 чел.)*	*Мини-проект по бизнес-аналитике* *(команда 3-5 чел.)*	*Письменный онлайн-тест по теории* *(индивидуально)*	*Презентация и защита результатов мини-проекта* *(команда 3-5 чел.)*
0-10 баллов	*0-10 баллов*	*0-10 баллов*	*0-10 баллов*
Результирующая оценка (Р)			
Р= 0,2*А + 0,2*О + 0,3*Т +0,3*З			

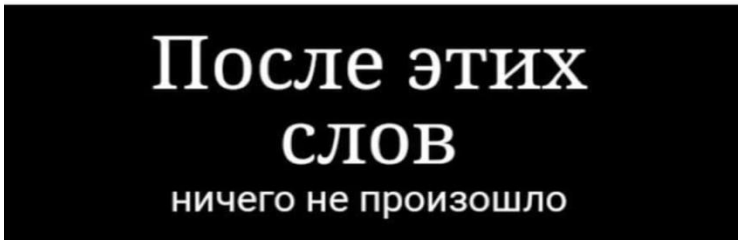
О курсе (3/3)

Регистрация проектных команд:

- **Количество:** 3-5 человек
- **Участники**: только из 22ой группы
- **Дедлайн регистрации**: 20.01
- **Способ регистрации**: Отправим гугл таблицу для распределения
- **За пропуск дедлайна**: -1 балл за каждый день просрочки

Введение

А теперь пора позаниматься



Зачем вам знать про DS?

- Понять, как варится аналитический борщ
- Поможет отличить инсайты от "я тут покликал в отчете, вроде норм"
- Ты заказал торт, а привезли пиццу
- "Дашборд красивый, а толку?" – чтобы оценивать, полезна ли аналитика для принятия решений, а не просто для красоты
- Поднять свою стоимость на рынке труда

Цели

- **Разобраться в основах аналитики и ML**
  что это такое и как работает.
- **Понять бизнес-ценность данных**
  как аналитика помогает принимать решения и повышать прибыль.
- **Изучить ключевые инструменты и подходы**
  от отчетов до прогнозов и ML в реальной работе
- **Научиться задавать правильные вопросы аналитикам**
  чтобы получать нужные и применимые ответы.

Примеры ML задач в бизнесе

Рекомендательные системы

Поисковые системы

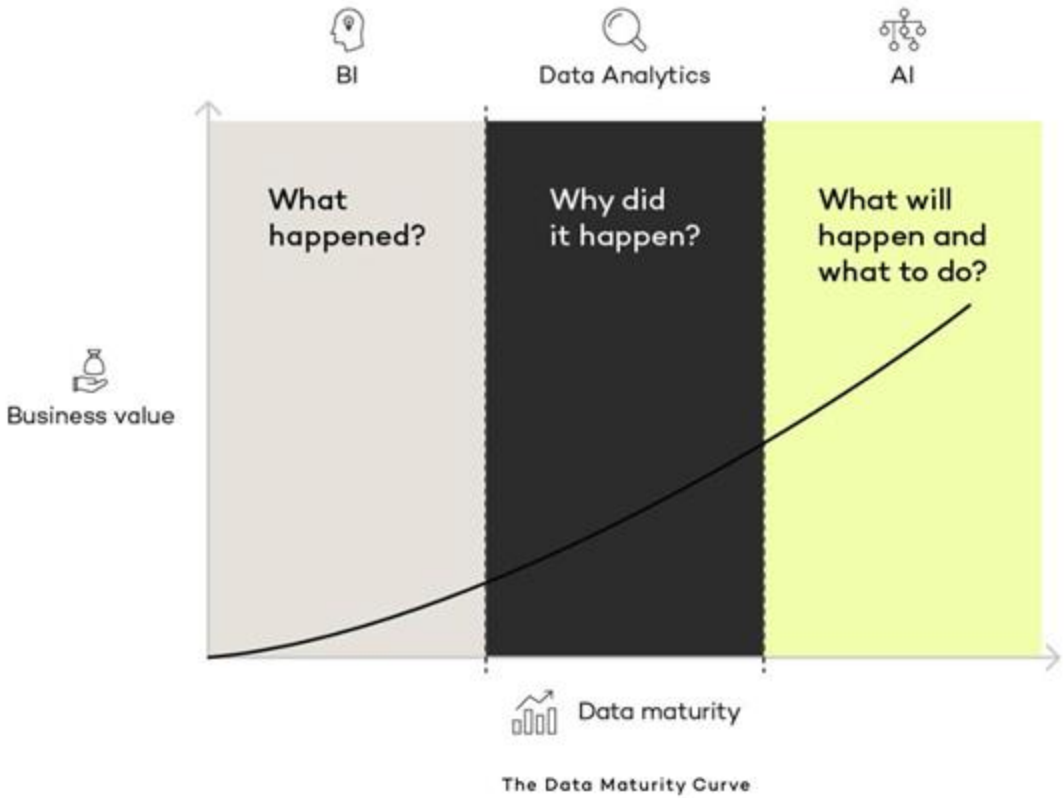
Предсказать с кем вы будете спать

Антифрод
Блокировки по 115 ФЗ

Предсказание модных трендов

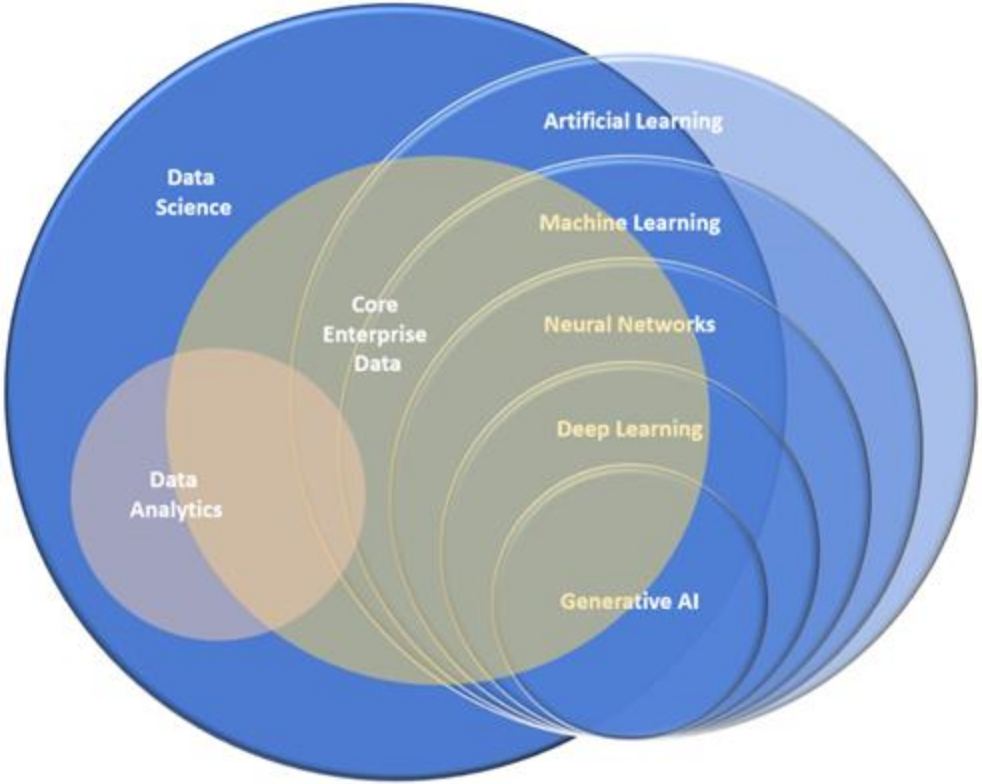
Автоматизация первых этапов отбора кандидатов

Что такое AI и BI?

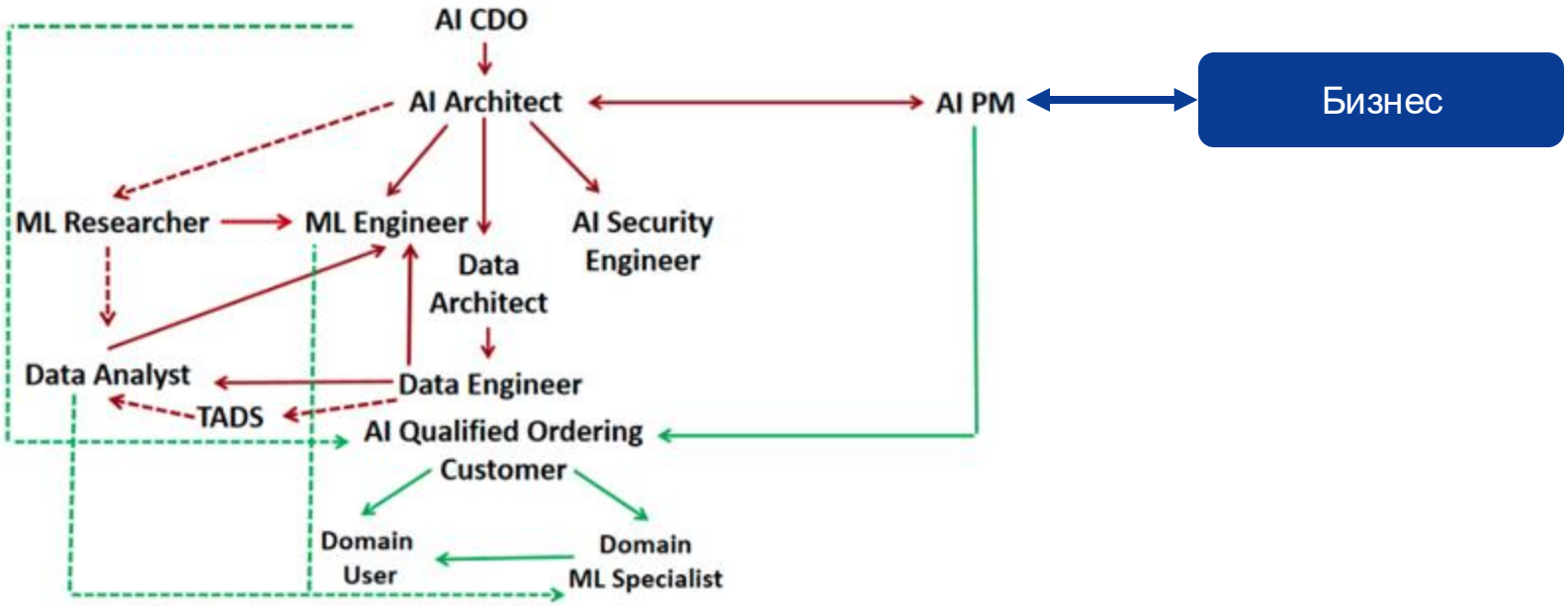


https://afry.com/en/competence/data-analytics-bi-ai

Что такое AI и BI?



Кто работает в этой сфере



https://github.com/aimclub/ai-competency-model
Подробнее о навыках спецов можно посмотреть туть: https://github.com/avito-tech/playbook

Цикл проекта в сфере Data Science

1. **Manager**: Постановка бизнес-задачи
2. **Data Analytic**: Сбор и подготовка данных
3. **Data Analytic:** Анализ данных (EDA) – разведывательный анализ
4. **ML Engineer:** Построение/обучение модели
5. **Data Analytic:** Оценка качества и интерпретация результатов
6. **Manager**: Принятие решения о проведении теста
7. **Data Analytic + Manager:** А/Б тестирование на куске бизнеса
8. **Manager + Бизнес:** Внедрение в бизнес-процесс

https://madewithml.com/courses/mlops/product-design/#timeline
https://github.com/IrinaGoloshchapova/ml_system_design_doc_ru/tree/main

Ключевые задачи и типы моделей в ML

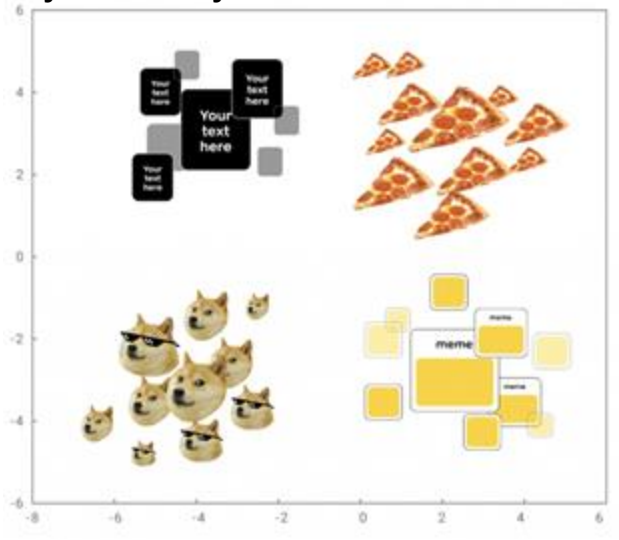
Классификация основных задач ML

1. **Обучение с учителем:** классификация, регрессия
2. **Обучение без учителя:** кластеризация, понижение размерности
3. **Обучение с подкреплением:** более сложные кейсы (роботы, игры и т.п.)

Обучение с учителем

	Regression	Classification
Outcome	Continuous	Class
Examples	Linear regression	Logistic regression, SVM, Naive Bayes

Обучение без учителя



https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-supervised-learning

Классификация

Определение:
прогнозирование категории (пример: понравится фильм или нет)

Примеры в бизнесе:
1. Фрод-мониторинг (мошенничество)
2. Отток клиентов (churn)
3. Сегментация по отклику на рекламную кампанию

Основные метрики
Acuracy, Precision, Recall, F1, ROC-AUC

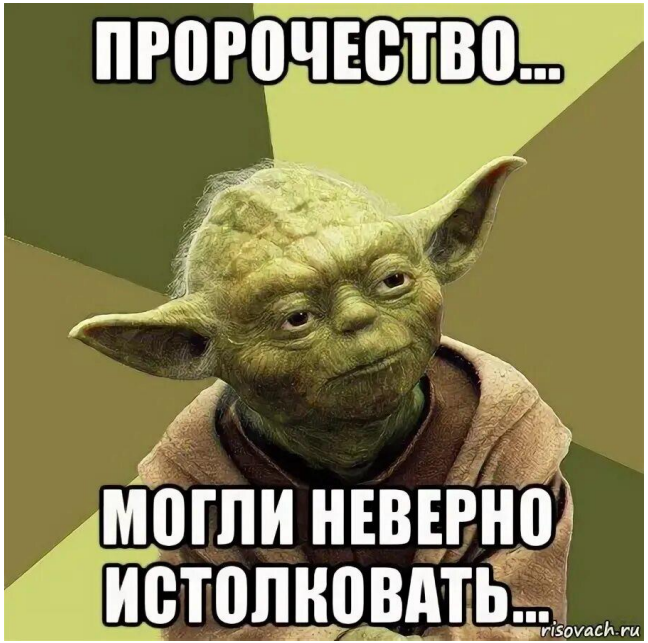
https://education.yandex.ru/handbook/ml

Регрессия

Определение:
прогнозирование числового значения (пример: будущие продажи)

Примеры в бизнесе:
1. Прогнозирование спроса, выручки, доходности
2. Оценка стоимости недвижимости

Основные метрики
MAE (Mean Absolute Error),
MSE (Mean Squared Error), RMSE и т.д.



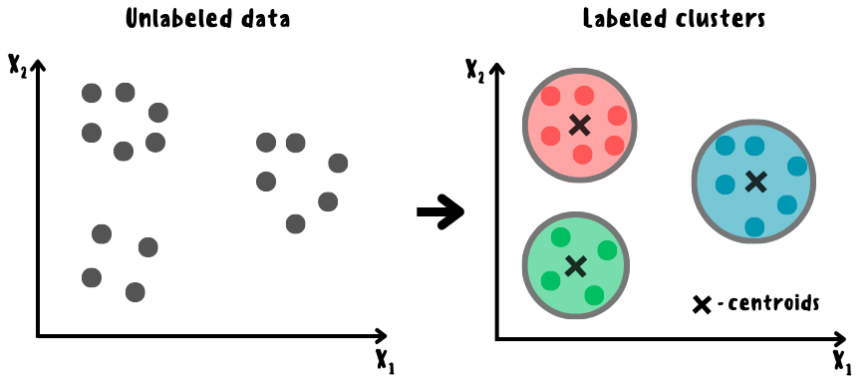
https://education.yandex.ru/handbook/ml

Кластеризация

Определение:
группировка объектов по схожести признаков (пример: сегментация клиентов)

Примеры в бизнесе:

1. Сегментация клиентов для персонализированных предложений каждой группе
2. Группировка товаров по схожести спроса
3. Анализ аномалий в транзакциях



Основные метрики
Внутрикластерное расстояние (Inertia)
Коэффициент силуэта (Silhouette Score)
и т.д.

https://education.yandex.ru/handbook/ml

Выбор подхода под бизнес-задачу

Связь с бизнес-целями:
прогнозирование числового значения (пример: будущие продажи)

Примеры в бизнесе:

1. Нужно предсказать факт (да/нет)?
   → Классификация
2. Нужно предсказать число (продажи)?
   → Регрессия
3. Нужно сгруппировать клиентов?
   → Кластеризация

Важно:
всегда начинаем с вопроса “зачем?”,
а не с “какой метод круче”

https://education.yandex.ru/handbook/ml

Основные шаги EDA (Exploratory Data Analysis)

Что такое EDA и зачем он нужен

Определение:
Процесс исследования данных, чтобы понять их структуру, найти закономерности, проверить гипотезы

Задачи EDA

1. Проверка **качества данных** (пропуски, дубликаты, выбросы)
2. **Анализ распределений** (графики, статистики)
3. Выявление **взаимосвязей** (корреляции, группировки)
4. **Формулирование гипотез** для дальнейшего моделирования

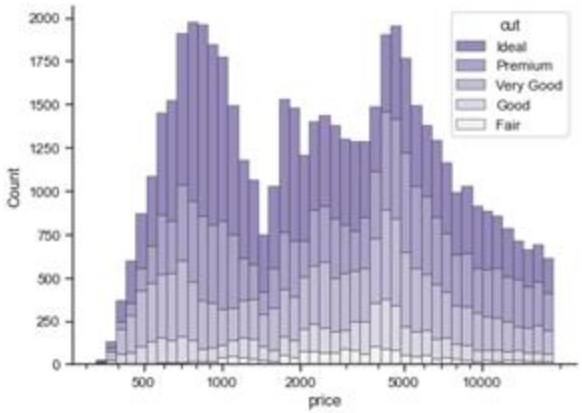
1. Загрузка и первичный осмотр

1. **Импорт данных**
   смотрим на размер (строки, столбцы)
2. **Описание данных**
   df.head(), df.info(), df.describe() (или аналоги)
3. **Пропуски**
   где, сколько, почему
4. **Проверка типов признаков**
   числовые, категориальные, даты, тексты

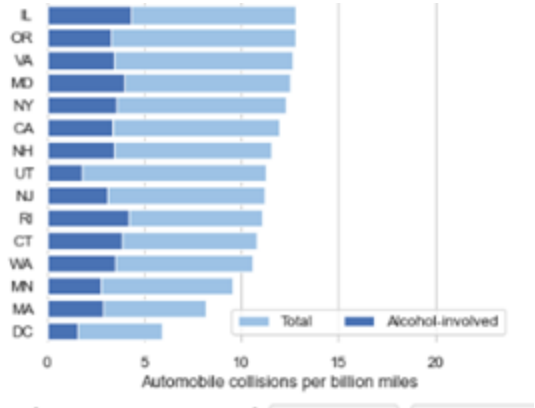
2. Анализ распределений

1. **Гистограммы (Histogram, Distplot):** понять распределение числовых признаков
2. **Barplot / Countplot** для категориальных признаков
3. **Выводы:**
   есть ли сильный перекос (skewness), выбросы, какие категории доминируют

Histogram



Barplot

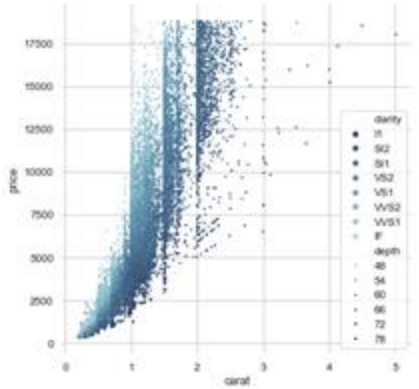


https://seaborn.pydata.org/examples/index.html

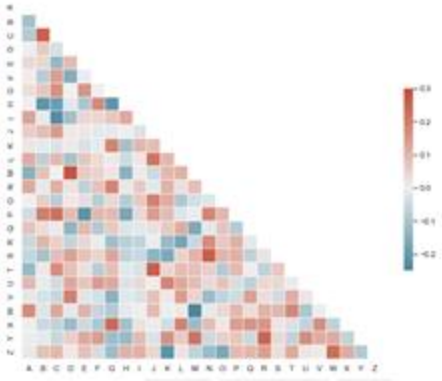
3. Взаимосвязи признаков

1. **Scatterplot (диаграмма рассеяния):** поиск зависимостей между числовыми признаками
2. **Heatmap (корреляционная матрица):** выявить сильные корреляции
3. **Boxplot:** сравнение распределений для разных категорий (например, доход по полу/возрастной группе)

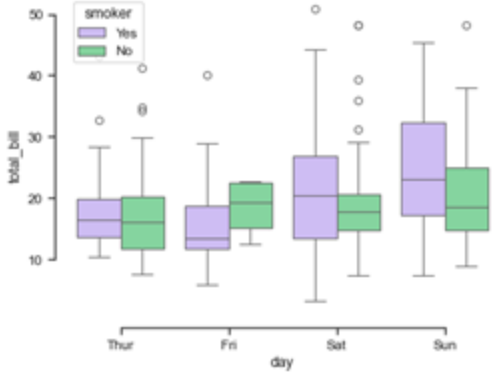
Scatterplot



Heatmap



Boxplot



4. Формулирование гипотез и подготовка к моделированию

1. **Признаки-кандидаты:** какие переменные могут влиять на целевую (из предыдущих анализов)
2. **Проверка мультиколлинеарности:** очень похожих друг на друга признаков и исключение их из модели (не считать 2 раза одно и то же)

Прилетели

Ключевые выводы EDA

1. **Общие закономерности:**
   что мы узнали?
2. **Проблемные места:**
   выбросы, пропуски, дисбаланс классов?
3. **Направления для ML:**
   какие гипотезы проверять, какие модели могут быть релевантны
4. **Связь с бизнесом:**
   какие решения можно принять на основе этих выводов

Заключение и Q&A

1. Важность ML-задачи всегда привязана к **бизнес-результату**
   *возможно задачу можно решить и не через ML
2. **Ключевые типы задач**: классификация, регрессия, кластеризация
3. **EDA** – фундаментальный этап для понимания и подготовки данных

Shit in shit out