АНАЛИТИЧЕСКАЯ КУЛЬТУРА

Как работать с данными правильно

О ЧЕМ СЕГОДНЯ ПОГОВОРИМ?

УПРАВЛЕНИЕ НА ОСНОВЕ ДАННЫХ: Зачем работать с данными и как работать с данными правильно?

ЦИФРОВИЗАЦИЯ ДРП: Как оценить уровень развития аналитики в компании и как управление на основе данных применяется в Дирекции региональных продаж?

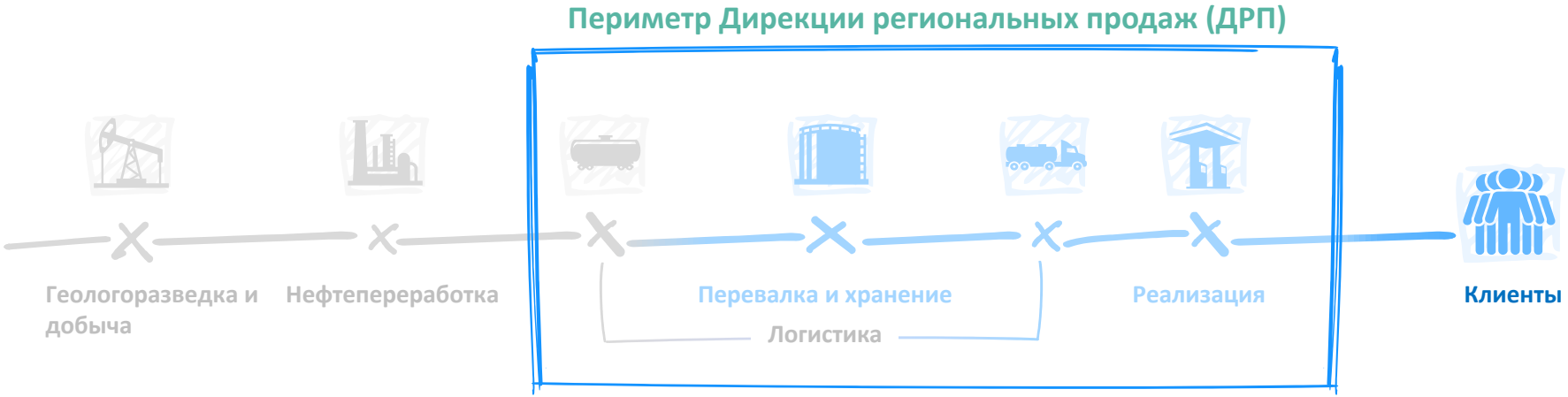
ПРОЕКТ ПО АНАЛИЗУ ДАННЫХ: Что нужно делать с данными и как? Какой жизненный цикл у аналитического проекта? Как собрать аналитическую команду?

ИНТЕГРАЦИЯ АНАЛИТИКИ В БИЗНЕС-ПРОЦЕСС: Какие Data Science проекты есть в периметре Дирекции региональных продаж?

ЗОНА ИССЛЕДОВАНИЯ ДАННЫХ: Что есть в Дирекции региональных продаж для анализа данных?

СТАЖИРОВКИ: Как попасть на стажировку и какие направления есть?

СТРУКТУРА БИЗНЕСА КОМПАНИИ «ГАЗПРОМ НЕФТЬ»



~17 млн т в год - объём реализации нефтепродуктов

ПАРАМЕТРЫ БИЗНЕСА ДИРЕКЦИИ РЕГИОНАЛЬНЫХ ПРОДАЖ

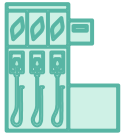
29 регионов присутствия в России + 4 страны СНГ

> 20 тысяч сотрудников

Клиентов-участников бонусной программы > 11,4 млн. человек

- Биржевые продажи нефтепродуктов
- Реализация нефтепродуктов крупным и мелким оптом
- Хранение нефтепродуктов
- Управление > 50 собств. нефтебазами

- Розничная реализация нефтепродуктов
- Продажи корпоративным клиентам
- Управление АЗС \ АСК > 1800 объектов



- > 800 розничных магазинов и кафе при АЗС
- Оказание услуг моек, СТО

- Управление автотранспортом: бензовозы/газовозы
- Услуги метрологии

УПРАВЛЕНИЕ НА ОСНОВЕ ДАННЫХ

УПРАВЛЕНИЕ НА ОСНОВЕ ДАННЫХ

1

Данные

Данные - основной фактор, обусловлива-ющий стратегию и влияющий на нее

2

Отчеты

Данные ложатся в основу отчетов, отражающих текущее положение бизнеса.

3

Анализ

Мало создавать отчеты, нужно проводить анализ, чтобы выявить основные проблемы и закономерности бизнеса

4

Действие

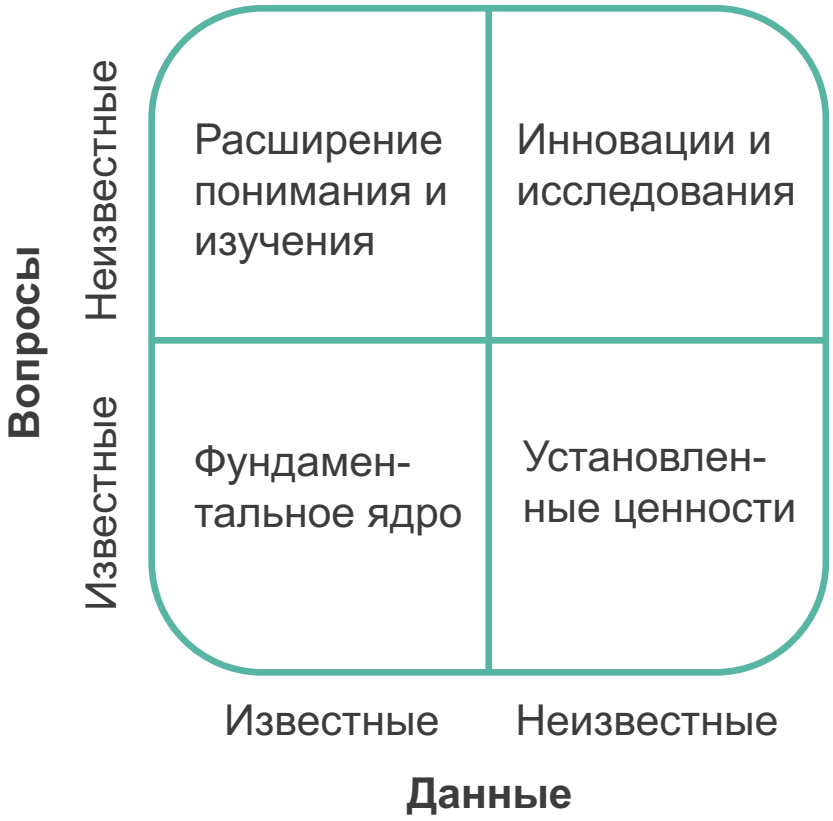
Результаты анализа ложатся в основу процесса принятия решений

5

Ценность

Данные и результаты анализа способны повлиять на стратегию компании или ее развитие

ВИДЫ ЗАДАЧ АНАЛИТИКИ ДАННЫХ



Вызовы и новые виды задач для аналитиков данных требуют выхода за пределы используемых инструментов и аналитических пакетов

- Все более сложные данные – много источников, слабая структурированность
- Кризис достоверности
- Требуются все более стратегические выводы на основе данных
- Все меньше времени на поиск решений и новые идеи

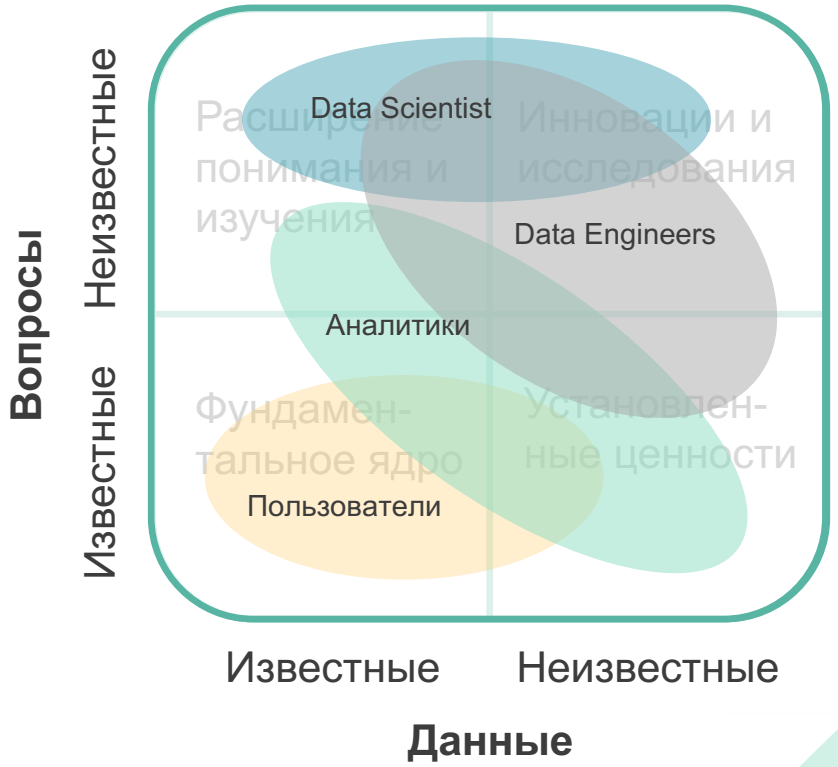
Новые задачи требуют новых решений!

ПРОДВИНУТАЯ АНАЛИТИКА ДАННЫХ

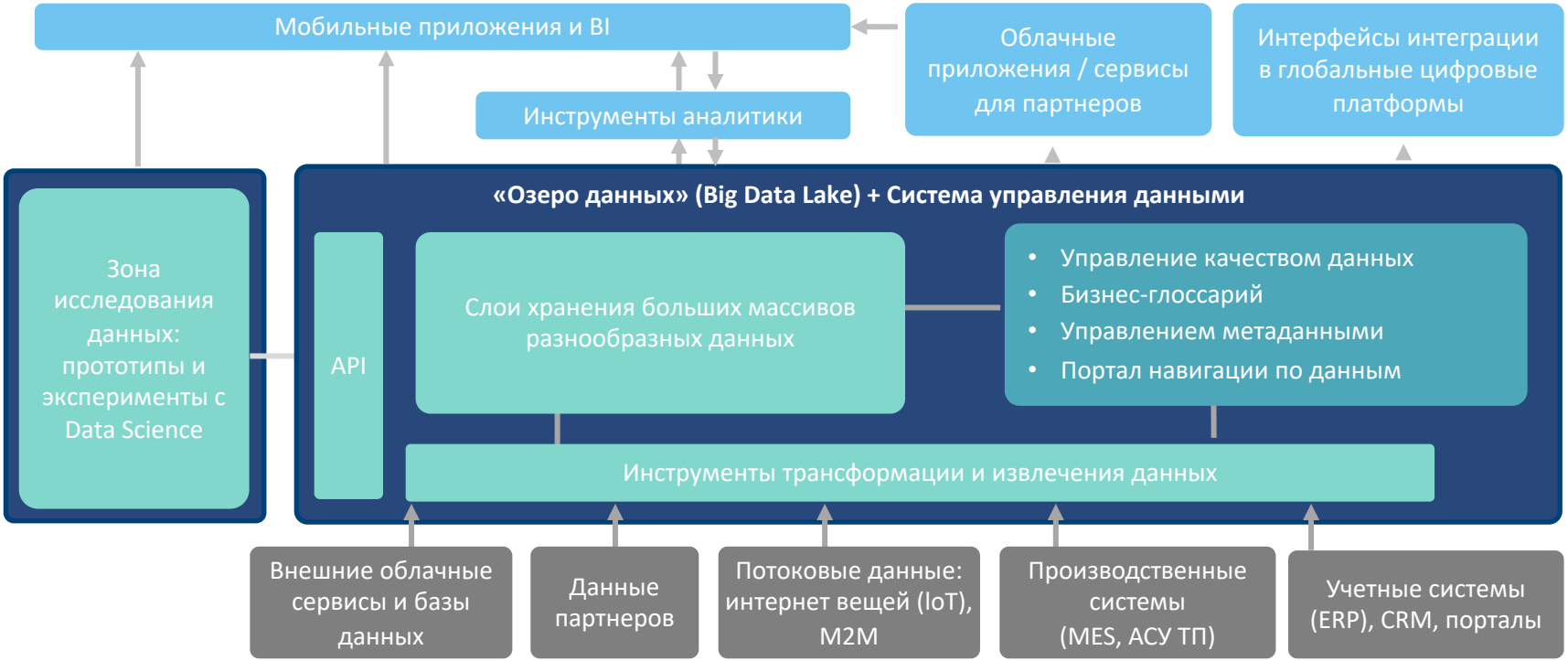
Продвинутая аналитика - технология автоматического, либо полуавтоматического изучения данных или контента с помощью техник, обычно не реализуемых стандартной бизнес-аналитикой, с целью получения глубинных знаний, подготовки прогнозов, выработки рекомендаций.

Техники продвинутой аналитики включают в себя разработку данных, машинное обучение, сопоставление образцов, предсказание, визуализацию, семантический анализ, анализ тональности высказываний, сетевой и кластерный анализ, многомерную статистику, анализ графов, имитационное моделирование, обработку событий, нейронные сети и пр.

Дифференциация ролей аналитика данных



ПЛАТФОРМА УПРАВЛЕНИЯ ДАННЫМИ ДРП



ЦИФРОВИЗАЦИЯ ДРП

АНАЛИТИЧЕСКИЕ ДОМЕНЫ

Аналитика от BI до AI

Информационный портал

Отчеты

Дашборды

- Достоверная
- Содержательная

Наблюдение

Аналитическая мастерская

Первичная аналитика и подготовка данных

Гражданская аналитика

- Гибкая
- Глубокая

Исследование

Лаборатория Data Science

Машинное обучение

Глубокое обучение

- Продвинутая
- Комплексная

Расследование

Центр искусственного интеллекта

Персональный цифровой помощник

Аналитика видео и изображений

- Самообучаемая
- Автономная

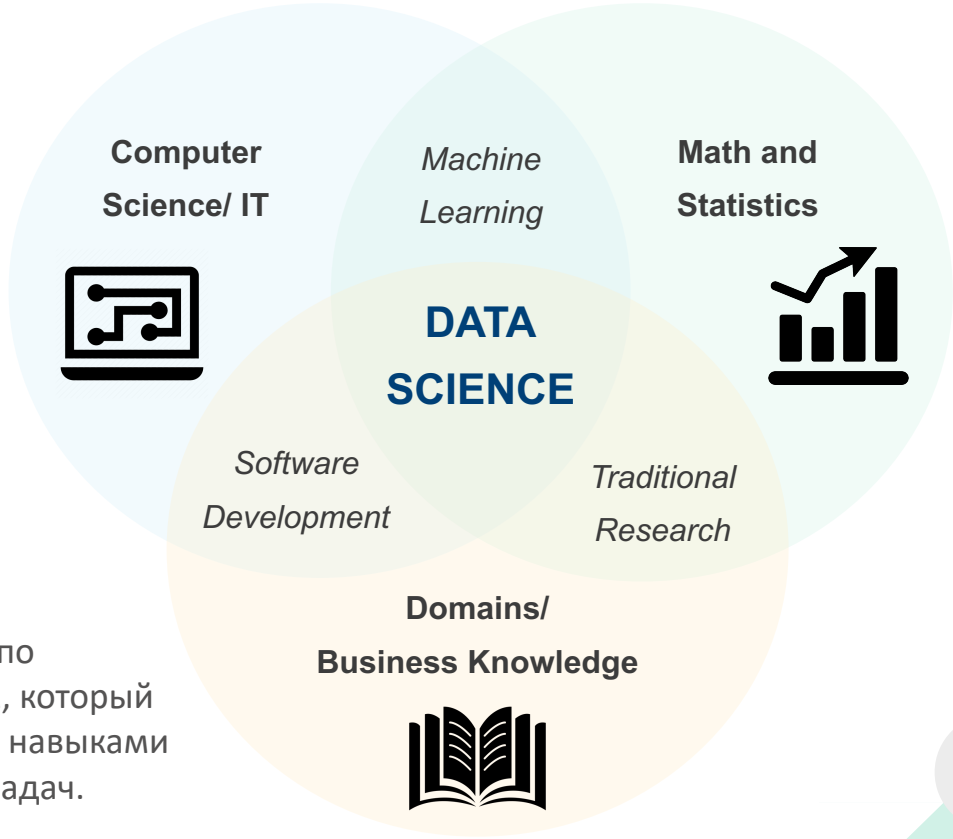
Исполнение

DATA SCIENCE ЛАБОРАТОРИЯ

Data Science Лаборатория – третий уровень зрелости компании.

Data Science (наука о данных) – наука, изучающая проблемы анализа, обработки и представления данных в цифровой форме.

Data Scientist – эксперт по аналитическим данным, который обладает техническими навыками для решения сложных задач.



ЦИФРОВИЗАЦИЯ ОУД ДРП

Созданы цифровые профили клиентов (сегмент, характер покупок, вовлеченность в цифровые продукты). Начато обогащение профиля клиента данными по марке, классу автомобиля, полу, возрасту клиента.

Проведена сегментация клиентов на основе методов машинного обучения и продвинутой аналитики.

Успешно завершен и введен в эксплуатацию пилотный проект для клиентской аналитики на компонентах «озера данных».

ЦИФРОВИЗАЦИЯ ОУД ДРП

В конце 2018 года офис управления данными запустил первые очные вечерние программы обучения нескольких уровней подготовки по языкам

аналитической обработки данных python , SQL

методам прикладной статистики и машинного обучения. Программы обучения проводят сотрудники офиса, и для участников они бесплатны.

В ходе серий занятий коллеги своими руками создают с нуля продвинутые аналитические модели, знакомятся со спектром методов обработки данных, инструментами, которые будут доступны в контуре «озера данных». Объем программ обучения постепенно расширяется.

ПРОЕКТ ПО АНАЛИЗУ ДАННЫХ

ЭТАПЫ ПРОЕКТА ПО АНАЛИЗУ ДАННЫХ

Business Understanding/ Бизнес-анализ	Data Understanding/ Понимание данных	Data Preparation/ Подготовка данных	Modeling/ Моделиро-вание	Evaluation/ Оценка решения	Deployment/ Внедрение
Определение бизнес-целей	Сбор данных	Выборка данных	Выбор алгоритмов	Оценка результатов	Внедрение
Оценка текущей ситуации	Описание данных	Очистка данных	Подготовка плана тестирования	Оценка процесса	Планирование мониторинга и поддержки
Определение целей аналитики	Изучение данных	Генерация данных	Обучение моделей	Определение следующих шагов	Подготовка отчета
Подготовка плана проекта	Проверка качества данных	Интеграция данных	Оценка качества моделей		Ревью проекта
		Форматирова-ние данных			

ПОСТАНОВКА ЗАДАЧИ

Бизнес-анализ и постановка задачи

Определение аспектов, которые необходимо проанализировать, выдвижение гипотез для проверки.

- Оценка возможного экономического эффекта
- Оценка реализуемости проекта

РЕГРЕССИЯ

Чему будет равно значение показателя?

- Какая выручка будет на следующей неделе?
- Сколько человек придет в магазин завтра?
- Сколько лет прослужит оборудование?

КЛАССИФИКАЦИЯ

К какому классу относится объект?

- Выдавать ли клиенту кредит (да/нет)?
- Болен ли человек (да/нет)?
- Тип сообщения (спам/не спам)?

КЛАСТЕРИЗАЦИЯ

На сколько и на какие группы делятся данные?

- Выделение сегментов потребителей
- Группировка стран по экономическому развитию
- Группировка АЗС по схожим показателям

ДАННЫЕ: ИЗМЕРЕНИЯ

Под **измерением** понимается процесс приписывания явлениям чисел так, чтобы в отношениях (в широком смысле) чисел отображались отношения между измеряемыми явлениями.

Неметрические (качественные):

- Номинальная (или номинативная, именная)
- Порядковая (или ординальная, шкала рангов)

Метрические (количественные):

- Интервальная (или шкала равных интервалов)
- Отношений (или пропорциональности)

При описании любых явлений необходимо всегда отдавать себе отчет в том, какая именно шкала используется, поскольку каждый способ обработки экспериментальных данных рассчитан на определенный тип шкал.

ДАННЫЕ: ИЗМЕРЕНИЯ

Шкала	Определение	Информация, допустимые операции	Примеры	Основные числовые характеристики
Номинальная	Классификация по наименованию	1) Отношение эквивалентности: сравнение на сходство – различие (= ; ≠)	• Пол • Профессия • Цвет глаз • Автомобильные номера	P – доли (относительные частоты) Mo – мода
Порядковая (ранговая)	Порядок в степени выраженности измеряемого свойства	1) … 2) Отношение порядка (предпочтения): сравнение больше – меньше, лучше – хуже (< ; >)	• Школьные оценки • Научные звания • Места в спортивных соревнованиях	P – доли Mo – мода Md – медиана
Интервальная	Отражение равенства разности степени выраженности свойства у двух объектов разности двух чисел, приписанных этим объектам по данному свойству	1) … 2) … 3) Знание расстояния между интервалами: больше или меньше на … (+, -)	• Баллы • Календарное время • Температура по Фаренгейту и Цельсию	P – доли Mo – мода Md – медиана Xcp – среднее арифметическое d – размах D – дисперсия σ– СКО V – коэффициент вариации
Шкала отношений	Отражение равенства отношения степени выраженности свойства у двух объектов отношению двух чисел, приписанных этим объектам по данному свойству	1) … 2) … 3) … 4) Знание отношения между любыми двумя значениями: больше или меньше в … раз (*, /)	• Рост • Вес • Время • Расстояние • Сопротивление	

ДАННЫЕ: ИЗМЕРЕНИЯ

Упражнение «Оценка авиадиспетчеров»

В исследовании, моделирующем деятельность авиадиспетчера, группа испытуемых (студентов физического факультета ЛГУ) проходила подготовку перед началом работы на тренажере. Испытуемые должны были решать задачи по выбору оптимального типа взлетно-посадочной полосы для данного типа самолета. Показатели количества ошибок в тренировочной сессии приведены в таблице:

Код испытуемого	1	2	3	4	5	6	7	8	9	10
Количество ошибок	29	54	13	8	14	26	9	20	2	17

1) Поставьте оценки за тренировочную сессию
2) Определите, кого можно допускать к работе, а какого – нельзя
3) Определите шкалы, по которым проведены измерения:
    a) количество ошибок
    b) оценка (за тренировочную сессию)
    c) допуск к работе

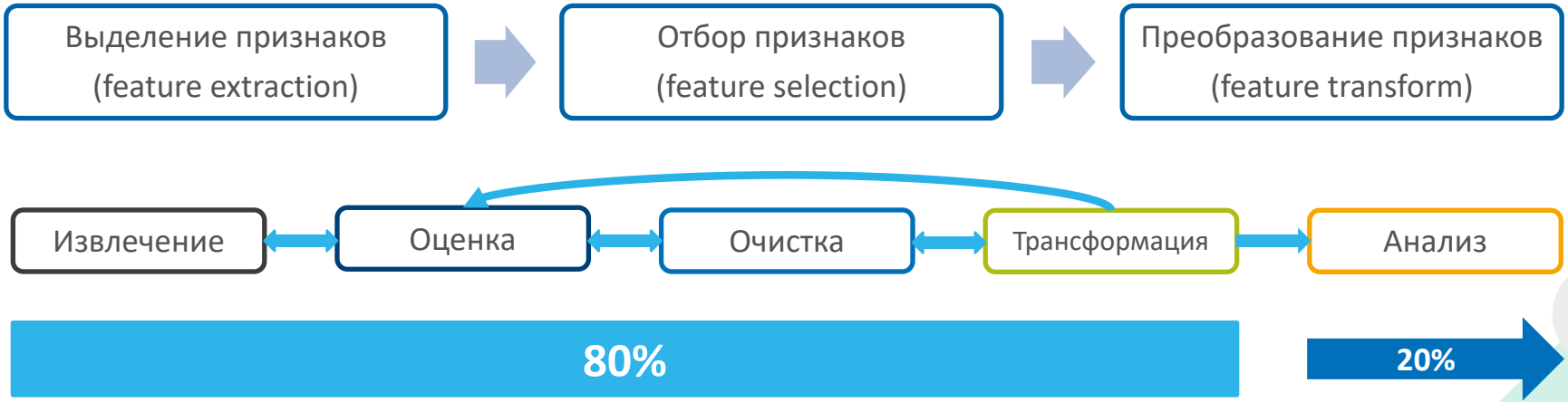
ДАННЫЕ: ЭТАПЫ РАБОТЫ С ДАННЫМИ

Идентификация данных – определение доступных и подходящих данных.

Извлечение данных – извлечение данных из источников с использованием существующих инструментов.

Подготовка данных – определение и осуществление необходимых шагов по трансформации и подготовке данных, очистка данных.

Агрегация данных – проведение необходимых расчетов, расчет необходимых агрегатов.



ДАННЫЕ: ТИПЫ ПЕРЕМЕННЫХ

Бинарная	Это простейший тип переменных только с двумя вариантами значения.
Категориальная	Если вариантов больше двух, информация может быть представлена категориальной переменной.
Целочисленная	Такой тип используется, когда информация может быть представлена целым числом.
Непрерывная	Переменная, содержит числа со знаками после запятой

Номер АЗС	Количество посетителей	Выручка	Автоматическая АЗС	Город
1	132	123,45	нет	Москва
2	145	150,34	нет	Санкт-Петербург
3	89	87,62	да	Екатеринбург

Вопрос: Как кодируем категориальные признаки?

ДАННЫЕ: КАТЕГОРИАЛЬНЫЕ ПАРЕМЕННЫЕ

Когда есть бинарные и категориальные переменные – встает вопрос об их кодировке: модель понимает только числовые значения признаков.

Категориальные признаки (город):

- средние продажи в городе
- сколько складов в городе

Номер Склада	Город	…	Сколько продано товара
1213	Москва	…	111 000
232	Санкт-Петербург	…	456 000
345	Новосибирск	…	34 000
3636	Екатеринбург	…	54 000

ДАННЫЕ: КАТЕГОРИАЛЬНЫЕ ПАРЕМЕННЫЕ

Как еще можно заменить **категориальные признаки**?

One-Hot Encoding.

Предположим, что некоторый признак может принимать 10 разных значений. В этом случае **One-Hot Encoding** подразумевает создание 10 признаков, все из которых равны нулю за исключением одного. На позицию, соответствующую численному значению признака мы помещаем 1.

Номер склада	Город Москва	Город Санкт-Петербург	Город Новосибирск	Город Екатеринбург	…	Сколько продано товара
1213	1	0	0	0	…	111 000
232	0	1	0	0	…	456 000
345	0	0	1	0	…	34 000
3636	0	0	0	1	…	54 000

ДАННЫЕ: ОБРАБОТКА ПРОПУСКОВ

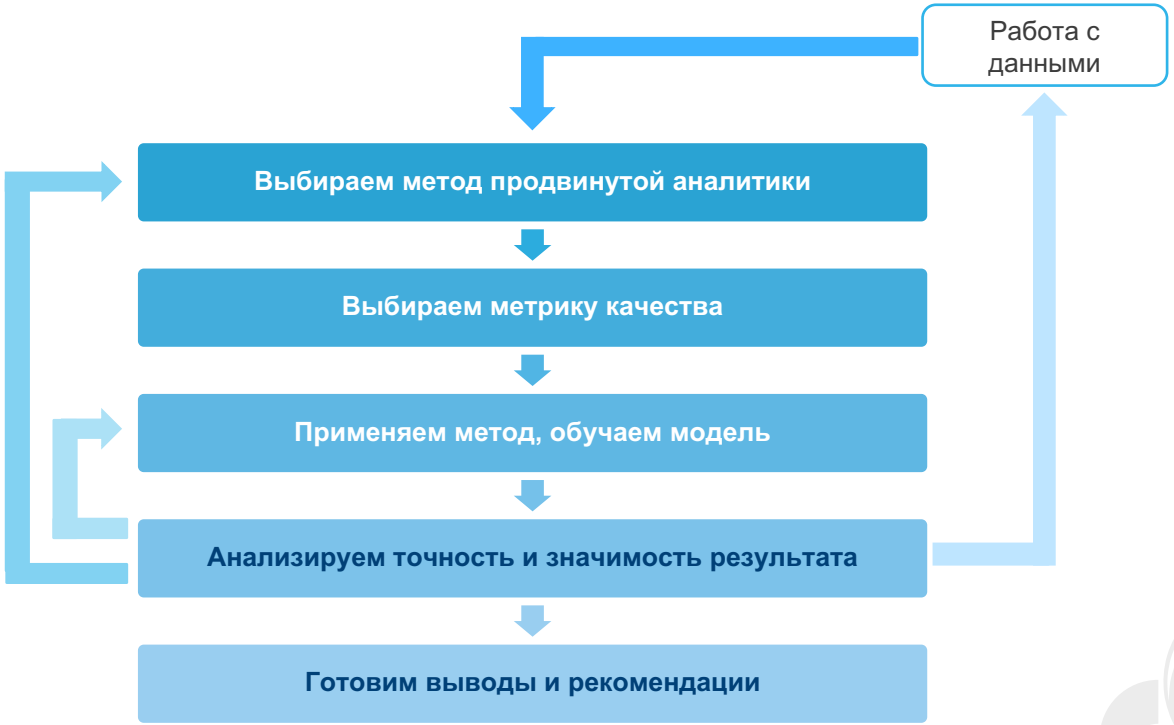
Приближение	❖ Если пропущено значение бинарного или категориального типа, его можно заменить самым типичным значением (мода, медиана, среднее арифметическое, усеченное среднее) переменной.
Вычисление	❖ Пропущенные значения также могут быть вычислены с применением более продвинутых алгоритмов обучения с учителем. Хотя такие вычисления требуют времени, они обычно приводят к более точным оценкам неполных значений.
Удаление	❖ В качестве последнего средства строки с неполными значениями могут быть удалены. Тем не менее этого обычно избегают, чтобы не уменьшать объем данных, доступных для анализа (если данных немного).

Вопрос: Когда какой способ используем?

МОДЕЛИРОВАНИЕ

Предположим:

1. Задача поставлена, определены аспекты, которые необходимо проанализировать, выдвинуты гипотезы для проверки.
2. Работа с данными проведена – отобраны и сконструированы необходимые факторы.
3. Начинаем моделирование:



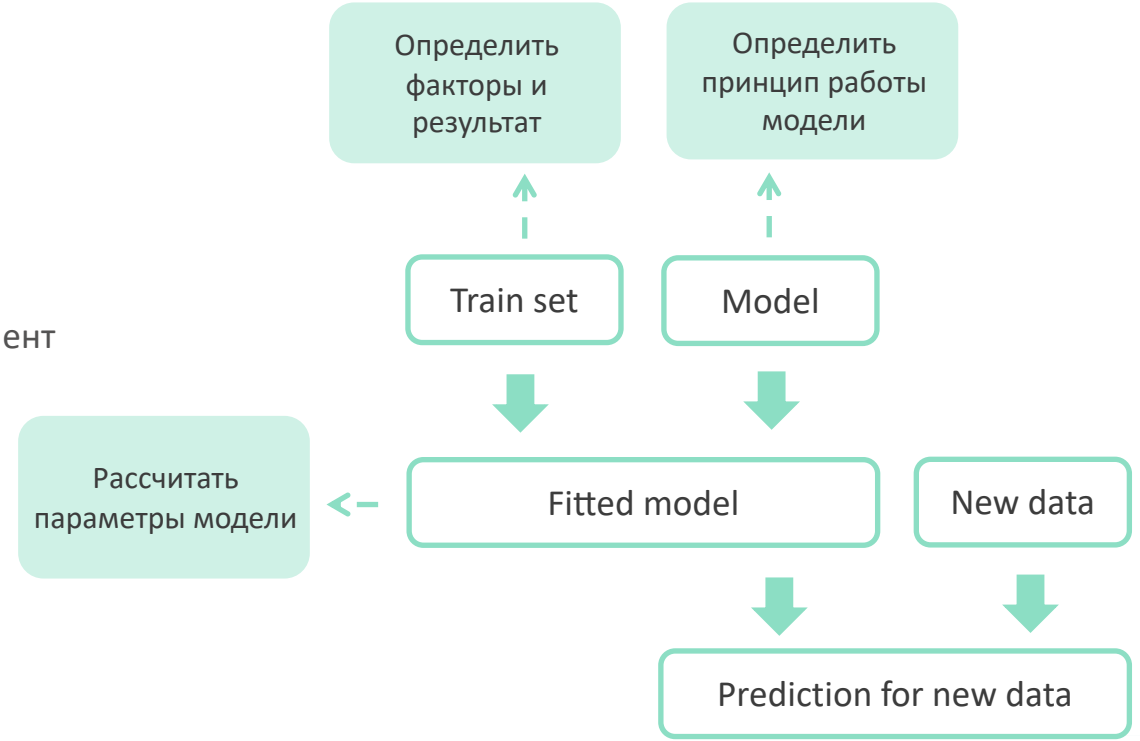
Вопрос: Какие методы и метрики Вам известны?

МОДЕЛИРОВАНИЕ: ML

Пример задачи ML:
Кредитный скоринг:

Задача:
предсказать класс: вернет клиент кредит или нет (0 или 1).

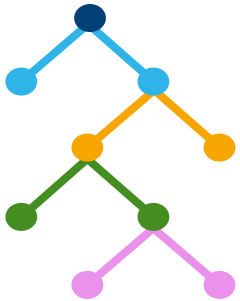
Более глобальная задача:
придумать алгоритм классификации.



МОДЕЛИРОВАНИЕ: ОЦЕНКА ТОЧНОСТИ

Определите, что Вам важнее **точность** или **интерпретируемость**, в зависимости от этого выбирайте модель:

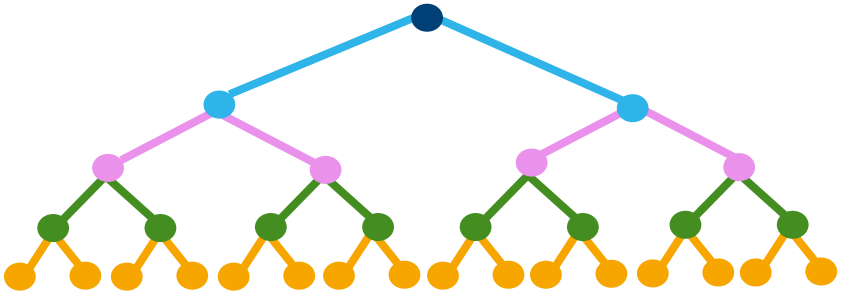
Простая модель



- Легко интерпретировать
- Низкая или средняя точность

VS

Сложная модель

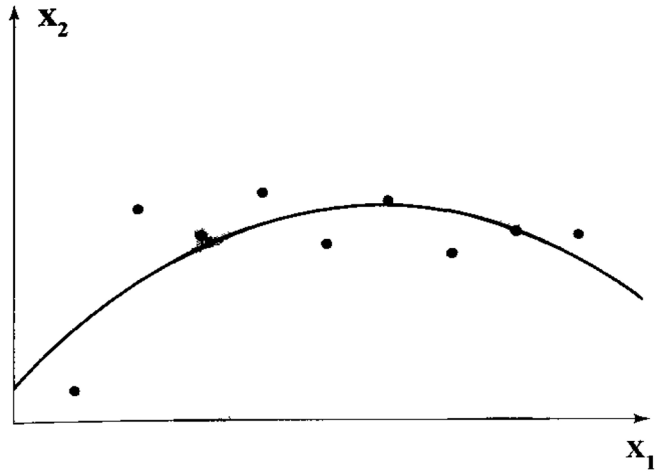


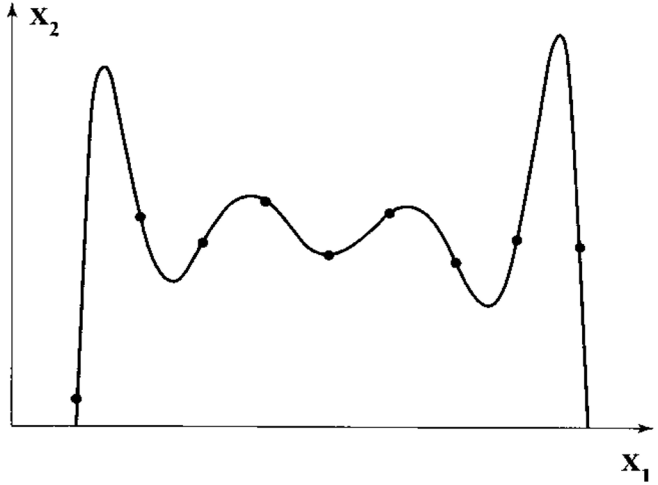
- Сложно интерпретировать
- Высокая точность

МОДЕЛИРОВАНИЕ: UNDERFITTING, OVERFITTING

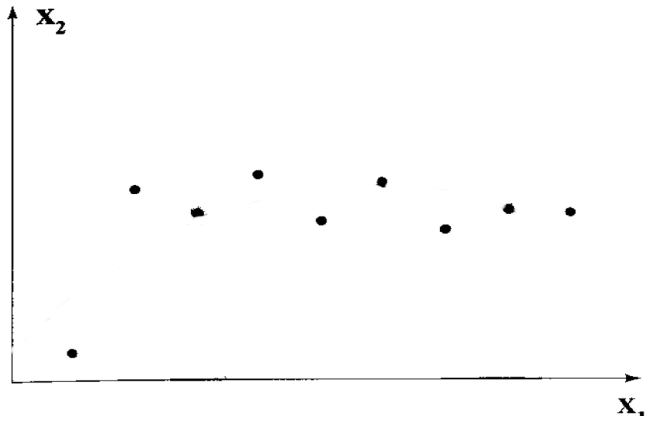
Недообучение (underfitting) – явление, когда ошибка на обучающей выборке достаточно большая, часто говорят «не удаётся настроиться на выборку».

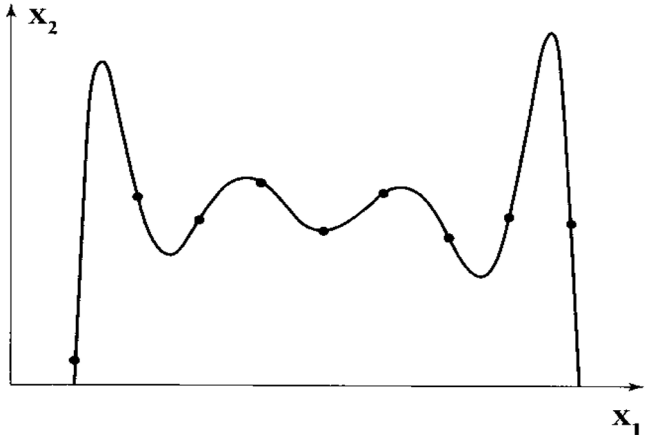
Переобучение (overfitting) – явление, когда ошибка на тестовой выборке заметно больше ошибки на обучающей. Это одна из главных проблем машинного обучения:

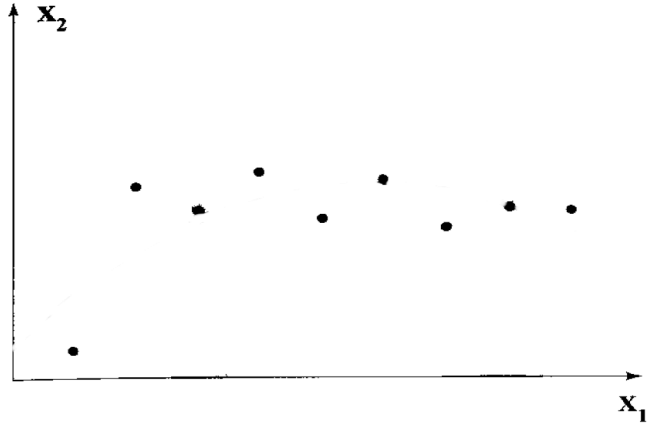


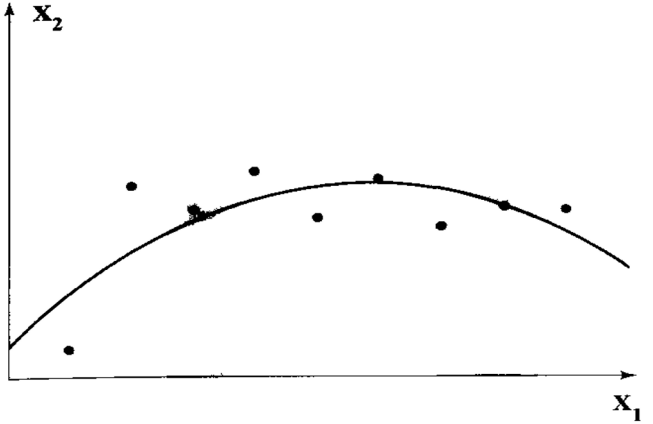


МОДЕЛИРОВАНИЕ: OVERFITTING

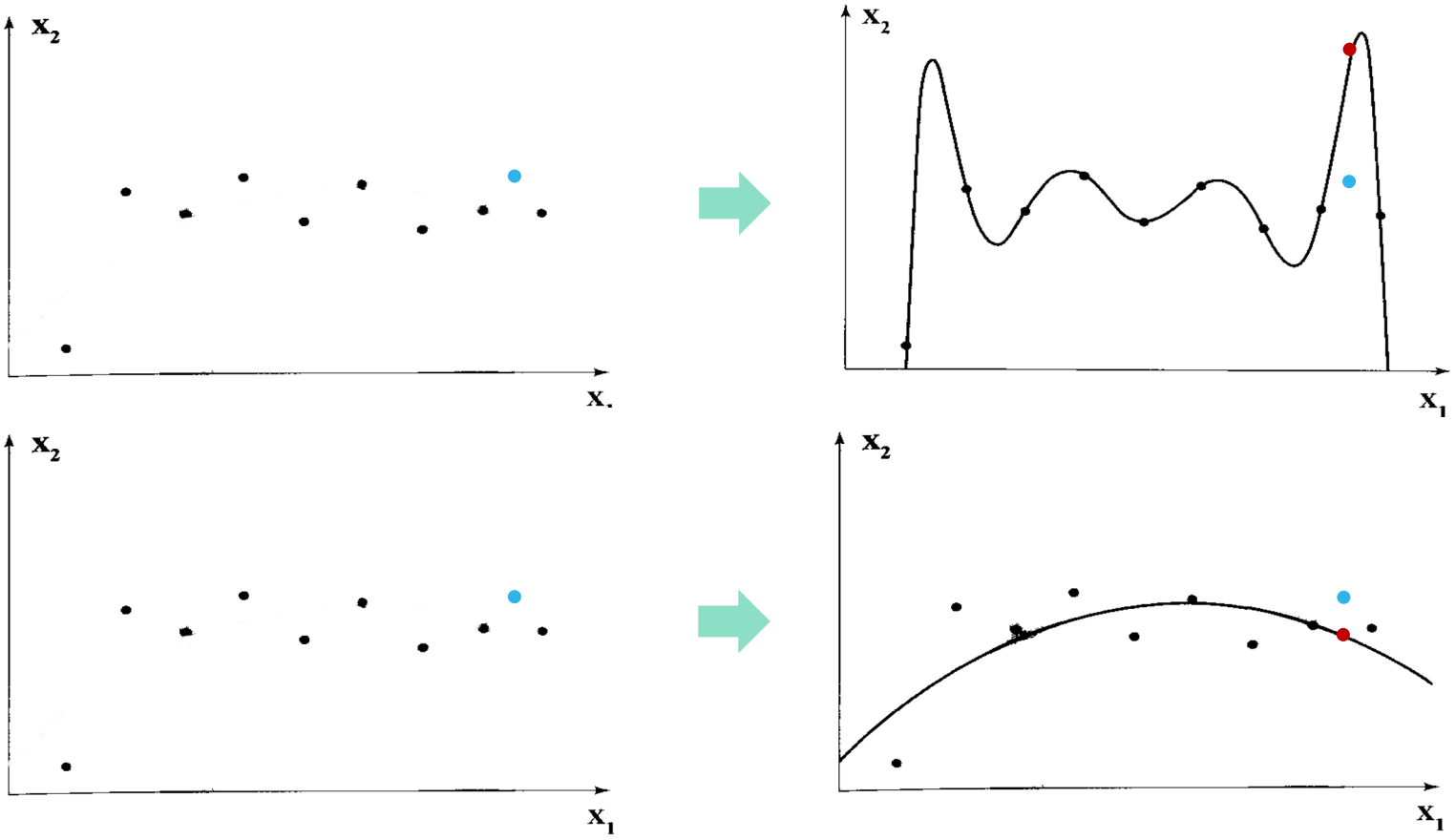




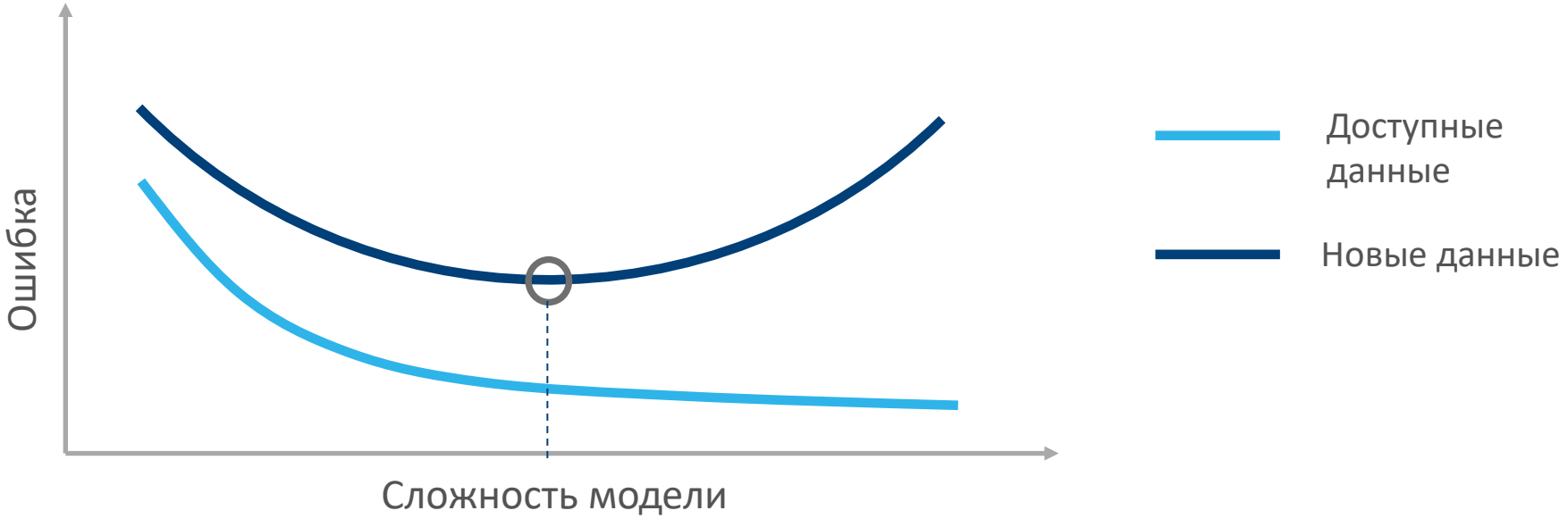




МОДЕЛИРОВАНИЕ: OVERFITTING



МОДЕЛИРОВАНИЕ: OVERFITTING



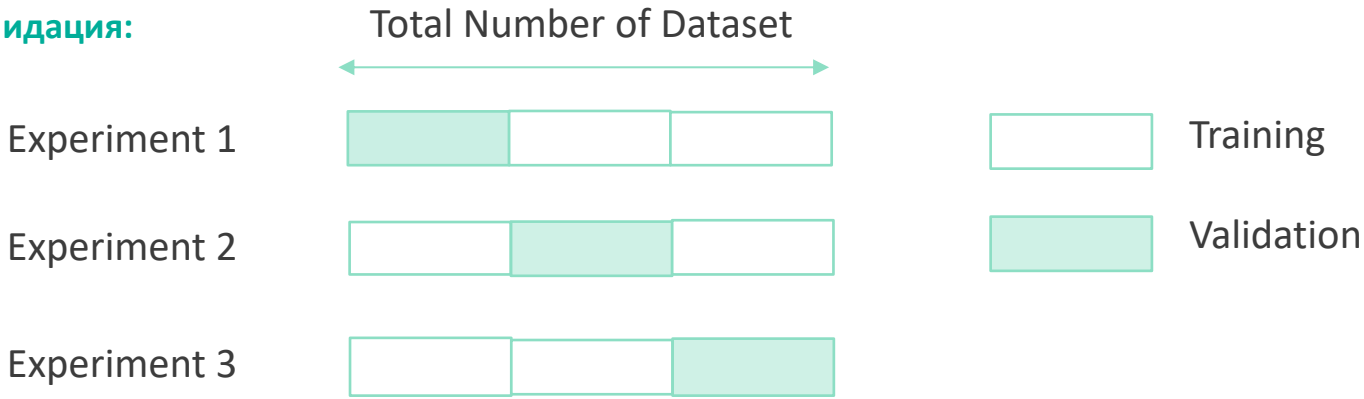
- Модель может идеально описывать доступные данные, но потерять при этом обобщающие свойства
- В итоге это приводит к снижению ожидаемой точности на новых данных
- Это всегда плохо

МОДЕЛИРОВАНИЕ: OVERFITTING

Как бороться с переобучением?

Разбиение на **тренировочную** и **тестовую** выборку:

Кросс-валидация:



МОДЕЛИРОВАНИЕ: OVERFITTING

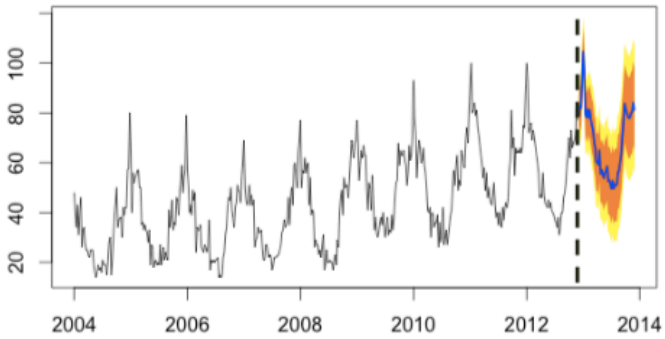
Вопрос: Есть ли танк на фото?

МОДЕЛИРОВАНИЕ: OVERFITTING

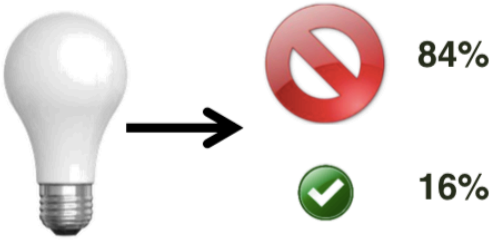
Вопрос: Есть ли танк на фото?

МОДЕЛИРОВАНИЕ: РЕЗУЛЬТАТ

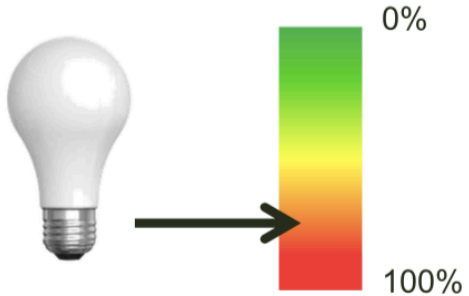
Результат работы модели – ВЕРОЯТНОСТЬ!



95% доверительный интервал прогноза



Решение о выдаче кредита



Вероятность ухода клиента
в следующем месяце

ЖИЗНЕННЫЙ ЦИКЛ АНАЛИТИКИ ДАННЫХ

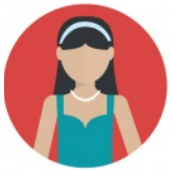
АНАЛИТИЧЕСКАЯ КОМАНДА

Заказчик

лицо (физическое или юридическое), заинтересованное в выполнении исполнителем работ, оказании им услуг или приобретении какого-либо продукта (в широком смысле).

Бизнес - аналитик

специалист, использующий методы бизнес-анализа для исследования потребностей деятельности организаций с целью определения проблем бизнеса и предложения их решения.



Руководитель проекта

специалист, отвечающий за успешное выполнение проекта: в указанные сроки, с необходимым качеством, при фиксированном бюджете, ограниченных человеческих ресурсах и в соответствии с требованиями заказчика.

АНАЛИТИЧЕСКАЯ КОМАНДА

Data Engineers

специалист, который делает процесс анализа данных в компании удобным для аналитиков, обеспечивает их очищенными данными в должном количестве и в должный срок.

Архитектор данных

специалист, который выбирает технологии для хранения информации, создает и оптимизирует запросы, занимается проектированием баз данных, контролирует их безопасность.

Data Scientist

эксперт по аналитическим данным, который обладает техническими навыками для решения сложных задач.

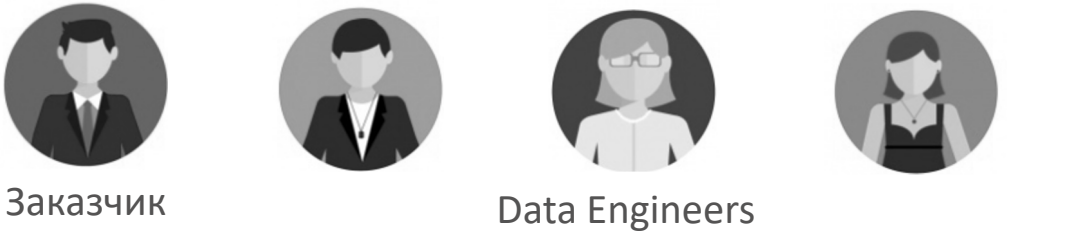
Разработчик ПО

специалист, заинтересованный в аспектах процесса разработки ПО, включая исследования, разработку, программирование и тестирование.

АНАЛИТИЧЕСКАЯ КОМАНДА

Задача 1:

Оценка эффективности маркетинговой акции.



АНАЛИТИЧЕСКАЯ КОМАНДА

Задача 2:

Прогнозирование дефолта контрагентов.

Заказчик

Data Scientist

Data Engineers

Бизнес - аналитик

Разработчик ПО

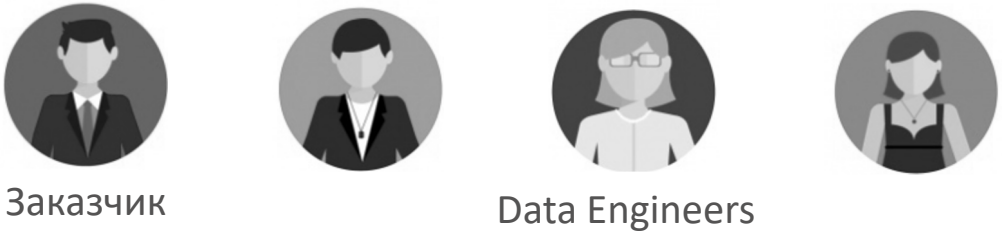
Руководитель проекта

Архитектор данных

АНАЛИТИЧЕСКАЯ КОМАНДА

Задача 3:

Распознавание клиента (аудио/видео – аналитика и пр.).



АНАЛИТИЧЕСКАЯ КОМАНДА

Задача 1:

Оценка эффективности маркетинговой акции.

Заказчик

Data Scientist

Data Engineers

Бизнес - аналитик

Руководитель проекта

Разработчик ПО

Архитектор данных

АНАЛИТИЧЕСКАЯ КОМАНДА

Задача 2:

Прогнозирование дефолта контрагентов.

Заказчик

Data Scientist **x2**

Data Engineers

Бизнес - аналитик

Руководитель проекта

Разработчик ПО

Архитектор данных

АНАЛИТИЧЕСКАЯ КОМАНДА

Задача 3:

Распознавание клиента (аудио/видео – аналитика и пр.).

Заказчик

Data Scientist **x3**

Data Engineers

Бизнес - аналитик

Руководитель проекта

Разработчик ПО

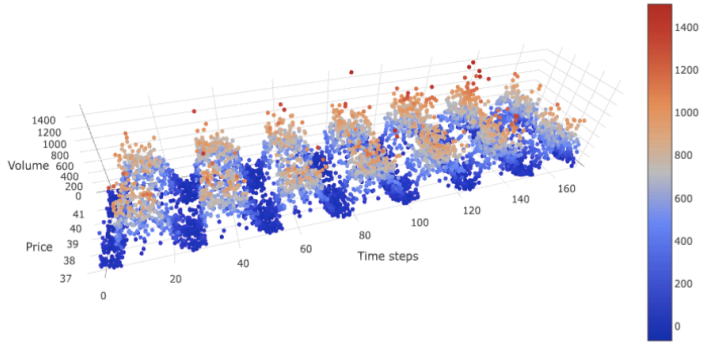
Архитектор данных

ИНТЕГРАЦИЯ АНАЛИТИКИ В БИЗНЕС-ПРОЦЕСС

DATA SCIENCE ПРОЕКТЫ

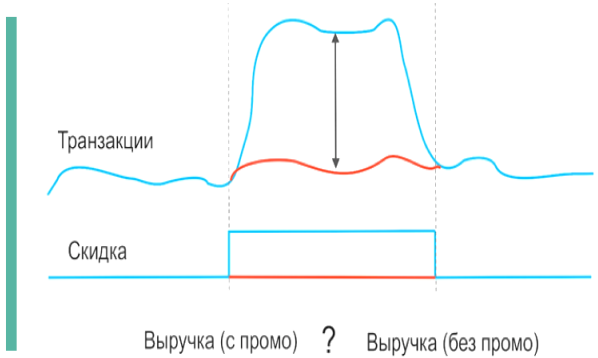
Динамическое ценообразование

- Адаптивная управление ценой на стеле на базе обучения с подкреплением
- Автоматизация принятия решений до 80%, повышение эффективности ценообразования



Оценка эффективности промо-акций

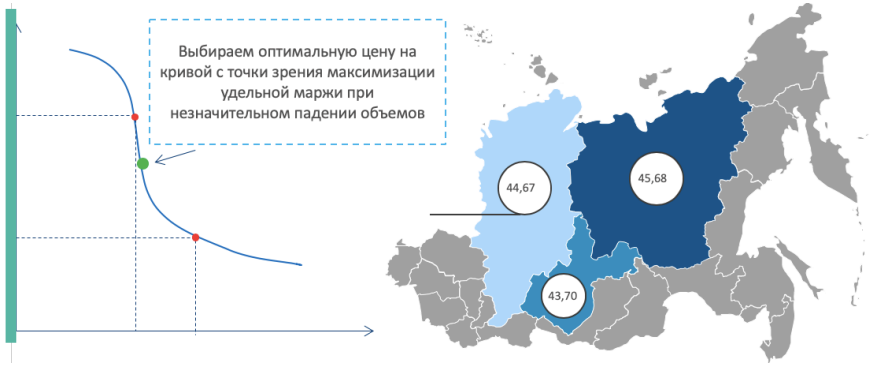
- Оценка эффекта с учетом тренда, сезонности, продаж и цен в категории, характеристик АЗС
- WHAT-IF анализ для акций (прошедшие и будущие)



DATA SCIENCE ПРОЕКТЫ

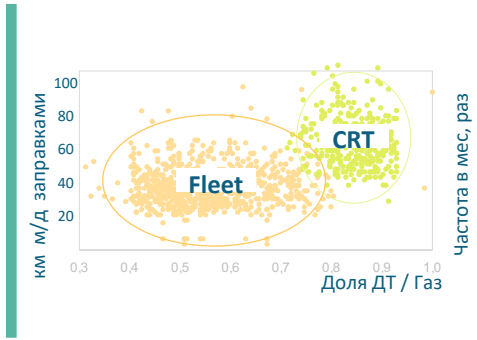
Central Pricing Desk

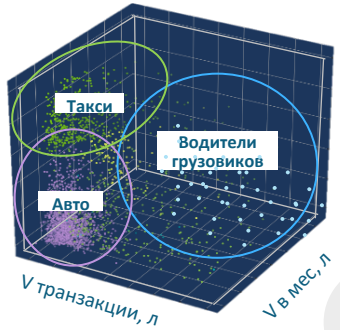
- Ценообразование с учетом чувствительности корпоративных клиентов к цене
- Точность прогнозирования ~95%
- Повышение маржинального дохода, упрощение контроля расходов для клиента



Сегментация клиентов

- Продвинутая сегментация клиентов с использованием ML
- Рост отклика на целевые предложения в 2 раза
- Увеличение удельного маржинального дохода на клиента по НП и СТиУ

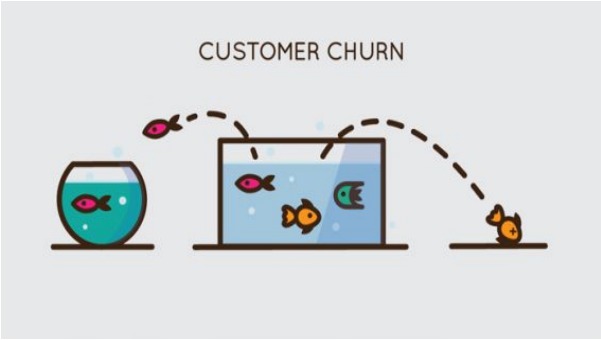




DATA SCIENCE ПРОЕКТЫ

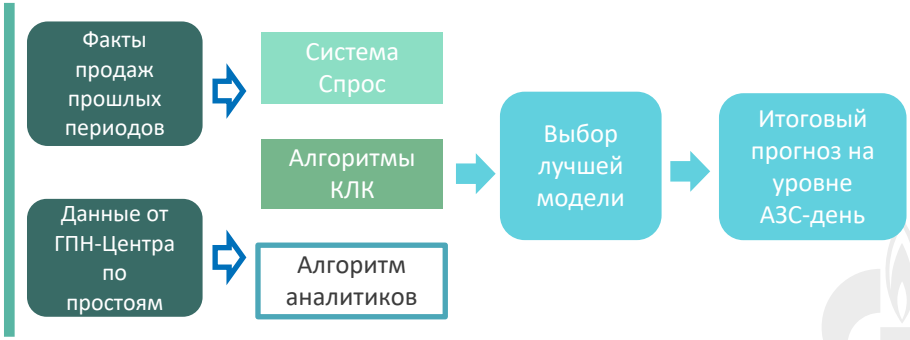
Сервис предотвращения оттока

- Определение вероятности оттока и экономической целесообразности возврата клиента
- В 2 раза эффективнее текущей механики
- Автоматический возврат клиентов



Планирование продаж АЗС

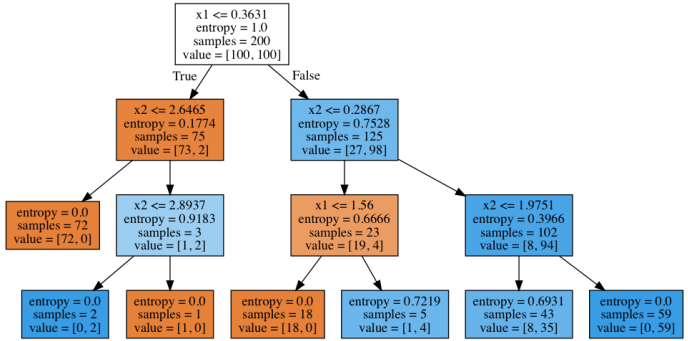
- Применение машинного обучения для построения плана продаж
- Точность прогнозирования 96% (план-10)
- Автоматизация планирования, повышение точности



DATA SCIENCE ПРОЕКТЫ

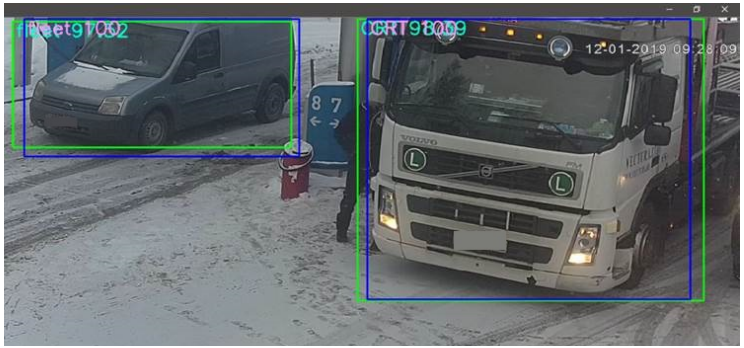
Кредитный рейтинг контрагентов

- Риск дефолта по данным СПАРК и внутренней отчетности
- Точность в 2.5 раза выше базовой
- Возможность уменьшения просроченной задолженности



Видео/аудио аналитика

- Детекция и классификация объектов по видео
- Обогащение клиентских профилей



ЗОНА ИССЛЕДОВАНИЯ ДАННЫХ

ЗОНА ИССЛЕДОВАНИЯ ДАННЫХ: ПЕСОЧНИЦА

«Песочница» Озера данных (ОД) ДРП

специально выделенная (изолированная) среда для проведения разработок и исследования данных, которая предназначена для тех, кто:

- является экспертом по аналитическим данным
- хочет программировать, придумывать и решать аналитические задачи

ЗОНА ИССЛЕДОВАНИЯ ДАННЫХ: ИНСТРУМЕНТЫ

В «песочнице» установлены следующие инструменты **анализа данных**:

SQL – язык структурированных запросов, применяемый для создания, модификации и управления данными в базе данных

R – язык программирования для статистической обработки данных и работы с графикой

Python – высокоуровневый язык программирования общего назначения, который широко применяется при решении задач продвинутой аналитики

Power BI – платформа бизнес-аналитики, с помощью которой можно визуализировать данные и получать полезные сведения для быстрого принятия взвешенных решений

ЗОНА ИССЛЕДОВАНИЯ ДАННЫХ: ИНСТРУМЕНТЫ

В «песочнице» установлены следующие инструменты **анализа данных**:

Confluence – вики-система для внутреннего использования с целью создания единой базы знаний по проектам

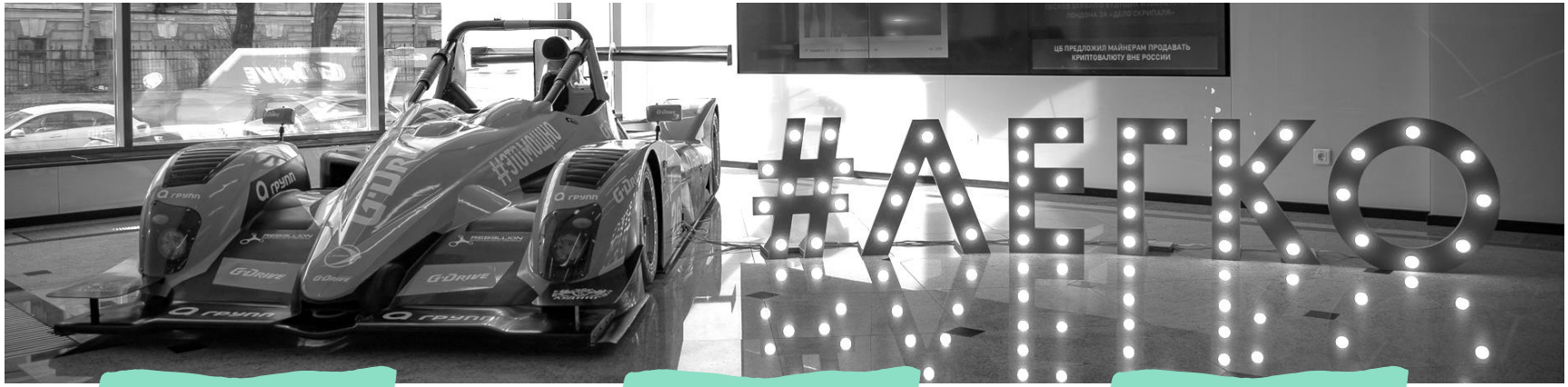
Jira – инструмент для отслеживания ошибок и управления проектами

Gitlab –система управления репозиториями кода для Git с собственной вики, системой отслеживания ошибок

EDC, Axon – программные продукты для управления каталогом данных и бизнес-глоссарием с возможностью контроля качества данных

СТАЖИРОВКИ

СТАЖИРОВКИ



Реальные задачи

Ты получишь уникальные возможности поработать с текущими задачами компании

Передовые знания

Ты узнаешь о новых подходах и технологиях, которые применяются в компании

Старт успешной карьеры

Ты наберешься опыта в компании, а также у тебя будет возможность присоединиться к команде

НАПРАВЛЕНИЯ СТАЖИРОВОК

GPN Intelligence Cup

ежегодный онлайн
кейс-чемпионат
ДРП «Газпром нефть»
для студентов ВУЗов
технологических и
экономических специальностей

http://gpn-cup.ru/

Анализ и инжиниринг данных

1. Инжиниринг данных
2. Разработка BI-приложений
3. Продвинутая аналитика
4. Системный анализ

Бизнес и стратегия

1. Бизнес-аналитика
2. Анализ эффективности
3. Инфографика

СПАСИБО ЗА ВНИМАНИЕ

Пожалуйста, ответьте на несколько вопросов о вебинаре, анкету можно открыть, если навести на QR-код: