Практикум

# Кто такие аналитики данных

Продолжительность основной программы — 7 месяцев

## 01. В чём суть профессии

Аналитики данных помогают компаниям принимать решения на основе данных. Они ищут в данных закономерности, чтобы понять, почему происходят те или иные события, или предугадать, как всё может поменяться в будущем.

## 02. В каких сферах работают

Анализ данных нужен везде: в маркетинге, финансах, промышленности, разработке любых новых продуктов и технологий.

## 03. Какие задачи решают

Ритейл-сеть хочет найти районы с высокой плотностью населения и небольшим количеством супермаркетов. Нужно проанализировать данные, которые есть в компании, а также информацию из внешних источников.

Маркетплейсу нужно сравнить два варианта текста рекламной рассылки и определить, на какой из них лучше реагируют пользователи, чтобы увеличить продажи.

Отделу маркетинга магазина одежды нужно определить причины провала рекламной кампании и оценить факторы, которые к этому привели.

## 04. Карьерные перспективы

Согласно исследованию «Анкора», около 45% всех компаний в России ищут специалистов по работе с данными.

## Уровни зарплат

**69 000 ₽**
Джуниор аналитик данных

**130 000 ₽**
Мидл аналитик данных

**220 000 ₽**
Синьор аналитик данных

## На курсе — всё, что нужно, чтобы начать карьеру аналитика

- Диплом о переподготовке
- Непрерывная практика
- Портфолио из 13+ проектов
- Помощь в трудоустройстве

Аналитик данных

# Почему этот курс подойдёт студентам без опыта

| | |
|---|---|
| Свободное расписание | Читать теорию и практиковаться можно в любое время, главное — соблюдать сроки сдачи проектов. У нас есть мобильное приложение для учёбы, чтобы вы не были привязаны к компьютеру. |
| Понятная теория | Термины и правила подкреплены примерами из жизни. Сложность и длина программы курса рассчитаны так, чтобы каждую следующую главу вы понимали всё лучше. Мы постоянно актуализируем программу: последнее обновление на данный момент — июль 2024 года. |
| Практика в тренажёре | Учитесь читать, визуализировать и интерпретировать данные. Ошибайтесь, быстро получайте обратную связь и исправляйте ошибки. |
| Учёба на реальных задачах | Вас ждут типичные для аналитика задачи из разных сфер. Выполненные проекты вы сможете добавить в своё портфолио. |
| Команда сопровождения | Мы поддержим, объясним сложные темы, поможем улучшить проекты и не дадим сдаться на полпути. |

| | | |
|---|---|---|
| Основы анализа данных с помощью SQL и BI | Анализ данных с помощью Python | Расчёт и визуализация бизнес-метрик и показателей |
| Визуализация данных с помощью DataLens. Создание дашбордов | Развитие бизнес-мышления и работа с метриками продукта и бизнеса | Анализ результатов A/B-тестирования с помощью Python |

# Чтобы начать учиться, не нужен опыт в IT

Среди наших выпускников, которые стали аналитиками данных, есть спортсмены, экономисты, таксисты и разнорабочие. После наших курсов аналитики находят работу и в стартапах, и в корпорациях вроде Яндекса, «Леруа Мерлен», «Вкусвилла» и Сбера.

Аналитик данных

Яндекс Практикум

# Аналитик данных, Аналитик данных расширенный

Основная программа — 7 месяцев

Расширенная программа — 11 месяцев

01

Основы анализа данных с помощью SQL и BI

12 недель

02

Анализ данных с помощью Python

8 недель

03

Продвинутый анализ данных для бизнеса

10 недель

Блоки, входящие в тариф «Аналитик данных расширенный»

04

Расширенная аналитика: дашборды, продукты и алгоритмы

9 недель

05

Работа с большими данными

6 недель

01

# Основы анализа данных с помощью SQL и BI

Спринт 1

2 недели

## Введение в аналитику. Аналитический отчёт в Google Таблицах

Узнаете, кто такой аналитик данных и какие задачи он решает. Познакомитесь с пайплайном работы аналитика. Создадите свой первый аналитический отчёт в Google Таблицах.

**Проект**

Создание аналитического отчёта в Google Таблицах для небольшого салона для кудрявых

**Темы**

1. **Использование данных в бизнесе**
   Структурированные и неструктурированные данные. Роль данных в бизнесе.
2. **Процесс анализа данных и задачи аналитика**
   Задачи аналитика данных. Пайплайн работы аналитика. Специализации в аналитике.
3. **Excel как инструмент аналитика. Основы Google Таблиц**
   Табличные редакторы, начало работы в Google Таблицах. Константы и формулы.
4. **Предобработка данных в Google Таблицах**
   Типы данных: числовые, текстовые. Форматирование данных. Очистка данных. Использование панели автоподсчёта. Сортировка и фильтрация данных.
5. **Использование формул и функций**
   Формулы и функции. Обзор базовых функций, синтаксис, использование. Математические функции (SUM, COUNT, ROUND, MIN, MAX, AVERAGE). Логические функции AND, OR, NOT, функции с условиями (IF, SUMIF и др.). Абсолютные и относительные ссылки. Функции даты и времени. Использование VLOOKUP (ВПР). Сводные таблицы.
6. **Презентация данных**
   Построение простых визуализаций. Как поделиться отчётом.

Инструменты и технологии

Пайплайн работы аналитика
Google Таблицы
Формулы
Функции
Отчёты

Спринт 2

2 недели

# Основы SQL. Извлечение данных для анализа

Узнаете, как могут храниться данные, и познакомитесь с языком запросов SQL для работы с базами данных. Напишете первые запросы на SQL и научитесь извлекать данные под задачу с фильтрацией, группировкой, сортировкой.

Инструменты и технологии

БД и СУБД

SQL

PostgreSQL

Типы данных

Группировка данных

Сортировка данных

Проект

Выгрузка данных для проведения статистики прослушиваний музыкального стримингового сервиса Поток

Темы

1. **Работа с базами данных**
   Откуда аналитик получает данные. Что такое база данных. Как управляют базами данных с помощью СУБД. Первый запрос: выгружаем данные с помощью SQL.

2. **Типы данных и их преобразования**
   Какими бывают данные. Меняем тип данных. Округляем данные. Арифметические операции.

3. **Фильтрация данных и агрегация**
   Когда нужна фильтрация данных. Фильтрация по условиям. Агрегация данных.

4. **Группировка и сортировка данных**
   Группируем данные. Группировка по нескольким полям. Фильтрация после агрегации. Сортировка данных по одному полю. Сортировка данных по нескольким полям.

Спринт 3

2 недели

# SQL. Обработка данных

Продолжите знакомиться с инструментами SQL и научитесь обрабатывать данные для анализа: устранять дубликаты и работать с пропущенными значениями. Сможете извлекать данные из нескольких таблиц, используя JOIN-ы, использовать подзапросы и CTE.

Инструменты и технологии

SQL

Пропуски

Дубликаты

Подзапросы

CTE

Присоединения таблиц (JOIN)

Проект

Проанализируете данные о тарифных планах и активности клиентов федерального оператора сотовой связи «Мегасеть»

Темы

1. **Связи между таблицами**
   Как таблицы хранятся в базе данных. Для чего нужны связи в базе данных. ER-диаграммы.

2. **Работа с пропущенными значениями и дубликатами**
   Как работать с пропусками и дубликатами. Работа с пропусками. Заполнение пропусков. Работа с дубликатами.

3. **Присоединение таблиц**
   Присоединение таблиц. Разные типы присоединения таблиц (INNER JOIN, LEFT JOIN и RIGHT JOIN, FULL OUTER JOIN). Особенности присоединения таблиц.

4. **Операции множеств и подзапросы**
   Объединение множеств. Пересечение и вычитание, Подзапросы (в секции WHERE и в секции FROM). CTE (обобщённые табличные выражения).

5. **Категоризация значений. Создание новых столбцов**
   Операции со столбцами (вычитание, сложение, усреднение). Категоризация значений (CASE WHEN THEN END). Обработка неявных дубликатов.

6. **Работа с датой и временем**
   Типы данных для даты и времени (TIMESTAMP, DATE, INTERVAL). Функции  EXTRACT() и DATE_TRUNC(). Фильтрация по дате и работа с интервалами.

Спринт 4

2 недели

# SQL. Анализ данных и решение ad hoc задач

Научитесь применять продвинутые инструменты SQL (оконные функции) для решения ad hoc задач аналитика разной сложности. Познакомитесь с необходимыми для решения таких задач понятиями описательной статистики.

Проект

Кейс-проект с ревью: проанализируете данные о продажах внутри онлайн-игры «Секреты Темнолесья»

Темы

1. **Знакомство с базой данных**
   Как изучать БД самостоятельно. Выводы об устройстве БД.

2. **Оконные функции. Агрегирующие функции**
   Назначение оконных функций, их классификация. Агрегирующие оконные функции  (SUM(), COUNT(), AVG(), MIN(), MAX()). Предложения PARTITION BY и ORDER BY.

3. **Оконные ранжирующие функции**
   Назначение функций ROW_NUMBER(); RANK() и DENSE_RANK(), NTILE(). Особенности ранжирующих оконных функций.

4. **Оконные функции смещения**
   Назначение и особенности функций LEAD(), LAG(), FIRST_VALUE(), и LAST_VALUE().

5. **Описательная статистика. Аналитические функции**
   Категориальные и количественные переменные. Меры центральной тенденции (среднее значение, мода, медиана, различие среднего и медианы, перцентили). Меры разброса. Аналитические функции PERCENTILE_DISC(), PERCENTILE_CONT(), оператор WITHIN, функция STDDEV(). Назначение и использование.

Инструменты и технологии

Решение ad hoc задач

Декомпозиция

SQL

Агрегирующие оконные функции

Ранжирующие оконные функции

Оконные функции смещения

Аналитические оконные функции

Мода

Медиана

Среднее

Перцентиль

Размах

6. **Практика решения ad hoc задач**
   Что такое ad hoc запросы. Алгоритм решения ad hoc запроса. Декомпозиция. Решение ad hoc запросов повышенной сложности.

Спринт 5

2 недели

# Визуализиция данных с помощью DataLens. Создание дашбордов

Разберётесь с основами визуализации данных в BI-инструменте DataLens. Научитесь подбирать тип визуализации под задачу. Изучите основы создания и настройки дашбордов.

Инструменты и технологии

SQL

DataLens

BI-инструменты

Чарт

Визуализация данных

Дашборд

Проект

Кейс-проект с ревью: разработка дашборда по чёткому ТЗ на данных конференция TED

Темы

1. **Визуализация в работе аналитика. Знакомство с DataLens**
   Визуализация как задача аналитика. BI-инструменты. Порядок работы в BI-инструментах. Интерфейс DataLens. Подключения в DataLens. Данные (credentials) для подключения к базе данных. Датасеты в DataLens. Типы данных в DataLens.

2. **Основы визуализации. Чарты**
   Виды визуализаций. Типы графиков. Линейная диаграмма, столбчатая диаграмма, линейчатая диаграмма, кольцевая диаграмма, круговая диаграмма, накопительная диаграмма с областями, таблица, сводная таблица, индикатор. Элементы визуализации. Создание чартов в DataLens. Оформление графиков. Графики для визуализации сравнения, соотношения части и целого, отображения изменений во времени.

3. **Вычисляемые поля**
   Вычисляемые поля на уровне датасета и на уровне чарта. Формулы. Агрегирующие функции: MIN(), MAX(), AVG(), AVG_IF(), COUNT(), COUNT_IF(), SUM(), SUM_IF(), COUNTD() и другие. Логические функции: IF(), CASE(). Функции для работы со строками: REPLACE(), CONCAT(), STARTSWITH() / ENDSWITH(), CONTAINS() и другие. Функции для работы с датами: DATEADD(), DATETRUNC(), DATEPART(), YEAR(), MONTH() и другие.

4. **Дашборды**
   Назначение дашбордов. Прототипирование дашбордов. Виджеты. Чарты. Добавление чартов на дашборд. Селекторы. Связи: входящие и исходящие. Настройка селекторов. Тексты. Заголовки. Композиция дашборда.

5. **Параметры**
   Использование параметров для дашборда. Использование параметров для чарта. Ситуации, в которых используются параметры. Специальные параметры.

6. **Интерпретация данных из дашбордов**
   Задачи визуализации и её аудитория. Принципы хорошей визуализации. Работаем с дашбордом.

1 неделя

# Итоговый проект модуля

Познакомитесь с БД через SQL и создадите дашборд с использованием связки SQL и BI.

Проект

Проанализируете объявления о продаже жилой недвижимости с помощью SQL-запросов и доработаете существующий дашборд по требованиям заказчика

Каникулы ⟶ 1 неделя

02

# Анализ данных с помощью Python

Спринт 6

## Основы Python

2 недели

Начнёте знакомство с языком программирования Python. Изучите основы синтаксиса, необходимые для последующего написания кода.

Инструменты и технологии

Python

Переменные

Типы данных

Строки

Списки

Циклы

Условный оператор

Функции

Множества

Словари

Проект

Решение проверочных заданий на знание синтаксиса Python

Темы

1. **Основы синтаксиса Python**
   Что такое Python. Почему аналитики пишут на Python. Что такое переменная. Как выбрать имя переменной. Арифметические операции. Используем переменные.

2. **Определение данных и их типы**
   Типы данных. Преобразование типов. Как работать со строками. Заполнение списков. Операции со списками. Строки и списки.

3. **Условные выражения**
   Логические выражения и операторы сравнения. Сложные логические выражения. Возвращаемые элементы True и False. Условная конструкция. Дополнительные ветки кода.

4. **Циклы и их организация**
   Понятие цикла. Элементы цикла. Перебор элементов цикла. Циклы и строки. Вложенные циклы. Обработка списков в цикле. Функция enumerate() в заголовке цикла. Функция zip() в заголовке цикла. Форма алгоритма. Управление циклом. Операторы break, continue и pass.

5. **Функции в Python**
   Встроенные функции и методы. Пользовательская функция и её синтаксис. Аргументы функции. Тело функции. Вывод результатов с помощью return. Математические функции.

6. **Словари и множества**
   Создание словаря. Поиск в словаре. Изменение словарей. Словари и циклы. Множества.

7. **Работа с вложенными структурами**
   Вложенные структуры: списки. Списковые включения. Работа с вложенными списками. Вложенные словари. Словари с вложениями. Классы, объекты, конструкторы.

Спринт 7

# Python. Предобработка данных

2 недели

Начнёте работу с библиотекой pandas. Научитесь предобрабатывать данные с помощью Python: очищать данные от выбросов, пропусков и дубликатов и преобразовывать разные форматы данных.

Инструменты и технологии

Python

Pandas

Предобработка данных

Обработка пропусков

Обработка дубликатов

Категоризация данных

Проект

Кейс-проект с ревью: проведёте предобработку данных в Python онлайн-игры «Секреты Темнолесья» для последующего анализа

Темы

1. **Основы библиотеки pandas. Обзор данных**
   Что такое библиотека в Python. Библиотека pandas. Объекты DataFrame и Series. Работа с файлами в датафрейме. Обзор данных в датафрейме. Сортировка данных.

2. **Типы данных. Работа с датой и временем**
   Типы данных pandas. Тип данных object. Пропуски. Преобразование типов данных. Типы данных для даты и времени. Операции с датой и временем. Интервал времени timedelta64. Среда Jupyter Notebook.

3. **Индексация в датафреймах. Фильтрация данных**
   Индексация и фильтрация в pandas. Индексация с помощью loc и iloc. Индексация с помощью логических операторов.

4. **Работа с пропущенными значениями**
   Методы работы с пропусками. Заполнение пропусков. Удаление пропусков. Явные и неявные дубликаты. Работа со строками.

5. **Обработка дубликатов**
   Дубликаты в данных. Виды дубликатов: явные и неявные дубликаты. Влияние дубликатов на дальнейший анализ и визуализации. Нахождение дубликатов в датафрейме. Удаление дубликатов. Подсчёт количества дубликатов.

6. **Категоризация данных**
   Категоризация данных. Функция apply() для Series. Функция apply() для датафреймов. Группировка данных. Агрегация с помощью метода agg().

Спринт 8

3 недели

# Исследовательский анализ данных и визуализация с помощью Python

Научитесь использовать Python для исследования и визуализации данных. Разберётесь с основами описательной статистики на примерах.

Инструменты и технологии

Python

Pandas

Matplotlib

Seaborn

Jupyter Notebook

Проект

Кейс-проект с ревью: проведение исследовательского анализа данных рынка общественных заведений Москвы для инвестиционного фонда

Темы

1. **Объединение датафреймов**
   Присоединение датафреймов. Присоединяем датафреймы по столбцам. Присоединяем датафреймы по индексам. Методы join() и merge().

2. **Описательная статистика**
   Меры центральной тенденции. Меры разброса. Строим гистограмму. Строим диаграмму размаха. Изучаем меры размаха. Оформляем графики.

3. **Взаимосвязь переменных**
   Знакомство с корреляцией. Виды корреляции. Считаем и оцениваем корреляцию. Визуализируем корреляцию.

4. **Визуализация для изучения данных**
   Столбчатая диаграмма. Линейный график. Тепловая карта. График распределения. Как аналитик исследует графики.

5. **Сводные таблицы**
   Сводные таблицы в pandas. Работа с мультииндексами в сводной таблице. Фильтрация и работа с пропусками в сводной таблице.

6. **Пример исследовательского анализа данных**
   Разбор исследовательского анализа данных в среде Jupyter Notebook.

1 неделя

# Итоговый проект модуля

Под запрос бизнеса выполните исследовательский анализ данных с последующей визуализацией с помощью инструментов Python. Предоставите рекомендации бизнесу по итогам исследования.

Проект

Исследование данных с помощью Python о стартапах для финансовой компании

Каникулы ⟶ 1 неделя

03

# Продвинутый анализ данных для бизнеса

Спринт 9

2 недели

## Расчёт и визуализация бизнес-метрик и показателей

Погрузитесь глубже в контекст бизнеса и продукта. Рассчитаете и визуализируете важные для бизнеса показатели с помощью SQL и разберётесь с основами когортного анализа.

Инструменты и технологии

SQL

DataLens

Проект

Кейс-проект с ревью: расчёт, визуализация и анализ основных метрик сервиса доставки еды

Темы

1. **Что такое бизнес-метрики и зачем они нужны**
   Различия бизнес-анализа и продуктового анализа. Продуктовый подход к анализу метрик. Классификация метрик. Data-driven-подход и генерация гипотез для анализа метрик.

2. **Расчёт, визуализация и интерпретация продуктовых метрик**
   Расчёт DAU. Визуализация DAU. Расчёт конверсии. Визуализация и интерпретация конверсии. Расчёт среднего чека. Визуализация и интерпретация среднего чека. Расчёт LTV. Визуализация и интерпретация LTV.

3. **Продуктовая воронка**
   Что такое продуктовая воронка. Расчёт воронки. Визуализация воронки и формулировка рекомендаций для бизнеса.

4. **Расчёт метрик конверсии и LTV по когортам**
   Что такое когортный анализ. Расчёт Retention Rate по когортам. Визуализация и интерпретация Retention Rate по когортам. Расчёт LTV по когортам. Визуализация и интерпретация LTV по когортам. Юнит-экономика: окупаемость пользователя или продукта.

5. **Выбор ключевых и кастомных метрик**
   Что такое ключевые метрики. Что такое кастомные метрики. Выбор ключевых метрик для разных сфер.

Спринт 10

3 недели

# Формулировка и проверка гипотез. Статистический анализ данных

Изучите основы статистического анализа данных и сможете применять статистику для проверки продуктовых гипотез.

Проект

Проверка гипотез в бизнесе для сервиса проката самокатов GoFast в тренажёре

Темы

1. **Основы теории вероятностей**
   Эксперимент, исходы, вероятностное пространство и события. Пересекающиеся и взаимоисключающие события. Диаграмма Эйлера — Венна. Сложение и умножение вероятностей. Вероятность: классическое и геометрическое определения. Закон больших чисел.

2. **Случайные величины**
   Случайные величины и их виды. Вероятность попасть в интервал для дискретной случайной величины. Функция распределения дискретной случайной величины. Математическое ожидание дискретной случайной величины. Дисперсия дискретной случайной величины.

3. **Распределения**
   Эксперимент Бернулли. Биномиальный эксперимент. Биномиальное распределение. Распределение Пуассона. Нормальное распределение и плотность вероятности. Параметры нормального распределения. Функция нормального распределения. Percent Point Function для нормального распределения.

4. **Основы статистики**
   Генеральная совокупность и выборки. Точечные оценки. Выборочное распределение. Центральная предельная теорема. Стандартная ошибка среднего.

5. **Проверка гипотез**
   Гипотезы. Логика проверки гипотез. P-value. Проверка гипотез с одной выборкой в Python. Проверка гипотез с двумя выборками в Python. T-тест Стьюдента. Ошибки I и II рода. Проблема множественной проверки гипотез. Поправки на множественную проверку гипотез.

Инструменты и технологии

Scipy

Statsmodels

Нулевая гипотеза

Альтернативная гипотеза

Распределения метрик

Ошибка первого рода

Ошибка второго рода

Статистический тест

Множественная проверка гипотез

T-тест

Биномиальный тест

Спринт 11

2 недели

# Анализ результатов A/B-тестирования с помощью Python

Познакомитесь с дорожной картой A/B-тестирования и сможете самостоятельно анализировать результаты A/B-теста и формулировать выводы для бизнеса.

Инструменты и технологии

SQL

Python

Pandas

Scipy

Statsmodels

Jupyter Notebook

T-тест

Тест Манна — Уитни

Z-тест

Проект

Кейс-проект с ревью: разработка A/B-тестирования и анализ результатов для новой рекомендательной системы развлекательного приложения

Темы

1. **Что такое A/B-тесты и зачем они нужны**
   Знакомство с A/B-тестированием. Особенности проведения A/B-тестирования. Преимущества и ограничения A/B-тестирования. Процесс проведения A/B-тестирования.

2. **Выбор метрики для проверки гипотезы**
   Классификация метрик эксперимента. Выбор метрик.

3. **Расчёт размера выборки. Валидация результатов**
   Расчёт размера выборки и длительности эксперимента. Минимальный размер эффекта. Расчёт размера аудитории с помощью калькуляторов. Проверка данных. Расчёт метрик по результатам эксперимента.

4. **Проверка результатов A/B-теста. Тест Манна — Уитни. Z-тест пропорций**
   Алгоритм проверки результатов A/B-теста. T-тест и t-тест Уэлча. Тест Манна — Уитни. Z-тест пропорций.

1 неделя

# Итоговый проект модуля

Рассчитаете несколько метрик с помощью SQL и интерпретируете полученные результаты. Затем проверите гипотезу с помощью Python и сформулируете выводы о результатах эксперимента для бизнеса.

Проект

На основании данных сервиса Яндекс Книги рассчитаете и интерпретируете метрики о ежедневной активности пользователей в SQL, после чего проверите гипотезу в Python.
Проанализируете результаты A/B-тестирования новой версии сайта интернет-магазина BitMotion Kit

Каникулы ⟶ 1 неделя

2 недели

## Дипломный проект курса «Аналитик данных»

Потренируетесь использовать все полученные на курсе навыки на большом объёме реальных данных, используя SQL (PostgreSQL), Yandex DataLens, Python (pandas, scipy, statsmodels).

Проект

Изучите данные Яндекс Афиши и рассчитаете метрики с помощью SQL-запросов, подготовите дашборд в Yandex Datalens с описанием основных трендов. Проведёте исследовательский анализ данных и проверите гипотезы статистическими методами с помощью Python, сформулируете основные выводы и рекомендации

3–5 часов на кейс

## Плюс 10 кейсов от реальных работодателей

Мы попросили у наших партнёров реальные задачи, проработали решение и добавили в курс. Вы сможете дополнительно выбрать, сколько задач взять в работу, необязательно решать их все.
Благодаря этим задачам вы с первых дней обучения будете видеть, с чем сталкиваются аналитики на работе.

В начале курса будут более простые задачи, адаптированные под ваш текущий уровень, но с каждым модулем они будут становиться сложнее и сложнее.

# 04 Расширенная аналитика: дашборды, продукты и алгоритмы

Спринт 12

2 недели

## Углубленное изучение DataLens

Научитесь строить продвинутые визуализации в DataLens и углубите свои знания в SQL, оптимизируете запросы, чтобы улучшить производительность дашбордов.

**Проект**

Кейс-проект с ревью: оптимизация запросов для дашборда и создание дашборда с продвинутыми визуализациями в DataLens

**Темы**

1. **Markdown и улучшение визуализации**
   Основы Markdown. Применение Markdown в DataLens. Условное форматирование — индикаторы с функциями разметки (COLOR, SIZE, BR, BOLD), условное форматирование таблиц, градиенты. Сортировка (через агрегированные поля, CASE, оконные функции RANK), ссылки между дашбордами.
2. **Продвинутая визуализация в DataLens**
   Сложные визуализации и зачем они нужны. Точечная диаграмма, пузырьковая диаграмма, древовидная диаграмма. Геоаналитика. Фильтры. Иерархии.
3. **Оптимизация SQL-запросов**
   Оптимизация запросов в аналитике данных. План запроса. Партиционирование. Практика.
4. **Оптимизация на уровне дашборда**
   Регулярность обновления дашборда. Кэширование.

Инструменты и технологии

- Datalens
- SQL
- Визуализация данных
- Markdown
- Оптимизация запросов

Спринт 13

3 недели

## Как измерить продукт?

Погрузитесь в работу с метриками. Научитесь строить систему метрик продукта. Углубитесь в расчёты метрик юнит-экономики. Научитесь использовать наиболее популярные методики поиска точек роста для обнаружения потенциальных точек улучшения продукта.

**Проект**

Кейс-проект с ревью: анализ метрик продукта с целью обнаружения точек улучшения продукта

Инструменты и технологии

- Python
- Метрики
- Юнит-экономика

Темы

1. **Анализ работы продукта**
   Исследование работы продукта и его экономики. Анализ особенностей продукта глазами пользователя. Выявление ключевых точек взаимодействия. Монетизационные модели продуктов.

2. **Работа в команде в больших компаниях и корпорациях**
   Кросс-функциональная команда. Роли в кросс-функциональной команде. ETL-процесс. Сбор требований у заказчика.

3. **Система метрик в продукте**
   Основные метрики продукта (DAU, ARPU, удержание и другие). Система метрик продукта, связь между метриками и их взаимное влияние. Классификация метрик. Иерархия метрик, главные метрики продуктов в различных областях бизнеса. Кастомные метрики.

4. **Основы юнит-экономики**
   Базовые понятия юнит-экономики. Юнит-экономика продукта и юнит-экономика клиента. Расчет юнит-экономики в примерах.
   Анализ сходимости юнит-экономики для различных когорт с помощью Python. Разбор кейсов по анализу юнит-экономики. Методы самопроверки.

5. **Продуктовый аналитик и задача поиска точек роста**
   Методики поиска точек роста продукта. Роль аналитика в процессе поиска точек роста.

Спринт 14

2 недели

## Основы машинного обучения в работе аналитика данных

Познакомитесь с основными моделями машинного обучения и научитесь их использовать.

Инструменты и технологии

ML

Модель оттока

K-Means

Регрессионный анализ

Проект

Кейс-проект с ревью: построение модели машинного обучения в помощью Python

Темы

1. **Введение в машинное обучение**
   Что такое ML. Бизнес-задачи, которые решает ML. Основные типы обучения: с учителем (Supervised) и без учителя (Unsupervised). Роль аналитика в ML-проектах: формулировка задачи, подготовка данных, интерпретация результатов.

2. **Процесс обучения модели**
   Полный цикл ML-проекта: от бизнес-задачи до внедрения. Подготовка данных для ML: фичи, кодирование категориальных переменных, масштабирование. Разделение данных на обучающую и тестовую выборки. Переобучение (Overfitting) и методы борьбы с ним.

3. **Виды моделей обучения**
   Классификация моделей: логистическая регрессия, Decision Trees. Регрессия: линейная регрессия, Random Forest. Кластеризация: K-Means, иерархическая кластеризация. Критерии выбора типа модели в зависимости от задачи.

Темы

4. **Как можно сегментировать автоматически**
   Задачи кластеризации в бизнесе: сегментация клиентов, пользователей, товаров. Подготовка данных для кластеризации. Методы определения оптимального числа кластеров. Интерпретация результатов кластеризации.

5. **Алгоритм K-Means**
   Принцип работы K-Means. Визуализация результатов кластеризации. Анализ характеристик кластеров. Ограничения и особенности алгоритма.

6. **Регрессионный анализ**
   Линейная регрессия для прогнозирования. Интерпретация коэффициентов регрессии. Метрики качества регрессии. Анализ остатков и выбросов.

1 неделя

## Итоговый проект модуля

С помощью системы продуктовых метрик и когортного анализа на Python поможете продуктовой команде идентифицировать проблемные группы пользователей.

Проект

Кейс-проект с ревью: проведёте полный цикл анализа продукта: рассчитаете ключевые продуктовые метрики, создадите интерактивные дашборды для мониторинга, сегментируете пользователей с помощью ML

Каникулы

# 05 Работа с большими данными

Спринт 15

2 недели

## ClickHouse как аналитическая СУБД

Научитесь особенностям синтаксиса в СУБД ClickHouse, узнаете, как извлечь и обработать большие данные с помощью ClickHouse.

**Проект**

Кейс-проект с ревью: решение нескольких SQL-запросов с помощью ClickHouse

**Темы**

1. **Хранение и обработка больших данных**
   Как работать с большими данными. Хранение данных. Обработка данных. Автоматизация данных.
2. **Что такое ClickHouse и как он устроения в DataLens.**
   Особенности СУБД ClickHouse. Особенности синтаксиса ClickHouse. Движки таблиц (table engines).
3. **Типы данных в ClickHouse**
   Типы данных в ClickHouse. Что такое массив. Работа с массивами в ClickHouse.
4. **Комбинаторы и аналитические функции в ClickHouse**
   Комбинаторы в ClickHouse. Специфичные агрегирующие функции и комбинаторы. Аналитические функции описательной статистики в ClickHouse. Функции — статистические тесты. Оконные функции в ClickHouse.
5. **Работа с таблицами в ClickHouse**
   Создание, переименование и удаление таблиц. Действия со столбцами и строками.
6. **Масштабирование данных в СУБД**
   Для чего нужно масштабирование данных. Механизмы масштабирования данных.

Инструменты и технологии

Большие данные

ETL

DWH

ClickHouse

Массив

Комбинатор

Спринт 16

2 недели

## Обработка данных с помощью PySpark

Научитесь извлекать и обрабатывать большие данные с помощью PySpark.

**Проект**

Кейс-проект с ревью: извлечение и обработка сырых данных из хранилища больших данных с помощью PySpark

Инструменты и технологии

PySpark

S3

RDD

Датафрейм

Ленивые вычисления

Кэширование

Темы

1. **Хранилище данных S3 и PySpark**
   Что такое хранилище данных S3. Знакомство с файлами с помощью S3. Назначение PySpark. Основные форматы данных и компоненты PySpark. Архитектура распределённых вычислений Spark. Создание объектов SparkSession.

2. **Основные структуры данных в PySpark**
   RDD. Датафрейм (DataFrame). Схема датафрейма и типы данных.

3. **Трансформации датафреймов и действия**
   Трансформации датафреймов. Фильтрация и группировка данных в PySpark. Сортировка данных и создание новых столбцов. Удаление полей и удаление строк-дубликатов. Присоединение данных и операции множеств. Действия в PySpark.

4. **Кэширование и запись данных в другие источники**
   Кэширование. Запись датафреймов в ClickHouse.

5. **Пример обработки данных с помощью PySpark**
   Работа с PySpark в Jupyter Notebook.

Спринт 17

2 недели

## Автоматизация ETL-процессов с помощью Airflow

Научитесь автоматизировать процессы с помощью Airflow.

Проект

Кейс-проект с ревью: автоматизация процесса извлечения и переноса данных из S3 в ClickHouse

Темы

1. **Автоматизация ETL-процессов с помощью оркестраторов**
   ETL-процессы. Оркестраторы. Apache Airflow.

2. **Создание DAG в Airflow**
   Класс DAG и его объекты. Создание простого DAG. Расписание запуска DAG.  Зависимости между задачами. Запуск DAG и проверка.

3. **Операторы и сенсоры**
   Оператор PythonOperator(). Другие операторы. Сенсоры.

4. **Автоматизация запуска PySpark-приложения**
   Дата выполнения DAG. Выполнение в DAG PySpark-приложения. Уведомления при ошибках.

Инструменты и технологии

Airflow

DAG

Формат cron

Оркестраторы

Операторы

Сенсоры

2 недели

## Финальный проект

Поработаете с большими данными через PySpark и ClickHouse.

Проект

Проект: модель прогнозирования оттока клиентов с автоматизированным пайплайном обработки, обучения и визуализации. От сбора фич до развёртывания в Airflow и обновляемого дашборда в Yandex DataLens