Специалист по Data Science

13 месяцев

продолжительность курса

17 проектов

в портфолио

9 НЕДЕЛЬ | 90 ЧАСОВ

00

Вводный модуль

9 НЕДЕЛЬ | 90 ЧАСОВ

01

- Основы SQL. Извлечение данных
- SQL. Обработка данных
- SQL. Анализ данных и решение ad hoc задач
- Итоговый проект модуля 1

10 НЕДЕЛЬ | 100 ЧАСОВ

02

- Основы Python
- Python. Предобработка данных
- Исследовательский анализ данных и визуализация с помощью Python
- Инструменты разработки для Data Science
- Итоговый проект модуля 2

8 НЕДЕЛЬ | 80 ЧАСОВ

03

- Знакомство с машинным обучением. Линейная регрессия
- Логистическая регрессия
- Метод опорных векторов
- Итоговый проект модуля 3

10 НЕДЕЛЬ | 100 ЧАСОВ

04

- Решающее дерево
- Метод случайного леса
- Бустинг и градиентный бустинг
- Обучение без учителя
- Итоговый проект модуля 4

10 НЕДЕЛЬ |100 ЧАСОВ

05

- Внедрение и мониторинг моделей
- Формулировка гипотез
- Анализ результатов A/B тестирования с помощью Python
- Итоговый проект модуля 5
- Дополнительный спринт: подготовка к собеседованию

3 НЕДЕЛИ | 45 ЧАСОВ

06

Выпускной проект

4 НЕДЕЛИ

07

- Мастерская
- Карьерный трек

Специалист по Data Science [расширенный тариф]

17 месяцев

продолжительность курса

22 проекта

в портфолио

00

Вводный модуль

9 НЕДЕЛЬ | 90 ЧАСОВ

01

- Основы SQL. Извлечение данных
- SQL. Обработка данных
- SQL. Анализ данных и решение ad hoc задач
- Итоговый проект модуля 1

10 НЕДЕЛЬ | 100 ЧАСОВ

02

- Основы Python
- Python. Предобработка данных
- Исследовательский анализ данных и визуализация с помощью Python
- Инструменты разработки для Data Science
- Итоговый проект модуля 2

8 НЕДЕЛЬ | 80 ЧАСОВ

03

- Знакомство с машинным обучением. Линейная регрессия
- Логистическая регрессия
- Метод опорных векторов
- Итоговый проект модуля 3

10 НЕДЕЛЬ | 100 ЧАСОВ

04

- Решающее дерево
- Метод случайного леса
- Бустинг и градиентный бустинг
- Обучение без учителя
- Итоговый проект модуля 4

10 НЕДЕЛЬ | 100 ЧАСОВ

05

- Внедрение и мониторинг моделей
- Формулировка гипотез
- Анализ результатов A/B тестирования с помощью Python
- Итоговый проект модуля 5
- Дополнительный спринт: подготовка к собеседованию

11 НЕДЕЛЬ | 110 ЧАСОВ

06

- Обработка больших данных
- Трекинг экспериментов в ML flow
- Введение в глубинное обучение. Нейронные сети для изображений
- Нейронные сети для текстов
- Рекомендательные системы

4 НЕДЕЛИ

07

- Мастерская
- Карьерный трек

3 НЕДЕЛИ | 45 ЧАСОВ

08

Выпускной проект

Основы анализа данных с помощью SQL

01

9 недель | 90 часов
3 спринта
+ 1 итоговый проект

Познакомитесь с Data Science, как областью знаний и узнаете, какие задачи решает специалист по Data Science. 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. Научитесь применять продвинутые инструменты SQL (оконные функции) для решения ad hoc задач аналитика разной сложности. В конце модуля выполните итоговый проект по созданию витрины данных.

Содержание модуля

01.	Основы SQL. Извлечение данных	• Работа с базами данных. СУБД • Типы данных и их преобразования • Извлечение данных из таблиц и фильтрация • Группировка и сортировка данных	**Домашнее задание** Решение задач на извлечение данных с помощью SQL-запросов в тренажёре **Встреча с куратором** Подготовка к обучению и знакомство со студентами **Воркшоп** Как решать задачи на SQL	2 недели
02.	SQL. Обработка данных	• Работа с пропущенными значениями и дубликатами • Присоединение таблиц • Операции множеств и подзапросы • Категоризация значений. Создание новых столбцов • Работа с датой и временем	**Домашнее задание** Написание SQL-запросов в рамках задачи подготовки данных **Вебинар** Решение задач SQL в DBeaver	2 недели
03.	SQL. Анализ данных и решение ad hoc задач	• Оконные функции. Агрегирующие функции • Оконные функции. Ранжирующие функции и функции смещения • Исследовательский анализ данных. Аналитические функции • Решение ad-hoc задач	**Проект** Решение ad-hoc аналитических запросов различной сложности на SQL в рамках задачи исследовательского анализа данных **Воркшоп** Решение ad-hoc запросов в SQL	3 недели

Основы анализа данных с помощью SQL

01

9 недель | 90 часов
3 спринта
+ 1 итоговый проект

Познакомитесь с Data Science, как областью знаний и узнаете, какие задачи решает специалист по Data Science. Узнаете, как могут хранится данные и познакомитесь с языком запросов SQL для работы с базами данных. Напишете первые запросы на SQL и научитесь извлекать данные. Научитесь применять продвинутые инструменты SQL (оконные функции) для решения ad hoc задач аналитика разной сложности. В конце модуля выполните итоговый проект по созданию витрины данных.

Содержание модуля

Итоговый проект модуля	**Итоговый проект** Создание витрины данных с помощью SQL-запросов и решение нескольких аналитических ad-hoc задач на её основе **Сессия вопросов и ответов с наставником.**	2 недели
Каникулы		1 неделя

Анализ данных с помощью Python 02

10 недель | 100 часов
4 спринта
+ 1 итоговый проект

Начнёте знакомство с языком программирования Python: изучите основы синтаксиса, работу с библиотекой Pandas. Научитесь предобрабатывать данные и использовать Python для исследования и визуализации данных. Разберётесь с основами описательной статистики и основами объектно-ориентированного программирования. Создадите своё виртуальное окружение и Git-репозиторий.

Содержание модуля

05.	Основы Python	• Знакомство с Python • Строки. Списки • Логические выражения. Условный оператор • Циклы. Списковые включения • Вложенные конструкции • Функции • Множества. Словари	**Домашнее задание** Решение задач для отработки навыка работы с Python в тренажёре **Вебинар** Решение задач на Python	2 недели
06.	Python. Предобработка данных	• Основы библиотеки pandas. Обзор данных • Типы данных. Работа с датой и временем • Индексация в датафреймах. Фильтрация данных. • Работа с пропущенными значениями. • Обработка дубликатов • Категоризация данных	**Проект** Подготовка "сырых" данных для последующего анализа с помощью Python **Воркшоп** Предобработка данных в pandas	2 недели
07.	Исследовательский анализ данных и визуализация с помощью Python	• Объединение датафреймов. Срезы данных • Описательная статистика • Взаимосвязь переменных • Визуализация для изучения данных • Сводные таблицы • Пример исследовательского анализа данных	**Проект** Исследовательский анализ данных для решения бизнес-кейса и подготовка отчёта по исследованию **Воркшоп** Презентации результатов заказчику	2 недели
08.	Инструменты разработки для Data Science + ООП	• Исследовательский анализ данных, сборка витрины данных и решение ad hoc задач на основе собранной витрины	**Домашнее задание** Создание класса предобработки данных, используя принципы ООП. Размещение первого проекта в Git репозитории **Вебинар** Возможности использования IDE	2 недели

Анализ данных с помощью Python 02

10 недель | 100 часов
4 спринта
+ 1 итоговый проект

Начнёте знакомство с языком программирования Python: изучите основы синтаксиса, работу с библиотекой Pandas. Научитесь предобрабатывать данные и использовать Python для исследования и визуализации данных. Разберётесь с основами описательной статистики и основами объектно-ориентированного программирования. Создадите своё виртуальное окружение и Git-репозиторий.

Содержание модуля

Итоговый проект модуля	**Итоговый проект** Исследовательский анализ данных под запрос бизнеса с последующей визуализацией с помощью инструментов Python. Предоставите рекомендации бизнесу по итогам исследования **Сессия вопросов и ответов с наставником.**	2 недели
Каникулы		1 неделя

Основы машинного обучения. Линейные модели 03

8 недель | 80 часов
3 спринта
+ 1 итоговый проект

Начнете своё знакомство с машинным обучение. Разберетесь, что такое модель и как она обучается. Познакомитесь с линейными моделями: линейная регрессия, логистическая регрессия, метод опорных векторов. Начнете осваивать библиотеку scikit-learn. Научитесь рассчитывать метрики в задачах регрессии и классификации. Обучите минимум 6 моделей за этот модуль.

Содержание модуля

10.	Знакомство с МО. Первая модель - Линейная регрессия	• Что такое Машинное обучение • Критерии качества моделей • Данные в машинном обучении • Как обучается модель • Линейные модели. Линейная регрессия • Библиотека scikit-learn	**Проект** Обучение линейной регрессии и расчёт метрик этой модели на готовых данных **Воркшоп** Перевод требований бизнеса на язык задач машинного обучения	2 недели
11.	Логистическая регрессия	• Предобработка данных для обучения моделей • Кросс валидация • Задачи классификации. Линейная классификация • Логистическая регрессия • Метрики классификации	**Проект** Решение задачи классификации с подготовкой данных и экспериментами с гиперпараметрами **Вебинар** Предобработка данных для линейных моделей	2 недели
12.	Метод опорных векторов	• Метод опорных векторов, как задача классификации • Калибровка классификаторов • Многоклассовая классификация • Отбор признаков	**Проект** Построение нескольких линейных моделей и сравнение их с точки зрения предсказания вероятности с последующей провести калибровкой модели **Вебинар** Многоклассовая классификация	2 недели

Основы машинного обучения. Линейные модели 03

8 недель | 80 часов
3 спринта
+ 1 итоговый проект

Начнете своё знакомство с машинным обучение. Разберетесь, что такое модель и как она обучается. Познакомитесь с линейными моделями: линейная регрессия, логистическая регрессия, метод опорных векторов. Начнете осваивать библиотеку scikit-learn. Научитесь рассчитывать метрики в задачах регрессии и классификации. Обучите минимум 6 моделей за этот модуль.

Содержание модуля

Итоговый проект модуля	**Итоговый проект** Для одной бизнес задачи обучите модель регрессии и классификации. Сравните и интерпретируете метрики, выбрать лучшее решение. Подготовите отчёт по итогам экспериментов с моделями. **Сессия вопросов и ответов с наставником.**	2 недели
Каникулы		1 неделя

Модели на основе деревьев 04

10 недель | 100 часов
4 спринта
+ 1 итоговый проект

Продолжите осваивать классические модели машинного обучения: познакомитесь с метрическими алгоритмами (KNN) и решающими деревьями, разберётесь в подходах ансамблирования (бэггинг, стекинг), изучите метод случайного леса (Random Forest). Освоите алгоритмы бустинга и градиентного бустинга. Научитесь использовать популярные библиотеки (LightGBM, XGBoost, CatBoost).
Обучите минимум 8 моделей.

Содержание модуля

14.	Решающее дерево	• Нелинейный модели. KNN • Нелинейные модели. Решающие деревья • Как строить решающее дерево • Проблема обобщающей способности деревьев • Предобработка данных для решающих деревьев • Гиперпараметры решающих деревьев. Optuna	**Проект** Решение задачи регрессии с помощью knn и решающего дерева с обработкой через кастомный пайплайн. **Воркшоп** Как писать кастомный класс	2 недели
15.	Метод случайного леса	• Ансамблирование • Бэггинг • Random forest • Стэкинг • Валидация данных с временной структурой • Определение важности признаков • Дисбаланс классов	**Проект** Обучение одного решающего дерева (Decision Tree) и ансамбля деревьев (Random Forest) на данных с дисбалансом классов **Вебинар** Работа с дисбалансом классов: методы работы и валидация при дисбалансе классов	2 недели
16.	Бустинг и градиентный бустинг	• Бустинги • Градиентный бустинг • Реализация градиентого бустинга в разных библиотеках • Гиперпараметры в бустингах • Интерпретация модели градиентного бустинга • Векторизация текстовых данных	**Проект** Решение задачи классификации, используя градиентный бустинг **Вебинар** Мастер-класс по библиотеке CatBoost	2 недели

Модели на основе деревьев 04

10 недель | 100 часов
4 спринта
+ 1 итоговый проект

Продолжите осваивать классические модели машинного обучения: познакомитесь с метрическими алгоритмами (KNN) и решающими деревьями, разберётесь в подходах ансамблирования (бэггинг, стекинг), изучите метод случайного леса (Random Forest). Освоите алгоритмы бустинга и градиентного бустинга. Научитесь использовать популярные библиотеки (LightGBM, XGBoost, CatBoost).
Обучите минимум 8 моделей.

Содержание модуля

17.	Обучение без учителя	• Задачи обучения без учителя • Задача снижения размерности. pca • Задача снижения размерности. tsne • Задача кластеризации. k means • Задача кластеризации. dbscan	**Проект** Решение задачи кластеризации и визуализация с помощью tsne **Воркшоп** Сбор требований бизнеса	2 недели
	Итоговый проект модуля		**Итоговый проект** Реализация решения на основе деревьев, задача усложнится требованием достичь целевых метрики модели **Сессия вопросов и ответов с наставником.**	2 недели
	Каникулы			1 неделя

Валидация, тестирование моделей 05

10 недель | 100 часов
4 спринта
+ 1 итоговый проект

Освоите полный цикл работы с моделями в реальном бизнесе — от внедрения и мониторинга до оценки гипотез и анализа результатов A/B-тестов. Научитесь применять статистику на практике, уверенно работать с экспериментами и принимать решения на основе данных. Разберете основные вопросы, с которыми столкнетесь на собеседованиях, подготовитесь к техническим собеседованиям.

Содержание модуля

19.	Внедрение и мониторинг моделей	• Внедрение, как этап жизненного цикла моделей • Мониторинг, как этап жизненного цикла моделей • Знакомство с Airflow • Батч инференс в Airflow • Мониторинг в Airflow	**Домашнее задание** Реализация мониторинга стабильности работы модели через AirFlow **Воркшоп** Мастер-класс по AirFlow	2 недели
20.	Формулировка гипотез	• Основы теории вероятностей • Случайные величины. Распределения • Проверка гипотез. Т-тест Стьюдента • Множественная проверка гипотез	**Домашнее задание** Проверка статистической значимости данных	2 недели
21.	Анализируем результаты A/B тестирования с помощью Python	• Что такое A/B-тесты и зачем они нужны • Выбор метрики для проверки гипотезы • Расчёт размера выборки. Валидация результатов • Проверка результатов A/B-теста. Тест Манна-Уитни. Z-тест пропорций • Анализ результатов A/B-теста: примеры	**Проект** Анализ результатов проведенного A/B-теста моделей с помощью стат. тестов **Воркшоп** Разбор кейсов Как анализировать AB-тесты	2 недели
	Итоговый проект модуля		**Итоговый проект** Доработка готового кода предобработки признаков и реализация инференса модели через Airflow **Сессия вопросов и ответов по итоговому проекту и модулю**	2 недели

Валидация, тестирование моделей 05

10 недель | 100 часов
4 спринта
+ 1 итоговый проект

Освоите полный цикл работы с моделями в реальном бизнесе — от внедрения и мониторинга до оценки гипотез и анализа результатов A/B-тестов. Научитесь применять статистику на практике, уверенно работать с экспериментами и принимать решения на основе данных. Разберете основные вопросы, с которыми столкнетесь на собеседованиях, подготовитесь к техническим собеседованиям.

Содержание модуля

23.	Подготовка к собеседованиям	• Алгоритмические задачи на Python • SQL-задачи • Статистика и ab-тесты • Вопрос по теории ML	**Домашнее задание** Решение SQL-задач и алгоритмических задач на Python с автопроверкой **Воркшоп** Советы по прохождению технических собеседований	2 недели
	Каникулы			1 неделя

Выпускной проект

3 недели

Дипломный проект охватывает весь жизненный цикл модели машинного обучения, начиная от сбора данных и заканчивая развертыванием инференса в Airflow.

Содержание модуля

Выпускной проект	• Выбор модели, исходя из бизнес-требований и ограничений • Работа с сырыми SQL-данными из нескольких таблиц • Разработка пайплайна предобработки, экспериментов и инференса. • Загрузка кода и модели в GitLab для версионирования и автоматизации	**Сессия вопросов и ответов по выпускному проекту**	3 недели

Специалист по Data Science расширенный состоит из базового курса + дополнительных тем

Продвинутые инструменты Data Science [расширенный тариф]

06

11 недель | 110 часов
5 спринтов

Погрузитесь в продвинутые темы: работа с большими данными, логирование экспериментов с параметрами модели, архитектура нейросетей. Также вы изучите подходы к построению рекомендательных систем и научитесь разрабатывать решения, готовые к запуску в продакшн.

Содержание модуля

24.	Обработка больших данных	• Введение в большие данные • Архитектура pyspark • Хранение больших данных • Предобработка с Spark Structured API • Работа с SQL в Pyspark • Оптимизация обработки данных	**Проект** Выполнение обработки больших данных в pyspark. Подготовка данных для обучения модели. **Вебинар** Как устроены распределенный системных хранения и обработки больших данных	2 недели
25.	Трекинг экспериментов в ML flow	• Знакомство с ML flow • Логирование параметров и метрик • Логирование модели • Сравнение экспериментов • Создание проектов	**Домашнее задание** Выполнение несколько практических заданий в MLflow с автопроверкой: логирование параметров и метрик, логирование модели, сравнение экспериментов, версионирование. **Вебинар** Пример деплоя в ml flow	2 недели
26.	Введение в глубинное обучение. Нейронные сети для изображений	• Задачи глубинного обучения • Архитектура нейронной сети • Полносвязные нейронные сети • Сверточные нейронные сети • Фреймворк Pytorch	**Проект** Решение задачи многоклассовой классификации двумя подходами: через полносвязные и через сверточные нейронные сети. Сравнение качества моделей и выбор лучшего решения **Вебинар** Разбор архитектуры сверточной сети	2 недели

Продвинутые инструменты Data Science [расширенный тариф]

06

11 недель | 110 часов
5 спринтов

Погрузитесь в продвинутые темы: работа с большими данными, логирование экспериментов с параметрами модели, архитектура нейросетей. Также вы изучите подходы к построению рекомендательных систем и научитесь разрабатывать решения, готовые к запуску в продакшн.

Содержание модуля

27.	Нейронные сети для текстов	• Векторное представление слов (Word2vec) • Рекуррентные нейронные сети • Архитектура трасформеров • Модели на архитектуре трасформеров: bert, gpt • Использование предобученных моделей с hugging face	**Итоговое задание спринта** Решение задачи классификации с текстовыми данными с использованием предобученных моделей **Воркшоп** Использование предобученных моделей: разбор кейсов	2 недели
28.	Рекомендательные системы	• Задача рекомендательных систем • Метрики рекомендаций • Контентные рек.системы • Коллаборативная фильтрация • Гибридные рек.системы • Инференс рек. систем	**Проект** Создание рекомендательной системы через коллаборативную и гибридную модель. Сравнение качества моделей и выбор лучшего решения **Воркшоп** Обучению рекомендательных систем с использованием ligth fm	3 недели
	Мастерская		Выполнение обязательного проекта в Мастерской	4 неделя
	Каникулы			1 неделя

Выпускной проект расширенного тарифа

3 недели

Пройдете весь жизненный цикл DS-проекта и столкнётесь с новыми техническими задачами: обработкой больших данных на PySpark и управлением экспериментами через MLflow.
У вас будет выбор из двух датасетов, на которых выполнять проект.

Содержание модуля

Выпускной проект	• Сбор данных (данные в SQL базе или pyspark) • Предобработка • EDA для рекомендаций • Генерация признаков из сырых данных • Эксперименты с моделями + логирование экспериментов в ML flow • Расчёт метрик • Выбор лучшего решения • Валидациях • Подготовка артефактов • Подготовка инференса модели • Оформление проекта в Git lab	**Сессия вопросов и ответов по выпускному проекту**	3 недели

Медведев Кирилл
Директор АНО ДПО «Образовательные технологии Яндекса»