Общество с ограниченной ответственностью «1Т»

ДОКУМЕНТ ПОДПИСАН УСИЛЕННОЙ КВАЛИФИЦИРОВАННОЙ ЭЛЕКТРОННОЙ ПОДПИСЬЮ

ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "1Т"

СВЕДЕНИЯ О СЕРТИФИКАТЕ ЭП

Сертификат: 02652A0E01D2B0339841DC93BA5D5602F8

Владелец: Кармаза Валентина Васильевна

Действителен: с 08.12.2023 по 08.03.2025

УТВЕРЖДАЮ

Генеральный директор ООО «1Т»

_____ (В.В. Кармаза)

« 08 » апреля 2024 г.

ДОПОЛНИТЕЛЬНАЯ ПРОФЕССИОНАЛЬНАЯ ПРОГРАММА ПОВЫШЕНИЯ КВАЛИФИКАЦИИ

«Аналитик данных»

Специальность: Аналитик данных (Data Scientist)

Целевое назначение: Разработка и использование технологий искусственного интеллекта

Срок обучения: 260 академических часов.

Москва 2024

ОГЛАВЛЕНИЕ

ОБЩИЕ ДАННЫЕ
О ДОПОЛНИТЕЛЬНОЙ ПРОФЕССИОНАЛЬНОЙ ПРОГРАММЕ
ПОВЫШЕНИЯ КВАЛИФИКАЦИИ «АНАЛИТИК ДАННЫХ»

№	Название	Описание
1.1	Название программы	Аналитик данных
1.2	Цель обучения	Получение слушателями компетенций, необходимых для профессиональной деятельности аналитика данных для разработки и применения технологических решений в области искусственного интеллекта и в смежных областях
1.3	Специальность	Аналитик данных (Data Scientist)
1.4	Форма обучения	Очно-заочная форма обучения, осуществляемая с применением электронного обучения и дистанционных образовательных технологий (онлайн-вебинары и т. п.) без отрыва от производства
1.5.	Количество академических часов	260
1.6.	Количество слушателей, которое может быть обеспечено обучением Провайдер по Образовательной программе по одному потоку в срок до 25 ноября	210
1.7.	Стоимость обучения	80 000

Дополнительная профессиональная программа (программа повышения квалификации) «Аналитик данных» (далее – Программа) разработана:

– в соответствии с нормами Федерального закона РФ от 29 декабря 2012 года № 273-ФЗ «Об образовании в Российской Федерации»;

– с учетом постановления Правительства Российской Федерации от 13 мая 2021 г. № 729 «О мерах по реализации программы стратегического лидерства «Приоритет-2030» (в редакции постановления Правительства Российской Федерации от 14 марта 2022 г. № 357 «О внесении изменений в постановление Правительства Российской Федерации от 13 мая 2021 г. № 729»);

– с учетом требований приказа Минобрнауки России от 1 июля 2013 г. № 499 «Об утверждении Порядка организации и осуществления образовательной деятельности по

дополнительным профессиональным программам», с изменениями, внесенными приказом Минобрнауки России от 15 ноября 2013 г. № 1244 «О внесении изменений в Порядок организации и осуществления образовательной деятельности по дополнительным профессиональным программам, утвержденный приказом Министерства образования и науки Российской Федерации от 1 июля 2013 г. № 499»;

– с учетом приказа Министерства образования и науки РФ от 23 августа 2017 г. № 816 «Об утверждении Порядка применения организациями, осуществляющими образовательную деятельность, электронного обучения, дистанционных образовательных технологий при реализации образовательных программ»;

– с учетом паспорта федерального проекта «Искусственный интеллект» национальной программы «Цифровая экономика Российской Федерации»;

– с учетом Методических рекомендаций-разъяснений Минобрнауки России по разработке дополнительных профессиональных программ на основе профессиональных стандартов от 22 апреля 2015 года № ВК-1030/06;

– с учетом приказа Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации от 28 февраля 2022 г. № 143 «Об утверждении методик расчета показателей федеральных проектов национальной программы «Цифровая экономика Российской Федерации» и признании утратившими силу некоторых приказов Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации об утверждении методик расчета показателей федеральных проектов национальной программы «Цифровая экономика Российской Федерации»;

– с учетом федеральных государственных образовательных стандартов (далее вместе – ФГОС ВО) по направлению подготовки 38.03.05 Бизнес-информатика (приказ Минобрнауки России от 29.07.2020), 09.04.04. Программная инженерия (Приказ Минобрнауки России от 19.09.2017 N 932 (ред. от 08.02.2021);

– на основе анализа требований рынка труда в сфере искусственного интеллекта и анализа данных.

В результате освоения данной программы выпускник программы «Аналитик данных» должен:

Знать:

− основные определения искусственного интеллекта и больших данных;

− различия между машинным обучением, нейронными сетями, глубоким обучением и EDA;

− основные конструкции языка Python;

− основы Hadoop;

- основы теории баз данных;
- принципы работы NoSQL баз данных;
- язык запросов к СУБД - SQL;
- основные уровни представления данных;
- основные ETL процессы и инструменты;
- особенности организации СУБД в MPP-системе;
- основные типы данных в СУБД Postgres;
- принципы построения дашбордов;
- основные понятия теории вероятности;
- основы комбинаторики;
- понятие A/B-тестирования;
- особенности продуктовой аналитики;
- существующие и перспективные методы и программный инструментарий технологий больших данных.

Уметь:

- производить аналитику для интеллектуального отслеживания ресурсов/процессов;
- применять SQL базы данных для прикладных решений;
- применять язык программирования Python и библиотеки при разработке решений на основе ИИ;
- осуществлять поиск и структурирование данных;
- визуализировать анализируемые данные;
- применять методы анализа на графах;
- создавать собственные модели данных с использованием UML-диаграмм;
- производить расчет вероятностных показателей с использованием языка Python;
- проводить аналитические и поисковые исследования по тематике информационных технологий, технологий больших данных;
- осуществлять математическое и информационное моделирование;
- разрабатывать технические проекты в сфере информационных технологий;
- осуществлять массово-параллельную обработку и анализ данных;
- оценивать результаты моделирования и определять критерии качества построенных моделей.

Владеть:

- навыками решения базовых аналитических кейсов с использованием инструментов визуализации;
- навыками использования статистических методов исследования;
- навыками расчета статистических показателей с использованием языка Python;
- методами разработки моделей машинного обучения и нейронных сетей;
- математическими методами анализа данных;
- навыками создания нескольких таблиц в СУБД Postgres посредствам Dbeaver;
- навыками интеллектуального анализа данных с помощью языка программирования PYTHON, R;
- навыками построения полносвязной нейронной сети для задачи классификации;
- навыками обучения нейронных сетей с помощью PyTorch, TensorFlow и Keras;
- навыками расчета ключевых метрик роста продукта с помощью Python;
- навыками настраивания кластеров Apache Spark и Hive на Hadoop;
- инструментами Weka,RapidMiner, Knime, Orange IBM SPSS Modeler, Tableau и др.;
- навыками использования баз данных (MongoDB, Clickhouse и др.).

Разработчики программы:

Борисов Вадим Владимирович, профессор кафедры вычислительной техники, филиал НИУ «МЭИ» в г. Смоленске, д.т.н., профессор,

*Хусаинов Наиль Шавкятович, з*аведующий кафедрой Института компьютерных технологий и информационной безопасности, ФГАОУ ВО «Южный федеральный университет», к.т.н.,

Санников Даниил Александрович, главный аналитик данных ПАО «Сбербанк»,

Кропивный Дмитрий Алексеевич, ведущий аналитик данных ООО «1Т»,

Жукова Людмила Вячеславовна, доцент кафедры «Магистерская школа информационных бизнес-систем», НИТУ МИСИС, к.э.н.,

Клавдеев Александр Владимирович, старший аналитик данных, ООО «1Т»,

Шарапов Никита Александрович, аналитик-исследователь ООО «1Т»,

Кулакова Надежда Сергеевна, старший аналитик данных ООО «1Т»,

Зиновьев Дмитрий Владимирович, системный аналитик ООО «1Т»,

Лашков Дмитрий Юрьевич, старший аналитик данных ООО «1Т»,

Костин Алексей Николаевич, ведущий преподаватель по ИИ, ООО «1Т»,

Королева Диана Олеговна, заведующая лабораторией инноваций в образовании, НИУ ВШЭ.

1. ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ.

1.1. Актуальность.

Один из актуальных трендов развития экономики и общества на сегодняшний день – это цифровизация. Актуальность цифровой экономики обусловлена тем, что произошли качественные изменения в этих сферах.

С 2017 года началось активное стимулирование процессов цифровизации экономики, осуществляемое со стороны государства. Об этом свидетельствуют многочисленные принятые в данном периоде нормативно-правовые акты в данной сфере. Прежде всего, была создана Автономная некоммерческая организация «Цифровая экономика», а также принят Указ Президента РФ «О стратегии развития информационного общества в РФ на 2017–2030 годы». В 2018 году был разработан нацпроект «Цифровая экономика РФ», а в 2019 году принят план информатизации Министерства цифрового развития, связи и массовых коммуникаций РФ, что стало отправной точкой в области цифровизации экономики России. В дальнейшем, в 2020 году, цифровизация экономики была простимулирована пандемией коронавируса, которая сопровождалась вынужденным режимом самоизоляции граждан, что повлекло массовый переход на дистанционные формы взаимодействия, требующие цифровых технологий и соответствующей инфраструктуры. В этом же году цифровизация была объявлена национальной целью развития России.

Указом Президента Российской Федерации от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» утверждена Национальная стратегия развития искусственного интеллекта на период до 2030 года.

Стратегия является основным программным документом, направленным на развитие и внедрение отечественных решений, формирующих внедрение инноваций во все сферы экономической деятельности и повседневной жизни граждан.

В развитие Национальной стратегии утвержден федеральный проект «Искусственный интеллект» (ИИ) сроком реализации до конца 2024 года.

Федеральный проект «Искусственный интеллект» предусматривает повышение уровня обеспечения российского рынка технологий ИИ квалифицированными кадрами и уровня информированности населения о возможных сферах использования ИИ.

Новой мерой в 2022 году стал запуск дополнительного профессионального образования граждан в области ИИ, в котором одной из актуальных специальностей является аналитик данных.

Аналитики данных востребованы, например, в сфере экономики и финансов. Они исследуют рынок с помощью инструментов анализа данных и строят свой прогноз на основе

результатов анализа. Кроме того, аналитик оценивает бизнес и приводит технические требования в соответствие с бизнес-проектами и целями.

Образовательная программа в рамках специальности «Аналитик данных», направленная на обеспечение получения гражданами дополнительного профессионального образования в области искусственного интеллекта и в смежных областях с использованием механизма персональных цифровых сертификатов, обеспечит приобретение следующих профессиональных компетенций:

- способность классифицировать и идентифицировать задачи искусственного интеллекта, выбирать адекватные методы и инструментальные средства решения задач искусственного интеллекта;
- способность разрабатывать и применять методы машинного обучения для решения задач;
- способность использовать инструментальные средства для решения задач машинного обучения;
- способность осуществлять сбор и подготовку данных для систем искусственного интеллекта;
- способность выполнять анализ больших данных;
- способность использовать одну или несколько сквозных цифровых субтехнологий искусственного интеллекта.

1.2. Категория слушателей. Требования к уровню подготовки слушателя.

К обучению на программе допускаются: предприниматели, работники и владельцы компаний IT-сектора, имеющие высшее образование или среднее профессиональное (либо получающие высшее или среднее профессиональное образование), а также мотивированные специалисты из других профессиональных сфер и студенты, обучающиеся в области информационных технологий, а также по иным специальностям, которые заинтересованы в получении новых компетенций по специальности «Аналитик данных».

Наличие опыта профессиональной деятельности: без опыта.

Требования к уровню подготовленности, определяемому контрольно-измерительными материалами.

Слушатели должны обладать следующими знаниями, умениями и владеть навыками:

PYTHON:

Знание синтаксиса языка.

Понимание базовых структур данных.

Владение основами ООП (класс, объект).

Владение базовыми знаниями в алгоритмах (“O” большое и т.д.).

SQL:

Знание базового синтаксиса SQL.

Умение составлять подзапросы, временные таблицы.

Навык работы с оконными функциями.

ИНФРАСТРУКТУРА:

Навыки работы с Docker.

Знание базовых команд Linux.

Навыки работы с системами контроля версий Git.

АНАЛИТИКА:

Понимание основ математической статистики.

Понимание основ теории вероятностей.

Базовые знания в выборе необходимого метода визуализации данных.

Требования к компетенциям, которыми должен обладать гражданин при поступлении на Образовательную программу:

Способен использовать и адаптировать существующие математические методы и системы программирования для разработки и реализации алгоритмов решения прикладных задач;

Способен применять и модифицировать математические модели для решения задач в области профессиональной деятельности;

Владеет широкой общей подготовкой (базовыми знаниями) для решения практических задач в области информационных систем и технологий;

Способен использовать современные компьютерные технологии поиска информации для решения поставленной задачи, критического анализа этой информации и обоснования принятых идей и подходов к решению;

Способен выбирать и оценивать способ реализации информационных систем и устройств (программно-, аппаратно- или программно-аппаратно-) для решения поставленной задачи.

1.3 Область профессиональной деятельности.

06 Связь и информационно-коммуникационные технологии (в сферах: анализа, моделирования и формирования интегрального представления стратегий и целей, бизнес-процессов и информационно-технологической инфраструктуры предприятий различной отраслевой принадлежности и различных форм собственности, а также учреждений государственного и муниципального управления; стратегического планирования и

управления развитием информационных систем и информационно-коммуникационных технологий управления предприятием; аналитической поддержки процессов принятия решений для управления предприятием).

1.4. Планируемые результаты обучения.

Программа повышения квалификации разработана с учетом профессиональных стандартов:

«Специалист по большим данным», утвержденный приказом Министерства труда и социальной защиты РФ от 6 июля 2020 года № 405н;

«Программист», утвержденный приказом Министерства труда и социальной защиты РФ от 20 июля 2022 № 424н;

«Системный аналитик», утвержденный приказом Министерства труда и социальной защиты РФ от 27.04.2023 № 367н.

Программа повышения квалификации разработана с учетом:

ФГОС 38.03.05 Бизнес-информатика, ФГОС 09.04.04. Программная инженерия.

По данной программе приобретаются компетенции универсальной модели компетенций в сфере искусственного интеллекта, разработанной РЭУ им. Г.В. Плеханова в 2021 году в рамках результата Федерального проекта «Искусственный интеллект».

Совершенствуемые и/или формируемые компетенции	Тип компетенции	Планируемые результаты обучения (знать, уметь, владеть - использовать конкретные инструменты)
Способен классифицировать и идентифицировать задачи искусственного интеллекта, выбирать адекватные методы и инструментальные средства решения задач искусственного интеллекта	профессиональная	Знать: − основные определения искусственного интеллекта и больших данных; − понятие A/B-тестирования; − особенности продуктовой аналитики. Уметь: − проводить аналитические и поисковые исследования по тематике информационных технологий, технологий больших данных; − строить несколько моделей и выбирать лучшую модель на данных. Владеть:

		– методами и инструментальными средствами решения задач искусственного интеллекта; – навыками расчета ключевых метрик роста продукта с помощью Python
Способен разрабатывать и применять методы машинного обучения для решения задач	профессиональная	Знать: – существующие и перспективные методы и программный инструментарий технологий больших данных; – математические основы машинного обучения (линейная алгебра, статистика, оптимизация). Уметь: – применять язык программирования Python и библиотеки при разработке решений на основе ИИ; – осуществлять массово параллельную обработку и анализ данных; – строить модели машинного обучения (регрессия, классификация, кластеризация, нейросети); – оценивать результаты моделирования и определять критерии качества построенных моделей. Владеть: – методами разработки моделей машинного обучения и нейронных сетей
Способен использовать инструментальные средства для решения задач машинного обучения	профессиональная	Знать: – основные конструкции языка Python, библиотеки; – системы обработки и анализа больших массивов данных (Hadoop, ETL, Spark и др.); – платформы и базы данных Уметь: – осуществлять парсинг интернет-данных; – применять SQL базы данных для прикладных решений; – производить расчет вероятностных показателей с

		использованием языка Python; − разрабатывать модели машинного обучения для решения задач. Владеть: − навыками расчета статистических показателей с использованием языка Python; − навыками создания нескольких таблиц в СУБД Postgres посредствам Dbeaver; − навыками интеллектуального анализа данных с помощью языка программирования R; − навыками обучения нейронных сетей с помощью PyTorch, TensorFlow и Keras; − навыками расчета ключевых метрик роста продукта с помощью Python; − навыками настраивания кластеров Apache Spark и Hive на Hadoop; − владение инструментами инструменты Weka,RapidMiner, Knime, Orange IBM SPSS Modeler, Tableau и др.; − использовать базы данных (MongoDB, Clickhouse и др.
Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта	профессиональная	Знать: − различия между машинным обучением, нейронными сетями, глубоким обучением и EDA; − язык запросов к СУБД; − основные ETL процессы и инструменты. Уметь: − осуществлять поиск и структурирование данных; − осуществлять подготовку и разметку структурированных и неструктурированных данных для машинного обучения. Владеть: − навыками решения базовых аналитических кейсов с использованием инструментов визуализации; − навыками поиска аномалий

		в данных, сегментации PCA, уменьшения размерности данных
Способен выполнять анализ больших данных	профессиональная	Знать: − основы теории баз данных; − принципы работы NoSQL баз данных; − основные уровни представления данных; − особенности организации СУБД в MPP-системе; − основные типы данных в СУБД Postgres; − особенности колоночного формата хранения данных; − принципы построения дашбордов; − основные понятия теории вероятности; − основы комбинаторики. Уметь: − производить аналитику для интеллектуального отслеживания ресурсов/процессов; − визуализировать анализируемые данные; − применять методы анализа на графах; − создавать собственные модели данных с использованием UML-диаграмм. Владеть: − навыками использования статистических методов исследования; − математическими методами анализа данных; − навыками интеллектуального анализа данных с помощью языка программирования R
Способен использовать одну или несколько сквозных цифровых субтехнологий искусственного интеллекта	профессиональная	Знать: − представление о сквозных цифровых субтехнологиях искусственного интеллекта; Уметь: − осуществлять математическое и информационное моделирование; − решать прикладные задачи и участвовать в реализации

		проектов в области сквозной цифровой субтехнологии «Компьютерное зрение». Владеть: − навыками использования предварительно обученных моделей для классификации изображений и других задач; − навыками обучения нейронной сети Keras многоклассовой классификации изображений на малом количестве данных

Критерии для оценки уровня сформированности указанных компетенций и соответствия индикаторам достижения компетенций.

Компетенция	Критерии для оценки уровня сформированности компетенций
Способен классифицировать и идентифицировать задачи искусственного интеллекта, выбирать адекватные методы и инструментальные средства решения задач искусственного интеллекта	Классифицирует и идентифицирует задачи систем искусственного интеллекта в зависимости от особенностей проблемной и предметной областей; Собирает исходную информацию и формирует требования к решению задач с использованием методов искусственного интеллекта
Способен разрабатывать и применять методы машинного обучения для решения задач	Проводит анализ требований и определяет необходимые классы задач машинного обучения; Определяет метрики оценки результатов моделирования и критерии качества построенных моделей; Принимает участие в оценке и выборе используемых методов машинного обучения
Способен использовать инструментальные средства для решения задач машинного обучения	Осуществляет оценку и выбор инструментальных средств для решения поставленной задачи; Разрабатывает модели машинного обучения для решения задач
Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта	Выполняет подготовку и разметку структурированных и неструктурированных данных для машинного обучения
Способен выполнять анализ больших данных	Использует знания о вариантах использования больших данных, определениях, словарях и эталонной архитектуре больших данных для

	эффективного извлечения, хранения, подготовки больших данных
Способен использовать одну или несколько сквозных цифровых субтехнологий искусственного интеллекта	Решает прикладные задачи и участвует в реализации проектов в области сквозной цифровой субтехнологии «Компьютерное зрение»

2. УЧЕБНЫЙ (ТЕМАТИЧЕСКИЙ) ПЛАН.

№ п/п	Наименование модулей/ тем программы	Всего, час	Виды учебных занятий			Формы контроля
			лекции	практические занятия	самостоятельная работа	
1	Модуль 1. Базовый					
2	Раздел 1.1 Введение в анализ данных					
3	Тема 1. Введение в анализ данных. Профессия Аналитик данных	6	2	2	2	Тест + Практическая работа
4	Тема 2. Определение искусственного интеллекта и больших данных. Применение искусственного интеллекта в различных областях	6	2	2	2	Тест + Практическая работа
5	Раздел 1.2 Инструменты, библиотеки и технологии анализа данных					
6	Тема 1. Знакомство с синтаксисом языка Python и средой разработки Jupyter Notebook. Обзор языка R	10	2	6	2	Тест + практическая работа
7	Тема 2. Работа с библиотеками Python	10	2	6	2	Практическая работа
8	Тема 3. Работа с внешними API и протоколом http. Парсинг Интернет-данных	8	2	4	2	Тест + Практическая работа
9	Тема 4. Системы обработки и анализа больших массивов данных (Hadoop, ETL, Spark и др.)	10	4	4	2	Тест + Практическая работа
10	Раздел 1.3 Технологии хранения и обработки больших данных					
11	Тема 1. Виды представления данных: табличные, графовые, временные ряды. Качество данных, подходы и инструменты	8	2	4	2	Тест + Практическая работа
12	Тема 2. Платформы данных	4	2	0	2	Тест + Практическая работа
13	Тема 3. Введение в теорию БД. Основы языка SQL	8	2	4	2	Тест + практичес

						кая работа
14	Тема 4. SQL базы данных	8	2	4	2	Тест + Практическая работа
15	Тема 5. Базы данных NoSQL	8	2	4	2	Тест + Практическая работа
16	Тема 6. Массово-параллельная обработка и анализ данных	6	2	2	2	Тест + Практическая работа
17	Тема 7. Знакомство с СУБД Postgres. Обзор GreenPlum	8	2	4	2	Тест + Практическая работа
18	Тема 8. Моделирование данных	8	2	4	2	Тест + Практическая работа
19	Тема 9. Построение дашбордов с помощью Superset	8	2	4	2	Практическая работа
20	Промежуточная аттестация	4		4		Решение кейса
21	Модуль 2. Профильный					
22	Раздел 2.1. Математическое моделирование, машинное обучение и оптимизация					
23	Тема 1. Введение в теорию вероятностей. Базовые понятия	8	2	4	2	Тест + практическая работа
24	Тема 2. Введение в математическую статистику. Статистические методы анализа данных	8	2	4	2	Тест + практическая работа
25	Тема 3. Основные понятия и термины в машинном обучении и нейронных сетях	8	2	4	2	Тест + практическая работа
26	Тема 4. Математические основы машинного обучения (линейная алгебра, статистика, оптимизация)	8	2	4	2	Тест + Практическая работа
27	Тема 5. Основные метрики оценки качества моделей машинного обучения	8	2	4	2	Тест + Практическая работа
28	Тема 6. Выбор и обработка данных для машинного обучения. Методы машинного обучения	8	2	4	2	Тест + Практическая работа

29	Тема 7. Построение моделей машинного обучения (регрессия, классификация, кластеризация, нейросети)	10	2	6	2	Тест + Практическая работа
30	Тема 8. Инструменты анализа данных и Machine Learning (Rapid Miner)	8	2	4	2	Тест + Практическая работа
31	Раздел 2.2. Глубокое обучение и нейронные сети					
32	Тема 1. Введение в глубокое обучение и нейронные сети	8	2	4	2	Тест + Практическая работа
33	Тема 2. Обзор основных архитектур нейронных сетей. Сверточные и рекуррентные сети	8	2	4	2	Тест + Практическая работа
34	Тема 3. Обучение нейронных сетей с помощью TensorFlow и Keras	8	2	4	2	Тест + Практическая работа
35	Тема 4. Использование предварительно обученных моделей для классификации изображений и других задач	8	2	4	2	Тест + Практическая работа
36	Тема 5. Small Data Learning и Сиамские нейронные сети	8	2	4	2	Тест + Практическая работа
37	Раздел 2.3. Продвижение продукта. Бизнес-метрики					Тест + Практическая работа
38	Тема 1. Введение в продуктовую аналитику.	8	2	4	2	Тест + практическая работа
39	Тема 2. Ключевые метрики роста продукта	8	2	4	2	Тест + практическая работа
40	Тема 3. A/B-тестирование	6	2	2	2	Тест + практическая работа
41	Промежуточная аттестация	4		4		Решение кейса
42	Итоговая аттестация	8		8		Решение кейсов
	Всего часов	260	68	130	62	

3. КАЛЕНДАРНЫЙ УЧЕБНЫЙ ГРАФИК.

Объем программы – 260 часов.

Продолжительность обучения – 3 месяца

Форма обучения – очно-заочная с применением электронного обучения и дистанционных образовательных технологий.

Режим занятий: 3-4 часа в день.

Завершение обучения: 25.11.2024.

№ п/п	Наименование учебных модулей/ практик/ аттестации	Трудоёмкость (час)	Срок освоения (кол-во учебных дней)
1	Модуль 1. Раздел 1.1 Введение в анализ данных	12	4
2	Раздел 1.2. Инструменты, библиотеки и технологии анализа данных	38	12
3	Раздел 1.3 Технологии хранения и обработки больших данных	66	22
4	Промежуточная аттестация. Решение практико-ориентированного кейса	4	1
5	Модуль 2. Раздел 2.1. Математическое моделирование, машинное обучение и оптимизация	66	22
6	Раздел 2.2. Глубокое обучение и нейронные сети	40	12
7	Раздел 2.3. Продвижение продукта. Бизнес-метрики	22	6
8	Промежуточная аттестация. Решение практико-ориентированного кейса	4	1
9	Итоговая аттестация	8	2

4. РАБОЧИЕ ПРОГРАММЫ МОДУЛЕЙ УЧЕБНОГО КУРСА ПРОГРАММЫ ПОВЫШЕНИЯ КВАЛИФИКАЦИИ «АНАЛИТИК ДАННЫХ».

Программа повышения квалификации состоит из 2 учебных модулей:

Модуль 1. Базовый.
Раздел 1.1. Введение в анализ данных.
Раздел 1.2. Инструменты, библиотеки и технологии анализа данных.
Раздел 1.3. Технологии хранения и обработки больших данных.

Модуль 2. Профильный.
Раздел 2.1. Математическое моделирование, машинное обучение и оптимизация.
Раздел 2.2. Глубокое обучение и нейронные сети.
Раздел 2.3. Продвижение продукта. Бизнес-метрики.

4.1 Рабочая программа модуля 1. Базовый.

Введение в анализ данных. Инструменты, библиотеки и технологии анализа данных. Технологии хранения и обработки больших данных.

Цель освоения модуля 1 – приобретение слушателями профессиональных компетенций в области анализа данных, инструментов, библиотек и технологий хранения и обработки больших данных.

Профессиональные компетенции, совершенствуемые и приобретаемые слушателями в процессе освоения модуля 1:

ПК-1.р. Способен классифицировать и идентифицировать задачи искусственного интеллекта, выбирать адекватные методы и инструментальные средства решения задач искусственного интеллекта.

ПК-4.р. Способен разрабатывать и применять методы машинного обучения для решения задач.

ПК-5.р. Способен использовать инструментальные средства для решения задач машинного обучения.

Планируемые результаты обучения по модулю 1.

По итогам освоения модуля слушатели должны:

Знать:

- основные определения искусственного интеллекта и больших данных;
- принципы работы NoSQL баз данных;
- основные уровни представления данных;
- основные типы данных в СУБД Postgres;
- основные конструкции языка Python, библиотеки;

- системы обработки и анализа больших массивов данных (Hadoop, ETL, Spark и др.);
- платформы и базы данных.

Уметь:

- проводить аналитические и поисковые исследования по тематике информационных технологий, технологий больших данных;
- строить несколько моделей и выбирать лучшую модель на данных;
- применять язык программирования Python и библиотеки при разработке решений на основе ИИ;
- осуществлять массово-параллельную обработку и анализ данных;
- строить модели машинного обучения (регрессия, классификация, кластеризация, нейросети).

Владеть:

- методами и инструментальными средствами решения задач искусственного интеллекта;
- навыками расчета ключевых метрик роста продукта с помощью Python;
- оценивать результаты моделирования и определять критерии качества построенных моделей;
- осуществлять парсинг Интернет-данных;
- применять SQL базы данных для прикладных решений;
- производить расчет вероятностных показателей с использованием языка Python;
- разрабатывать модели машинного обучения для решения задач.
- навыками расчета статистических показателей с использованием языка Python;
- навыками создания нескольких таблиц в СУБД Postgres посредствам Dbeaver;
- навыками интеллектуального анализа данных с помощью языка программирования R;
- навыками обучения нейронных сетей с помощью PyTorch, TensorFlow и Keras;
- навыками расчета ключевых метрик роста продукта с помощью Python;
- навыками настраивания кластеров Apache Spark и Hive на Hadoop;
- владение инструментами инструменты Weka,RapidMiner, Knime, Orange IBM SPSS Modeler, Tableau и др.;
- использовать базы данных (MongoDB, Clickhouse и др.).

Учебно-тематический план модуля 1 Базовый.

№ № п/п	Наименование дисциплины, модуля, темы	Трудоемкость		В том числе				Форма контроля
				Контактная работа[1]			Самостоятельная работа	
					из них			
		В зачетных единицах	В часах	Всего	Лекции	Практические занятия		
1	2	3	4	5	6	7	8	9
Раздел 1.1. Введение в анализ данных								
1	Тема 1. Введение в анализ данных. Профессия Аналитик данных		6	4	2	2	2	Тест + Практическая работа
2	Тема 2. Определение искусственного интеллекта и больших данных. Применение искусственного интеллекта в различных областях		6	4	2	2	2	Тест + Практическая работа
Раздел 1.2. Инструменты, библиотеки и технологии анализа данных								
3	Тема 1. Знакомство с синтаксисом языка Python и средой разработки Jupyter Notebook. Обзор языка R		10	8	2	6	2	Тест + Практическая работа
4	Тема 2. Работа с библиотеками Python		10	8	2	6	2	Практическая работа
5	Тема 3. Работа с внешними API и протоколом http. Парсинг Интернет-данных		8	6	2	4	2	Тест + Практическая работа
6	Тема 4. Системы обработки и анализа больших массивов данных (Hadoop, ETL, Spark и др.)		10	8	4	4	2	Тест + Практическая работа
Раздел 1.3. Технологии хранения и обработки больших данных								
7	Тема 1. Виды представления данных: табличные, графовые, временные ряды. Качество данных, подходы и инструменты		8	6	2	4	2	Тест + Практическая работа
8	Тема 2. Платформы данных		4	2	2	0	2	Тест + Практическая работа
9	Тема 3. Введение в теорию БД. Основы языка SQL		8	6	2	4	2	Тест + Практическая работа
10	Тема 4. SQL базы данных		8	6	2	4	2	Тест + Практическая работа
11	Тема 5. Базы данных NoSQL		8	6	2	4	2	Тест + Практическая работа
12	Тема 6. Массово-параллельная		6	4	2	2	2	Тест +

[1] *С применением дистанционных образовательных технологий и электронного обучения*

	обработка и анализ данных							Практическая работа
13	Тема 7. Знакомство с СУБД Postgres. Обзор GreenPlum		8	6	2	4	2	Тест + Практическая работа
14	Тема 8. Моделирование данных		8	6	2	4	2	Тест + Практическая работа
15	Тема 9. Построение дашбордов с помощью Superset		8	6	2	4	2	Практическая работа
16	Промежуточная аттестация		4	4		4		Решение кейса
17	Итого по модулю 1		120	90	32	58	30	

Раздел 1.1. Введение в анализ данных.

Тема 1. Введение в анализ данных. Профессия Аналитик данных.

Аналитик данных: потребность и ценность. Задачи, навыки, инструменты в классификации данных. Обязанности и функция в команде.

Тема 2. Определение искусственного интеллекта и больших данных. Применения искусственного интеллекта в различных областях.

Основные определения искусственного интеллекта и больших данных. Сущность и использование больших данных. Взаимосвязь больших данных и искусственного интеллекта.

Раздел 1.2. Инструменты, библиотеки и технологии анализа данных.

Тема 1. Знакомство с синтаксисом языка Python и средой разработки Jupyter Notebook. Обзор языка R.

Введение в программирование на языке Python. Основные конструкции языка. Переменные, объекты, типы данных. Базовые структуры данных. Основные операторы и управляющие конструкции – операторы условий, циклы и т.д. Знакомство со средой разработки Jupyter Notebook. Обзор языка R.

Тема 2. Работа с библиотеками Python.

Знакомство с библиотекой для анализа данных Pandas. Работа с файлами формата xlsx и csv. Понятие датасета. Базовые операции работы с табличными данными. Основные формы визуализации данных: таблицы, диаграмма, гистограмма, график. Знакомство с библиотеками для визуализации данных языка Python: matplotlib и seaborn. Решение базовых аналитических кейсов с использованием инструментов визуализации.

Тема 3. Работа с внешними API и протоколом http. Парсинг Интернет-данных.

Рассмотрение возможностей языка Python для получения данных из внешних API.

Краткое введение в протоколы http и rest. Погружение в библиотеку requests. Введение в понятие парсинга и веб-скрейпинга. Разбор библиотеки lxml.

Тема 4. Системы обработки и анализа больших массивов данных (Hadoop, ETL, Spark и др.).

Основы Hadoop. Архитектура Apache Hadoop. Apache Spark. Использование Hadoop. Создание масштабируемых решения для обработки данных с помощью Apache Hive и Apache Spark. Работа с распределенной кластерной системой. ETL процессы и инструменты.

Раздел 1.3. Технологии хранения и обработки больших данных.

Тема 1. Виды представления данных: табличные, графовые, временные ряды. Качество данных, подходы и инструменты.

Виды представления данных: табличные, графовые, временные ряды. Data quality качество данных, подходы и инструменты Weka, RapidMiner, Knime, Orange IBM SPSS Modeler, Tableau и др.

Тема 2. Платформы данных.

Платформы данных (облачные и внутрикорпоративные). Цифровая платформа анализа данных.

Тема 3. Введение в теорию БД. Основы языка SQL.

Основы теории баз данных. Разновидности баз данных: реляционные, графовые, колоночные сетевые и т.д. Ключевые особенности реляционных баз данных. Введение в язык SQL для работы с реляционными БД. Понятие СУБД. Отличие СУБД от БД. Знакомство с приложением dbeaver для доступа к различным реляционным СУБД.

Тема 4. SQL базы данных.

Погружение в язык SQL. Операторы условия, сортировки, удаления дубликатов, группировки, соединения и т.д. Оконные функции в языке SQL. Порядок вызова операторов в запросе select.

Тема 5. Базы данных NoSQL.

Хранилища данных NoSQL, назначение и особенности. Отличия от реляционных баз данных. NoSQL в больших данных. Классы NoSQL-СУБД с точки зрения CAP-теоремы и их значимость для больших данных. Использование базы данных NoSQL: MongoDB.

Тема 6. Массово-параллельная обработка и анализ данных.

Массово-параллельные базы данных в больших данных. Что MPP-СУБД и как это работает. MPP-СУБД для хранения и аналитики больших данных на примере GreenPlum.

Тема 7. Знакомство с СУБД Postgres. Обзор GreenPlum.

Подробное знакомство с СУБД Postgres. Введение понятий: ограничения, первичный ключ, внешний ключ. Основные типы данных в СУБД Postgres. Ключевые особенности PSQL. Обзор БД GreenPlum.

Тема 8. Моделирование данных.

Введение в моделирование данных. Понятие модели данных. Концептуальный, логический и физический уровень моделирования. Базовые модели данных: звезда и снежинка. Знакомство с UML-диаграммами.

Тема 9. Построение дашбордов с помощью Superset.

Введение в принципы построения дашбордов. Обзор различных BI-инструментов. Какие ключевые задачи решает визуализация с помощью BI-систем. Введение в Apache Superset. Ключевые преимущества и недостатки.

Содержание практических занятий по модулю 1.

№ темы	Наименование (содержание) темы, по которой предусмотрено практическое занятие	Формы и методы проведения
Раздел 1.1. Введение в анализ данных		
1	Тема 1. Решить кейс по анализу условного продукта на базе имеющихся знаний. Предложить список необходимых данных для анализа и продумать ключевые метрики	Решение практических заданий
2	Тема 2. Произвести аналитику для интеллектуального отслеживания ресурсов/процессов	Решение практических заданий
Раздел 1.2. Инструменты, библиотеки и технологии анализа данных		
1	Тема 1. Установить интерпретатор Python. Запустить программы на языке Python. Установить Jupyter Notebook. Решить задачу на проработку базовых навыков программирования на языке Python	Решение практических заданий
2	Тема 2. Практика использования возможностей библиотеки Pandas на примере данных о поездках в такси. Визуализация данных о поездках в такси для решения аналитических задач поиска аномалий, стандартного отклонения и кластеризации	Решение практических заданий
3	Тема 3. Разработать скрипт для получения данных из открытого источника с помощью подключения к API. Разработать дополнительный скрипт для парсинга веб-страницы	Решение практических заданий
4	Тема 4. Настроить кластер Apache Spark и Hive на Hadoop	Решение практических заданий
Раздел 1.3. Технологии хранения и обработки больших данных		
1	Тема 1.Решить практические задачи на Python, используя библиотеки Matplotlib, seaborn, plotly. Загрузить датасет, сделать визуализацию разных признаков, их распределения, корреляции и взаимосвязей	Решение практических заданий
3	Тема 3. Установить dbeaver. Подключить к БД. Создать базу данных и таблицы с	Решение практических заданий

	помощью языка SQL	
4	Тема 4. Выполнить запросы на закрепление базовых операторов языка SQL на примере PostgreSQL	Решение практических заданий
5	Тема 5. Разработать консольную утилиту для преобразования лога веб-сервера в формате CSV (Comma Separated Values) в формат JSON. Лог должен содержать поля со следующими названиями: URL, IP, timeStamp, timeSpent	Решение практических заданий
6	Тема 6. Провести анализ данных транспорта в системах мониторинга логистики с помощью гео-расширения (используя MPP-СУБД)	Решение практических заданий
7	Тема 7. Создать несколько таблиц в СУБД Postgres с помощью Dbeaver. Таблицы должны быть связаны внешними ключами, при этом сохраняя ссылочную целостность. На данные в таблицах должны быть наложены ограничения. Наполнить таблицы данными с учетом типов данных. Отработать запросы на языке PSQL на созданных таблицах	Решение практических заданий
8	Тема 8. Создать собственную модель данных с использованием UML-диаграмм на примере данных о студентах в университете. Практическая работа должна содержать все 3 уровня моделирования и обоснование выбора модели данных	Решение практических заданий
9	Тема 9. Построить дашборд с помощью Apache Superset. Подключить к СУБД Postgres. Проработать стиль и оформление дашборда	Решение практических заданий

Содержание самостоятельной работы слушателей по модулю 1.

Основная цель самостоятельной работы слушателей – закрепление знаний, полученных в ходе лекционных и практических занятий.

Индивидуальная консультационная работа преподавателей со слушателями осуществляется весь период обучения.

№ темы	Наименование (содержание) темы, по которой предусмотрена самостоятельная работа	Формы и методы проведения
Раздел 1.1 Введение в анализ данных		
1	Тема 1. Введение в анализ данных. Профессия Аналитик данных	Самостоятельно разобрать типовые кейсы, которые решают аналитики данных в практике. Тест 1: Задачи аналитика данных
2	Тема 2. Определение искусственного интеллекта и больших данных. Применение искусственного	Изучить материал: Различия между Data Science, машинным обучением, ИИ, глубоким обучением и Data

	интеллекта в различных областях	Mining Тест 2: Искусственный интеллект и большие данные
Раздел 1.2. Инструменты, библиотеки и технологии анализа данных		
1	Тема 1. Знакомство с синтаксисом языка Python и средой разработки Jupyter Notebook. Обзор языка R	Изучить материал: синтаксис языка Python и его основных конструкций. Тест 1: Базовый синтаксис языка Python
2	Тема 2. Работа с библиотеками Python	Изучить материал: Знакомство с библиотекой Numpy. Изучение различных форматов хранения данных: json, tsv, avro, xml. Тест 2: Работа с библиотеками Python
3	Тема 3. Работа с внешними API и протоколом http. Парсинг Интернет-данных	Изучить материал: Дополнительные возможностей библиотек requests и lxml. Знакомство с другими решениями для подключения к API и парсинга веб-ресурсов – BeautifulSoup, Selenium и другими. Поиск и попытка получения данных из альтернативных открытых источников. Тест 3: Работа с внешними API и протоколом http
4	Тема 4. Системы обработки и анализа больших массивов данных (Hadoop, ETL, Spark и др.)	Изучить материал: Бакетирование vs партиционирование в Apache Hive и Spark. Тест 4: Системы обработки и анализа больших массивов данных
Раздел 1.3 Технологии хранения и обработки больших данных		
1	Тема 1. Виды представления данных: табличные, графовые, временные ряды. Качество данных, подходы и инструменты	Изучить материал: Методы анализа на графах. Случайные графы, безмасштабные графы, социальные сети – сети тесного мира. Тест 1: Виды представления данных
2	Тема 2. Платформы данных	Изучить материал: On premises /Cloud solutions. Облака в сравнении с on-premises инфраструктурой: возможности, преимущества, особенности. Экосистема Hadoop и элементы Системы Обработки Данных. Аналоги из экосистем GCP, AWS. Тест 2: Платформы данных
3	Тема 3. Введение в теорию БД. Основы языка SQL	Изучить материал: Совместное использование базы данных. Безопасность данных. Тест 3: Введение в теорию БД
4	Тема 4. SQL базы данных	Изучить дополнительные возможности языка SQL: триггеры, хранимые процедуры, функции, индексы и т.д. Алгоритмическая сложность вызова

		разных типов операторов join. Тест 4: SQL базы данных
5	Тема 5. Базы данных NoSQL	Изучить материал: Документно-ориентированная модель данных MongoDB. Тест 5: Базы данных NoSQL
6	Тема 6. Массово-параллельная обработка и анализ данных	Изучить материал: Особенности организации СУБД в MPP-системе Тест 6: Массово-параллельная обработка и анализ данных
7	Тема 7. Знакомство с СУБД Postgres. Обзор GreenPlum	Провести сравнение: Postgres с СУБД MS SQL, OracleDB, MySQL. Тест 7: СУБД Postgres
8	Тема 8. Моделирование данных	Изучить материал: Модели данных: созвездие, data vault и anchor modeling. Тест 8: Моделирование данных
9	Тема 9. Построение дашбордов с помощью Superset	Изучить материал: Возможностей BI-системы Apache Superset. Изучение альтернативных BI-систем

Рекомендуемый перечень вопросов для отработки в часы самостоятельной работы, подготовки к промежуточной аттестации.

1. Задачи аналитика данных.
2. Искусственный интеллект и большие данные.
3. Базовый синтаксис языка Python.
4. Работа с библиотеками Python.
5. Работа с внешними API и протоколом http.
6. Системы обработки и анализа больших массивов данных.
7. Виды представления данных.
8. Платформы данных.
9. Введение в теорию БД.
10. SQL базы данных.
11. Базы данных NoSQL.
12. Массово-параллельная обработка и анализ данных.
13. СУБД Postgres.
14. Моделирование данных.

Учебно-методическое обеспечение и информационное сопровождение.

Обучающие материалы представлены в виде видеолекций, текстовых и графических материалов, размещенных на платформе https://data.1t.ru/.

Перечень основной и дополнительной учебной литературы.

Нормативно-правовые акты:

1. Национальная программа «Цифровая экономика Российской Федерации», утв. распоряжением Правительства Российской Федерации от 28 июля 2017 г. № 1632-р // Электронный фонд правовых и нормативно-технических документов. – Режим доступа: https://docs.cntd.ru/document/436754837 (Дата обращения 27.03.2024).

2. Паспорт федерального проекта «Искусственный интеллект» национальной программы «Цифровая экономика Российской Федерации» (приложение № 3 к протоколу президиума Правительственной комиссии по цифровому развитию, использованию информационных технологий для улучшения качества жизни и условий ведения предпринимательской деятельности от 27.08.2020 № 17). – Код доступа: https://spa.msu.ru/wp-content/uploads/5-1.pdf, дата обращения 27.03.2024.

Основная литература:

1. Петров, А. Распределенные данные. Алгоритмы работы современных систем хранения информации. – Санкт-Петербург: Питер, 2021. − 336 с.

2. Карпова, И.П. Базы данных. Учебное пособие. − Санкт-Петербург: Питер, 2021. - 240 с.

3. Куницын, А.П. Технический анализ: Полный курс / А.П. Куницын, Б. Зуев – Москва: Альпина Паблишер, 2017. − 880 с. − ISBN 978-5-9614-3737-9. – Текст: непосредственный.

4. Ния, Н. Apache Kafka. Потоковая обработка и анализ данных / Н. Ния, Ш. Гвен, П. Тодд. − Санкт-Петербург: Питер, 2021. − 320 с.

5. Силен, Д. Основы Data Science и Big Data. Python и наука о данных: илюстрации – (Серия «Библиотека программиста») / Д. Силен, А. Мейсман, М. Али – Санкт-Петербург: Питер, 2017. – 336 с.

Дополнительная литература:

1. Шеннон, Б., Йон Б, Ходоров К. MongoDB: полное руководство. – Москва: ДМК Пресс, 2020. – 540 с.

2. Деревянко, М.Э., Нилова, Н.М. Обзор современных информационных систем управления бизнес-процессами – 2021.

3. Еременко, К. Работа с данными в любой сфере: как выйти на новый уровень, используя аналитику – Москва: Альпина Паблишер, 2019.

4. Поляков, В.М., Агаларов З.С. Методы оптимизации. Учебное пособие. – Москва: Дашков и К., 2022. – 86 с.

5. Цветков, А.А. Теория и практика бизнес-анализа в ИТ – Москва: ООО «ДиректМедиа», 2020.

Учебно-методические и информационные материалы:

1. Аналитика Больших данных как инструмент бизнес-инноваций. /[Электронный ресурс] – Режим доступа: https://filearchive.cnews.ru/img/files/2019/05/27/20190424idchitachiwpbdafin.pdf (Дата обращения 27.03.2024).

2. Большие данные в социальных и гуманитарных науках: Сборник обзоров и рефератов / РАН. ИНИОН. Центр научно-информационных исследований по науке, образованию и технологиям; отв. ред. – Гребенщикова Е.Г. – Москва, 2019. – 193 с.

3. Миронов, В. Профессия «бизнес-аналитик». Краткое пособие для начинающих – Москва: Litres, 2021.

4. Понкин, И.В., Лаптева А.И. Методология научных исследований и прикладной аналитики: Учебник. Издание 2-е, дополн. и перераб. / Консорциум «Аналитика. Право. Цифра». – Москва: Буки Веди, 2021. – 567 с.

5. Рафалович, В. Data mining, или интеллектуальный анализ данных для занятых. Практический курс – Москва: Litres, 2022.

6. Системный анализ: учебник и практикум для вузов / В.В. Кузнецов [и др.]; под общей редакцией В.В. Кузнецова. – Москва: Издательство «Юрайт», 2023. – 270 с.

7. Цифровая экономика от теории к практике: как российский бизнес использует искусственный интеллект / исслед. РАЭК / НИУ ВШЭ при поддержке Microsoft. – 2019. – 66 с. – Код доступа: http://raec.ru/upload/files/190715-ii.pdf (Дата обращения 27.03.2024).

Информационное сопровождение.

Электронные образовательные ресурсы:

1. Проектирование баз данных: Распределенные базы и хранилища данных. Агрегирование // Национальный Открытый Университет «ИНТУИТ». – Режим доступа: http://www.intuit.ru/studies/professional_retraining/953/courses/214/lecture/5508/ (Дата обращения 27.03.2024).

2. Бесплатные материалы по Data Engineering от преподавателей МФТИ. – Режим доступа: https://fpmi-edu.ru/free-de (Дата обращения 27.03.2024).

3. Курс «Big Data и Data Science: начни погружение с нуля». – Режим доступа: https://stepik.org/course/101687/promo (Дата обращения 27.03.2024).

Электронные информационные ресурсы:

1. Сайт УНТИ 2035 «Обучение в области искусственного интеллекта» / [Электронный ресурс] – Режим доступа: https://ai.2035.university/ (Дата обращения 27.03.2024).

2. Сайт Национального проекта «Цифровая экономика» / [Электронный ресурс] – Режим доступа: https://национальныепроекты.рф/projects/tsifrovaya-ekonomika (Дата обращения 27.03.2024).

3. Сайт федерального проекта «Искусственный интеллект» Национального проекта «Цифровая экономика» / [Электронный ресурс] – Режим доступа: https://национальныепроекты/projects/tsifrovaya-ekonomika/p-iskusstvennyy-intellekt-p (Дата обращения 27.03.2024).

4. Сайт образовательной платформы ООО «1Т» / [Электронный ресурс] – Режим доступа: https://data.1t.ru/ (Дата обращения 27.03.20244).

5. Раздел «Искусственный интеллект» на сайте РБК / [Электронный ресурс] – Режим доступа: https://trends.rbc.ru/trends/tag/ai (Дата обращения 27.03.2024).

Описание системы оценки качества освоения модуля 1.

Контроль результатов освоения модуля 1 осуществляется в ходе текущего контроля успеваемости и промежуточной аттестации.

Текущий контроль предусмотрен в ходе изучения каждой темы.

Формами текущего контроля являются тесты и выполнение практических работ.

Тесты содержат не менее 5 вопросов с одним или несколькими правильными ответами. За каждый правильный ответ ставится 1 балл. Критерием прохождения теста является получение не менее 75% правильных ответов.

Примеры тестов.

Модуль 1.

Раздел 1.2. Инструменты, библиотеки и технологии анализа данных

Тема: Системы обработки и анализа больших массивов данных (Hadoop, ETL, Spark и др.)

Вопросы:

1. Что такое Hadoop?

a) Система анализа малых данных.

b) Фреймворк для обработки больших объемов данных.

c) Библиотека для машинного обучения.

d) Фреймворк для разработки мобильных приложений.

2. Как называется файловая система, используемая в Hadoop?

a) Hadoop Distributed File System (HDFS)

b) Apache File System (AFS)

c) Hadoop File System (HFS)

d) High-performance File System (HFS)

3. Какой язык программирования может использоваться для работы со Spark?

a) Python

b) Java

c) Scala

d) Все вышеперечисленные

4. Каким образом Spark обрабатывает большие объемы данных?

a) Обрабатывает данные на одном компьютере.

b) Распределяет данные на несколько компьютеров и обрабатывает их параллельно.

c) Использует реляционную SQL-базу данных для обработки данных.

d) Использует квантовые компьютеры для обработки данных.

5. Какой пакет в Spark используется для машинного обучения?

a) Spark SQL

b) Spark Streaming

c) GraphX

d) MLlib

Выполнение практических работ оценивается в бинарной системе: зачтено / не зачтено.

Зачтено: задача решена, могут быть недочеты и неточности в решении.

Не зачтено: задача не решена.

Примеры практических работ.

Модуль 1.

Раздел 1.2. Инструменты, библиотеки и технологии анализа данных

Тема: 3. Работа с внешними API и протоколом http. Парсинг Интернет-данных

Задача: Парсинг Интернет данных с сайта hh.ru

Необходимо спарсить данные о вакансиях аналитика с сайта hh.ru, введя в поиск “python разработчик” и указав, что мы рассматриваем все регионы. Необходимо спарсить:

Название вакансии

Требуемый опыт работы

Заработную плату

Регион

И сохранить эти данные в формате json.

Решения практических заданий отправьте в файле формата ".ipynb" используя редакторы кода Jupyter / Google Colab.

Промежуточная аттестация.

Промежуточная аттестация проводится в формате решения практико-ориентированной задачи (кейса).

Примеры практико-ориентированных задач (кейсов).

Модуль 1.

Задание 1.

Компания заметила снижение продаж своих товаров в последние несколько месяцев и хотела бы проанализировать данные, чтобы выяснить, что вызвало это снижение. Ваша задача как аналитика данных – проанализировать данные и предоставить рекомендации, чтобы компания могла принять меры для увеличения продаж.

Вам доступны данные о продажах за последние 6 месяцев, включая количество проданных товаров, сумму продаж и каналы продаж. Также у вас есть данные о рекламных кампаниях, проведенных компанией в течение этого периода, включая бюджет, каналы и эффективность.

Ваша задача – провести анализ данных и ответить на следующие вопросы:

1. Каковы были общие продажи за последние 6 месяцев и как они сравниваются с аналогичным периодом годом ранее?
2. Какие каналы продаж показали хорошие результаты, а какие не очень?
3. Какие рекламные кампании были наиболее эффективными в привлечении клиентов, а какие не сработали?
4. Есть ли какие-либо другие факторы, которые могут объяснить снижение продаж, кроме рекламных кампаний?
5. Какие рекомендации вы можете дать компании для увеличения продаж в будущем?

Ссылку на репозиторий с кодом прислать в чат с преподавателем.

Задание 2.

Компания занимается онлайн-ритейлом и имеет множество клиентов со всего мира. Компания хранит данные о своих клиентах, включая информацию о покупках, демографических данных и информацию о сессиях на сайте. Общий объем данных, которые компания обрабатывает и хранит, составляет несколько петабайт.

Компания хочет оптимизировать свою инфраструктуру хранения данных, чтобы улучшить производительность и снизить затраты на хранение данных. Ваша задача как аналитика данных – проанализировать текущую инфраструктуру хранения данных компании и предложить решения для ее оптимизации.

Для выполнения задачи вам необходимо выполнить следующие шаги:

1. Проведите анализ данных о текущей инфраструктуре хранения данных компании. Изучите объем и типы данных, используемые технологии и проблемы, с которыми компания столкнулась в процессе обработки и хранения данных.

2. Определите ключевые метрики производительности инфраструктуры хранения данных, такие как время доступа к данным, время резервного копирования, скорость передачи данных и использование хранилища. Сравните эти метрики с аналогичными метриками других компаний в вашей отрасли и определите возможные проблемы в инфраструктуре хранения данных компании.

3. Определите области, где можно сократить объем данных, не ухудшая качество анализа данных. Рассмотрите возможность сжатия данных, удаления неиспользуемых данных, агрегирования данных и использования других техник для уменьшения объема данных.

4. Разработайте план оптимизации инфраструктуры хранения данных, основанный на результатах анализа. Включите в план предложения по замене текущих технологий на более эффективные, оптимизации процессов резервного копирования, улучшения скорости передачи данных и использования хранилища.

5. Оцените стоимость внедрения рекомендованных изменений и прогнозирование. Оцените стоимость внедрения рекомендованных изменений, включая затраты на оборудование, программное обеспечение и обучение персонала. Спрогнозируйте потенциальные экономические выгоды от оптимизации инфраструктуры хранения данных, такие как снижение затрат на хранение или повышение производительности.

Ссылку на репозиторий с кодом прислать в чат с преподавателем.

4.2. Рабочая программа модуля 2. Профильный.

Математическое моделирование, машинное обучение и оптимизация. Глубокое обучение и нейронные сети. Продвижение продукта. бизнес-метрики.

Цель освоения модуля 2 – приобретение слушателями профессиональных компетенций в области математического моделирования, машинного обучения и оптимизации, глубокого обучения и нейронных сетей, продвижения продукта, бизнес-метрики.

Профессиональные компетенции, совершенствуемые и приобретаемые слушателями в процессе освоения модуля 2:

ПК-.6.п. Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта.

ПК-7.п. Способен выполнять анализ больших данных.

ПК-8.п. Способен использовать одну или несколько сквозных цифровых субтехнологий искусственного интеллекта.

Планируемые результаты обучения по модулю 2.

По итогам освоения модуля слушатели должны:

Знать:

- нейросетевые модели и методы;
- сверточные и рекуррентные сети;
- основы теории баз данных;
- понятие A/B-тестирования;
- особенности продуктовой аналитики;
- существующие и перспективные методы и программный инструментарий технологий больших данных;
- математические основы машинного обучения (линейная алгебра, статистика, оптимизация);
- принципы построения дашбордов;
- основные понятия теории вероятности;
- основы комбинаторики;
- представление о сквозных цифровых субтехнологиях искусственного интеллекта;
- Small Data Learning и Сиамские нейронные сети.

Уметь:

- разрабатывать системы искусственного интеллекта на основе моделей искусственных нейронных сетей и инструментальных средств;

- строить модели машинного обучения;
- производить аналитику для интеллектуального отслеживания ресурсов/процессов;
- визуализировать анализируемые данные;
- применять методы анализа на графах;
- создавать собственные модели данных с использованием UML-диаграмм;
- осуществлять математическое и информационное моделирование.

Владеть:

- методами разработки моделей машинного обучения и нейронных сетей;
- навыками построения полносвязной нейронной сети для задачи классификации;
- навыками обучения нейронных сетей с помощью TensorFlow и Keras;
- навыками использования статистических методов исследования;
- математическими методами анализа данных;
- навыками интеллектуального анализа данных с помощью языка программирования PYTHON, R.
- навыками использования предварительно обученных моделей для классификации изображений и других задач;
- навыками обучения нейронной сети Keras многоклассовой классификации изображений на малом количестве данных;
- навыками внедрения сиамских нейронных сетей для задач биометрии и распознавания образов.

Учебно-тематический план модуля 2 Профильный.

№ № п/п	Наименование дисциплины, модуля, темы	Трудоем-кость		В том числе				Форма контроля
				Контактная работа[2]			Самостоятельная работа	
					из них			
		В зачетных единицах	В часах	Всего	Лекции	Практические занятия		
1	2	3	4	5	6	7	8	9
Раздел 2.1. Математическое моделирование, машинное обучение и оптимизация								
1	Тема 1. Введение в теорию вероятностей. Базовые понятия		8	6	2	4	2	Тест + практическая работа
2	Тема 2. Введение в математическую статистику. Статистические методы анализа данных		8	6	2	4	2	Тест + практическая работа
3	Тема 3. Основные понятия и термины в машинном обучении и нейронных сетях		8	6	2	4	2	Тест + практическая работа
4	Тема 4. Математические основы машинного обучения (линейная алгебра, статистика, оптимизация)		8	6	2	4	2	Тест + практическая работа
5	Тема 5. Основные метрики оценки качества моделей машинного обучения		8	6	2	4	2	Тест + практическая работа
6	Тема 6. Выбор и обработка данных для машинного обучения. Методы машинного обучения		8	6	2	4	2	Тест + практическая работа
7	Тема 7. Построение моделей машинного обучения (регрессия, классификация, кластеризация, нейросети)		10	8	2	6	2	Тест + практическая работа
8	Тема 8. Инструменты анализа данных и Machine Learning (Rapid Miner)		8	6	2	4	2	Тест + практическая работа
Раздел 2.2. Глубокое обучение и нейронные сети								
9	Тема 1. Введение в глубокое обучение и нейронные сети		8	6	2	4	2	Тест + практическая работа
10	Тема 2. Обзор основных архитектур		8	6	2	4	2	Тест +

[2] *С применением дистанционных образовательных технологий и электронного обучения*

	нейронных сетей. Сверточные и рекуррентные сети							практическая работа
11	Тема 3. Обучение нейронных сетей с помощью TensorFlow и Keras		8	6	2	4	2	Тест + практическая работа
12	Тема 4. Использование предварительно обученных моделей для классификации изображений и других задач		8	6	2	4	2	Тест + практическая работа
13	Тема 5. Small Data Learning и Сиамские нейронные сети		8	6	2	4	2	Тест + практическая работа
Раздел 2.3. Продвижение продукта. Бизнес-метрики								
14	Тема 1. Введение в продуктовую аналитику		8	6	2	4	2	Тест + практическая работа
15	Тема 2. Ключевые метрики роста продукта		8	6	2	4	2	Тест + практическая работа
16	Тема 3. A/B-тестирование		6	4	2	2	2	Тест + практическая работа
17	Промежуточная аттестация		4	4		4		Решение кейса
18	Итого по модулю 2		132	100	32	68	32	Решение кейсов

Раздел 2.1. Математическое моделирование, машинное обучение и оптимизация.

Тема 1. Введение в теорию вероятностей. Базовые понятия.

Введение в теорию вероятностей. Разбор основных понятий теории вероятностей: случайные события, алгебра событий вероятность, условная вероятность, зависимые и независимые события, формула полной вероятности и теорема Байеса, статистическое определение вероятности.

Тема 2. Введение в математическую статистику. Статистические методы анализа данных.

Введение в математическую статистику. Основные понятия: выборка, генеральная совокупность, вариационный ряд, выборочное среднее, мода, медиана, среднеквадратичное отклонение, дисперсия и т.д. Статистические методы исследования: Описательная статистика, статистические критерии: Крамера-Уэлча, Вилкоксона-Манна-Уитни, хи–квадрат, Фишера и др.

Тема 3. Основные понятия и термины в машинном обучении и нейронных сетях.

Терминология различных структур и процессов в машинном обучении. Виды обучения в МО (Обучение с учителем (supervised learning), Обучение без учителя (unsupervised learning), Обучение с подкреплением (reinforcement learning). Основные компоненты нейронных сетей. Типы нейронных сетей. Принципы обучения нейронных сетей.

Тема 4. Математические основы машинного обучения (линейная алгебра, статистика, оптимизация).

Для чего нужна линейная алгебра в МО. Матрицы и работа матриц в классических моделях. Математическое моделирование. Методы и модели классификации: логистическая регрессия, деревья решений. Методы и модели регрессии: линейная регрессия, деревья решений. Методы оценки моделей: оценка качества построенной модели по тестовой выборке и анализ обобщающих способностей алгоритма. Статистика и оптимизация в машинном обучении.

Тема 5. Основные метрики оценки качества моделей машинного обучения.

Изучение основных метрик оценки качества моделей. Важность выбора метрики, в зависимости от поставленной задачи. Интерпретация результатов обучения модели: как оценивать полученные метрики. Рекомендации по выбору подходящих метрик.

Тема 6. Выбор и обработка данных для машинного обучения. Методы машинного обучения.

Роль данных в процессе машинного обучения: оценка их качества для успешного обучения моделей. Основные этапы предобработки. Методы машинного обучения для решения различных типов задач: обучение с учителем, обучение без учителя. Валидация данных.

Тема 7. Построение моделей машинного обучения (регрессия, классификация, кластеризация, нейросети).

Основные современные и популярные алгоритмы моделей машинного обучения - градиентные бустинги на деревьях и нейронные сети. Алгоритмы кластеризации.

Тема 8. Инструменты анализа данных и Machine Learning (Rapid Miner).

Интеллектуальный анализа данных с помощью языка программирования R.

Манипуляция, анализ и моделирование данных с помощью RapidMiner. Инструменты анализа данных и ML Rapid Miner.

Раздел 2.2. Глубокое обучение и нейронные сети.

Тема 1. Введение в глубокое обучение и нейронные сети.

Введение в глубокое обучение и нейронные сети. Как работает глубокое обучение. Революция в глубоком обучении. Как устроены многослойные нейронные сети. Как обучают нейросети. Алгоритм обратного распространения ошибки для обучения нейронных сетей. Основные архитектуры и задачи нейронных сетей. Производная, оптимизация.

Тема 2. Обзор основных архитектур нейронных сетей. Сверточные и рекуррентные сети.

Продвинутое обучение нейронных сетей.

Многослойный перцептрон.

Архитектура CNN, свёрточные нейронные сети. Разница между моделями и как устроена архитектура. Комбинирование нескольких алгоритмов нейронных сетей.

Тема 3. Обучение нейронных сетей с помощью TensorFlow и Keras.

Искусственная нейронная сеть в TensorFlow. Как обучить нейронную сеть с помощью TensorFlow и Keras. Чем отличается TensorFlow от PyTorch.

Тема 4. Использование предварительно обученных моделей для классификации изображений и других задач.

Классификация, сегментация изображений. Распознавание объектов на изображении. Распознавание текста на изображении. Способы оптимизации. Анализ изображений и видео с помощью методов искусственного интеллекта.

Тема 5. Small Data Learning и Сиамские нейронные сети.

Обучение нейронных сетей на малом количестве данных. Аугментация данных. Использование сиамских нейронных сетей обучения на малых данных и для задач биометрии и распознавания образов.

Раздел 2.3. Продвижение продукта. Бизнес-метрики.

Тема 1. Введение в продуктовую аналитику.

Ключевые задачи продуктовой аналитики. Как анализ данных может влиять на развитие продукта. Какие знания о продукте важны для анализа.

Тема 2. Ключевые метрики роста продукта.

Обзор ключевых метрик роста продукта: метрики привлечения (CPI, LTV и т.д.), метрики вовлеченности (DAU/WAU/MAU и т.д.) и метрики производительности.

Тема 3. A/B-тестирование.

Понятие A/B-тестирования. Где и зачем применяется. Критерии оценки качества проведенного A/B-тестирования. Математические основы A/B-тестирования.

Содержание практических занятий по модулю 2.

№	Наименование (содержание) темы, по которой	Формы и методы

темы	предусмотрено практическое занятие	проведения
Раздел 2.1. Математическое моделирование, машинное обучение и оптимизация		
1	Произвести расчет вероятностных показателей с использованием языка Python. Модуль random	Решение практических заданий
2	Практика расчета статистических показателей с использованием языка Python. Введение в модуль scipy	Решение практических заданий
3	Основные понятия и термины в машинном обучении и нейронных сетях	Решение практических заданий
4	Создание собственной модели линейной регрессии, сравнение работы с готовым решением в Scikit-learn	Решение практических заданий
5	Предобработка данных и построение нескольких моделей. Выбор лучшей модели на основе метрик машинного обучения	Решение практических заданий
6	Произвести поиск аномалий в данных, сегментация PCA, уменьшение размерности данных	Решение практических заданий
7	Построить модель машинного обучения для поиска мест залегания полезных ископаемых В роли признаков здесь выступают сведения, добытые при помощи геологической разведки: наличие на территории местности каких-либо пород (и это будет признаком бинарного типа), их физические и химические свойства (которые раскладываются на ряд количественных и качественных признаков)	Решение практических заданий
8	Ознакомление с синтаксисом языка R для анализа данных. Обработка данных с помощью библиотеки tidyverse. Статистический анализ данных в R	Решение практических заданий
Раздел 2.2. Глубокое обучение и нейронные сети		
1	Построение собственной небольшой модели линейной регрессии на основе библиотеки PyTorch.	Решение практических заданий
2	Построение полносвязной нейронной сети для задачи классификации на основе датасета MNIST	Решение практических заданий
3	Изменить нейронную сеть (ссылка). Вычислить средние и дисперсии по выборке обучающего датасета и применить их для того, чтобы производить более качественную предобработку данных	Решение практических заданий
4	Построение сверточной нейронной сети для задачи многоклассовой классификации	Решение практических заданий
5	Обучение нейронной сети Keras многоклассовой классификации изображений на малом количестве данных. Сравнение метрик качества с использованием и без использования аугментации изображений.	Решение практических заданий
Раздел 2.3. Продвижение продукта. Бизнес-метрики		
1	Решение кейса по анализу продукта с разработкой полного цикла от получения данных, хранения,	Решение практических заданий

	обработки и вывода результатов – в виде схемы, ключевая задача – определить метрики для анализа качества продукта	
2	Реализация расчет ключевых метрик роста продукта с помощью Python	Решение практических заданий
3	A/B-тестирование с помощью языка Python	Решение практических заданий

Содержание самостоятельной работы слушателей по модулю 2.

Основная цель самостоятельной работы слушателей – закрепление знаний, полученных в ходе лекционных и практических занятий.

Индивидуальная консультационная работа преподавателей со слушателями осуществляется весь период обучения.

№ темы	Наименование (содержание) темы, по которой предусмотрена самостоятельная работа	Формы и методы проведения
Раздел 2.1 Математическое моделирование, машинное обучение и оптимизация		
1	Тема 1. Введение в теорию вероятностей. Базовые понятия	Изучить материал: Основы комбинаторики. Изучение модуля math, itertools. Тест 1. Введение в теорию вероятностей
2	Тема 2. Введение в математическую статистику. Статистические методы анализа данных	Изучить материал: Проверка статистических гипотез. Тест 2: Введение в математическую статистику
3	Тема 3. Основные понятия и термины в машинном обучении и нейронных сетях	Изучить материал: Наиболее актуальные в настоящее время алгоритмы машинного обучения и архитектуры нейронных сетей. Понятия, входящие в данные архитектуры Тест 3: Основные понятия и термины в машинном обучении и нейронных сетях
4	Тема 4. Математические основы машинного обучения (линейная алгебра, статистика, оптимизация)	Изучить материал: Генетические алгоритмы, эволюционное программирование Тест 4: Математические основы машинного обучения
5	Тема 5. Основные метрики оценки качества моделей машинного обучения	Изучить материал: Метрики оценки LLM моделей. Математическое обоснование целесообразности использования изученных метрик. Тест 5: Основные метрики оценки качества моделей машинного обучения
6	Тема 6. Выбор и обработка данных для машинного обучения. Методы машинного обучения	Изучить материал: Подготовка датасета для машинного обучения: 10 базовых способов совершенствования

		данных. Тест 6: Выбор и обработка данных для машинного обучения
7	Тема 7. Построение моделей машинного обучения (регрессия, классификация, кластеризация, нейросети)	Изучить тему: Применение алгоритмов кластеризации данных. Тест 7: Построение моделей машинного обучения
8	Тема 8. Инструменты анализа данных и Machine Learning (Rapid Miner)	Изучить материал: Применение Rapid Miner для анализа временных рядов и анализа выделенных параметров. Тест 8: Инструменты анализа данных и Machine Learning
Раздел 2.2 Глубокое обучение и нейронные сети		
1	Тема 1. Введение в глубокое обучение и нейронные сети	Изучить материал: Преимущества глубокого обучения в облаке. Тест 1: Введение в глубокое обучение и нейронные сети
2	Тема 2. Обзор основных архитектур нейронных сетей. Сверточные и рекуррентные сети	Изучить материал: Комбинирование нескольких алгоритмов нейронных сетей Тест 2: Обзор основных архитектур нейронных сетей
3	Тема 3. Обучение нейронных сетей с помощью TensorFlow и Keras	Изучить материал: Классификация изображений с помощью TensorFlow и Keras Тест 3: Обучение нейронных сетей с помощью TensorFlow и Keras
4	Тема 4. Использование предварительно обученных моделей для классификации изображений и других задач	Тест 4: Использование предварительно обученных моделей для классификации изображений и других задач
5	Тема 5. Small Data Learning и Сиамские нейронные сети	Изучить материал: Как использовать сиамские нейронные сети для задач биометрии и распознавания образов. Тест 5: Small Data Learning и Сиамские нейронные сети
Раздел 2.3. Продвижение продукта. Бизнес-метрики		
1	Тема 1. Введение в продуктовую аналитику	Изучить материал: Особенности продуктовой аналитики. Изучить реальные кейсы как анализ данных помог вывести продукт на новый уровень. Тест 1: Основам продуктовой аналитики
2	Тема 2. Ключевые метрики роста продукта	Изучить материал: Когортный анализ: метрики роста против метрик продукта. Тест 2: Ключевые метрики роста продукта
3	Тема 3. A/B-тестирование	Изучить материал: Альтернативные способы тестирования гипотез.

		Практика методов тестирования с помощью языка Python. Тест 3: A/B-тестирование

Рекомендуемый перечень вопросов для отработки в часы самостоятельной работы, подготовки к промежуточной аттестации.

1. Введение в теорию вероятностей.

2. Введение в математическую статистику.

3. Основные понятия и термины в машинном обучении и нейронных сетях.

4. Математические основы машинного обучения.

5. Основные метрики оценки качества моделей машинного обучения.

6. Выбор и обработка данных для машинного обучения.

7. Построение моделей машинного обучения.

8. Инструменты анализа данных и Machine Learning.

9. Введение в глубокое обучение и нейронные сети.

10. Обзор основных архитектур нейронных сетей.

11. Обучение нейронных сетей с помощью TensorFlow и Keras.

12.Использование предварительно обученных моделей для классификации изображений и других задач.

13. Small Data Learning и Сиамские нейронные сети.

14. Основам продуктовой аналитики.

15. Ключевые метрики роста продукта.

16. A/B-тестирование.

Учебно-методическое обеспечение и информационное сопровождение.

Обучающие материалы представлены в виде видеолекций, текстовых и графических материалов, размещенных на платформе https://data.1t.ru/.

Перечень основной и дополнительной учебной литературы.

Нормативно-правовые акты:

1. Национальная программа «Цифровая экономика Российской Федерации», утв. распоряжением Правительства Российской Федерации от 28 июля 2017 г. № 1632-р // Электронный фонд правовых и нормативно-технических документов. – Режим доступа: https://docs.cntd.ru/document/436754837 (Дата обращения 27.03.2024).

2. Паспорт федерального проекта «Искусственный интеллект» национальной программы «Цифровая экономика Российской Федерации» (приложение № 3 к протоколу президиума Правительственной комиссии по цифровому развитию, использованию

информационных технологий для улучшения качества жизни и условий ведения предпринимательской деятельности от 27.08.2020 № 17). – Код доступа: https://spa.msu.ru/wp-content/uploads/5-1.pdf, дата обращения 27.03.2024.

Основная литература:

1. Болотова, Ю.А., Друки, А.А., Спицын, В.Г. Методы и алгоритмы интеллектуальной обработки цифровых изображений. – Томск: Томский политехнический университет, 2016. – 208 с.

2. Каменнова, М. С. Моделирование бизнес-процессов. В 2 ч. Часть 2: учебник и практикум для вузов / М. С. Каменнова, В. В. Крохин, И. В. Машков. – Москва: Издательство Юрайт, 2023. – 228 с. – (Высшее образование). – ISBN 978-5-534-09385-8. – Текст: электронный // Образовательная платформа Юрайт [сайт]. – URL: https://urait.ru/bcode/517266 (дата обращения: 27.03.2024).

3. Фальк, К. Рекомендательные системы на практике. Практическое пособие. – Москва: ДМК Пресс, 2020. – 448 с.

4. Клеппман, М. Высоконагруженные приложения. Программирование, масштабирование, поддержка. – Санкт-Петербург: Питер, 2018. − 740 с.

5. Ковалев, С.М.; Ковалев, В.М. Настольная книга аналитика //Практическое руководство по проектированию бизнес-процессов и организационной структуры: Практическое руководство. – Москва: 1С-Паблишинг, 2020.

6. Куницын, А.П.; Зуев Б. Технический анализ: Полный курс – Москва: Альпина Паблишер, 2017. − 880 с. − ISBN 978-5-9614-3737-9.

7. Петров, А. Распределенные данные. Алгоритмы работы современных систем хранения информации. − Санкт-Петербург: Питер, 2021. − 336 с.

Дополнительная литература:

1. Брокман, Д. Что мы думаем о машинах, которые думают: Ведущие мировые учёные об искусственном интеллекте – Москва: Альпина Паблишер, 2017.

2. Бруссард М. Искусственный интеллект: Пределы возможного – Москва: Альпина Паблишер, 2020.

3. Деревянко, М.Э.; Нилова Н.М. Обзор современных информационных систем управления бизнес-процессами – 2021.

4. Еременко, К. Работа с данными в любой сфере: как выйти на новый уровень, используя аналитику – Москва: Альпина Паблишер, 2019.

5. Дэвенпорт, Т. Внедрение искусственного интеллекта в бизнес-практику. Преимущества и сложности. − Москва: Альпина Паблишер, 2021. – 316 с.

6. Конвински Энди, Венделл Патрик, Захария Матей, Карау Холден. Изучаем Spark. Молниеносный анализ данных. – Москва: ДМК Пресс, 2015. – 304 с.

7. О'Коннелл, М. Искусственный интеллект и будущее человечества – Москва: Litres, 2019.

8. Лекторский В.А., Васильев С.Н., Макаров В.Л., Хабриева Т.Я., Кокошин А.А. и др. Человек и системы искусственного интеллекта. – Санкт-Петербург: Общество с ограниченной ответственностью «Издательство «Юридический центр», 2022. – 328 с.

9. Лукьянова, Н.Ю.; Галицкая Е.Г. Аналитические методы исследований в цифровой экономике – 2019.

10. Маркус Г., Дэвис Э. Искусственный интеллект: Перезагрузка. Как создать машинный разум, которому действительно можно доверять – Москва: Альпина Паблишер, 2021.

11. Мартин, Ф. Архитекторы интеллекта: вся правда об искусственном интеллекте от его создателей – Санкт-Петербург: Издательский дом «Питер», 2019.

12. Нархид Ния, Шапира Гвен, Палино Тодд. Apache Kafka. Потоковая обработка и анализ данных. – Санкт-Петербург: Питер, 2021. – 320 с.

13. Пиковер, К. Искусственный интеллект. Иллюстрированная история. От автоматов до нейросетей. – Москва: Litres, 2022.

14. Поляков В.М., Агаларов З.С. Методы оптимизации. Учебное пособие. – Москва: Дашков и К., 2022. – 86 с.

15. Цветков, А.А. Теория и практика бизнес-анализа в ИТ. – Москва: ООО «ДиректМедиа», 2020.

16. Цзэн Мин. Как Alibaba использует искусственный интеллект в бизнесе: Сетевое взаимодействие и анализ данных. – Москва: Альпина Паблишер, 2022. – 360 с.

Учебно-методические и информационные материалы:

1. Аналитика Больших данных как инструмент бизнес-инноваций. / [Электронный ресурс]. – Режим доступа: https://filearchive.cnews.ru/img/files/2019/05/27/20190424idchitachiwpbdafin.pdf (Дата обращения 27.03.2024).

2. Большие данные в социальных и гуманитарных науках: Сборник обзоров и рефератов / РАН. ИНИОН. Центр научно-информационных исследований по науке, образованию и технологиям; отв. ред. – Гребенщикова Е.Г. – Москва: 2019. – 193 с. – (Сер.: Наука, образование и технологии).

3. Миронов, В. Профессия «бизнес-аналитик». Краткое пособие для начинающих – Москва: Litres, 2021.

4. Понкин, И.В.; Лаптева А.И. Методология научных исследований и прикладной аналитики: учебник – 2 изд., доп. и перераб. – 2021.

5. Рафалович В. Data mining, или интеллектуальный анализ данных для занятых. Практический курс – Москва: Litres, 2022.

6. Системный анализ: учебник и практикум для вузов / В.В. Кузнецов [и др.]; под общей редакцией В.В. Кузнецова. – Москва: Издательство «Юрайт», 2023. – 270 с.

7. Цифровая экономика от теории к практике: как российский бизнес использует искусственный интеллект / исследования РАЭК / НИУ ВШЭ при поддержке Microsoft. – 2019. – 66 с. – Код доступа: http://raec.ru/upload/files/190715-ii.pdf (Дата обращения 27.03.2024).

8. Бизнес переходит на искусственный интеллект. / РБК+. Решения. #1 Искусственный интеллект, 5 декабря 2022. – Режим доступа: https://plus.rbc.ru/news/638ce98f7a8aa9f3126daaa2 (Дата обращения 27.03.2024).

9. Расставить нейросети. / РБК+. Инновации. #1 Искусственный интеллект, 5 декабря 2022. – Режим доступа: https://plus.rbc.ru/news/638ce67f7a8aa9e27b22f26e (Дата обращения 27.03.2024).

10. «Технологии позволяют учитывать специфику каждой отрасли». / РБК+. От первого лица. #1 Искусственный интеллект, 5 декабря 2022. – Режим доступа: https://plus.rbc.ru/news/638ced6c7a8aa9e28f7bf148 (Дата обращения 27.03.2024).

Статьи:

1. Акулин, Е.В. Специфика и особенности задач системного анализа //Актуальные проблемы теории и практики развития научных – 2022. – С. 20.

2. Алфимов, В.А. Использование R/S-анализа и фрактальной теории при анализе финансовых временных рядов //Современные наука и образование: достижения и перспективы развития – 2021. – С. 8–13.

3. Будасова, В.А. Методы технического анализа рынка //Цифровая экономика-инструмент и среда общественного развития – 2021. – С. 18–21.

4. Городнова, Н.В. Применение искусственного интеллекта в бизнес-сфере: современное состояние и перспективы // Вопросы инновационной экономики. – 2021. – Том 11. – № 4. – С. 1473–1492.

5. Доржиева, В.В. Цифровизация промышленности: роль искусственного интеллекта и возможности для России // Вопросы инновационной экономики. – 2022. – Т. 12. № 4. – С. 2383–2394.

6. Еременко, К. Работа с данными в любой сфере. Как выйти на новый уровень, используя аналитику – Москва, 2018. – С. 20–58.

7. Звягин, Л.С. Использование прикладного системного анализа как инструмента моделирования для управления бизнесом //Хроноэкономика – 2019. – №. 7 (20). – С. 26–31.

8. Кондуров, И.В.; Тушев А.Н. Лидерство бизнес-и системного аналитика на IT-рынке //Программно-техническое обеспечение автоматизированных систем – 2021. – С. 17–22.

9. Кондуров, И.В.; Тушев А.Н. Технические основы системного аналитика для успешной коммуникации с командой разработки //Высокопроизводительные вычислительные системы и технологии – 2020. – Т. 4. – №. 2. – С. 91–95.

10. Красов А.В., Штеренберг С.И., Фахрутдинов Р.М., Рыжаков Д.В., Пестов И.Е. – Текст: электронный. //Анализ информационной безопасности предприятия на основе сбора данных пользователей с открытых ресурсов и мониторинга информационных ресурсов с использованием машинного обучения // T-Comm: Телекоммуникации и транспорт – 2018. Том 12. №10. – Код доступа: https://cyberleninka.ru/article/n/analiz-informatsionnoy-bezopasnosti-predpriyatiya-na-osnove-sbora-dannyh-polzovateley-s-otkrytyh-resursov-i-monitoringa (Дата обращения 27.03.2024).

11. Львович, И.Я. Проблемы методологии проектирования интеллектуальных информационных систем // Информационные технологии в управлении, автоматизации и мехатронике. – 2020. – С. 120–123.

12. Люкевич, И.Н.; Горбатенко, И.И.; Пынзарь, Е.Г. Цифровые технологии финансовых рынков: платформы технического анализа //Фундаментальные и прикладные исследования в области управления, экономики и торговли – 2021. – С. 101–112.

13. Панкратова, Н.Д.; Панкратов В.А. Роль и место системного анализа в практической деятельности //Системный анализ в проектировании и управлении – 2019. – Т. 23. – №. 1. – С. 31–40.

14. Родионова, П.Д. Применение цифровых технологий на рынке ценных бумаг //Кластеризация цифровой экономики: Глобальные вызовы – 2020. – С. 205–209.

15. Садовский, Г.Л. Применение больших данных и систем аналитики для эффективного управления проектами //Управление научно-техническими проектами – 2020. – С. 225–228.

16. Субботин, А.В.; Тагирова, Л.Ф. Математическое моделирование информационных процессов проектирования интеллектуальных систем на основе использования метода Мамдани // Информационные технологии как основа прогрессивных научных исследований. – 2020. – С. 95–99.

17. Чижик, В.П. Сравнительная характеристика методов фундаментального и технического анализа финансовых активов //Сибирский торгово-экономический журнал. – 2013. – №. 1 (17). – С. 49.

Информационное сопровождение.

Электронные образовательные ресурсы:

1. Воронцов К. В. Машинное обучение: курс лекций // MachineLearning.ru. - Режим доступа:http://www.recognition.su/wiki/index.php?title=Машинное_обучение_(курс_лекций%2С_К.В.Воронцов). (Дата обращения 27.03.2024).

2. Бесплатные материалы по Data Engineering от преподавателей МФТИ. – Режим доступа: https://fpmi-edu.ru/free-de (Дата обращения 27.03.2024).

3. Курс «Big Data и Data Science: начни погружение с нуля». – Режим доступа: https://stepik.org/course/101687/promo (Дата обращения 27.03.2024).

4. Open Machine Learning Course. – Режим доступа: https://mlcourse.ai/book/index.html (Дата обращения 27.03.2024).

Электронные информационные ресурсы:

1. Сайт УНТИ 2035 «Обучение в области искусственного интеллекта» / [Электронный ресурс] – Режим доступа: https://ai.2035.university/ (Дата обращения 27.03.2024).

2. Сайт Национального проекта «Цифровая экономика» / [Электронный ресурс] – Режим доступа: https://национальныепроекты.рф/projects/tsifrovaya-ekonomika (Дата обращения 27.03.2024).

3. Сайт федерального проекта «Искусственный интеллект» Национального проекта «Цифровая экономика» / [Электронный ресурс] – Режим доступа: https://национальныепроекты.рф/projects/tsifrovaya-ekonomika/p-iskusstvennyy-intellekt-p (Дата обращения 27.03.2024).

4. Сайт образовательной платформы ООО «1Т» / [Электронный ресурс] – Режим доступа: https://data.1t.ru/ (Дата обращения 27.03.2024).

5. Раздел «Искусственный интеллект» на сайте РБК / [Электронный ресурс] – Режим доступа: https://trends.rbc.ru/trends/tag/ai (Дата обращения 27.03.2024).

Описание системы оценки качества освоения модуля 2.

Контроль результатов освоения дисциплины осуществляется в ходе текущего контроля успеваемости и промежуточной аттестации.

Текущий контроль предусмотрен в ходе изучения каждой темы.

Формами текущего контроля являются тесты и выполнение практических работ.

Тесты содержат не менее 5 вопросов с одним или несколькими правильными ответами. За каждый правильный ответ ставится 1 балл. Критерием прохождения теста является получение не менее 75 % правильных ответов.

Примеры тестов.

Модуль 2.

Раздел 2.1. Математическое моделирование, машинное обучение и оптимизация.

Тема: Выбор и обработка данных для машинного обучения. Методы машинного обучения.

Вопросы:

1. Какой алгоритм машинного обучения используется для задачи классификации?

a) Градиентный бустинг

b) Решающее дерево

c) Линейная регрессия

d) Все вышеперечисленные.

2. Какой метод Машинного Обучения используется для уменьшения размерности пространства признаков?

a) Случайный лес

b) k-ближайших соседей

c) Метод главных компонент

d) Линейная регрессия.

3. Какой метод обучения используется для задачи кластеризации?

a) Метод k-средних

b) Логистическая регрессия

c) Дерево решений

d) Градиентный бустинг.

4. Что такое функция потерь (loss function) в машинном обучении?

a) Функция, которая считает расстояние между признаками

b) Функция, которая определяет, насколько предсказание модели отличается от правильного ответа

c) Функция, которая подбирает оптимальные веса для модели

d) Функция, которая оценивает скорость работы модели.

5. Что такое переобучение (overfitting) в машинном обучении?

a) Когда модель слишком проста и не может обработать большое количество данных

b) Когда модель слишком сложная и слишком хорошо подстраивается под обучающие данные, что приводит к плохой работе на новых данных

c) Когда модель не учитывает взаимосвязь между признаками

d) Когда модель не может обработать большое количество данных из-за ограниченности вычислительных ресурсов.

Выполнение практических работ оценивается в бинарной системе: зачтено / не зачтено.

Зачтено: задача решена, могут быть недочеты и неточности в решении.

Не зачтено: задача не решена.

Примеры практических работ.

Модуль 2.

Раздел 2.2. Глубокое обучение и нейронные сети.

Тема 4. Использование предварительно обученных моделей для классификации изображений и других задач.

Задача: Распознавание объектов на фотографиях (Object Recognition in Photographs).

CIFAR-10 (классификация небольших изображений по десяти классам: самолет, автомобиль, птица, кошка, олень, собака, лягушка, лошадь, корабль и грузовик).

Задачи:

1. Ознакомиться со сверточными нейронными сетями.
2. Изучить построение модели в Keras в функциональном виде.
3. Изучить работу слоя разреживания (Dropout).

Требования:

1. Построить и обучить сверточную нейронную сеть.
2. Исследовать работу сеть без слоя Dropout.
3. Исследовать работу сети при разных размерах ядра свертки.

Решения практических заданий отправьте в файле формата ".ipynb", используя редакторы кода Jupyter / Google Colab.

Промежуточная аттестация.

Промежуточная аттестация проводится в формате решения практико-ориентированной задачи (кейса).

Примеры практико-ориентированных задач (кейсов).

Модуль 2.

Задание 1.

Компания-застройщик планирует построить новый жилой комплекс и нуждается в предсказании цен на недвижимость для разных типов квартир. Для этого они собирают

данные о продажах аналогичных недвижимостей в округе за последние несколько лет, а также описательные характеристики каждой квартиры, такие как количество комнат, этаж, площадь, наличие балкона и т.д.

Для предсказания цен на недвижимость мы можем использовать математическую модель, такую как линейная регрессия, которая может установить связь между описательными характеристиками квартиры и ее стоимостью. Однако, также может быть эффективным использование алгоритмов машинного обучения, таких как случайный лес или градиентный бустинг, которые могут учитывать нелинейные взаимодействия между характеристиками квартиры и ее стоимостью.

Для решения этой задачи нам необходимо:

1. Собрать данные о продажах аналогичных недвижимостей и их описательных характеристиках.
2. Подготовить данные, очистив их от выбросов, заполнив пропущенные значения и преобразовав категориальные признаки в числовые.
3. Выбрать модель машинного обучения или математическую модель, которая будет использоваться для предсказания цен на недвижимость.
4. Обучить модель на тренировочных данных и проверить ее на тестовых данных.
5. Оценить точность и качество модели и, если необходимо, произвести ее настройку.
6. Применить обученную модель для предсказания цен на недвижимость для новых квартир, которые войдут в состав нового жилого комплекса.

Ссылку на репозиторий с кодом прислать в чат с преподавателем.

Задание 2.

Детектирование объектов на изображении с помощью сверточных нейронных сетей для автоматической проверки качества продукции.

Во многих производственных процессах требуется контролировать качество продукции, чтобы избежать выпуска некачественных изделий и увеличить эффективность производства. Одним из способов автоматического контроля качества может быть использование методов глубокого обучения, таких как сверточные нейронные сети, для детектирования дефектов на изображении продукции.

Для выполнения задачи требуется собрать данные изображений продукции, на которых присутствуют дефекты, и провести их предварительную обработку и очистку, например, изменить размер изображений, привести к одному формату и т.д. Затем

необходимо применить сверточные нейронные сети для создания модели детектирования дефектов на изображении.

Модель должна быть обучена на исторических данных с отметками о наличии или отсутствии дефектов на изображении, а затем протестирована на новых данных для оценки ее точности и эффективности. Результаты работы модели могут использоваться для автоматического контроля качества продукции и сокращения количества некачественных изделий, что позволит увеличить эффективность производства и снизить затраты на ремонт и замену дефектных изделий.

Ссылку на репозиторий с кодом прислать в чат с преподавателем.

Критерии и шкала оценивания:

0-4 балла: имеются содержательные и логические ошибки, решение кейса не найдено.

5-6 баллов: решение кейса в целом найдено, но оно неоптимально и/или имеются логические ошибки.

7-8 баллов: решение кейса найдено, но имеются неточности в решении.

9-10 баллов: решение кейса найдено, ошибки отсутствуют.

Максимально возможное число баллов за работу – 10.

не менее 9 баллов – «отлично».

7-8 баллов – «хорошо».

5-6 баллов – «удовлетворительно».

0-4 балла – «неудовлетворительно».

5. ОРГАНИЗАЦИОННО-ПЕДАГОГИЧЕСКИЕ УСЛОВИЯ РЕАЛИЗАЦИИ ПРОГРАММЫ.

Специфика организационных действий и педагогических условий.

Для достижения планируемых результатов обучение строится с использованием следующих:

методов: case-study, метод проектов, модульное обучение, проблемное обучение, контекстное обучение;

форм: лекции с использованием мультимедиа, практические занятия, самостоятельная работа.

Обучение строится с применением **технологий** электронного обучения и дистанционных образовательных технологий на образовательной платформе https://data.1t.ru/**.**

Кадровое обеспечение программы (преподавательский состав).

К реализации программы привлечены представители образовательных организаций высшего образования и представители компаний со стажем работы в области искусственного интеллекта и в смежных областях.

Представители образовательных организаций высшего образования имеют высшее образование, ученую степень кандидата или доктора наук, стаж научно-педагогической работы более трех лет, а также не выполняют функции иностранного агента.

Представители компаний со стажем работы в области искусственного интеллекта и в смежных областях имеют опыт решения практических задач с использованием технологий искусственного интеллекта более 3 лет в течение последних 10 лет в профильной компании или в профильном подразделении, а также не выполняют функции иностранного агента.

Члены преподавательского состава имеет за последние 3 года научные публикации, соответствующие направлению данной программы, в журналах, включенных в перечень ВАК (К1 и К2), а также в журналах, включенных в «Белый список» Минобрнауки РФ 1-го квартиля.

№ п/п	Фамилия, имя, отчество (при наличии)	Место основной работы и должность, ученая степень и ученое звание (при наличии)	Ссылки на веб-страницы с портфолио (при наличии)
1.	Борисов Вадим Владимирович *«Белый список» – 4 шт, из них 1-ого квартиля – 1шт.* *ВАК К1 –2 шт.*	Профессор кафедры вычислительной техники, филиал НИУ «МЭИ» в г. Смоленске, д.т.н., профессор	
2.	Санников Даниил Александрович	Главный аналитик данных, ПАО «Сбербанк»	
3.	Кропивный Дмитрий Алексеевич *ВАК К1 – 1 шт.*	Ведущий аналитик данных, ООО «1Т»	https://data.1t.ru/kropivnyy
4.	Жукова Людмила Вячеславовна *«Белый список» – 1 шт.* *ВАК К1 – 3 шт.* *ВАК К2 – 1 шт.*	Доцент кафедры «Магистерская школа информационных бизнес-систем», НИТУ МИСИС, к.э.н.	https://data.1t.ru/zhukova
5.	Хусаинов Наиль Шавкятович *ВАК К2 – 2 шт.*	Заведующий кафедрой, Институт компьютерных технологий и информационной безопасности, ФГАОУ ВО «Южный федеральный университет», к.т.н.	
6.	Клавдеев Александр Владимирович *ВАК К1 – 1 шт.*	Старший аналитик данных, ООО «1Т»	https://data.1t.ru/klavdeev
7.	Шарапов Никита Александрович	Аналитик-исследователь, ООО «1Т»	
8.	Кулакова Надежда Сергеевна	Старший аналитик данных, ООО «1Т»	
9.	Зиновьев Дмитрий Владимирович	Системный аналитик, ООО «1Т»	
10.	Лашков Дмитрий Юрьевич	Старший Аналитик данных, ООО «1Т»	
11.	Костин Алексей Николаевич *ВАК К1 – 1 шт.*	Ведущий преподаватель по ИИ, ООО «1Т»	https://data.1t.ru/kostin
12.	Королева Диана Олеговна *«Белый список» – 3 шт., из них 1-го квартиля – 2 шт.* *ВАК К1 – 3 шт.*	Заведующая лабораторией инноваций в образовании, НИУ ВШЭ	

Материально-технические условия реализации программы

Вид занятий	Наименование оборудования, программного обеспечения
Лекционные занятия	Персональный компьютер с установленным на нем: Windows 10-11, x64/x86; от 8 Gb RAM; от 128 Gb SSD/HDD, монитор от 15”; сетевой интерфейс Fast Ethernet 100Мбит; веб-браузеры Google Chrome, Mozila Firefox, Opera, Microsoft Edge, Яндекс.Браузер и др.
Практические занятия, самостоятельная работа, промежуточная и итоговая аттестация	Персональный компьютер с установленным на нем: Windows 10 и выше, x64/x86; от 8 Gb RAM; от 128 Gb SSD/HDD, монитор от 15”; сетевой интерфейс Fast Ethernet 100Мбит; веб-браузеры Google Chrome, Mozila Firefox, Opera, Microsoft Edge, Яндекс.Браузер и др. Anaconda 2.7 или 3.5 Доступ к облачным вычислительным ресурсам

Материально-технические условия соответствуют действующим санитарным и противопожарным правилам и нормам.

При проведении учебных занятий с применением дистанционных образовательных технологий (ДОТ) у слушателя должен быть персональный компьютер, оснащенный аудиоколонками, с доступом в сеть интернет и установленным видеоплеером, способным воспроизводить видеофайлы.

Выдаваемый документ при успешном освоении программы

Удостоверение о повышении квалификации ООО «1Т».

6. СИСТЕМА ОЦЕНКИ КАЧЕСТВА ОСВОЕНИЯ ПРОГРАММЫ.

В систему оценки качества освоения программы входят:

1) текущий контроль;

2) промежуточная аттестация;

3) итоговая аттестация.

Формы, методы проведения и оценочные материалы текущего контроля и промежуточной аттестации представлены в соответствующих рабочих программах модулей.

Для зачисления на программу потенциальному слушателю необходимо пройти входную диагностику (вступительное испытание).

Входная диагностика (вступительное испытание).

Входная диагностика (вступительное испытание) проводится в тестовой форме.

1. Какая команда используется для создания новой папки?

a) mkdir

b) rd

c) cd

d) dir

2. Какая команда используется для вывода содержимого текстового файла в консоль?

a) cat

b) grep

c) head

d) tail

3. Какая команда используется для перемещения файла в другую директорию?

a) cp

b) move

c) mv

d) rename

4. Какая команда используется для поиска файлов в Linux?

a) find

b) search

c) locate

d) lookup

5. Какая команда используется для вывода процесса в консоль?

a) ps

b) top

c) ls

d) dir

6. Какая команда используется для выбора данных из таблицы в SQL?

a) SELECT

b) DELETE

c) INSERT

d) UPDATE

7. Какая команда используется для удаления данных из таблицы в SQL?

a) SELECT

b) DELETE

c) INSERT

d) UPDATE

8. Какая команда используется для создания таблицы в SQL?

a) CREATE

b) DELETE

c) INSERT

d) UPDATE

9. Какая команда используется для выбора уникальных значений из таблицы в SQL?

a) UNIQUE

b) SELECT DISTINCT

c) SELECT UNIQUE

d) DIST

10. Какая команда используется для упорядочивания результатов запроса по возрастанию в SQL?

a) ORDER ASC

b) ASCENDING

c) SORT

d) ORDER BY

11. Какая команда используется в Python для определения функций?

a) def

b) fun

c) func

d) function

12. Как обратиться к последнему элементу в списке my_list?

a) my_list[-1]

b) my_list[:1]

c) my_list[-2]

d) my_list[1:]

13. Что выведется на экран в результате выполнения следующего кода?

```
a = 2
b = 3
print(a ** b)
```

a) 5

b) 8

c) 6

d) 2

14. Что произойдет при выполнении следующего кода?

```
my_str = "Hello, World!"
print(my_str[7:])
```

a) На экран будет выведено "World!"

b) На экран будет выведено "Hello,"

c) На экран будет выведено "orld!"

d) На экран будет выведено "ello, World!"

15. Какая функция используется в Python для получения длины строки, списка или кортежа?

a) size()

b) len()

c) length()

d) count()

16. Что такое Git?

a) Онлайн-сервис для общения с людьми.

b) Система управления версиями.

c) База данных.

d) Текстовый редактор.

17. Как создать новый репозиторий Git?

a) git new

b) git create

c) git init

d) git start

18. Как добавить файлы в индекс Git?

a) git stage

b) git add

c) git commit

d) git push

19. Что такое команда Git clone?

a) Копирует содержимое удаленного репозитория на локальный компьютер.

b) Позволяет создать новый репозиторий.

c) Отменяет последний коммит.

d) Удаляет локальный репозиторий и его историю.

20. Какая команда используется для создания новой ветки в Git?

a) git branch

b) git checkout

c) git commit

d) git merge.

Пороговое значение для успешного прохождения вступительного испытания - не менее 65% от общего количества результатов выполнения заданий.

Итоговая аттестация.

Итоговая аттестация проходит в форме решения практико-ориентированных задач (кейсов).

Практико-ориентированные задачи (кейсы) для итоговой аттестации предоставлены организациями, работающими в области искусственного интеллекта и смежных областях (ООО «Альмира»), а также организациями, применяющими технологии интеллектуальной обработки данных для решения своих производственных задач (ООО «Верконт Сервис», ООО СП «Содружество», ООО ЭЦ «Социология и аналитика», ПК «ИТ Союз»).

Практико-ориентированная задача (кейс) 1. Исследование надёжности заёмщиков – анализ банковских данных.

Проблема:

Заказчик – кредитный отдел банка. Нужно разобраться, влияет ли семейное положение и количество детей клиента на факт погашения кредита в срок. Входные данные

от банка — статистика о платёжеспособности клиентов. Результаты исследования будут учтены при построении модели кредитного скоринга — специальной системы, которая оценивает способность потенциального заёмщика вернуть кредит банку.

Задача:

Ответить на следующие вопросы:

– Есть ли зависимость между наличием детей и возвратом кредита в срок?

– Есть ли зависимость между семейным положением и возвратом кредита в срок?

– Есть ли зависимость между уровнем дохода и возвратом кредита в срок?

– Как разные цели кредита влияют на его возврат в срок?

Описание данных:

– children – количество детей в семье

– days_employed – общий трудовой стаж в днях

– dob_years – возраст клиента в годах

– education – уровень образования клиента

– education_id – идентификатор уровня образования

– family_status – семейное положение

– family_status_id – идентификатор семейного положения

– gender – пол клиента

– income_type – тип занятости

– debt – имел ли задолженность по возврату кредитов

– total_income – ежемесячный доход

– purpose – цель получения кредита

Практико-ориентированная задача (кейс) 2. Анализ рынка недвижимости.

Проблема:

В вашем распоряжении данные сервиса недвижимости – архив объявлений о продаже квартир в Санкт-Петербурге и соседних населённых пунктах за несколько лет. Нужно научиться определять рыночную стоимость объектов недвижимости. Ваша задача – установить параметры. Это позволит построить автоматизированную систему: она отследит аномалии и мошенническую деятельность. По каждой квартире на продажу доступны два вида данных. Первые вписаны пользователем, вторые – получены автоматически на основе картографических данных. Например, расстояние до центра, аэропорта, ближайшего парка и водоёма.

Задача:

– Изучите следующие параметры: площадь, цена, число комнат, высота потолков. Постройте гистограммы для каждого параметра.

– Изучите время продажи квартиры. Постройте гистограмму. Посчитайте среднее и медиану. Опишите, сколько обычно занимает продажа. Когда можно считать, что продажи прошли очень быстро, а когда необычно долго?

– Уберите редкие и выбивающиеся значения. Опишите, какие особенности обнаружили.

– Какие факторы больше всего влияют на стоимость квартиры? Изучите, зависит ли цена от площади, числа комнат, удалённости от центра. Изучите зависимость цены от того, на каком этаже расположена квартира: первом, последнем или другом. Также изучите зависимость от даты размещения: дня недели, месяца и года.

– Выберите 10 населённых пунктов с наибольшим числом объявлений. Посчитайте среднюю цену квадратного метра в этих населённых пунктах. Выделите среди них населённые пункты с самой высокой и низкой стоимостью жилья. Эти данные можно найти по имени в столбце 'locality_name'.

– Изучите предложения квартир: для каждой квартиры есть информация о расстоянии до центра. Выделите квартиры в Санкт-Петербурге ('locality_name'). Ваша задача – выяснить, какая область входит в центр. Создайте столбец с расстоянием до центра в километрах: округлите до целых значений. После этого посчитайте среднюю цену для каждого километра. Постройте график: он должен показывать, как цена зависит от удалённости от центра. Определите границу, где график сильно меняется – это и будет центральная зона.

– Выделите сегмент квартир в центре. Проанализируйте эту территорию и изучите следующие параметры: площадь, цена, число комнат, высота потолков. Также выделите факторы, которые влияют на стоимость квартиры (число комнат, этаж, удалённость от центра, дата размещения объявления). Сделайте выводы. Отличаются ли они от общих выводов по всему городу?

Описание данных:

– airports_nearest – расстояние до ближайшего аэропорта в метрах (м)

– balcony – число балконов

– ceiling_height – высота потолков (м)

– cityCenters_nearest – расстояние до центра города (м)

– days_exposition – сколько дней было размещено объявление (от публикации до снятия)

– first_day_exposition – дата публикации

– floor – этаж

– floors_total – всего этажей в доме

– is_apartment – апартаменты (булев тип)

– kitchen_area – площадь кухни в квадратных метрах (м²)

– last_price – цена на момент снятия с публикации

– living_area – жилая площадь в квадратных метрах(м²)

– locality_name – название населённого пункта

– open_plan – свободная планировка (булев тип)

– parks_around3000 – число парков в радиусе 3 км

– parks_nearest – расстояние до ближайшего парка (м)

– ponds_around3000 – число водоёмов в радиусе 3 км

– ponds_nearest – расстояние до ближайшего водоёма (м)

– rooms – число комнат

– studio – квартира-студия (булев тип)

– total_area – площадь квартиры в квадратных метрах (м²)

– total_images – число фотографий квартиры в объявлении

Практико-ориентированная задача (кейс) 3. Аналитика данных для iot на колесах.

Описание проекта

В последнее время автомобили становятся все более умными и связанными с интернетом. В этом проекте вы будете работать с данными, собранными с IOT устройств на автомобиле, и производить анализ данных, который поможет оптимизировать эксплуатацию автомобиля и улучшить удобство его использования.

Основные задачи

Сбор и обработка данных: Написание скрипта для сбора данных с IOT устройств в автомобиле, а также обработка данных для последующего анализа.

Визуализация данных: Создание интерактивной дашборды, которая позволит пользователям просматривать данные, получаемые с автомобиля в режиме реального времени, а также проводить анализ данных за определенный период.

Анализ данных: Использование алгоритмов машинного обучения для определения оптимального времени для прохождения техобслуживания автомобиля, а также анализ данных о расходе топлива для определения оптимальных условий вождения, которые позволят сократить расход топлива и увеличить срок службы автомобиля.

Технологии

— Python для сбора, обработки и анализа данных.

— Flask для создания веб-приложения.

— Библиотеки matplotlib, plotly и seaborn для визуализации данных.

— Библиотеки Scikit-Learn для анализа данных и построения моделей машинного обучения.

Возможные улучшения

— Добавление функционала, который позволит автомобилю автоматически записывать данные и отправлять их на сервер для дальнейшего анализа.

— Создание еще более точных моделей машинного обучения, которые могут давать предупреждения в режиме реального времени, если автомобиль находится в критическом состоянии.

— Создание мобильного приложения, которое позволит водителям видеть данные об их автомобиле и его производительности в режиме реального времени.

Практико-ориентированная задача (кейс) 4. Анализ данных в авиакомпании.

Проблема:

Вам предстоит изучить базу данных и проанализировать спрос пассажиров на рейсы в города, где проходят крупнейшие фестивали для российской авиакомпания.

Авиакомпания выполняет внутренние пассажирские авиаперевозки. Сотни перелётов каждый день. Важно понять предпочтения пользователей, покупающих билеты на те или иные направления.

Задача:

Выбрать топ-10 городов по количеству рейсов;

Построить графики: модели самолетов и количество рейсов, города и количество рейсов, топ-10 городов и количество рейсов;

Сделать выводы по каждому из графиков, пояснить результат.

Описание данных:

Таблица airports – информация об аэропортах:

airport_code – трёхбуквенный код аэропорта

airport_name – название аэропорта

city – город

timezone – временная зона

Таблица aircrafts – информация об самолётах:

aircraft_code – код модели самолёта

model – модель самолёта

range – количество самолётов

Таблица tickets – **информация о билетах:**

ticket_no – уникальный номер билета

passenger_id – персональный идентификатор пассажира

passenger_name – имя и фамилия пассажира

Таблица flights – **информация о рейсах:**

flight_id – уникальный идентификатор рейса

departure_airport – аэропорт вылета

departure_time – дата и время вылета

arrival_airport – аэропорт прилёта

arrival_time – дата и время прилёта

aircraft_code – id самолёта

Таблица ticket_flights – **стыковая таблица «рейсы-билеты»**

ticket_no – номер билета

flight_id – идентификатор рейса

Таблица festivals – **информация о фестивалях**

festival_id – уникальный номер фестиваля

festival_date – дата проведения фестиваля

festival_city – город проведения фестиваля

festival_name – название фестиваля

Практико-ориентированная задача (кейс) 5. Оптимизация маркетинговых затрат.

Проблема:

Вас пригласили на стажировку в отдел маркетинговой аналитики в компанию по продаже билетов. Первое задание: помочь маркетологам снизить расходы – отказаться от невыгодных источников трафика и перераспределить бюджет.

Есть данные Компании с июня 2017 по конец мая 2018 года:

– лог сервера с данными о посещениях сайта компании,

– выгрузка всех заказов за этот период,

– статистика рекламных расходов.

Вам предстоит изучить:

– как клиенты пользуются сервисом,

– когда делают первые покупки на сайте,

– сколько денег приносит компании каждый клиент,

– когда расходы на привлечение клиента окупаются.

Задача:

Рассчитайте DAU, WAU и MAU. Вычислите средние значения этих метрик за весь период. Отобразите изменения метрик во времени на графиках;

Определите, сколько раз за день пользователи в среднем заходят на сайт. Постройте график, отражающий изменения метрики во времени;

Исследуйте, сколько времени пользователи проводят на сайте. Узнайте продолжительность типичной пользовательской сессии за весь период. Чтобы выбрать подходящую среднюю меру, постройте график распределения.

Рассчитайте Retention Rate, применяя когортный анализ. Покажите изменения метрики во времени на графике. Найдите средний Retention Rate на второй месяц «жизни» когорт.

Описание данных:

Файл visits_log.csv хранит лог сервера с информацией о посещениях сайта,

orders_log.csv – информацию о заказах,

a costs.csv – информацию о расходах на маркетинг.

Структура visits_log.csv

Uid – уникальный идентификатор пользователя,

Device – категория устройства пользователя,

Start Ts – дата и время начала сессии,

End Ts – дата и время окончания сессии,

Source Id – идентификатор источника перехода на сайт.

Структура orders_log.csv

Uid – уникальный идентификатор пользователя,

Buy Ts – дата и время заказа,

Revenue – сумма заказа.

Структура costs.csv

source_id – идентификатор рекламного источника,

dt – дата проведения рекламной кампании,

costs – расходы на эту кампанию.

Практико-ориентированная задача (кейс) 6. Проверка гипотез по увеличению выручки в интернет-магазине – оценка результатов a/b теста.

Проблема:

Вы аналитик крупного интернет-магазина. Вместе с отделом маркетинга вы подготовили список гипотез для увеличения выручки и провели A/B-тест.

Задача:

Сделайте выводы и предположения:

Постройте график кумулятивной выручки по группам.

Постройте график кумулятивного среднего чека по группам.

Постройте график относительного изменения кумулятивного среднего чека группы B к группе A.

Постройте график кумулятивной конверсии по группам.

Постройте график относительного изменения кумулятивной конверсии группы B к группе A.

Постройте точечный график количества заказов по пользователям.

Посчитайте 95-й и 99-й перцентили количества заказов на пользователя. Выберите границу для определения аномальных пользователей.

Постройте точечный график стоимостей заказов.

Посчитайте 95-й и 99-й перцентили стоимости заказов. Выберите границу для определения аномальных заказов.

Посчитайте статистическую значимость различий в конверсии между группами по «сырым» данным.

Посчитайте статистическую значимость различий в среднем чеке заказа между группами по «сырым» данным.

Посчитайте статистическую значимость различий в конверсии между группами по «очищенным» данным.

Посчитайте статистическую значимость различий в среднем чеке заказа между группами по «очищенным» данным.

Описание данных:

Файл /datasets/orders.csv. Скачать датасет

transactionId – идентификатор заказа;

visitorId – идентификатор пользователя, совершившего заказ;

date – дата, когда был совершён заказ;

revenue – выручка заказа;

group – группа A/B-теста, в которую попал заказ.

Файл /datasets/visitors.csv. Скачать датасет

date – дата;

group – группа A/B-теста;

visitors – количество пользователей в указанную дату в указанной группе A/B-теста.

Практико-ориентированная задача (кейс) 7. Исследования рынка общепита в Москве для принятия решения об открытии нового заведения.

Проблема:

Вы решили открыть небольшое кафе в Москве. Оно оригинальное – гостей должны обслуживать роботы. Проект многообещающий, но дорогой. Вместе с партнёрами вы решились обратиться к инвесторам. Их интересует текущее положение дел на рынке – сможете ли вы снискать популярность на долгое время, когда все зеваки насмотрятся на роботов-официантов? Вы гуру аналитики, и партнёры просят вас подготовить исследование рынка. У вас есть открытые данные о заведениях общественного питания в Москве.

Задача:

Исследуйте соотношение видов объектов общественного питания по количеству. Постройте график.

Исследуйте соотношение сетевых и несетевых заведений по количеству. Постройте график.

Для какого вида объекта общественного питания характерно сетевое распространение?

Что характерно для сетевых заведений: много заведений с небольшим числом посадочных мест в каждом или мало заведений с большим количеством посадочных мест?

Для каждого вида объекта общественного питания опишите среднее количество посадочных мест. Какой вид предоставляет в среднем самое большое количество посадочных мест? Постройте графики.

Выделите в отдельный столбец информацию об улице из столбца address .

Постройте график топ-10 улиц по количеству объектов общественного питания. Воспользуйтесь внешней информацией и ответьте на вопрос – в каких районах Москвы находятся эти улицы.

Найдите число улиц с одним объектом общественного питания. Воспользуйтесь внешней информацией и ответьте на вопрос – в каких районах Москвы находятся эти улицы.

Посмотрите на распределение количества посадочных мест для улиц с большим количеством объектов общественного питания. Какие закономерности можно выявить?

Сделайте общий вывод и дайте рекомендации о виде заведения, количестве посадочных мест, а также районе расположения. * Прокомментируйте возможность развития сети.

Описание данных

Таблица rest_data:

id – идентификатор объекта;

object_name – название объекта общественного питания;

chain – сетевой ресторан;

object_type – тип объекта общественного питания;

address – адрес;

number – количество посадочных мест.

Практико-ориентированная задача (кейс) 8. Анализ данных социальных медиа для оценки влияния музыкальных фестивалей на туризм.

Цель проекта.

Целью проекта является анализ влияния музыкальных фестивалей на туризм в разных регионах мира. Выполнение проекта потребует использования навыков сбора, обработки и анализа данных социальных медиа, проведения статистического анализа результатов и презентации полученных выводов.

Описание данных.

Для проекта необходимо собрать данные социальных медиа с использованием API, а также данные о музыкальных фестивалях с помощью веб-скрейпинга. Для каждого региона мира будут рассмотрены фестивали, проведенные за последние несколько лет.

Задачи проекта.

1. Собрать данные о музыкальных фестивалях в различных регионах мира.
2. Собрать данные социальных медиа для каждого фестиваля.
3. Использовать инструменты обработки данных для анализа активности в социальных медиа.
4. Оцените влияние фестивалей на туризм в каждом регионе.
5. Построить графики и графики для иллюстрации результатов.
6. Подготовить отчет о результате и презентовать его перед публикой.

Используемые инструменты.

— Язык программирования Python для сбора, обработки и анализа данных.

— Библиотеки Python, такие как pandas, numpy, matplotlib, seaborn, scipy, для обработки и визуализации данных.

— API различных социальных медиа для сбора данных.

— Библиотеки для веб-скрейпинга данных фестивалей.

Важные замечания.

1. При сборе данных социальных медиа необходимо соблюдать правила использования API социальных медиа.

2. При использовании данных социальных медиа необходимо учитывать этические соображения в отношении конфиденциальности пользователей.

3. Результаты должны быть представлены в понятном и доступном формате.

4. Необходимо использование аналитических и статистических методов для обработки данных.

В результате успешного выполнения проекта будет получена ценная информация о влиянии музыкальных фестивалей на туризм в различных регионах мира. Эта информация может быть использована для принятия решений при планировании туристических мероприятий и для развития туризма в регионах, где проводятся музыкальные фестивали.

Практико-ориентированная задача (кейс) 9. Анализ пользовательского поведения в мобильном приложении.

Проблема.

Вы работаете в стартапе, который продаёт продукты питания. Нужно разобраться, как ведут себя пользователи вашего мобильного приложения.

Изучите воронку продаж. Узнайте, как пользователи доходят до покупки. Сколько пользователей доходит до покупки, а сколько «застревает» на предыдущих шагах? На каких именно?

После этого исследуйте результаты A/A/B-эксперимента. Дизайнеры захотели поменять шрифты во всём приложении, а менеджеры испугались, что пользователям будет непривычно. Договорились принять решение по результатам A/A/B-теста. Пользователей разбили на 3 группы: 2 контрольные со старыми шрифтами и одну экспериментальную — с новыми. Выясните, какой шрифт лучше.

Создание двух групп A вместо одной имеет определённые преимущества. Если две контрольные группы окажутся равны, вы можете быть уверены в точности проведенного тестирования. Если же между значениями A и A будут существенные различия, это поможет обнаружить факторы, которые привели к искажению результатов. Сравнение контрольных групп также помогает понять, сколько времени и данных потребуется для дальнейших тестов.

В случае общей аналитики и A/A/B-эксперимента работайте с одними и теми же данными. В реальных проектах всегда идут эксперименты. Аналитики исследуют качество работы приложения по общим данным, не учитывая принадлежность пользователей к экспериментам.

Задача.

Ответьте на следующие вопросы:

Сколько всего событий в логе?

Сколько всего пользователей в логе?

Сколько в среднем событий приходится на пользователя?

Сколько пользователей в каждой экспериментальной группе?

Есть 2 контрольные группы для A/A-эксперимента, чтобы проверить корректность всех механизмов и расчётов. Проверьте, находят ли статистические критерии разницу между выборками 246 и 247.

Выберите самое популярное событие. Посчитайте число пользователей, совершивших это событие в каждой из контрольных групп. * Посчитайте долю пользователей, совершивших это событие. Проверьте, будет ли отличие между группами статистически достоверным. * Проделайте то же самое для всех других событий (удобно обернуть проверку в отдельную функцию). Можно ли сказать, что разбиение на группы работает корректно?

Аналогично поступите с группой с изменённым шрифтом. Сравните результаты с каждой из контрольных групп в отдельности по каждому событию. Сравните результаты с объединённой контрольной группой. Какие выводы из эксперимента можно сделать?

Какой уровень значимости вы выбрали при проверке статистических гипотез выше? Посчитайте, сколько проверок статистических гипотез вы сделали. При уровне значимости 0.1 каждый десятый раз можно получать ложный результат. Какой уровень значимости стоит применить? Если вы хотите изменить его, проделайте предыдущие пункты и проверьте свои выводы.

Описание данных.

Каждая запись в логе – это действие пользователя, или событие.

EventName – название события;

DeviceIDHash – уникальный идентификатор пользователя;

EventTimestamp – время события;

ExpId – номер эксперимента: 246 и 247 – контрольные группы, а 248 – экспериментальная.

Практико-ориентированная задача (кейс) 10. Создание дашборда по пользовательским событиям для агрегатора новостей.

Проблема.

Вы работаете аналитиком в Яндекс.Дзене. Почти всё ваше время занимает анализ пользовательского взаимодействия с карточками статей. Каждую карточку определяют её тема и источник (у него тоже есть тема). Примеры тем: «Красота и здоровье», «Россия»,

«Путешествия». Пользователей системы характеризует возрастная категория. Скажем, «26-30» или «45+».

Есть три способа взаимодействия пользователей с системой: Карточка отображена для пользователя (show); Пользователь кликнул на карточку (click); Пользователь просмотрел статью карточки (view).

Процесс пора автоматизировать – нужно сделать дашборд.

Задача.

Нужно сделать дашборд. Дашборд будет основываться на пайплайне, который будет брать данные из таблицы, в которых хранятся сырые данные, трансформировать данные и укладывать их в агрегирующую таблицу. Пайплайн будет разработан для вас дата-инженерами.

Каждую неделю менеджеры задают вам одни и те же вопросы: Сколько взаимодействий пользователей с карточками происходит в системе с разбивкой по темам карточек? Как много карточек генерируют источники с разными темами? Как соотносятся темы карточек и темы источников?

Описание данных

record_id 30745 non-null int64

item_topic 30745 non-null object

source_topic 30745 non-null object

age_segment30745 non-null object

dt 30745 non-null datetime64[ns]

visits

Практико-ориентированная задача (кейс) 11. Прогнозирование вероятности оттока пользователей для фитнес-центров.

Проблема.

Сеть фитнес-центров «Культурист-датасаентист» разрабатывает стратегию взаимодействия с клиентами на основе аналитических данных. Распространённая проблема фитнес-клубов и других сервисов – отток клиентов. Для фитнес-центра можно считать, что клиент попал в отток, если за последний месяц ни разу не посетил спортзал. Конечно, не исключено, что он уехал на Бали и по приезде обязательно продолжит ходить на фитнес. Однако чаще бывает наоборот. Если клиент начал новую жизнь с понедельника, немного походил в спортзал, а потом пропал – скорее всего, он не вернётся.

Чтобы бороться с оттоком, отдел по работе с клиентами «Культуриста-датасаентиста» перевёл в электронный вид множество клиентских анкет.

Задача

Провести анализ и подготовить план действий по удержанию клиентов.

Научиться прогнозировать вероятность оттока (на уровне следующего месяца) для каждого клиента.

Сформировать типичные портреты клиентов: выделить несколько наиболее ярких групп и охарактеризовать их основные свойства.

Проанализировать основные признаки, наиболее сильно влияющие на отток.

Сформулировать основные выводы и разработать рекомендации по повышению качества работы с клиентами:

— выделить целевые группы клиентов;

— предложить меры по снижению оттока;

— определить другие особенности взаимодействия с клиентами.

Описание данных

'Churn' – факт оттока в текущем месяце;

Данные клиента за предыдущий до проверки факта оттока месяц:

'gender' – пол;

'Near_Location' – проживание или работа в районе, где находится фитнес-центр;

'Partner' – сотрудник компании-партнёра клуба (сотрудничество с компаниями, чьи сотрудники могут получать скидки на абонемент – в таком случае фитнес-центр хранит информацию о работодателе клиента);

Promo_friends – факт первоначальной записи в рамках акции «приведи друга» (использовал промо-код от знакомого при оплате первого абонемента);

'Phone' – наличие контактного телефона;

'Age' – возраст;

'Lifetime' – время с момента первого обращения в фитнес-центр (в месяцах).

Информация на основе журнала посещений, покупок и информация о текущем статусе абонемента клиента:

'Contract_period' – длительность текущего действующего абонемента (месяц, 3 месяца, 6 месяцев, год);

'Month_to_end_contract' – срок до окончания текущего действующего абонемента (в месяцах);

'Group_visits' – факт посещения групповых занятий;

'Avg_class_frequency_total' – средняя частота посещений в неделю за все время с начала действия абонемента;

'Avg_class_frequency_current_month' – средняя частота посещений в неделю за предыдущий месяц;

'Avg_additional_charges_total' – суммарная выручка от других услуг фитнес-центра: кафе, спорттовары, косметический и массажный салон.

Практико-ориентированная задача (кейс) 12. Анализ поведения пользователей в мобильном приложении.

Проблема.

Ваша задача – помочь команде маркетинга лучшим образом подобрать целевую аудиторию для баннерной рекламы в мобильном приложении «Встречайте». Для того, чтобы рекламная кампания прошла наиболее эффективно, вам необходимо на основе действий пользователей в мобильном приложении подобрать наиболее активную группу пользователей.

Задача.

Проанализируйте связь целевого события – просмотра контактов – и других действий пользователей.

Оцените, какие действия чаще совершают те пользователи, которые просматривают контакты.

Проведите исследовательский анализ данных.

Проанализируйте влияние событий на совершение целевого события.

Проверьте статистические гипотезы:

– Одни пользователи совершают действия tips_show и tips_click, другие — только tips_show. Проверьте гипотезу: конверсия в просмотры контактов различается у этих двух групп.

– Сформулируйте собственную статистическую гипотезу. Дополните её нулевой и альтернативной гипотезами. Проверьте гипотезу с помощью статистического теста.

Описание данных

Колонки в "mobile_sources.csv":

userId – идентификатор пользователя;

source – источник, с которого пользователь установил приложение.

Колонки в "mobile_dataset.csv":

event.time – время совершения,

user.id – идентификатор пользователя,

event.name – действие пользователя.

Виды действий:

advert_open – открыл карточки объявления,

photos_show – просмотрел фотографий в объявлении,

tips_show – увидел рекомендованные объявления,

tips_click – кликнул по рекомендованному объявлению,

contacts_show и show_contacts – посмотрел номер телефона,

contacts_call – позвонил по номеру из объявления,

map – открыл карту объявлений,

search_1 – search_7 – разные действия, связанные с поиском по сайту,

favorites_add – добавил объявление в избранное.

Практико-ориентированная задача (кейс) 13. Анализ программы лояльности.

Проблема.

Менеджер магазина строительных материалов "Строили, строили и наконец построили", отвечающий за программу лояльности клиентов, хочет оценить её эффективность. Цель проекта - на основе анализа программы лояльности магазина сформулировать предложения о повышении эффективности программы.

Задача.

Графики количества уникальных покупателей в день, неделю и месяц с разбивкой на две категории (с и без программы лояльности)

Графики среднего чека в день, неделю и месяц с разбивкой на две категории (с и без программы лояльности)

Графики среднего размера корзины в день, неделю и месяц с разбивкой на две категории (с и без программы лояльности)

Топ 10 самых часто продаваемых товаров с картой и без

Магазины, в которых пользуются картой лояльности

Количество возвратов с картой лояльности и без неё

Описание данных.

Таблица retail_dataset содержит следующие данные:

purchaseId – id чека

item_ID – id товара

purchasedate – дата покупки

Quantity – количество товара

CustomerID – id покупателя

ShopID – id магазина

loyalty_program – участвует ли покупатель в программе лояльности

Таблица product_codes содержит два столбца:

productID – id товара

price_per_one – стоимость одной единицы товара.

Практико-ориентированная задача (кейс) 14. Анализ данных билетного агрегатора.

Проблема.

Мы стажируемся на работу в отдел аналитики **J.K**. Было получено задание помочь маркетологам оптимизировать маркетинговые затраты.

В распоряжении есть данные от J.K. с июня 2017 по конец мая 2018 года:

лог сервера с данными о посещениях сайта J.K.;

выгрузка всех заказов за этот период;

статистика рекламных расходов.

Задача.

Сколько денег потратили? Всего / на каждый источник / по времени?

Сколько стоило привлечение одного покупателя из каждого источника?

Помесячная выручка с отображением вклада каждой когорты

Описание данных.

Таблица **visits** (лог сервера с информацией о посещениях сайта):

Uid – уникальный id пользователя;

Device – тип устройства пользователя;

Start Ts – дата и время начала сессии;

End Ts – дата и время окончания сессии;

Source Id – id рекламного источника, из которого пришел пользователь.

Таблица **orders** (информация о заказах):

Uid – уникальный id пользователя, который сделал заказ;

Buy Ts – дата и время заказа;

Revenue – выручка J.K. с этого заказа.

Таблица **costs** (информация о затратах на маркетинг):

source_id – id рекламного источника;

dt – дата;

costs – затраты на этот рекламный источник в этот день.

Практико-ориентированная задача (кейс) 15. Анализ бизнес-показателей развлекательного приложения Procrastinate pro+.

Проблема.

Вам необходимо провести полное исследование пользователей вашего мобильного приложения для грамотного планирования маркетинговой кампании.

Задача.

На основе данных, предоставленных компанией, необходимо провести анализ и ответить на вопросы:

откуда приходят пользователи и какими устройствами они пользуются;

сколько стоит привлечение пользователей из различных рекламных каналов;

сколько денег приносит каждый клиент;

когда расходы на привлечение клиента окупаются;

какие факторы мешают привлечению клиентов.

Описание данных.

Данные о пользователях, привлечённых с 1 мая по 27 октября 2019 года:

лог сервера с данными об их посещениях (файл visits_info_short.csv)

User Id – уникальный идентификатор пользователя,

Region – страна пользователя,

Device – тип устройства пользователя,

Channel – идентификатор источника перехода,

Session Start – дата и время начала сессии,

Session End – дата и время окончания сессии;

выгрузка их покупок за этот период (файл orders_info_short.csv)

User Id – уникальный идентификатор пользователя,

Event Dt – дата и время покупки,

Revenue – сумма заказа;

рекламные расходы (файл costs_info_short.csv)

Channel – идентификатор рекламного источника,

Dt – дата проведения рекламной кампании,

Costs – расходы на эту кампанию.

Практико-ориентированная задача (кейс) 16. Банки — сегментация пользователей по потреблению продуктов.

Задача.

Анализ и сегментация клиентов регионального банка по количеству потребляемых продуктов:

проведение исследовательского анализа данных;

сегментация пользователей на основе данных о количестве потребляемых продуктов,

формулировка и проверка статистических гипотез:

– гипотеза различия дохода между теми клиентами, которые пользуются двумя продуктами банка, и теми, которые пользуются одним;

– гипотеза о различии в скоринговых баллах действующих и ушедших клиентов банка.

Описание данных.

Датасет содержит данные о клиентах банка, файл banc_dataset.csv.

Колонки:

userid – идентификатор пользователя,

score – баллы кредитного скоринга,

City – город,

Gender – пол,

Age – возраст,

Objects – количество объектов в собственности,

Balance – баланс на счёте,

Products – количество продуктов, которыми пользуется клиент,

CreditCard – есть ли кредитная карта,

Loyalty – активный клиент,

estimated_salary – заработная плата клиента,

Churn – ушёл или нет.

Практико-ориентированная задача (кейс) 17. Выявление профилей потребления в e-commerce.

Проблема.

Сегментация покупателей по истории их покупок для создания специальных предложений.

Задача.

Разобьем товары на категории.

ТОП 5 продаваемых товаров в каждой категории.

Графики выручки и количества покупателей по категориям.

Сегментируем наших покупателей по истории покупок.

Сегментируем покупателей по числовым признакам.

Смотрим сезонность товаров по категориям.

Сравниваем полученные сегменты по средним значениям метрик.

График средней выручки с покупателя по сегментам покупателей.

График количества заказов по сегментам покупателей.

Описание данных.

Файл ecommerce_dataset.csv содержит колонки:

date – *Дата заказа*

customer_id – *Идентификатор покупателя*

order_id – *Идентификатор заказа*

product – *Наименование товара*

quantity – *Количество товара в заказе*

price – *Цена товара*

Практико-ориентированная задача (кейс) 18. Анализ данных о продажах магазина одежды.

Задача.

Вы получили информацию о продажах одежды по месяцам.

Необходимо ответить на вопросы:

Назовите общую сумму продаж в рублях за июль?

Какая общая сумма прибыли (продажи-себестоимость) в рублях в октябре?

Назовите артикул (новый), который в рейтинге продаж (в рублях) был на 5 месте в августе?

Какой остаток на складе ТОП-10 артикулов по продажам (в рублях) в августе?

Определите, нормальное ли распределение ‘Продаж, шт’? Вычислите квадратное отклонение выборки, а также в каком диапазоне лежат 95 % значений.

Постройте график распределения количества буквенных размеров ('M', 'L', 'S', 'XL', 'XS', 'XXL', 'XXXL', 'XXS') в выборке.

Описание данных.

Артикул новый - код цветомодели (например, код красной футболки), у разных размеров одной цветомодели один код. Состоит из 16 символов (12 + 4 через пробел).

Размер – размер модели (или его отсутствие "no size", если вещь без размера).

Остаток на складе, шт. – остатки товара на складе.

Себестоимость, руб. – себестоимость одной единицы товара.

Цена продажи, руб. – цена, за которую товар продавался в прошлом году.

Практико-ориентированная задача (кейс) 19. Анализ сервиса электронных книг.

Проблема.

Вы – один из членов молодой команды по разработке сервиса продажи и чтения электронных книг. Вашу компанию можно назвать не иначе как «стартап». Вам предстоит тщательно проанализировать всю доступную в данный момент информацию по вашему продукту для улучшения пользовательского опыта и повышения базы клиентов.

Задача.

проанализировать информацию о книгах, издательствах, авторах;

проанализировать пользовательские обзоры книг.

Необходимо рассчитать для каждой книги:

количество обзоров;

среднюю оценку.

Определить издательство, которое выпустило наибольшее число книг толще 50 страниц (в целях исключения из анализа брошюр).

Определить автора с самой высокой средней оценкой книг (необходимо учитывать только книги с 50 и более оценками).

Рассчитать среднее количество обзоров от пользователей, которые поставили больше 50 оценок.

Описание данных

Таблица books – содержит информацию о книгах

book_id – идентификатор книги;

author_id – идентификатор автора;

title – название книги;

num_pages – количество страниц;

publication_date – дата публикации книги;

publisher_id – идентификатор издателя.

Таблица authors – содержит информацию об авторах

author_id – идентификатор автора;

author – имя автора.

Таблица publishers – содержит информацию об издательствах

publisher_id – идентификатор издательства;

publisher – название издательства.

Таблица ratings – содержит информацию о пользовательских оценках

rating_id – идентификатор оценки;

book_id – идентификатор книги;

username – имя пользователя, оставившего оценку;

rating – оценка книги.

Таблица reviews – содержит информацию о пользовательских обзорах

review_id – идентификатор обзора;

book_id – идентификатор книги;

username – имя пользователя, написавшего обзор;

text – текст обзора.

Практико-ориентированная задача (кейс) 20. Исследование эффективности операторов кол-центров.

Проблема.

Ваш заказчик – Продуктовый менеджер из кол-центра, который оказывает услуги бизнесу. Ваша задача – определить критерии неэффективности оператора для дальнейшей разработки функционала автоматизации сервиса, влияющего на повышение качества оказываемой услуги.

Задачи.

Определить критерии неэффективности оператора.

Выявить проблемы клиента кол-центра, вызванные неэффективной работой оператора.

Выявить кол-центры с наибольшим числом неэффективных операторов и сформировать.

Сформировать рекомендации для автоматизации, позволяющие улучшить сервис.

Описание данных.

Данные – Для исследования взяты следующие данные:

Таблица telecom_dataset (информация о звонках, принимаемых и совершаемых операторами кол-центра):

user_id – Идентификатор клиентского аккаунта в сервисе

date – Дата статистики

direction – Направление вызовов (out - исходящий вызов, in — входящий вызов)

internal – Является ли звонок внутренним звонком между операторами клиента

operator_id – Идентификатор оператора

is_missed_call – Является ли звонок пропущенным

calls_count – Количество звонков

call_duration – Длительность звонка (без учета времени ожидания)

total_call_duration – Длительность звонка (с учетом времени ожидания)

Таблица telecom_clients (информация о клиентах кол-центра):

user_id – Идентификатор клиентского аккаунта в сервисе

tariff_plan – Текущий тарифный план клиента

date_start – Дата регистрации клиентов в сервисе.

Практико-ориентированная задача (кейс) 21. Разработка дорожной карты по внедрению программного обеспечения на стороне заказчика.

Проблема.

Вы ведущий аналитик крупной консалтинговой ИТ-компании.

Ваш заказчик – группа компаний, которая является одной из крупнейших инвестиционно-финансовых компаний России и предоставляет услуги по направлениям: инвестиционная деятельность, внешнеэкономическая деятельность и недвижимость.

ИТ-департамент заказчика обслуживает сложную территориально-распределенную структуру, состоящую из 10 филиалов на территории Российской Федерации, в которых на данный момент работает 2000 сотрудников.

Ваш проект – внедрение ПО «Автоматизированная система поддержки пользователей ИТ-услуг». В вашу компанию обратился вышеописанный заказчик с некоторым перечнем бизнес-проблем, касающихся своего ИТ-департамента. Вы посетили предприятие заказчика, провели переговоры с персоналом и изучили его бизнес-процессы.

Вам необходимо по итогам вашей работы на территории заказчика написать коммерческое предложение, в котором Вы изложите концептуальный подход к решению задач, стоящих перед ним, отразите ресурсы и возможности исполнителя (вашей консалтинговой компании) по решению указанных задач с использованием предлагаемой ПО «Автоматизированная система поддержки пользователей ИТ-услуг».

Задача.

Подробно описать ваш проект по внедрению специализированного ПО в следующем виде:

Обзор текущей ситуации: обзор текущей ситуации у заказчика.

Цели и задачи проекта: содержит цели и задачи проекта, сформулированные исполнителем на основании ожиданий заказчика от проекта.

Подход к решению задачи: описывает основные аспекты подхода исполнителя к решению задач заказчика.

Планируемый состав работ: содержит развернутый перечень планируемых мероприятий.

Бизнес-выгоды от внедрения решения: бизнес-выгоды, которые могут быть получены при решении задачи заказчика.

Команда проекта: описание таких важных параметров услуг Исполнителя, как имеющийся опыт, уровень качества и гарантия постоянного состава команды консультантов.

Критерии оценивания, шкала оценивания.

0-4 балла: имеются содержательные и логические ошибки, решение кейса не найдено.

5-6 баллов: решение кейса в целом найдено, но оно неоптимально и/или имеются логические ошибки.

7-8 баллов: решение кейса найдено, но имеются неточности в решении.

9-10 баллов: решение кейса найдено, ошибки отсутствуют.

Максимально возможное число баллов – 10.

не менее 9 баллов – «отлично».

7-8 баллов – «хорошо».

5-6 баллов – «удовлетворительно».

0-4 балла – «неудовлетворительно».

В ходе итоговой аттестации обучающиеся должны продемонстрировать следующие ключевые знания, умения и навыки:

Знать:

- основные определения искусственного интеллекта и больших данных;
- различия между машинным обучением, нейронными сетями, глубоким обучением и EDA;
- основные конструкции языка Python;
- основы Hadoop;
- основы теории баз данных;
- принципы работы NoSQL баз данных;
- язык запросов к СУБД - SQL;
- основные уровни представления данных;
- основные ETL процессы и инструменты;
- особенности организации СУБД в MPP-системе;
- основные типы данных в СУБД Postgres;
- особенности колоночного формата хранения данных;
- принципы построения дашбордов;
- основные понятия теории вероятности;

- основы комбинаторики;
- понятие A/B-тестирования;
- особенности продуктовой аналитики;
- существующие и перспективные методы и программный инструментарий технологий больших данных.

Уметь:

- производить аналитику для интеллектуального отслеживания ресурсов/процессов;
- применять SQL базы данных для прикладных решений;
- применять язык программирования Python и библиотеки при разработке решений на основе ИИ;
- осуществлять поиск и структурирование данных;
- визуализировать анализируемые данные;
- применять методы анализа на графах;
- создавать собственные модели данных с использованием UML-диаграмм;
- производить расчет вероятностных показателей с использованием языка Python;
- проводить аналитические и поисковые исследования по тематике информационных технологий, технологий больших данных;
- осуществлять математическое и информационное моделирование;
- разрабатывать технические проекты в сфере информационных технологий;
- решать прикладные задачи и участвовать в реализации проектов в области сквозной цифровой субтехнологии «Компьютерное зрение»;
- осуществлять массово параллельную обработку и анализ данных;
- оценивать результаты моделирования и определять критерии качества построенных моделей.

Владеть:

- навыками решения базовых аналитических кейсов с использованием инструментов визуализации;
- навыками использования статистических методов исследования;
- навыками расчета статистических показателей с использованием языка Python;
- методами разработки моделей машинного обучения и нейронных сетей;
- математическими методами анализа данных;

– навыками создания нескольких таблиц в СУБД Postgres посредствам Dbeaver;

– навыками интеллектуального анализа данных с помощью языка программирования R;

– навыками построения полносвязной нейронной сети для задачи классификации;

– навыками обучения нейронных сетей с помощью PyTorch, TensorFlow и Keras;

– навыками расчета ключевых метрик роста продукта с помощью Python;

– навыками настраивания кластеров Apache Spark и Hive на Hadoop;

– инструментами Weka,RapidMiner, Knime, Orange IBM SPSS Modeler, Tableau и др.;

– навыками использования баз данных (MongoDB, Clickhouse и др.).

Тугой И.А.	Академический директор ООО «1Т»
Разработчики программы:	
Борисов Вадим Владимирович	Профессор кафедры вычислительной техники, филиал НИУ «МЭИ» в г. Смоленске, д.т.н., профессор
Санников Даниил Александрович	Главный аналитик данных ПАО «Сбербанк»
Кропивный Дмитрий Алексеевич	Ведущий аналитик данных ООО «1Т»
Жукова Людмила Вячеславовна	Доцент кафедры «Магистерская школа информационных бизнес-систем», НИТУ МИСИС, к.э.н.
Хусаинов Наиль Шавкятович	Заведующий кафедрой Института компьютерных технологий и информационной безопасности, ФГАОУ ВО «Южный федеральный университет», к.т.н.
Клавдеев Александр Владимирович	Старший аналитик данных ООО «1Т»
Шарапов Никита Александрович	Аналитик-исследователь ООО «1Т»
Кулакова Надежда Сергеевна	Старший аналитик данных ООО «1Т»
Зиновьев Дмитрий Владимирович	Системный аналитик ООО «1Т»
Лашков Дмитрий Юрьевич	Старший аналитик данных ООО «1Т»
Костин Алексей Николаевич	Ведущий преподаватель по ИИ ООО «1Т»

Королева Диана Олеговна	Заведующая лабораторией инноваций в образовании НИУ ВШЭ