МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»  
(ФГБОУ ВО «ВГТУ», ВГТУ)

Факультет информационных технологий и компьютерной безопасности

Кафедра компьютерных интеллектуальных технологий проектирования

ОТЧЁТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №1

По дисциплине: «Системы хранения и обработки данных»

Тема: «Сравнительный анализ профессий в области искусственного интеллекта»

Выполнил работу студент группы мИИВТ-241: Бокарев С.А.

подпись, дата

Принял: Короленко В.В.

подпись, дата

Воронеж 2024

Цель лабораторной работы: Изучение основных профессий в области искусственного интеллекта.

Основные задачи: Описание профессии «Специалист по анализу данных (Data Scientist)».

1) Необходимо описать профессию, обратить внимание на следующие аспекты:

• типовые задачи, решаемые специалистом;

• требуемое образование, необходимые знания и навыки;

• средства, используемые специалистом в профессиональной деятельности (оборудование, программное обеспечение, методические материалы, информационные ресурсы);

• команда, в которой работает соответствующий специалист, роль в команде, описание взаимодействия с членами команды;

• соотношение со смежными профессиями;

• перспективы профессионального роста для соответствующего специалиста.

2) Подготовить отчёт о проделанной работе и доклад.

Специалист по анализу данных выполняет разнообразные задачи, которые могут варьироваться в зависимости от конкретной области и компании.

1. Сбор и подготовка данных:

- Сбор данных из различных источников (базы данных, API, веб-скрапинг).

- Очистка и предобработка данных (удаление дубликатов, обработка пропусков, нормализация).

2. Анализ данных:

- Проведение описательной статистики для понимания распределения и характеристик данных.

- Визуализация данных с помощью графиков и диаграмм для выявления закономерностей.

3. Моделирование:

- Построение предсказательных моделей с использованием методов машинного обучения (регрессия, классификация, кластеризация).

- Выбор и настройка алгоритмов для достижения наилучших результатов.

4. Оценка моделей:

- Оценка качества моделей с использованием метрик (точность, полнота, F1-мера и т.д.).

- Проведение кросс-валидации для проверки устойчивости моделей.

5. Интерпретация результатов:

- Интерпретация полученных результатов и их представление в понятной форме для заинтересованных сторон.

- Подготовка отчетов и презентаций для донесения выводов.

6. Оптимизация процессов:

- Использование аналитики для оптимизации бизнес-процессов и повышения эффективности.

- Разработка рекомендаций на основе анализа данных.

7. Работа с большими данными:

- Использование технологий обработки больших данных (Hadoop, Spark) для работы с объемными наборами данных.

8. Разработка и внедрение систем:

- Участие в разработке систем рекомендаций, прогнозирования спроса и других приложений на основе данных.

- Внедрение моделей в производственные системы и мониторинг их работы.

9. Сотрудничество с другими командами:

- Работа в команде с инженерами, аналитиками и бизнес-экспертами для решения комплексных задач.

10. Обучение и развитие:

- Постоянное обучение новым методам и технологиям в области анализа данных и машинного обучения.

Эти задачи требуют сочетания навыков программирования, статистики, аналитического мышления и понимания бизнес-процессов.

Для того чтобы стать специалистом по анализу данных, требуется сочетание образования, знаний и навыков. Обычно требуется степень бакалавра в области математики, статистики, компьютерных наук, инженерии или смежных областях. Многие Data Scientists имеют также степень магистра или доктора. Дополнительные курсы по машинному обучению, анализу данных, статистике и программированию могут быть полезны. Сертификаты от известных платформ (Coursera, edX, DataCamp) также добавляют ценности. К необходимым знаниям относятся:

1. Статистика и математика: Основы статистики, теории вероятностей, линейной алгебры и математического анализа.

2. Программирование: Умение программировать на языках, таких как Python или R. Знание SQL для работы с базами данных.

3. Машинное обучение: Понимание алгоритмов машинного обучения (регрессия, классификация, кластеризация) и их применение.

4. Обработка данных: Знания о библиотеках для обработки и анализа данных (Pandas, NumPy) и визуализации (Matplotlib, Seaborn).

5. Большие данные: Основы работы с технологиями для обработки больших данных (Hadoop, Spark).

Из необходимых навыков выделим:

1. Аналитическое мышление: Способность анализировать данные и выявлять закономерности.

2. Коммуникационные навыки: Умение объяснять сложные концепции простым языком и представлять результаты анализа.

3. Проблемное мышление: Способность подходить к задачам креативно и находить эффективные решения.

4. Работа в команде: Умение работать в междисциплинарной команде с различными специалистами.

5. Управление проектами: Основы управления проектами для эффективного выполнения задач.

Дополнительными навыками могут быть: опыт работы с инструментами BI (Tableau, Power BI), знания о том, как интегрировать данные из различных источников через API, а также знание специфики отрасли и бизнес-процессов для более точного анализа.

Специалист по анализу данных использует разнообразные средства и инструменты в своей профессиональной деятельности.

Компьютеры и ноутбуки: Мощные устройства с достаточным объемом оперативной памяти и производительными процессорами для обработки больших объемов данных.

Серверы: Для работы с большими данными и выполнения сложных вычислений могут использоваться серверы или облачные платформы.

Программное обеспечение:

1. Языки программирования:

- Python: Основной язык для анализа данных, включает библиотеки, такие как Pandas, NumPy, SciPy, Matplotlib и Scikit-learn.

- R: Популярен в статистическом анализе и визуализации данных.

2. Инструменты для работы с базами данных:

- SQL: Для извлечения и манипуляции данными в реляционных базах данных.

- NoSQL базы: MongoDB, Cassandra для работы с неструктурированными данными.

3. Средства для машинного обучения:

- TensorFlow и Keras: Для создания и обучения нейронных сетей.

- PyTorch: Альтернатива TensorFlow, популярная в научных кругах.

4. Инструменты визуализации данных:

- Tableau, Power BI: Для создания интерактивных дашбордов и отчетов.

- Matplotlib, Seaborn: Библиотеки для визуализации данных в Python.

5. Облачные платформы:

- AWS, Google Cloud, Microsoft Azure: Предлагают услуги хранения данных, вычислений и машинного обучения.

Методические материалы:

1. Книги: Литература по статистике, машинному обучению и анализу данных.

2. Онлайн-курсы: Платформы, такие как Coursera, edX, Udacity, предлагают курсы по анализу данных и машинному обучению.

3. Научные статьи: Исследования и публикации в области анализа данных и машинного обучения.

Информационные ресурсы:

1. Форумы и сообщества: Stack Overflow, Kaggle, Reddit для обмена опытом и решения проблем.

2. Блоги и подкасты: Ресурсы, посвященные аналитике данных, машинному обучению и новым технологиям.

3. Документация и учебные материалы: Официальная документация библиотек и инструментов (например, документация Pandas, Scikit-learn).

Эти средства помогают специалистам по анализу данных эффективно решать задачи, связанные с обработкой, анализом и визуализацией данных.

Специалист по анализу данных обычно работает в междисциплинарной команде, состоящей из различных профессионалов, каждый из которых вносит свой вклад в общий процесс анализа данных и принятия решений. Рассмотрим основные роли в команде и взаимодействие между членами команды.

1. Data Engineer:

- Роль: Отвечает за создание и поддержку инфраструктуры для сбора, хранения и обработки данных.

- Взаимодействие: Data Scientist работает с Data Engineer для получения необходимых данных, а также для оптимизации процессов извлечения и трансформации данных.

2. Аналитик данных (Data Analyst):

- Роль: Проводит первичный анализ данных, создает отчеты и визуализации.

- Взаимодействие: Data Scientist может сотрудничать с аналитиками для понимания бизнес-требований и использования их отчетов как основы для более глубокого анализа.

3. Инженер по машинному обучению (Machine Learning Engineer):

- Роль: Разрабатывает и внедряет модели машинного обучения в продуктивную среду.

- Взаимодействие: Data Scientist разрабатывает модели и передает их ML Engineer для развертывания и мониторинга.

4. Бизнес-аналитик (Business Analyst):

- Роль: Определяет бизнес-требования и помогает интерпретировать результаты анализа.

- Взаимодействие: Data Scientist работает с бизнес-аналитиком для понимания потребностей бизнеса и формулирования гипотез, которые могут быть протестированы с помощью анализа данных.

5. Продуктовый менеджер (Product Manager):

- Роль: Определяет стратегию продукта и отвечает за его развитие.

- Взаимодействие: Data Scientist предоставляет аналитические данные, которые помогают продуктовым менеджерам принимать обоснованные решения о функциональности и улучшениях продукта.

6. Разработчики (Software Developers):

- Роль: Создают программное обеспечение и приложения, которые используют данные.

- Взаимодействие: Data Scientist может работать с разработчиками для интеграции моделей в приложения или системы.

Взаимодействие в команде:

- Командные собрания: Регулярные встречи для обсуждения текущих задач, статуса проектов и обмена идеями.

- Совместная работа над проектами: Использование инструментов управления проектами (например, Jira, Trello) для отслеживания задач и совместной работы.

- Обмен знаниями: Проведение семинаров или презентаций, где члены команды делятся своими находками, методами и результатами.

- Обратная связь: Регулярное получение обратной связи от коллег для улучшения качества работы и моделей.

- Документация: Ведение документации по проектам, чтобы все члены команды могли легко получить доступ к необходимой информации.

Эффективное взаимодействие между членами команды позволяет специалисту по анализу данных не только глубже понимать данные, но и применять свои находки для решения реальных бизнес-задач.

Специалист по анализу данных имеет тесные связи с несколькими смежными профессиями, каждая из которых вносит свой вклад в процесс работы с данными:

1. Аналитик данных (Data Analyst)

- Обязанности: Анализирует данные, создает отчеты и визуализации, помогает принимать бизнес-решения.

- Соотношение: Data Analyst фокусируется на описательном анализе и отчетности, в то время как Data Scientist использует более сложные методы, такие как машинное обучение и предсказательная аналитика. Data Scientist может использовать отчеты аналитиков как основу для дальнейшего анализа.

2. Инженер по данным (Data Engineer)

- Обязанности: Создает и поддерживает инфраструктуру для обработки и хранения данных.

- Соотношение: Data Engineer обеспечивает Data Scientist необходимыми данными и инструментами. Без надежной инфраструктуры Data Scientist не сможет эффективно работать с данными.

3. Инженер по машинному обучению (Machine Learning Engineer)

- Обязанности: Разрабатывает, тестирует и внедряет модели машинного обучения в продуктивную среду.

- Соотношение: Data Scientist разрабатывает модели и алгоритмы, а ML Engineer отвечает за их внедрение и оптимизацию. Оба специалиста работают в тесной связке для достижения общих целей.

4. Бизнес-аналитик (Business Analyst)

- Обязанности: Определяет бизнес-требования и помогает интерпретировать результаты анализа.

- Соотношение: Бизнес-аналитик помогает Data Scientist понять, какие данные важны для бизнеса, а также интерпретировать результаты анализа в контексте бизнес-целей.

5. Продуктовый менеджер (Product Manager)

- Обязанности: Определяет стратегию и развитие продукта.

- Соотношение: Data Scientist предоставляет аналитические данные, которые помогают продуктовым менеджерам принимать решения о функциональности и улучшениях продукта.

6. Разработчик программного обеспечения (Software Developer)

- Обязанности: Создает приложения и системы, которые могут использовать данные.

- Соотношение: Data Scientist может работать с разработчиками для интеграции моделей в приложения или системы, обеспечивая функциональность на основе данных.

7. Специалист по визуализации данных (Data Visualization Specialist)

- Обязанности: Создает визуализации данных для облегчения понимания информации.

- Соотношение: Data Scientist может работать с визуализаторами данных для представления сложных результатов анализа в понятной форме.

Взаимодействие между специалистом по анализу данных и смежными профессиями является ключевым для успешного выполнения проектов. Каждая роль дополняет друг друга, создавая целостный процесс от сбора данных до принятия обоснованных бизнес-решений.

Перспективы профессионального роста специалиста по анализу данных (Data Scientist) достаточно многообещающие, особенно в условиях растущего спроса на данные в различных отраслях:

1. Углубление специализации

- Машинное обучение и искусственный интеллект: Углубление знаний в области алгоритмов машинного обучения, нейронных сетей и глубокого обучения.

- Обработка больших данных: Изучение технологий, таких как Hadoop, Spark и NoSQL базы данных.

2. Руководящие должности

- Главный аналитик данных (Chief Data Officer): Ответственность за стратегию работы с данными в компании.

- Руководитель команды аналитиков (Analytics Team Lead): Управление группой специалистов, распределение задач и координация работы.

3. Переквалификация в смежные области

- Инженер по машинному обучению (Machine Learning Engineer): Переход к более технической роли, связанной с разработкой и внедрением моделей.

- Бизнес-аналитик или продуктовый менеджер: Переход к более бизнес-ориентированным ролям, где важна интерпретация данных для принятия решений.

4. Консультирование

- Консультант по данным: Предоставление экспертных услуг другим компаниям по вопросам анализа данных и внедрения аналитических решений.

5. Научная деятельность и исследования

- Исследователь в области данных: Работа в академической среде или научных учреждениях, занимающихся передовыми исследованиями в области анализа данных.

6. Обучение и преподавание

- Преподаватель или тренер: Передача знаний другим, работа в университетах или обучающих центрах.

7. Собственный бизнес

- Создание стартапа: Разработка собственных продуктов или услуг на основе анализа данных, например, создание аналитических платформ или консалтинговых агентств.

Профессиональный рост специалиста по анализу данных зависит от желания учиться, адаптироваться к новым технологиям и расширять свои навыки. С учетом текущих трендов и потребностей рынка, перспективы для карьерного роста остаются весьма многообещающими.

**Вывод**: в данной лабораторной работе были изучены основные профессии в области искусственного интеллекта, в частности профессия «Аналитик данных».

**Контрольные вопросы**

1. Дайте определение термина «Искусственный интеллект».

Искусственный интеллект – технология, позволяющая системе, машине или компьютеру выполнять задачи, требующие разумного мышления, то есть имитировать поведение человека для постепенного обучения с использованием полученной информации и решения конкретных вопросов.

1. Дайте определение термина «Наука о данных».

Наука о данных – профессиональная деятельность, связанная с эффективным и максимально достоверным поиском закономерностей в данных, извлечение знаний из данных в обобщённой форме, а также их оформление в виде, пригодном для обработки заинтересованными сторонами (людьми, программными системами, управляющими устройствами) в целях принятия обоснованных решений.

1. Дайте определение термина «Машинное обучение».

Машинное обучение – это методы искусственного интеллекта, которые позволяют построить обучаемые модели для разных целей: например, автоматизации процессов, автоматического перевода текстов, распознавания изображений.

1. Дайте определение термина «Анализ данных».

Анализ данных – это область между математикой и информатикой, которая занимается построением и исследованием методов извлечения полезной информации из данных. Анализ данных позволяет преобразовать массив данных в выводы, на основе которых будут приниматься решения и строиться действия. Он охватывает большое количество методов из разных областей науки.

1. Что означает термин «Big Data» в информационных технологиях?

Big Data – это сложные и объёмные наборы разной информации. Они представлены в «сыром виде» и требуют предварительной обработки, чтобы получить из них ценные сведения, которые могут принести пользу предприятиям и организациям.

1. Какие главные характеристики Big Data?

Объем:

Количество данных – важный фактор. Располагая ими в больших количествах, Вам потребуется обрабатывать большие объемы неструктурированных данных низкой плотности. Ценность таких данных не всегда известна. Это могут быть данные каналов Twitter, данные посещаемости веб-страниц, а также данные мобильных приложений, сетевой трафик, данные датчиков. В некоторые организации могут поступать десятки терабайт данных, в другие — сотни петабайт.

Скорость:

Скорость в данном контексте – это скорость приема данных и, возможно, действий на их основе. Обычно высокоскоростные потоки данных поступают прямо в оперативную память, а не записываются на диск. Некоторые "умные" продукты, функционирующие на основе Интернета, работают в режиме реального или практически реального времени. Соответственно, такие данные требуют оценки и действий в реальном времени.

Разнообразие:

Разнообразие означает, что доступные данные принадлежат к разным типам. Традиционные типы данных структурированы и могут быть сразу сохранены в реляционной базе данных. С появлением Big Data данные стали поступать в неструктурированном виде. Такие неструктурированные и полуструктурированные типы данных как текст, аудио и видео, требуют дополнительной обработки для определения их значения и поддержки метаданных.

1. Соотношение бизнес-аналитики (Business Intelligence) и науки о данных (Data Science).

Data science и BI различаются по конечному результату, полученному характерными для них методами работы с данными. Для BI результирующим продуктом является информация, а для data science таким продуктом служат знания. Общим же является то, что как информация, так и знания извлекаются из данных с участием человека и предназначены для передачи другому человеку.

Второй фактор, определяющий различия – полнота используемых данных. Любые данные, в конечном счете, являются отражением окружающего мира, но эти отражения могут различаться по полноте представления. В BI используются только структурированные цифровые данные, дающие весьма ограниченную картину окружающего мира, а в data science могут быть использованы любые данные, достаточные для отражения картины окружающего мира с любой требуемой полнотой.

1. Что такое жизненный цикл анализа данных?

Жизненный цикл – это последовательность шагов и процессов, которые выполняются при проведении анализа данных от начала до включает в себя следующие этапы:

1. Постановка задач;

2. Сбор данных;

3. Подготовка данных;

4. Анализ данных;

5.Презентация результатов;

6. Применение результатов;

7. Мониторинг и обновление.

Жизненный цикл анализа данных позволяет систематизировать и организовать процесс работы с данными, что помогает создать надежные и точные результаты анализа.