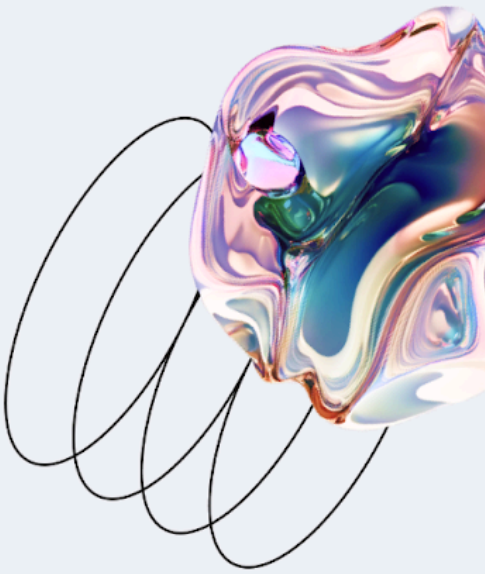
GeekBrains



Data Science. Обзор

Data Science

Оглавление

Введение

Всем привет! Мы начинаем курс по Data Science. Разберемся, что такое Data Science и данные, почему в современном мире им уделяют так много внимания, как с ними работать, как трансформировать и хранить.

Поговорим про визуализацию и представление — важный этап работы, который поможет лучше понять, какие данные у нас есть и что с ними делать — то есть выбрать стратегию, которая поможет получить пользу из данных. Обязательно затронем базы данных и научимся работать с реляционными базами данных.

В конце курса доберемся до трендовой темы — машинного обучения. Узнаем, что это и какую пользу приносит. Попытаемся понять, почему все крупные игроки на рынке инвестируют в развитие машинного обучения и стараются обогнать конкурентов. А еще разберем статистическую основу нейронных сетей и построим наши первые нейронные сети.

Словарь терминов

Данные — зарегистрированная информация, представление фактов, понятий или инструкций в форме, приемлемой для общения, интерпретации, обработки человеком или с помощью автоматических средств.

Data Science — наука о данных, цель которой — извлечение значимой информации для бизнеса. Это междисциплинарный подход, который сочетает в себе принципы и методы из математики, статистики, искусственного интеллекта и вычислительной техники для анализа больших объемов данных. Этот анализ помогает специалистам по работе с данными задавать вопросы — что произошло, почему это произошло, что произойдет и что можно сделать с результатами — и отвечать на них.

ETL — общий термин для процессов, которые происходят, когда данные переносят из нескольких систем в одно хранилище. Аббревиатура от Extract, Transform, Load — извлечение, преобразование, загрузка.

Регрессия — это метод нахождения взаимосвязи между двумя, казалось бы, не связанными между собой точками данных.

Кластеризация — это метод группировки тесно связанных данных для поиска закономерностей и аномалий.

Классификация — это сортировка данных по группам или категориям.

Что такое Data Science

Data Science (наука о данных) сочетает в себе математику и статистику, специализированное программирование, расширенную аналитику, искусственный интеллект (ИИ) и машинное обучение с конкретными предметными знаниями.

Цель Data Science — выявить значимую информацию и полезные идеи, скрытые в данных, чтобы использовать их для принятия решений и стратегического планирования.

Объем источников данных и самих данных активно растет. За счет этого Data Science стал одной из самых быстроразвивающихся областей. Компании все больше полагаются на Data Science в интерпретации данных, принятии решений, предсказании событий и улучшении бизнес-показателей в целом.

Жизненный цикл проектов по работе с данными состоит из четырех этапов. Для каждого этапа характерны свои роли, инструменты и процессы:

- **Прием данных** — первый этап жизненного цикла. Данные собирают из разных источников с помощью разных методов: ручного ввода, просмотра веб-страниц, потоковой передачи данных из систем и устройств в реальном времени. В источниках могут быть структурированные данные (например, данные клиентов) и неструктурированные (например, файлы журналов, видео, аудио, изображения, статистика интернета вещей и многое другое).

- **Хранение и обработка данных.** Форматы и структуры данных бывают разными. Поэтому компаниям нужно рассматривать разные системы хранения в зависимости от типов данных.

  Команды по управлению данными помогают установить стандарты в отношении хранения и структуры данных. Впоследствии они упростят рабочие процессы, связанные с моделями аналитики, машинного обучения и глубокого обучения. Весь процесс работы станет легче, если данные будут структурированными.

  Этап обработки данных включает в себя их очистку, дедупликацию (устранение дубликатов), преобразование и объединение данных с использованием заданий ETL (извлечение, преобразование, загрузка) или других технологий интеграции данных. Подготовка данных нужна, чтобы повысить их качество перед загрузкой в хранилище.

- **Анализ данных.** Специалисты по данным проводят предварительный анализ, чтобы изучить зависимости, закономерности, диапазоны и распределения значений в данных.

  Исследование данных стимулирует создание гипотез для A/B-тестирования. Позволяет аналитикам определять релевантность данных для построения моделей прогнозирования, аналитики, машинного обучения и глубокого обучения. В

зависимости от точности модели компании могут полагаться на эти данные при принятии бизнес-решений, что позволяет им масштабироваться.

- **Коммуникация.** В конце жизненного цикла данные представляют в виде отчетов и других визуализаций. Такой подход помогает специалистам, принимающим решения, лучше понимать данные и их влияние на бизнес.

Основные языки программирования для обработки данных — Python и R (менее популярный, отходит на второй план). Но можно использовать другие языки или инструменты с графическим интерфейсом.

Так как современные компании перегружены данными, собирать и хранить информацию помогают специальные инструменты. Например, онлайн-системы и платежные порталы собирают данные в области электронной коммерции, медицины, финансов и других аспектов жизни.

Значение термина «наука о данных» со временем менялось. Термин появился в 60-х годах как альтернативное название статистики. А в конце 90-х его формализовали профессионалы в области компьютерных наук. Предложенное определение рассматривало науку о данных как отдельную область с тремя аспектами: проектированием данных, сбором и анализом. Потребовалось еще одно десятилетие, чтобы термин стали использоваться за пределами академических кругов.

Инновации в области искусственного интеллекта и машинного обучения сделали обработку данных более быстрой и эффективной. Отраслевой спрос создал экосистему курсов, степеней и должностей в области науки о данных. Но потребность в дата-специалистах продолжает расти.

Виды анализа в Data Science

Data Science изучает данные четырьмя основными способами.

1. Описательный анализ направлен на исследование с целью получить представление о том, что произошло или что происходит в среде данных. Для него характерна визуализация: круговые диаграммы, гистограммы, линейные графики, таблицы или сгенерированные описания.

Например, служба бронирования авиабилетов может фиксировать количество билетов, забронированных каждый день. Описательный анализ выявит всплески и спады бронирований, а также месяцы с высокой востребованностью этой услуги.

2. Диагностический анализ — это глубокое изучение данных с целью понять, почему что-то произошло. Для диагностического анализа характерны детализация, обнаружение данных, интеллектуальный анализ данных и корреляции. Несколько операций с данными и преобразования могут быть выполнены с заданным набором данных, чтобы обнаружить уникальные закономерности в каждом из этих методов.

Например, служба полетов может детализировать особенно производительный месяц, чтобы лучше понять всплеск бронирования. В результате анализа можно выяснить, что многие клиенты посещают определенный город, чтобы попасть на ежемесячное спортивное мероприятие.

3. Прогностический анализ использует статистические данные, чтобы делать прогнозы. Для него характерны машинное обучение, прогнозирование, сопоставление с образцом и прогнозное моделирование. В каждом из этих методов компьютеры обучены анализировать причинно-следственные связи в данных.

Например, группа обслуживания полетов может использовать науку о данных для прогнозирования моделей бронирования рейсов на предстоящий год в начале года. Компьютерная программа или алгоритм могут анализировать прошлые данные и прогнозировать всплески бронирований для определенных направлений, например, в мае. Прогнозируя будущие потребности клиентов, компания может начать таргетированную рекламу рейсов в города, популярные у туристов в мае, уже с февраля.

4. Предписывающий анализ выводит прогностические данные на новый уровень. Позволяет не только предсказывать, что может произойти, но и предлагать оптимальную реакцию. Таким образом, можно анализировать потенциальные последствия вариантов выбора и рекомендовать лучший план действий. Метод основан на анализе графов, моделировании, обработке сложных событий, нейронных сетях и механизмах рекомендаций машинного обучения.

Вернемся к примеру с бронированием авиабилетов. Предписывающий анализ может рассмотреть исторические маркетинговые кампании, чтобы максимизировать преимущество предстоящего всплеска бронирования. Исследователь данных поможет прогнозировать результаты бронирования для разных уровней маркетинговых расходов по различным маркетинговым каналам. Эти прогнозы дали бы компании по бронированию авиабилетов большую уверенность в принятии маркетинговых решений.

Data Science и бизнес-аналитика

Может быть легко спутать термины «наука о данных» (Data Science) и «бизнес-аналитика» (Business intelligence, BI) — оба они относятся к данным организации и анализу этих данных, но различаются по направленности.

Бизнес-аналитика (BI) — общий термин для технологии, которая обеспечивает подготовку данных, их анализ, управление ими и визуализацию. Инструменты и процессы бизнес-аналитики позволяют конечным пользователям извлекать полезную информацию из необработанных данных, облегчать принятие решений.

В то время как инструменты обработки данных во многом пересекаются, бизнес-аналитика больше фокусируется на данных из прошлого, а информация, полученная с помощью BI-инструментов, носит более описательный характер. Бизнес-аналитика использует данные, чтобы понять, что произошло раньше, чтобы определить курс дальнейших действий.

BI ориентирован на статические (неизменные) данные, которые обычно структурированы. А наука о данных использует описательные данные для определения прогностических переменных, которые затем используются для классификации данных или для составления прогнозов.

Наука о данных и бизнес-аналитика не исключают друг друга — компании могут использовать и то, и то, чтобы полнее понять данные и извлечь из них ценность. Как правило, хороший дата-сайентист должен разбираться в процессах бизнес-аналитики и понимать базовую теорию.

Преимущества науки о данных для бизнеса

Наука о данных меняет методы работы компаний. Многим компаниям, независимо от их размера, нужна надежная стратегия обработки данных, чтобы стимулировать рост и поддерживать конкурентное преимущество.

Разберем главные преимущества.

1. Изучение новых моделей трансформации. Наука о данных позволяет предприятиям открывать новые закономерности и отношения, которые могут изменить организацию. Анализ поможет выявить малозатратные изменения в управлении ресурсами для максимального влияния на размер прибыли.

Например, компания электронной коммерции использует науку о данных, чтобы обнаружить, что слишком много запросов клиентов генерируется в нерабочее время. Исследования показывают, что клиенты с большей вероятностью совершат покупку, если получат быстрый ответ, а не ответ на следующий рабочий день. Внедряя круглосуточное обслуживание клиентов, бизнес увеличивает доход на 30%.

2. Инновация новых продуктов и решений. Наука о данных поможет выявить пробелы и проблемы, которые иначе остались бы незамеченными. Глубокое понимание решений о покупке, отзывов клиентов и бизнес-процессов может стимулировать инновации во внутренних операциях и внешних решениях.

Например, решение для онлайн-платежей использует науку о данных для сопоставления и анализа комментариев клиентов о компании в социальных сетях. Анализ показывает, что клиенты забывают пароли в пиковые периоды покупок и недовольны текущей системой поиска паролей. Компания может разработать лучшее решение и значительно повысить удовлетворенность клиентов.

3. Оптимизация в режиме реального времени. Предприятиям, особенно крупным, сложно реагировать на изменяющиеся условия в режиме реального времени. Это может привести к значительным потерям или сбоям в деловой активности. Наука о данных может помочь компаниям прогнозировать изменения и оптимально реагировать на обстоятельства.

Например, транспортная компания, использующая грузовики, с помощью науки о данных может решить свои логистические проблемы. Например, сократить простои, когда грузовики

ломаются. Они определяют маршруты и графики смен, которые приводят к более быстрым поломкам, и корректируют графики работы грузовиков. Они также создают резерв запасных частей, которые требуют частой замены, чтобы грузовики можно было ремонтировать быстрее.

Методы науки о данных

В этом разделе рассмотрим эффективные техники, которые используют специалисты по работе с данными.

Классификация — это сортировка данных по группам или категориям. Компьютеры обучены идентифицировать и сортировать данные. Известные наборы данных используются для построения алгоритмов принятия решений на компьютере, который быстро обрабатывает и классифицирует данные.

Примеры:

- сортировка товаров на популярные и непопулярные;
- сортировка заявок на страхование с высоким и низким риском;
- сортировка комментариев в социальных сетях на положительные, отрицательные и нейтральные.

Специалисты по науке о данных используют вычислительные системы для отслеживания процесса обработки данных.

Регрессия — это метод нахождения взаимосвязи между двумя, казалось бы, не связанными между собой точками данных. Связь обычно моделируется на основе математической формулы и представляется в виде графика или кривых. Когда значение одной точки данных известно, регрессия используется для прогнозирования другой точки данных.

Примеры:

- скорость распространения болезней, передающихся воздушно-капельным путем;
- взаимосвязь между удовлетворенностью клиентов и количеством сотрудников;
- зависимость между количеством пожарных депо и количеством пострадавших в результате пожара в конкретном месте.

Кластеризация — это метод группировки тесно связанных данных для поиска закономерностей и аномалий. Кластеризация отличается от классификации, поскольку данные нельзя точно классифицировать по фиксированным категориям. Следовательно, данные сгруппированы в наиболее вероятные отношения. С помощью кластеризации можно обнаружить новые закономерности и взаимосвязи.

Примеры:

- группировка клиентов с похожим покупательским поведением для улучшения обслуживания.
- группировка сетевого трафика, чтобы определять модели ежедневного использования и быстрее выявлять сетевые атаки;
- кластеризация статей по нескольким категориям новостей и использование этой информации для поиска поддельного новостного контента.

Основные принципы техник науки о данных. Детали могут разниться, однако основные принципы остаются неизменными.

- Научите машину сортировать данные на основе известного набора данных. Например, образцы ключевых слов передаются компьютеру с их значением сортировки. «Радоваться» — хорошо, а «ненавидеть» — плохо.
- Дайте машине неизвестные данные и позвольте самостоятельно сортировать набор данных.
- Допускайте неточности результатов и учитывайте фактор вероятности результата.

Кто такой Data Scientist

Data Science — раздел науки, Data Scientist (дата-сайентист) — практики в этой области.

Дата-сайентисты не обязательно несут прямую ответственность за все процессы, связанные с жизненным циклом науки о данных в компании. Например, конвейеры данных обычно обрабатывают дата-инженеры, но дата-сайентисты могут давать рекомендации о том, какие данные будут полезны.

Специалисты по данным могут создавать модели машинного обучения, но масштабирование этих усилий требует больше навыков разработки программного обеспечения. Для этих целей специалисты по данным часто сотрудничают с инженерами по машинному обучению.

Обязанности дата-сайентиста могут совпадать с обязанностями аналитика данных, особенно с исследовательским анализом и визуализацией. Однако набор навыков дата-сайентиста обычно шире, чем у аналитика данных. Например, специалисты по обработке и анализу данных используют языки программирования (такие как R и Python), чтобы получить дополнительные статистические выводы или визуализировать данные.

Чтобы выполнять эти задачи, дата-сайентистам нужны знания и навыки в области компьютерных наук, математики и математической статистики. Эти навыки выходят за рамки обычного анализа данных.

Кроме того, дата-сайентист должен понимать специфику бизнеса, для которого он разрабатывает решение — будь то автомобилестроение, электронная коммерция или здравоохранение.

Обобщим. Дата-сайентист должен:

- Достаточно знать о бизнесе, чтобы задавать уместные вопросы и определять его болевые точки.
- Применять статистику и информатику, обладать деловой хваткой для анализа данных.
- Использовать широкий спектр инструментов и методов для подготовки и извлечения данных — от баз данных и SQL до интеллектуального анализа данных и методов интеграции данных.
- Извлекать полезные сведения из больших данных с помощью прогнозной аналитики и искусственного интеллекта (ИИ), включая модели машинного обучения, обработку естественного языка и глубокое обучение.
- Писать программы, автоматизирующие обработку данных и расчеты.
- Рассказывать и визуализировать истории, которые ясно доносят значение результатов до заинтересованных лиц, даже если они плохо понимают технические аспекты.
- Объяснять, как результаты можно использовать для решения бизнес-задач.
- Сотрудничать с другими членами группы по обработке и анализу данных: аналитиками, ИТ-архитекторами, дата-инженерами и разработчиками.

Эти навыки пользуются большим спросом. Чтобы попасть в профессию и оставаться востребованными специалистами, дата-сайентистам приходится постоянно изучать смежные сферы: программирование, математику, статистику. Процесс обучения практически не останавливается.

Инструменты дата-сайентиста

Дата-сайентисты полагаются на популярные языки программирования для проведения исследовательского и статистического анализа данных. Эти инструменты с открытым исходным кодом поддерживают встроенное статистическое моделирование, машинное обучение и возможности визуализации.

К таким языкам относятся:

- **R Studio** — язык программирования и среда для разработки статистических вычислений и графики.
- **Python** — динамичный и гибкий язык программирования со множеством библиотек для быстрого анализа данных, например NumPy, Pandas, Matplotlib.

Чтобы облегчить обмен кодом и другой информацией, дата-сайентисты могут использовать блокноты GitHub и Jupyter — среды разработки с подключенными модулями визуализации.

SAS — комплексный набор инструментов для анализа, прогнозного моделирования, составления отчетов, визуализаций и интерактивных информационных панелей (дашбордов).

Инструменты статистического анализа (SAS, инструменты в R и Python) предлагают расширенный статистический анализ, большую библиотеку алгоритмов машинного обучения, анализ текста, расширяемость с открытым исходным кодом, интеграцию с большими данными и простое развертывание в приложениях.

Для дата-сайентистов также важны навыки использования платформ обработки больших данных — Apache Spark, Apache Hadoop и базы данных NoSQL. А также инструменты визуализации: от простых (например, Microsoft Excel) до коммерческих (Tableau, IBM Cognos, QuickSight) и инструментов с открытым исходным кодом (JavaScript-библиотека D3.js и RAW Graphs).

Для построения моделей машинного обучения специалисты по данным обращаются к фреймворкам, таким как PyTorch, TensorFlow, MXNet и Spark MLib.

Кривая обучения дата-сайентиста очень крутая. На рынке не так много глубоких специалистов, чтобы закрыть потребности компаний. Решить проблему могут Multipersona DSML — многопользовательские платформы обработки данных и машинного обучения.

Multipersona DSML — это инструменты с понятным пользовательским интерфейсом и возможностями автоматизации. Пользоваться ими можно с минимальным знанием кода или без него, так что люди с небольшим опытом в ИТ и обработке данных тоже смогут создавать ценность для бизнеса. А для опытных специалистов Multipersona DSML могут предоставить более технологичный и гибкий интерфейс.

Multipersona DSML помогут наладить совместную работу в компании.

Облачные технологии в Data Science

Облачные технологии масштабируют науку о данных: предоставляют доступ к дополнительной вычислительной мощности, хранилищу и другим необходимым инструментам.

Поскольку Data Science часто использует большие наборы данных, инструменты, которые могут масштабироваться вместе с их размером, важны, особенно для срочных проектов. Решения для облачного хранения, такие как озера данных (хранилища для структурных и неструктурных данных), обеспечивают доступ к инфраструктуре, способной принимать и обрабатывать большие объемы данных.

Эти системы хранения обеспечивают гибкость для конечных пользователей, позволяя им развертывать большие кластеры по мере необходимости. Они также могут добавлять вычислительные узлы для ускорения обработки данных, позволяя бизнесу идти на компромисс в краткосрочной перспективе для достижения большего долгосрочного результата.

У облачных платформ разные модели ценообразования, например плата за использование или подписка. Подходящую найдут и крупные компании и стартапы.

Технологии с открытым исходным кодом широко используются в наборах инструментов для обработки данных. Когда они размещены в облаке, командам не нужно устанавливать, настраивать, обслуживать или обновлять их локально. Некоторые облачные провайдеры также предлагают готовые наборы инструментов, которые позволяют дата-специалистам создавать модели без программирования, что еще больше упрощает доступ к технологическим инновациям и анализу данных.

Самые крупные игроки на рынке облачных решений — Amazon, Google Cloud и Microsoft Azure. Каждая из этих компаний пытается предложить что-то новое и лучшее как можно скорее, чтобы перехватить клиентский трафик и получать крупную сумму за использования своих решений.

Например, Microsoft заключили контракт с Open AI, создателями чат-бота ChatGPT — нейросети, которая дает полноценный ответ на поставленный вопрос.

ChatGPT использует технологию последовательного обучения — анализирует структуру вопроса и находит ответ в доступной базе знаний, используя передовые технологии NLP (Natural Language Processing, обработка естественного языка). В результате Microsoft предоставляют эксклюзивный доступ к API чат-бота клиентам Azure.

Сферы применения Data Science

Наука о данных нашла свое применение практически во всех отраслях.

- **Здравоохранение.** Медицинские компании используют науку о данных в создании сложных медицинских инструментов для диагностики и лечения заболеваний.
- **Игры.** Гейм-студии создают игры с помощью науки о данных, чтобы поднять игровой опыт на новый уровень.
- **Распознавание изображений.** Выявление закономерностей на изображениях и обнаружение объектов — одна из самых популярных областей применения науки о данных. Например, когда вы загружаете фотографию с кем-то из своих друзей в социальную сеть, алгоритмы могут распознать и отметить их.
- **Системы рекомендаций.** Рекомендательные сервисы Netflix и Amazon советуют фильмы в зависимости от того, что смотрел или покупал пользователь на их платформах.
- **Логистика.** Логические компании используют Data Science, чтобы оптимизировать маршруты, ускорить доставку и повысить эффективность работы.
- **Обнаружение мошенничества.** С помощью Data Science банки и финансовые организации обнаруживают мошеннические транзакции.

- **Интернет-поиск.** Google, Яндекс, Yahoo, Bing и другие поисковые системы используют алгоритмы обработки данных, чтобы в ответ на запрос пользователя за секунды предлагать лучшую выдачу. Например, Google обрабатывает более 20 петабайт данных в день. Без науки о данных он не стал бы тем сервисом, который мы знаем сегодня.
- **Распознавание речи.** Методы науки о данных преобладают в распознавании речи. Результат этой работы мы видим в повседневной жизни. Пример — голосовые помощники: Google Assistant, Alexa, Siri и другие.
- **Таргетированная реклама.** Компании используют данные о клиентах и их активности в интернете, чтобы сделать персонализированную рекламу. В то время, как вы видите рекламу образовательной программы по Data Science, другой человек в том же регионе может видеть рекламу бренда одежды. У полезных предложений более высокий CTR (Call-Through Rate, показатель кликабельности).
- **Авиакомпании.** Благодаря науке о данных стало легче прогнозировать задержки рейсов.
- **Дополненная реальность.** Крупнейшие игроки — Apple, Google, китайские компании — пытаются представить на рынок гарнитуры дополненной реальности и пока сталкиваются с рядом проблем. Внедрение Data Science в эту область очень актуально.

Примеры проектов Data Science

Использование Data Science дает компаниям много преимуществ. Примеры — оптимизация процессов за счет интеллектуальной автоматизации или улучшение качества обслуживания за счет лучшего таргетинга и персонализации.

Разберем несколько примеров использования науки о данных и искусственного интеллекта.

- **Правоохранительные органы.** Наука о данных помогла полиции Бельгии понять, когда и куда лучше направлять полицейских, чтобы предотвращать преступления. Несмотря на большую площадь покрытия и ограниченные ресурсы, отчеты и дашборды (лаконичные информационные панели с данными) помогли офицерам лучше ориентироваться в ситуации, рассредоточиваться, поддерживать порядок и предвидеть преступления.
- **Борьба с пандемией.** В штате Род-Айленд хотели вновь открыть школы, но, были осторожны, так как пандемия COVID-19 еще продолжалась. Государство использовало науку о данных, чтобы отследить цепочки заболевших, и понять, какие ученики могут посещать школы. В этом помогли звонки от граждан, обеспокоенных ситуацией, желающих узнать о своих дальнейших действиях в связи с карантином.

Эта информация помогла штату создать колл-центр и скоординировать профилактические меры.

- **Беспилотные транспортные средства.** Компания Lunewave, производитель датчиков, искал способ сделать свою технологию более рентабельной и точной. Наука о данных помогла сделать датчики более безопасными и надежными, а также оптимизировать производство с помощью 3D-печати.

- **Развлечения.** Наука о данных позволяет стриминговым сервисам отслеживать и оценивать то, что просматривают пользователи. Эти данные помогают создавать новые популярные сериалы и фильмы. Рекомендательные алгоритмы на основе данных советуют контент, который может понравиться пользователю, на основе истории его просмотров.

- **Производство.** Наука о данных находит применение на производстве: для управления цепочками поставок, оптимизации распределения, профилактического обслуживания и прогнозирования неисправностей оборудования.

- **Здравоохранение.** Модели машинного обучения и другие инструменты обработки данных используют больницы и медицинские компании. Они помогают диагностировать заболевания и планировать лечение на основе предыдущих результатов лечения пациентов.

- **Розничная торговля.** Продавцы оценивают поведение клиентов и тенденции рынка, чтобы делать индивидуальные предложения, настраивать таргетированную рекламу и планировать акции. Наука о данных помогает управлять запасами продуктов на складе и цепочками поставок.

- **Международный банк** предоставляет более быстрые кредитные услуги с помощью мобильного приложения, используя модели кредитного риска на основе машинного обучения и архитектуру гибридных облачных вычислений, мощную и безопасную.

- **Производитель электроники** разрабатывает сверхмощные датчики, напечатанные на 3D-принтере для управления беспилотными автомобилями. Решение основано на инструментах обработки данных и аналитики для расширения возможностей обнаружения объектов в реальном времени.

- **Поставщик решений для роботизированной автоматизации процессов** (RPA) разработал решение для интеллектуального анализа бизнес-процессов, которое сокращает время обработки инцидентов на 15−95% для компаний-клиентов. Решение обучено понимать содержание и настроение электронных писем клиентов, направляя сервисные службы на решение приоритетных задач.

Эти примеры — лишь малая часть применения Data Science. Рынок вакансий в этой области в постоянном дефиците, несмотря на то что планка зарплаты выше, чем в других сферах ИТ.

Почему данные — это важно

Сегодня сбор данных, полезных для бизнеса, относительно прост. Настолько прост, что есть риск того, что данных будет слишком много для обработки.

В статье How Can Small Businesses Use Big Data гуру данных и аналитики Бернард Марр сказал: «У среднего и малого бизнеса меньше самостоятельно сгенерированных данных, чем у крупных игроков, но это не значит, что данные запрещены. Во многих отношениях большие данные лучше подходят для малого бизнеса — как правило, он гибче и может быстрее действовать на основе данных».

В статье Forbes, посвященной опросу Deloitte, отмечается, что «49% респондентов заявили, что аналитика помогает им принимать более обоснованные решения, 16% — что она лучше способствует реализации ключевых стратегических инициатив, а 10% — что она помогает им улучшить отношения как с клиентами, так и с деловыми партнерами». Но чтобы в полной мере воспользоваться преимуществами данных и аналитики, вам нужно знать, как получить максимальную отдачу от данных.

Респонденты опроса Deloitte показали, что даже небольшие стартапы генерируют данные. Любой бизнес, у которого есть сайт и аккаунты в соцсетях, а также возможность принимать электронные платежи, собирает данные о клиентах — их привычках, веб-трафике, демографии и многом другом. У всех этих данных есть потенциал, его нужно суметь раскрыть.

На принятие решения влияет множество факторов — события в компании, глобальные новости, интуиция руководителя. Но нет ничего более убедительного, чем достоверные данные. Это — та сила для принятия решений и увеличения прибыли, от которой компания не может отказаться.

Малый и средний бизнес может получить те же преимущества, что и крупные организации, при правильном использовании данных. С помощью данных можно принимать решения о:

- поиске новых клиентов,
- повышении лояльности клиентов,
- улучшении качества обслуживания клиентов,
- лучшем управлении маркетингом,
- отслеживании взаимодействия в соцсетях,
- прогнозировании тенденций продаж.

Таким образом, данные помогают лидерам принимать более взвешенные решения о том, куда направлять свои компании.

Данные помогают решать проблемы

Как определить, что пошло не так после месяца медленных продаж или неэффективной маркетинговой кампании?

Отслеживание и анализ данных из бизнес-процессов поможет выявить проблемы с производительностью, чтобы вы могли лучше понять каждую часть процесса и узнать, какие шаги нужно исправить, а какие работают хорошо.

«Лучше всего управляемые компании управляются данными, и этот навык отличает их от конкурентов», — Томаш Тунгуз, автор статей и крупный инвестор в области DS.

Данные помогают оценить производительность

Спортивные команды — пример организаций, которые собирают данные об эффективности, чтобы стать лучше. Сегодня нет профессиональной команды, в которой не работала бы команда сборщиков данных и аналитиков, помогающих поддерживать и улучшать игру на поле. Они всегда обновляют данные о том, кто что делает хорошо, и как это может помочь команде преуспеть.

Задумывались ли вы когда-нибудь, как работает ваша команда, отдел, компания, маркетинговые усилия, обслуживание клиентов, доставка или другие подразделения? Все это может показать сбор и анализ данных.

Если руководство компании не уверено в эффективности сотрудников или в маркетинге, как инвесторы могут узнать, используются ли их деньги с пользой? Приносит ли компания больше денег, чем тратит?

Например, у компании есть высокоэффективный торговый представитель, к которому отправляют больше всего потенциальных клиентов. Однако, когда мы углубимся в данные, то увидим, что он закрывает сделки с более низкой скоростью, чем один из других торговых представителей, который получает меньше потенциальных клиентов, но закрывает сделки с более высоким процентом.

Данные о производительности могут повлиять на то, как компания распределяет потенциальных клиентов, и увеличить доход.

Данные помогают улучшить процессы

Данные помогают понять и улучшить бизнес-процессы, чтобы компании могли сократить напрасную трату денег и времени.

Например, плохие рекламные решения могут быть причиной больших потерь ресурсов в компании. С данными, которые показывают, как работают маркетинговые каналы, можно понять, какие из них более эффективны и сосредоточиться на них. А можно выявить, какие каналы неэффективны и поработать над их улучшением.

В результате можно получить больше потенциальных клиентов без увеличения расходов на рекламу.

Данные помогают понять потребителей

Без данных нельзя понять, кто клиенты компании, нравится ли им продукт, эффективны ли маркетинговые усилия, сколько денег тратит и зарабатывает компания. Данные — ключ к пониманию клиентов и рынка.

Однако можно легко потеряться во всех данных, если нет нужных инструментов для их понимания. Правильная работа с данными и использование специальных инструментов — лучший способ больше узнать о клиентах и улучшить бизнес-показатели компании.

Сегодня управление бизнесом с помощью данных — это инвестиции в будущее. Бизнес рискует остаться в прошлом, если не будет использовать данные. К счастью, достижения в области обработки и визуализации данных позволяют бизнесу развиваться быстрее, чем когда-либо.

Заключение

Сегодня мы узнали, что такое наука о данных и что нужно знать человеку, который хочет с ними работать. Разобрались, какие инструменты нужны дата-сайентисту. А также поговорили о том, почему данные — это важно и как они могут помочь бизнесу.

До встречи на следующем занятии!