УДК 004.584

Позднякова Анастасия Олеговна,
студентка бакалавриата 2 курс,
институт информационных технологий
МИРЭА-Российский технологический университет,
Россия, г. Москва
Научный руководитель: Свищёв Андрей Владимирович

BIG DATA, DATA SCIENCE И MACHINE LEARNING, КАК НОВЕЙШИЕ ТРЕНДЫ СОВРЕМЕННОСТИ

Аннотация: *В данной научной статье рассматриваются такие важнейшие области IT – Информационных технологий, как Big Data (Большие данные), Data Science (Наука о данных) и Machine Learning (Машинное обучение). Проводится ознакомление с данными областями и сферами, уточняется их актуальность и применение в современных бизнес-процессах. Каждая область подробно описывается и проводится рассуждение об актуальности и возможности применения в современном бизнесе на сегодняшний день, также всё подкрепляется примерами в различных IT-компаниях и не только. Подводятся итоги и описываются возможности дальнейшего применения в ближайшем будущем.*

Ключевые слова: *Big Data, Data Science, Data Scientist, Machine Learning, Artificial Intelligence, IT, нейронные сети, большие данные, наука о данных, машинное обучение, искусственный интеллект.*

Annotation: *This scientific article discusses such important areas of IT – Information technologies as Big Data, Data Science and Machine Learning. Familiarization with these areas and spheres is carried out, their relevance and application in modern business processes are clarified. Each area is described in detail and reasoned about the relevance and possibility of application in modern*

business today, everything is also supported by examples in various IT companies and not only. The results are summarized and the possibilities of further application in the near future are described.

Key words: *Big Data, Data Science, Data Scientist, Machine Learning, Artificial Intelligence, IT, neural networks, big data, data science, machine learning, artificial intelligence.*

На сегодняшний день, когда человечество перешло в эпоху четвертой промышленной революции, невозможно представить бизнес без квалифицированных IT-специалистов, особенно когда необходимо работать с огромными количествами информации. Наше современное общество, в эпоху информационного и глобального скачка, стремительно развивается, а объемы и потоки данных постоянно растут, что приводят нас к новым открытиям. В результате, возникают новые значения, новые термины, в научном и практическом мире получившие названия Big Data, Data Science и Machine Learning или переводя данные термины на русский язык – «большие данные», «наука о данных» и «машинное обучение».

Big Data – «большие данные»

Big Data – этим термином называют соединение множества различных технологий и методик сбора, обработки и анализа неструктурированных и структурированных данных в большом объеме. Несколько лет назад большие данные являлись инновационными тенденциями, которые использовались только в секторе высокотехнологичных разработок. На данный момент большие данные присутствуют во всех сферах и отраслях жизни человека и занимают огромное место в нашей ежедневной жизни, также помимо всего прочего их можно найти, использовать и применить как в коммерческих, так и не коммерческих целях и средах.

Технологии Big Data позволяют обрабатывать большие объемы данных, систематизировать их, анализировать и выявлять закономерности там, где

человеческий мозг бы их никогда не заметил. Это открывает совершенно новые возможности по использованию данных. Понятие Big Data означает не просто большие пакеты данных, это огромные хранимые и обрабатываемые массивы из сотен гигабайтов, и даже петабайтов данных. Говоря коротко, можно определить Big Data, как технологии обработки общего количества информации для получения определенной информации.

С развитием BigData развивались и технологии мировых компаний. На текущий момент, BigData удел не только гигантов IT мира. Это направление, благодаря набору облачных сервисов от IBM, Amazon, Google становится доступным практически любым компаниям, работающим в сфере ИТ. А такие решения как Clickhouse, Cassandra, InfluxDB позволяют войти в сферу работы с BigData даже отдельным разработчикам, желающим создать сови бизнес-проекты.

Грамотное использование BigData на сегодняшний день является обязательным условием для развития крупных IT компаний. Без анализа поведения своих пользователей, без возможности прогнозирования, руководствуясь только опытом и интуицией, в настоящее время крайне сложно оставаться конкурентоспособным с такими крупными компаниями, как Amazon и Google. Настроенная и работающая система BigData позволяет в секунды предоставить ценную информацию, полученную и составленную из анализа миллиардов действий клиентов компании.

В бизнесе на сегодняшний день уже зародилось понятие Data Driven Managment, которое означает управление компанией руководствуясь исключительно информацией, полученной из анализа данных. И такие способы управления компаниями показывают блестящие результаты. Facebook, Google, Mail.ru и Yandex уже давно используют аналитику для принятия решений. Также на сегодняшний день в BigData заинтересован и традиционный бизнес, представители которого нуждаются в новых инструментах повышения эффективности.

Основные принципы работы с BigData.

1. Горизонтальная масштабируемость.

   Так как при работе с данными их может быть большое количество, то и система, в которой они хранятся данные должна иметь возможность расширяться. Если объем данных вырос вдвое, то и количество кластеров должно увеличиваться в 2 раза, и по аналогии при увеличении не вдвое, а на другую определенную цифру.

2. Отказоустойчивость.

   Горизонтальная масштабируемость означает тот факт, что машин, работающих с данными, в кластере огромное количество. И соответственно нельзя исключать возможности того, что эти машины будут по тем или иным причинам выходить из строя. К примеру, Hadoop-кластер Yahoo насчитывает более 42000 машин. Методы работы с BigData должны учитывать эту возможность и продолжать работу без видимых потерь при выходе из строя определенного количества машин.

3. Локальность данных.

   В больших системах данные распределены на большом количестве машин. Если данные находятся на одной машине, а обрабатываются на другой, то расходы на передачу этих данных могут и вовсе превысить расходы на обработку. Поэтому важным вопросом в проектировании BigData стоит принцип локальности данных, или по-другому выражаясь, обработке информации там же, где она хранится изначально.

Глобальное использование Big Data стало причиной возникновения новых тенденций, одной из них можно назвать Data Science или переводя на русский – «наука о данных». Большинство крупнейших компании на сегодняшний день применяют Data Science, чтобы предоставлять своим клиентам персональные предложения. Ярким примером этого является Google

AdSense, который собирает информацию о пользователях и показывает контекстную рекламу.

Data Science – «Наука о данных».

Data Science (Наука о данных) – это раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме. Объединяет методы по обработке данных в условиях больших объемов и высокого уровня параллелизма, статические методы, методы интеллектуального анализа данных и приложения искусственного интеллекта (artificial Intelligence) для работы с данными, а также методы проектирования и разработки баз данных.

Термин Data Science впервые ввел и характеризовал в своей книге датский ученный Петер Наур в 1974 г., хотя существует мнение, что Наур употреблял данный термин ещё в 1960-х годах. Однако, свою известность термин Data Science получил лишь в первом десятилетии 21-го века, во многом благодаря популяризации концепции Big Data (больших данных).

Как следствие из появления данной науки, Data Science порождает спрос на новые профессии в этой сфере, одной из которых является Data Scientist.

Data Scientist – это специалист по данным или инженер, обладающий высокими навыками в математике, программировании и аналитики. Данную сферу и профессию можно охарактеризовать как ту, которая находится в топе и еще долгое время будет занимать лидирующие места, так как специалисты Data Science, обладающие высокими математическими и аналитическими качествами очень востребованы на рынке труда сейчас и по прогнозам аналитиков еще долгое время на них будет очень высокий спрос.

Если рассматривать данную сферу и профессию подробнее, то можно отметить, что Data Scientist является специалистом, который очень тесно работает в математической сфере, углубляясь в более сложные категории и подкатегории математики, такие как математическая статистика, теория

вероятностей и линейная алгебра, а также умеет применять математические знания в практическом плане, используя различные программные средства.

Всё вышенаписанное и является главным отличием Data Scientist от рядового математика. Данная профессия требует глубоких теоретических и реальных практических знаний методов статистического анализа данных, навыков построения математических моделей (к примеру, нейронных сетей), работы с большими массивами данных и уникальной способности находить закономерности.

Обобщая все вышенаписанное, нужно отметить, что Data Scientist – это специалист, который разбирается во многих областях и направлениях в сфере IT (информационных технологий), таких как, аналитике, бизнес-аналитике, искусственном интеллекте (artificial intelligence), машинном обучении (machine learning), глубоком обучении (deep learning) и во многом другом.

В процессе изучения особенностей концепций Big Data и перспектив развития Data Science нельзя не затронуть такое важное направление в IT, о котором уже упоминалось выше, как машинное обучение (machine learning).

Machine Learning – «Машинное обучение».

Не существует точного общепринятого определения Machine Learning, из-за этого ниже будут представлены трактовки машинного обучения от различных крупнейших представителей IT-индустрии и исследовательских компаний.

- «Практическое использование алгоритмов для анализа данных, изучения их и последующего прогнозирования какого-либо явления» (NVIDIA).
- «Наука о том, как научить компьютеры функционировать без явного программирования» (Стэндфортский университет).
- «Технология, основанная на алгоритмах, способных учиться на заложенных данных без помощи средств программирования» (McKinsey & Co).

- «Алгоритмы, способные самостоятельно выбирать метод решения важных задач путем обобщения заложенных в систему примеров» (Вашингтонский университет).
- «Сфера деятельности, функция которой состоит в поиске способов создания компьютерных систем, способных самообучаться и самостоятельно улучшаться по мере накопления опыта, а также в поиске фундаментальных закономерностей, по которым работают все процессы обучения» (Университет Карнеги Меллон).

Ознакомившись со всеми выше представленными определениями, ниже выдвинем свою обобщенную трактовку Machine Learning (машинное обучение).

Machine Learning (машинное обучение) – это подраздел Artificial Intelligence (искусственного интеллекта) и Data Science (науки о данных), специализирующийся на использовании данных и алгоритмов для имитации человеческой возможности обучения или, выражаясь по другому, построения обучаемых моделей для различных целей: например, автоматизации процессов, автоматического перевода текстов, распознания изображений. Именно такое направление, как машинное обучение помогает ранжировать контент в различных социальных сетях и создавать голосовых или текстовых помощников, которые общаются на естественном языке, создавая иллюзию реального собеседника, к примеру Siri от Apple или Алиса от Yandex.

Типы машинного обучения.

Machine learning (машинное обучение) можно разделить на два типа:

1. Дедуктивное обучение (экспертные системы).

В этом случае есть сформулированные и формализованные знания. К примеру, это может быть база данных, в которой указано, что если температура превышает 30 градусов, то нужно включить кондиционер, а если на улице идет дождь, то необходимо закрыть окна. Нужно вывести из них новое правило, которое можно применить к конкретному определенному случаю. Экспертные

системы чаще относят к ответвлению кибернетики — науки об управлении информацией в сложных системах, — чем к машинному обучению.

2. Индуктивное обучение, которое в свою очередь подразделяется на:

- Обучение с учителем.

  Пример возможных задач: по предыдущему курсу валют нужно предсказать курс на завтрашний день; отличить по изображениям кошек от собак (в этом случае изначально должна быть информация, на какой картинке и где изображены кошка и собака).

- Обучение без учителя.

  Пример возможной задачи: разделить группу пользователей сайта на основе их интересов или демографических характеристик. Обычно нужно знать, сколько групп уже имеется в данных.

- Обучение с подкреплением.

  Пример возможной задачи: серия игр Super Mario, в которых компьютер (агент) взаимодействует со средой (уровень игры) и получает либо положительные, либо отрицательные очки.

- Активное обучение.

  Пример возможной задачи: подсказка слов на раскладке клавиатуры смартфона.

Многие методы индуктивного обучения связаны не столько с обучением, сколько с извлечением информации.

Если подробнее понимать возможности алгоритмов машинного обучения с учителем, то стоит, к примеру, рассмотреть несколько задач или проблем, которые могут быть решены алгоритмами Machine Learning:

- Определение почтового индекса по рукописным цифрам на конверте;
- Нахождение доброкачественности опухоли на основе медицинских изображений;
- Обнаружение мошеннической деятельности в сделках по кредитным картам;

- Прогнозирование отказов высокотехнологичного и сложного промышленного оборудования;
- Идентификация и распознавание изображений, полученных беспилотными летательными аппаратами.

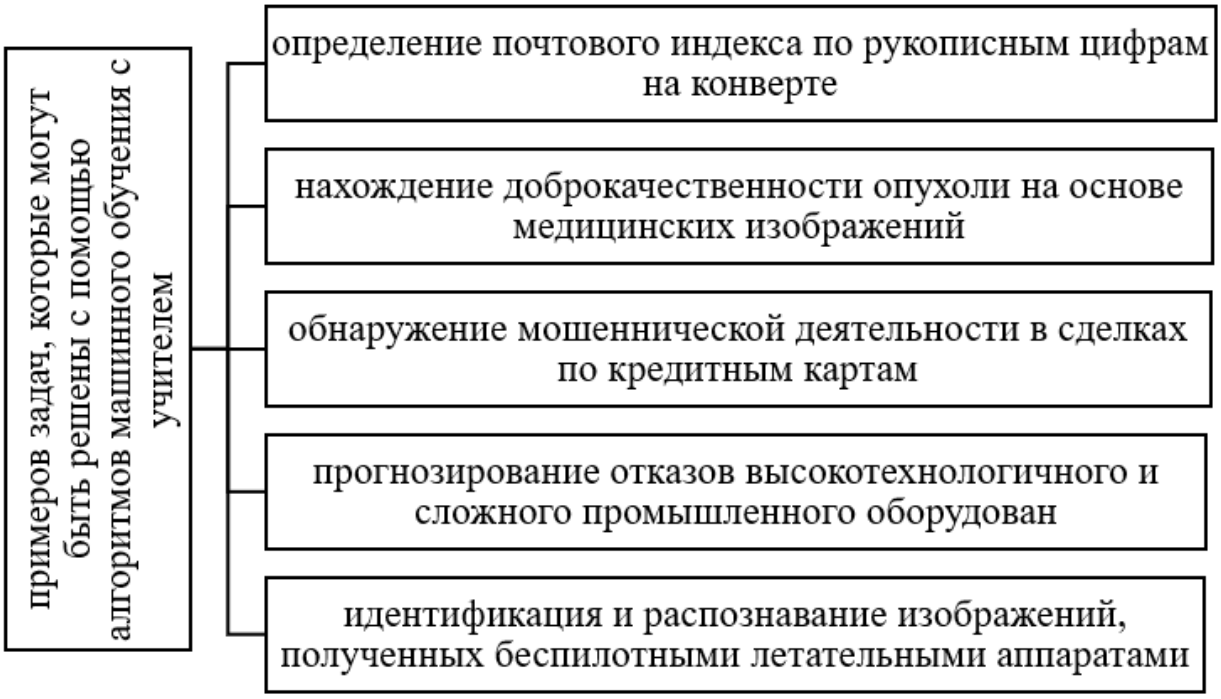

Рисунок 1. Примеры задач, решаемых ML

Подводя итоги выше представленного исследования, можно сделать следующие выводы. В эпоху развития информационных технологий и перехода человечеством к четвертой промышленной революции, появилось огромное количество данных в цифровом формате или больших данных – Big Data. Появилось множество различных технологий хранения, вычисления, математических инструментов анализа и обработки данных. Всё это приводит к появлению новых бизнес-процессов, научных областей и профессий. Одними из таких областей являются Data Science – «Наука о Данных» и Machine Learning – «Машинное обучение». На сегодняшний день рынок нуждается в высококвалифицированных специалистах, разбирающихся в данных сферах, это актуально как никогда и будет актуально еще очень долгое время. Применение данных и науки о данных на данный момент не

ограниченны одной лишь сферой IT, каждый бизнес нуждается в специалистах, разбирающихся в больших объемах информации и умеющих грамотно её проанализировать и работать с полученными данными. Именно эти факты делают Big Data, Data Science и Machine Learning очень важными и актуальными для современного мира.

Использованные источники:

1. Веретенников А.В. BigData: анализ больших данных сегодня — 2017. — № 32 (166). — С. 9-12.
2. Lee R. Big Data, Cloud Computing, and Data Science Engineering. — Cham: Springer. — 2020. — 214 p.
3. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных. — СПб.: Питер. — 2017. — 336 с.
4. Большой гид по Data Science для начинающих: термины, применение, образование и вход в профессию. [Электронный ресурс]. URL: https://netology.ru/blog/01-2020-gid-po-data-science (дата обращения: 4.01.2022).
5. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. — М.: Вильямс. — 2017. — 393 с.
6. Обзор профессии Data Scientist. [Электронный ресурс]. URL: https://habr.com/ru/company/netologyru/blog/329068/ (дата обращения: 5.01.2022).
7. Cheng Q., Li H., Wu Q., Ngan K. Hybrid-Loss Supervision for Deep Neural Network. — Neurocomputing. — 2020. — Vol. 388. — P. 78–89.
8. Big Data от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce // Хабрахабр. [Электронный ресурс]. URL: https://habrahabr.ru/company/dca/blog/267361/ (дата обращения: 4.01.2022).

9. Машинное обучение. [Электронный ресурс]. URL: https://blog.skillfactory.ru/glossary/mashinnoe-obuchenie/ (дата обращения: 5.01.2022).

10. Что такое машинное обучение. [Электронный ресурс]. URL: https://www.ibm.com/ru-ru/cloud/learn/machine-learning (дата обращения: 6.01.2022).