25 января 2021

Сколько нужно труда и денег, чтобы создать систему анализа данных

В условиях пандемии роль аналитики, основанной на больших данных, возросла. Однако четкого понимания, как именно ее можно использовать себе во благо, у бизнеса по-прежнему нет.
В России отсутствует законодательно закрепленное определение больших данных (big data), но обычно под ними понимают разнообразные инструменты, ручной труд, например деятельность андеррайтеров в банках. Раньше проверка занятости, валидация анкетных данных и прочие проверки делались полностью вручную. Сегодня значимая доля проверок поддается автоматизации за счет синергии технологий машинного обучения и поведенческих данных. В результате прибыль страхового портфеля увеличивается на 5-10%.

Примеры применения продуктов big data в других областях тоже впечатляют. На искусственный интеллект перешла сейчас, например, вся сортировка почты. Другой пример – автоматическое планирование маршрутов перевозок, причем от обычных навигаторов с учетом пробок до оптимизации логистической цепи и путей поставок крупных компаний доставки. Посылки стали приходить быстрее, а время маршрута можно планировать более точно.

Не менее наглядный пример – управление производством, выявление брака. Например, брак при производстве бутылок или же гнилые овощи раньше выявлял специальный человек, который сбрасывал бракованную продукцию с конвейера. Сейчас системы компьютерного зрения делают это автоматически и в десятки раз быстрее.

Отстает ли Россия?

Российская индустрия больших данных только формируется. В 2019 году рынок big data в России составлял 45 млрд рублей, тогда как в США – 3 трлн рублей, в Китае – 157 млрд рублей. Такие цифры содержатся в стратегии развития рынка больших данных до 2024 года, разработанной Ассоциацией больших данных совместно с The Boston Consulting Group. Примечательно, что Китай также является мировым лидером по количеству патентов, зарегистрированных в сфере искусственного интеллекта и машинного обучения: с 2013 года он оформил около ста тысяч патентов, в 125 раз больше, чем Россия.

Вместе с тем ежегодный темп роста рынка больших данных в России не-плохой. С 2015 года он составляет 12%, отмечает исполнительный директор Ассоциации больших данных Алексей Нейман. Все больше компаний начинают использовать сервисы на основе аналитики больших данных, говорит директор по аналитике и алгоритмам компании oneFactor Максим Воеводский.

Осложняют развитие в России рынка больших данных главным образом стоимость решений – десятки миллионов рублей, а также сложность

внедрения. Вместе с внедрением таких продуктов в компаниях идут обновление и замещение устаревших систем. Эти процессы влияют как на ИТ-ландшафт, так и на данные. В частности, появляются новые источники, что приводит к сложностям внедрения продуктов, основанных на характере данных.

Развитию big data в России мешает также недостаточная степень автоматизации бизнес-процессов в компаниях, а именно отсутствие некоторых базовых ИТ-решений, необходимых для эффективной работы решений с технологиями машинного обучения, например CRM-систем. Еще один барьер – нехватка квалифицированных кадров в области big data.

Есть ли у вас большие данные, которые могут быть обработаны

Абсолютно все данные могут быть использованы для построения моделей искусственного интеллекта. В роли источников информации могут выступать как структурированные (табличные) данные, так неструктурированные (фото, видео) и полуструктурированные (различные данные телеметрии, логи событий, геоданные и др.).

«Зачастую при решении задачи выбор данных определяется исключительно их доступностью и качеством, поскольку в любых данных благодаря использованию методов ИИ можно выделить разного рода взаимосвязи», – пояснил «Эксперту» руководитель направления «Большие данные» компании «Техносерв» Денис Рыбченко. В конечном счете все данные приводятся к некоторым цифрам, описывающим объект или ситуацию.

«Если это текст, то может браться как количество слов “машина”, так и более сложные, иногда не объяснимые словами свойства текста. Если говорят про изображение, то опять же берется числовое описание форм, цветов и фигур, которые на нем присутствуют. Данное числовое описание – его называют embedding – и позволяют получить, в частности, нейронные сети. Таким образом, любой оцифрованный вид данных потенциально подлежит анализу, вопрос только в том, как применительно к нему поставить задачу: соотнести изображения и класс, описать нужные свойства текстов, иными словами, разметить обучающую выборку», – рассказал «Эксперту» архитектор машинного обучения цифровой лаборатории Softline Николай Князев.

Руководитель направления продвинутой аналитики департамента анали тических решений ГК «КОРУС Консалтинг» Александр Зенькович привел нам конкретные примеры данных, которые могут быть проанализированы.

«Если мы хотим оптимизировать продажи в рознице, в дело идут чековые данные (для анализа покупательской картины) и фотографии продуктов на витрине (для оценки их качества здесь и сейчас, более точного прогнозирования спроса). Фото и видео помогают отсле- живать процессы, например видеть брак на производстве или отсутствие каски на голове рабочего. Поведение клиентов в интернете тоже используется во множестве задач: автоматическом поиске похожих товаров, подборе жилья, оценке музыкальных и видеопредпочтений», – рассказал **Александр Зенькович**.

В зависимости от «носителя» данных и целей аналитики последняя разделяется на несколько видов:

- обработка естественного языка (всем знакомая «Сири» в iOS), помогающая перевести аудиотрек в текст, выделить из него смыслы, эмоции, по необходимости сформулировать ответную по смыслу реакцию, облачить ее во фразу и воспроизвести;
- обработка фото и видео (FaceID в iOS, распознавание текста на фото страницы книги, номер автомобиля нарушителя в «письмах счастья», огромное количество фотоприложений с фильтрами, спецэффекты в фильмах и проч.);
- отбор брака на конвейере с камерой;
- обработка потока структурированных данных в реальном времени (системы безопасности, алгоритмы торговли, динамическое ценообразование, антифрод, системы активной стабилизации и проч.);
- обработка потока данных (цифровой след, датчики, IoT и проч.) для выявления тренда, экстраполяции наблюдаемого процесса, формирование реакции на тренд;
- обработка структурированных данных.

Таким образом, учитывая, насколько разнообразны сегодня поддающиеся аналитике данные, обладать ими могут очень многие предприятия.

С помощью каких инструментов можно работать с данными

Для работы с большими данными используются два типа решений: проприетарное программное обеспечение и открытое. К первым относятся аналитические решения SAS, Tableau, Qlik, SAP и другие. Они представляют

собой инструменты, дающие на выходе разнообразную инфографику, средства представления и визуализации, статистического анализа и поиска информации. К открытому программному обеспечению относится огромный набор программ для анализа данных средствами языков программирования Python, R.

«На базе комплекса этих решений ИТ-службы организуют наиболее подходящий для компании аналитический ландшафт, так как все решения уникальны для сферы бизнеса и инфраструктуры компании. Так, решения, созданные в разных компаниях, даже в разных департаментах одной организации, при помощи одних и тех же инструментов, могут разительно отличаться и не быть взаимозаменяемыми. В конце концов, и всем известный Excel – отличный инструмент для анализа начального уровня», – рассказал «Эксперту» Алексей Нейман.

Стоимость решений для создания аналитики может отличаться от заказчика к заказчику. «Если заказчик хочет получить максимально экономичный вариант и у него нет жестких ограничений или предпочтений по вендорам, то решение строится исключительно на опенсорс-компонентах. В таком случае можно уложиться в 15-20 миллионов рублей. Но среднюю стоимость, учиты вающую внедрение и поддержку в течение года, я бы оценил в 50 миллионов рублей. И еще обязательно нужно учитывать необходимость инвестиций не только в поддержку, но и в дальнейшее развитие решения», – отмечает Денис Рыбченко.

Таким образом, организовать работу с большими данными посильно практически для любого бизнеса, однако не так много организаций имеют

возможность получить качественную аналитику.

Какие технологии используются для обработки данных

Для хранения и обработки данных главным образом используют три технологии: классическую реляционную систему управления базами данных (СУБД), горизонтально масштабируемое хранилище данных и системы работы с данными in-memory.

Классическая реляционная СУБД – это система, данные в которой находятся и обрабатываются логически «на одном сервере» в одной базе с заданной структурой, и для работы с данными используется язык SQL. «Данная система – наиболее широко и давно применяемая технология работы с данными, где можно назвать таких производителей, как Oracle, Microsoft SQL и даже MS Access, входящий в пакет стандартных офисных приложений», – отмечает Алексей Нейман.

В случае с горизонтально масштабируемым хранилищем данных речь идет о системе, где данные распределяются по большому количеству серверов, те могут не иметь заранее определенной структуры, при этом в хранилище можно добавлять новые серверы. Анализ данных происходит на всех серверах параллель- но, результат параллельных вычислений консолидируется. «Так работает Hadoop, и для подобных вычислений была придумана технология MapReduce», – поясняет Алексей Нейман.

Системы работы с данными in-memory позволяют создавать и анализировать структурированные и слабоструктурированные данные в режиме реального

времени с высокой производительностью. «Это системы, сочетающие в себе достоинства OLTP (трансакционных) и OLAP (аналитических) систем, являющиеся обработчиком так называемых горячих, то есть наиболее актуальных на настоящий момент данных. Примером такой системы может служить SAP HANA», – говорит Алексей Нейман.

Все перечисленные системы, по его словам, используются лидерами как зарубежного, так и российского рынка. Hadoop считается наименее затратным способом организации неструктурированного хранилища данных, он практически не ограничен по объемам, но сильно ограничен по производительности вычислений. Реляционные базы данных отлично работают с хорошо структурированной индексированной информацией, но производительность падает при увеличении обрабатываемых объемов данных и количества пользователей, конкурирующих за запись и анализ данных. Вычисления in-memory – самый затратный способ хранения и обработки данных, но на порядок выигрывает в производительности.

«ИТ-службы умело комбинируют для аналитиков все эти технологии, обеспечивая максимум аналитических возможностей для своих сотрудников при минимуме затрат на это», – рассказывает Алексей Нейман.

Как устроен процесс обработки данных и что на него влияет

Работа с большими данными начинается с проработки запроса: определяется, какие процессы заказчик собирается менять с помощью аналитики и как намерен ее использовать. «Для этого мы сначала делаем

прототип решения, чтобы на нем собрать всевозможные вопросы и проблемы до внедрения основного решения. Для успешного построения прототипа необходима совместная работа специалиста с опытом внедрения систем машинного обучения и эксперта из бизнеса, именно от успеха их совместной работы зависит результат проекта», – говорит Николай Князев.

Далее с данными, по словам Алексея Неймана, происходит следующее: сохранение в хранилище; формирование на их основе бизнес-гипотезы; поиск подходящих под гипотезу данных; размещение данных; подготовка – чистка, обогащение, «причесывание», гармонизация данных для эксперимента; проведение эксперимента, подтверждение гипотезы; тестирование полученной модели (аналитики), валидация ее устойчивости; планирование и интеграция модели (аналитики) в бизнес-процесс; обеспечение необходимого потока данных в бизнес- процесс; внедрение модели (аналитики) в бизнес-процесс; периодическая валидация корректности работы модели.

Получение качественной модели зависит не только от наличия большого объема данных, но и от ошибок, возникающих на каждом из этапов обработки. «Речь, например, идет об ошибках выборки, связанных напрямую с источниками больших данных. Априори не существует идеальных источников данных, особенно большое количество пропусков/ошибок наблюдается при работе со слабоструктурированными источниками данных, такими как социальные сети, логи событий, click-streaming, фото, видео и так далее. При построении моделей для обработки подобных данных часто применяются подходы на базе нейронных сетей, которые, с одной стороны, показывают впечатляющую точность, а с другой – сильно зависимы от

стабильности и качества входного потока данных. При существенном изменении заполняемости входных параметров нейронные сети могут отрабатывать [данные на входе] непредсказуемым образом, что может повлечь за собой серьезные последствия», – пояснил Денис Рыбченко.

Сложности с извлечением и обработкой данных также возникают из-за их огромных объемов. «Обычно для этого используются распределенные системы обработки информации. При обработке информации подобным образом, с одной стороны, повышается скорость обработки информации, а с другой – повышается отказоустойчивость системы в целом, поскольку обработка производится на разных узлах (серверах), а значит, при выходе из строя отдельного узла обработка не остановится и возможные потери данных будут минимальны. Однако при постоянном росте объема хранимой базы требуется разработка дополнительных инженерных решений по оптимизации и масштабированию распределенной системы обработки информации», – отмечает Денис Рыбченко. При работе с данными также могут возникать ошибки, связанные с корреляцией информации из разных источников.

«Порой из-за запутанной причинно-следственной связи невозможно однозначно предсказать, как в результате поведет себя модель прогноза какого- либо события. И главное, при огромных объемах хранимых данных на источнике может случиться их частичная потеря, и, соответственно, выборка данных будет неполной и несбалансированной. Но выявить факты потери данных очень и очень сложно из-за их объема. Поэтому работа подобных решений зависит как от качества самих данных, включая их применимость и полноту, так и от корректности реализации алгоритмов их обработки, а это

уже задача инженеров данных, специалистов по работе с ними и аналитиков», – поясняет Денис Рыбченко.

По словам руководителя направления big data компании «Крок» Егора Осипова, под качеством данных подразумевают то, насколько бизнес может им доверять, насколько они отражают реальность: «Именно поэтому тематика качества данных сейчас так популярна – для большинства компаний, где нет выстроенной культуры работы с данными, это острый вопрос».

Однако не все игроки рынка придерживаются такого мнения. Так, руководитель ИТ-блока банка «Открытие» Сергей Русанов считает, что качественные или некачественные данные – это вопрос цели обработки. «Для одних моделей достаточен приблизительный подход, для других, например для выявления мошенничества или определения дефолта, нужны “чистые” данные, чтобы были точные предсказания и минимальное количество ложных срабатываний», – поясняет Сергей Русанов. Таким образом, то, насколько качественно собираются данные для аналитики, ничуть не менее важно, чем то, как она создается.

Что получают, обработав данные

В результате обработки больших данных можно обучать модели, позволяющие делать точные прогнозы для бизнеса. Эти прогнозы могут быть типовыми и индивидуальными. Более распространены последние.

В качестве примеров прогнозов можно привести прогнозы поведения пользователя на сайте, предпочтений покупателя на основе информации социальных сетей, количества проданной продукции, оттока клиентов. На

основе прогнозов строятся различные рекомендательные модели для клиентов и системы поддержки принятия решений для поставщиков.

Как работают рекомендательные системы можно понять на примере сервисов компании «Норбит». «Наши системы подсказывают, что еще можно предложить клиентам. Покупатель приходит на сайт ювелирного магазина за кольцом. В момент оплаты сервис предлагает дополнительные продукты, которые могут заинтересовать покупателя и которые с наибольшей вероятностью будут добавлены в корзину», – поясняет руководитель направления «Машинное обучение» «Норбит» Дмитрий Тимаков.

Аналогично работают рекомендательные сервисы банков. «Представьте, что банк запускает новый продукт. Чтобы рассказать о нем, компания проводит массовую рассылку с одним и тем же содержанием – описанием, стоимостью и так далее. Пользователи получают одинаковые письма, хотя для многих из них предложение даже нерелевантно. Аналитика, во-первых, позволяет точно определить сегмент аудитории, потенциально заинтересованный в новом продукте. Во-вторых, анализ позволяет выделить самые интересные для пользователей особенности предложения, адаптировать текст, иллюстрации и тональность сообщений – обратиться к клиентам на понятном языке. В итоге пользователи получают рассылку с релевантным предложением, в котором компания обращается персонально к ним. Конверсия и продажи продукта растут», – рассказывает руководитель аналитических сервисов Predict Mail.ru Group Роман Стятюгин.

Как помогают рекомендательные сервисы при технологическом процессе, можно понять на примере решения Softline. «Наш продукт AILine

прогнозирует результат технологического процесса и выдает рекомендации. Например, повар готовит борщ, он покупает свеклу, мясо, кладет соль, доводит все до определенной температуры. Система определяет, сколько нагревался суп, какая используется соль, какие продукты, и сообщает, сколько его оптимально варить, а если есть риск пригорания из-за плохой посуды или сильного огня, сообщит об этом», – рассказал «Эксперту» Николай Князев.

Цена вопроса

Стоимость решений big data сильно варьируется. Например, цена продуктов компании «Крок» колеблется в диапазоне от нескольких миллионов до десятков миллионов рублей. Полная стоимость big data проекта у «Крок» составляет от 10 млн до 50 млн рублей. «Конечно, бывают и более крупные проекты, и менее масштабные, но подавляющее большинство попадает в этот диапазон. К этой цифре также стоит прибавить стоимость лицензий и “железа”, которые варьируются в еще большем диапазоне. Но обычно эта цифра сравнима со стоимостью работ», – поясняет Егор Осипов.

Цена big data решений зависит от того, насколько детальны прогнозы, рекомендации и оценки, которые благодаря этому можно получить. Есть и другие компоненты. Все составляющие факторы игроки рынка не называют, однако суть ценообразования из их слов становится понятна.

«Аналитические продукты и сервисы призваны улучшить бизнес-показатели компании, которая их внедряет. Это может быть дополнительная выручка или существенная экономия, оптимизация затрат. Соответственно, ценность аналитических решений полностью за- вязана на той пользе, которую она

при- носит клиенту», – объясняет Роман Стятюгин.

Максим Воеводский говорит о формировании цены на решения big data более конкретно, отмечая, что у продуктов oneFactor она определяется как доля от эффекта использования сервиса компании на бизнес заказчика, но не ниже себестоимости.

«Для этого мы часто проводим бес- платное тестирование наших продуктов, чтобы заказчик мог точно оценить, насколько эффективны наши продукты для его бизнеса. Если заказчик не может оценить сам, тогда мы помогаем: для этого у нас в штате есть индустриальные консультанты с многолетним опытом работы в соответствующих отраслях. После завершения тестирования и оценки мы назначаем стоимость, в среднем равную двадцати процентам от эффекта для бизнеса заказчика. Бывают случаи, когда оказывается, что двадцать про- центов эффекта для бизнеса заказчика – это ниже нашей себестоимости, тогда мы в такие проекты не идем», – рассказал «Эксперту» Максим Воеводский.

По словам Дмитрия Тимакова, стоимость продуктов «Ланит» зависит от сложности интеграции с источниками данных, требований по быстродействию (например, модель должна дать ответ либо через сутки, либо за одну секунду), необходимой точности работы моделей. На цену продуктов big data также влияют количество данных заказчика и сложность разработки, отмечает **Александр Зенькович**.

Иногда, например в случае геоаналитики сотового оператора Tele2, стоимость продукта зависит от технического задания.

«Проанализировать загрузку авто- дорог большого региона, маятниковую миграцию жителей Подмосковья или по- мочь бизнесу построить оптимальную логистику внутри одного региона – совершенно разные задачи, которые требу- ют разных трудозатрат. Отсюда разная стоимость. Конечно, есть и массовые продукты, стоимость которых фиксирована. Один из них – платформа “SMS-Таргет”, которая дает бизнесу возможность на- страивать и отправлять таргетированные SMS-рассылки. Отправка одного сообщения для бизнес-клиента стоит в среднем два-три рубля и зависит от объема рассылки и выбранных таргетов», – рассказал директор по аналитике больших данных Tele2 Антон Мерзляков.

Помимо продуктов big data свою цену имеют и специалисты, необходимые для аналитики больших данных. Их нужно много. По оценке Google, для успешной работы отдела больших данных требуются следующие специалисты:

- аналитик данных (позволяющий найти инсайды в данных);
- инженер данных (организующий систему хранения);
- статистик (проверяющий гипотезы);
- исследователь (выдвигающий гипотезы в отношении данных);
- data scientist (проверяющий применимость алгоритмов машинного обучения);
- ML-инженер (внедряет алгоритмы для использования на практике);
- специалист по предметной области, ставящий задачу;
- представитель бизнеса, принимающий решения (product owner).

«На практике эти роли могут сходиться в одном человеке. Стоимость времени опытных специалистов составляет около пяти тысяч рублей в час», – проинформировал Николай Князев. «По моим оценкам, для Москвы стоимость средней команды data scientist из пяти человек составляет около миллиона рублей в месяц. Для Санкт-Петербурга эта цифра меньше на 20 процентов, а для регионов – на 50 процентов и более», – рассказал Денис Рыбченко.

Как видно, нанять специалистов для работы с данными может выйти для бизнеса в копеечку. Именно поэтому он зачастую стоит перед дилеммой: делать аналитику самому или покупать ее?

«Если компания сама генерирует огромное количество данных, на которых эту аналитику нужно делать, то стоит задуматься о внедрении у себя технологий их обработки и разработки моделей и аналитики. Если основные данные для повышения эффективности деятельности компании лежат за внутренним периметром, то стоит задуматься о покупке аналитических сервисов “под ключ” и минимизации затрат на собственную инфраструктуру работы с данными», – рассказал Алексей Нейман.

Об этом же говорят последние исследования и проекты компании oneFactor. Согласно им, малому и среднему бизнесу незачем вкладываться в аналитику больших данных: они могут воспользоваться уже готовыми сервисами дата-монополистов. Речь идет об интернет-площадках, операторах связи, а также банках.

Данные на вырост

По словам **Александра Зеньковича**, если в компании есть повторяющиеся процессы, оптимизация которых даже на небольшой процент способна приносить постоянную прибыль, ей стоит пробовать искусственный интеллект. Так же считает и Алексей Нейман: «Все так или иначе ее покупают или создают сами, если хотят быть конкурентоспособными. Создать собственную аналитику на больших данных – это технологический вызов, с одной стороны, но и серьезное конкурентное преимущество – с другой».

Денис Рыбченко солидарен с Алексеем Нейманом. По его словам, определить какую-либо индустрию или тип компании, для которой аналитика больших данных нецелесообразна или тем более невозможна, очень сложно: единственное необходимое условие применения аналитики – наличие истории наблюдений и надлежащее качество данных.

«В крупных компаниях зачастую выше потенциал для оптимизации: больше процессов и больше самих собираемых данных, что дает больше возможностей для анализа и построения сложных моделей с использованием методов ИИ. Однако и для небольшой компании аналитика больших данных будет полезна, особенно если она имеет большой объем данных, доступный для анализа», – отметил Денис Рыбченко.

Николай Князев также обращает внимание на необходимость для аналитики определенного объема данных. Такого, на котором будут работать закон больших чисел и другие статистические законы. «Например, если вы собираетесь разделять пользователей по категориям, вам необходимо хотя бы по сто клиентов каждой категории. С другой стороны, сейчас активно развивается направление transfer learning и использования схожих данных;

грубо говоря, обучив модель на данных одной ресторанной сети, ее возможно применить для не большого кафе», – пояснил Николай Князев.

Таким образом, и отсутствие необходимого объема данных для бизнеса уже не проблема: при необходимости он все равно сможет воспользоваться аналитикой big data.

Сможем ли мы жить без больших данных

Если вдруг станет невозможно использовать искусственный интеллект для аналитики больших данных, то бизнесу ничего не останется, как снова прибегнуть к естественному интеллекту. Уже сейчас для передовых компаний, у которых важные процессы строятся в том числе на обработке больших данных и с применением ИИ, потеря таких технологий станет катастрофой.

«Последует невозможность выполнения функций средствами автоматизации на требуемом уровне, потребуется наем огромного штата операторов для ручной обработки данных или принятия решений. При этом качество решений будет значительно хуже и, как результат, станет приводить к большим запасам/резервам, неэффективной логистике, снижению продаж и так далее», – поясняет Денис Рыбченко.

«В первую очередь рухнет мировая паутина сайтов, так как станет невозможно поддерживать механизмы поиска и загрузки информации. А без интернета мир изменится так кардинально, что остальные нюансы покажутся несущественными», – заверил Николай Князев.

«Инстаграм и другие социальные сети перестанут существовать сразу. Рынок рекламы очень сильно просядет, 80 процентов представителей интернет-маркетинга разорятся. Около 90 процентов иконок на смартфоне перестанут работать, а сам смартфон надо будет поменять на кнопочный телефон. Биржевая торговля уже немыслима без современной аналитики данных. Медицина также откатится к уровню прошлого века: половина современной медицинской техники использует эти технологии. Но не без хорошего: нам перестанут приходить “письма счастья” из ГИБДД, так как не будет технологичной фото- и видеофиксации с распознаванием символов», – отмечает Алексей Нейман.

Продукты на основе больших данных открывают перед бизнесом безграничные возможности, и современный мир уже трудно представить без них. Однако без труда получить желаемое от аналитики big data нельзя. Нужны не только довольно значительные финансовые средства, но и как минимум полный порядок с данными.

Источник: «Эксперт»