

Методологии для бизнес-аналитиков и аналитических проектов

Аналитические проекты уникальны тем, что итогом их реализации является функционирующее программное решение, которое в отрасли называют *дата-продуктом (data product)*. Это продукт на основе данных, который используется организацией в повседневных бизнес-операциях и создает измеримую ценность. Типичным примером дата-продукта является предиктивная модель, которая прогнозирует вероятность того, что клиент щелкнет на всплывающем рекламном объявлении. Методология, которая сегодня лежит в основе большинства аналитических проектов, называется Межотраслевым стандартом для интеллектуального анализа данных (Cross Industry Standard for Data Mining, CRISP-DM). Хотя сам стандарт сегодня упоминается реже из-за появления более современных методологий, эти методологии по-прежнему базируются на основных этапах, перечисленных в CRISP-DM (<https://oreil.ly/ESg-0>). У каждого этапа есть цель и результаты, и на каждом могут участвовать бизнес-аналитики. На рис. 2.1 показаны этапы и возможные итерации аналитического проекта.

В этой главе мы рассмотрим шесть этапов аналитического проекта: понимание бизнес-задачи, изучение и подготовку данных, моделирование и оценку, а также развертывание. Мы также разберемся, что происходит с дата-продуктом после его внедрения в бизнес-процессы. Когда дата-продукт используется в бизнес-операциях, он считается находящимся «в проде». Когда решение находится в продакшене, оно требует поддержки и мониторинга, — эта стадия называется эксплуатацией модели.

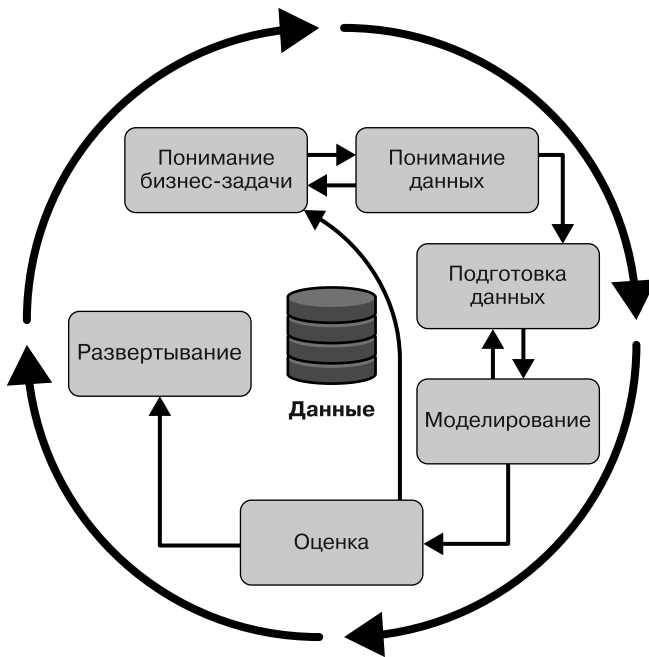


Рис. 2.1. Этапы и итерации CRISP-DM

Понимание бизнес-задачи

Начало аналитического проекта — самый важный этап, поскольку именно здесь определяется задача, которую необходимо решить. Бизнес-аналитики участвуют в этом процессе, чтобы изучить возникающие трудности, понять последствия и сформулировать задачу. Последнее включает в себя контекст, формулировку, допущения, сроки, ограничения и ожидаемые результаты. Бизнес-аналитикам важно участвовать в этом процессе, чтобы убедиться, что решается корректная, а не мнимая проблема. Слишком часто внимание уделяется не той задаче или решение предлагается до того, как станет понятна настоящая проблема. Основными результатами этого этапа будут определение бизнес-целей, оценку ситуации (контекста), определение ожидаемых результатов и подтверждение общего подхода к реализации проекта и его плана.

Определите бизнес-задачи

Хорошие задачи, как правило, можно четко сформулировать и решить, если определить желаемые результаты. Придерживаться популярного принципа «начинай с конца» — отличная стратегия для формулирования задачи. Четкое

определение бизнес-целей позволяет заинтересованным сторонам понять, чего они ожидают от проекта и что будет считаться успехом. Бизнес-цели должны быть четкими, конкретными и измеримыми. Рассмотрим пример.

Мы хотим снизить отток клиентов на 10 % в течение трех месяцев, прогнозируя его вероятность с помощью аналитической модели.

В данной формулировке есть информация, которая будет ценной при планировании аналитического проекта. Во-первых, результат может быть измерен в течение определенного периода. Во-вторых, основное внимание уделяется тому, что должно быть предсказано. Это обеспечивает детализацию, которую можно использовать для определения аналитического подхода, требуемых данных и специфики успеха. Хотя не все аналитические проекты будут основаны на предиктивной модели, постановка бизнес-целей на таком уровне делает описание задачи предельно ясным.

Оцените ситуацию

Как только задача полностью определена, бизнес-аналитики помогают сформулировать требования к аналитическому проекту. Требования аналитического проекта касаются не функциональности программного обеспечения, а того, какое поведение требуется от дата-продукта. Должен ли он что-то прогнозировать или как-то сегментировать клиентов? Требования к аналитическому проекту сосредоточены на информационных результатах и на их использовании в дальнейшем.

Следующий шаг — понять, какие ресурсы доступны для поддержки проекта. К ресурсам относятся технологические инструменты, вычислительные платформы, члены команды, обладающие необходимыми навыками, и бюджет. Аналитические проекты часто используют существующие технологические инструменты и вычислительные платформы, но что, если имеющейся инфраструктуры недостаточно для поддержки проекта? Кроме того, обладают ли сотрудники навыками, необходимыми для выполнения поставленной задачи? Прежде чем приступить к реализации проекта, необходимо устранить нехватку ресурсов.

Не всегда понятно, как результаты аналитического проекта повлияют на организацию. Иногда возникают непредвиденные последствия. Например, компания Target, как сообщается, предсказала беременность подростка (<https://oreil.ly/w7P7r>), что не входило в планы маркетингового алгоритма. Использование любого дата-продукта сопряжено с рисками, и важно заранее их обозначить.

Наконец, важно понимать, каков бюджет аналитического проекта. На это будет влиять то, как ваша организация распределяет средства на проекты, но не всегда понятно, какой бюджет необходим, из-за итеративного характера проекта.

Часто на начальном этапе выделяется фиксированная сумма, а после пилотного запуска дата-продукта предоставляются дополнительные средства.

Определите цели

Для достижения успеха важно наметить конкретные задачи. Бизнес-цели используются для определения и формулирования целей проекта. Как и бизнес-цели, цели проекта должны быть конкретными, своевременными и измеримыми. Критерии успеха напрямую зависят от целей, поставленных в рамках проекта. Мнение бизнес-аналитика здесь имеет решающее значение, поскольку критерии успеха должны четко показывать, какую пользу работа принесет бизнесу. Следует учитывать, что информационные продукты (например, модель прогнозирования) в первую очередь улучшают процесс принятия решений или создают новые возможности. Точные результаты могут быть недостижимы, поскольку в большинстве случаев неизвестно, как информационный продукт будет способствовать достижению конкретных результатов.

Определите подход и план

На данный момент должно быть собрано достаточно информации, чтобы можно было определить аналитический подход и план. Аналитические подходы — это то, на чем будет сосредоточено основное внимание в рамках проекта по созданию дата-продукта. Например, если информационным продуктом будет предиктивная модель, то основное внимание будет уделяться предсказательным подходам. Если целью является сегментация клиентов, то, скорее всего, здесь будут рассматриваться подходы к обучению модели без меток. Этот шаг, по сути, определяет рамки проекта. Необходимо проконсультироваться с бизнес-аналитиками, поскольку могут существовать ограничения, которые нужно учитывать, а также понять требования к срокам для составления графика.

Имейте в виду, что аналитические проекты не являются последовательными и не основаны на традиционном проектном подходе. План, как правило, представляет собой график ожидаемых этапов и временных рамок, на которые все будут ориентироваться. Из-за итеративного характера проекта можно вернуться к предыдущему этапу.

Оцените инструменты и методики

Последний шаг в понимании бизнес-задачи — определить, доступны ли необходимые технологические инструменты для поддержки проекта. Вероятно, потребуется новый подход, поскольку существующий не соответствует целям. Бизнес-аналитик может участвовать в оценке возможностей текущей или новой технологии. Цель — убедиться в наличии необходимой методологии до начала проекта.

Изучение и подготовка данных

Изучение и подготовка данных могут затянуться и занять значительную часть времени проекта. Сложность этого этапа заключается в том, что он ориентирован на поиск, а результатом являются очищенные, интегрированные и отформатированные данные с правильными атрибутами, которые можно использовать для построения аналитической модели. Непросто понять, какие данные следует выбирать и изучать, и бизнес-аналитик может с этим помочь. Ответ зависит от того, какую главную задачу необходимо решить. Например, если нужно создать предиктивную модель, то основное внимание будет уделяться выбору источников данных и атрибутов, которые следует учитывать.

Если на этапе изучения и подготовки данных не удастся получить полноценный набор данных, возможно, проект вернется на этап понимания бизнес-задачи. Иногда в принципе не существует данных для решения задач.

Оцените содержание и качество данных

Содержание данных можно анализировать с разных точек зрения. Одна из них подразумевает понимание содержания атрибутов данных. Чтобы понять содержание и способы использования данных, требуется ответить на базовые вопросы, выяснив тип данных, диапазон, структуру, допустимые значения и закономерности. Бизнес-аналитики играют ключевую роль в этом процессе, поскольку во многих организациях нет метаданных, которые можно было бы использовать для ответа на данный вопрос. Помимо понимания того, что представляют собой данные в бизнес-контексте, необходимо оценить их качество, чтобы определить, можно ли использовать их для аналитического проекта.

Проверка проводится по двум основным критериям: по полноте и корректности. Полнота означает наличие всех требуемых записей, а корректность проверяется применением специальных правил валидации. Сюда могут входить оценка точности, согласованности, детализации и своевременности информации.

Выберите и очистите данные

После оценки содержания и качества данных их нужно выбрать и очистить. Если данные не соответствуют критериям прогнозирования или имеют проблемы с качеством, их удаляют. Оставшиеся данные, необходимые для этапа моделирования, очищают и форматируют. Например, можно создать новые атрибуты на основе существующих или изменить формат даты с ММ-ДД-ГГГГ на ГГГГ-ММ-ДД. Способ очистки часто определяется предметной областью, а завершением этого этапа считается создание окончательного набора данных.

Подготовьте и интегрируйте данные

Для перехода к этапу моделирования и оценки необходимо создать окончательный набор данных. Для этого часто объединяют данные из нескольких источников. Интеграция обычно происходит путем сопоставления общих элементов в наборе данных перед их объединением. Например, данные могут быть сгруппированы по общему признаку, такому как город, или связаны общим ключом, в роли которого выступает идентификатор клиента. Финальная подготовка и объединение данных часто требуют навыков инженера по обработке данных. Такой специалист умеет очищать, формировать и интегрировать данные. При этом бизнес-аналитиков могут привлекать для проверки качества и согласованности итогового массива данных.

Создайте выборки данных для разработки модели

Окончательный набор данных обычно разделяют на части: обучающую, тестовую и проверочную выборки. Они формируются из одного и того же набора данных и являются взаимоисключающими (то есть один и тот же пример не может находиться в разных выборках). Подготовленные таким образом выборки обычно сохраняют во внешнем файле или в таблице базы данных.

Моделирование и оценка

На этапах моделирования и оценки основное внимание уделяется выбору аналитического метода, созданию модели на его основе и оценке модели с точки зрения того, как хорошо она решает поставленную бизнес-задачу. Бизнес-аналитики участвуют в выборе способа, чтобы убедиться в его соответствии ожидаемым результатам проекта, и непосредственно оценивают способность модели решать бизнес-задачу. Если оценка окажется неудовлетворительной, возможно, проект придется вернуть на этап изучения и подготовки данных, чтобы найти новые источники данных.

Выберите метод аналитики

В зависимости от решаемой бизнес-задачи можно рассмотреть сотни аналитических методов. Для задач прогнозирования обычно используют методы регрессии или классификации. Другие аналитические задачи могут не быть предсказательными, и тогда применяются методы обучения без учителя либо их комбинации с методами обучения с учителем. Различные подходы будут рассмотрены в последующих главах.

Создайте и оцените модель

Создание модели включает в себя индукцию (использование обучающих данных) и дедукцию (оценку с помощью тестовых данных). Этот процесс часто итеративен и направлен на непрерывное совершенствование для достижения максимально возможной точности. Часто строят несколько моделей на основе разных методов, чтобы сформировать эталон. Бизнес-аналитиков привлекают для анализа результатов и оценки каждой модели. Модель или модели могут быть выбраны в качестве кандидатов для развертывания. Оценка результатов подробнее рассматривается в последующих главах.

Развертывание

Развертывание модели — это этап, на котором модель тестируется в реальных бизнес-операциях. Оно может проходить в разных форматах, но часто начинается с пробного или пилотного проекта. Модели внедряются в промышленную среду, где программная версия модели обрабатывает реальные данные и выдает свой прогноз. Результаты работы затем используются для принятия решений в организации, а итоги принятых решений оцениваются. Например, если модель прогнозирует вероятность покупки, ее успех оценивается по коэффициенту эффективности — сравнению результатов, полученных с моделью и без нее. Бизнес-аналитики непосредственно участвуют в развертывании и оценке эффективности модели, поскольку именно на этом этапе проявляется ее ценность для бизнеса.

Оцените эффективность модели

Эффективность модели до развертывания оценивается по тому, насколько точно она предсказывает результаты на исторических данных. После внедрения она оценивается с точки зрения того, насколько хорошо решает поставленные бизнес-задачи. Бизнес-аналитики играют здесь ключевую роль, поскольку понимают контекст использования модели и могут оценить ее полезность для бизнеса. Модель сначала проверяется в пилотном режиме — это позволяет создать контролируемые условия и управлять влиянием на бизнес-процессы. Если она признается полезной для бизнеса, то развертывается в реальной среде.

Определите интервалы оценки

На начальном этапе модель может работать хорошо, но важно определить, как часто будет проводиться ее повторная оценка. В течение бизнес-циклов данные меняются. Например, следующая маркетинговая кампания может привести

к новым закономерностям в данных, которые повлияют на качество модели. Бизнес-аналитики понимают бизнес-процессы и должны предлагать, как часто оценивать качество модели.

Эксплуатация модели

Эксплуатация модели подразумевает постоянную поддержку и мониторинг после ее внедрения в производственную среду, что гарантирует ее эффективное функционирование и пользу для бизнес-операций. Не все аналитические решения доходят до этого этапа, поскольку некоторые не проходят оценку. Модели, которые достигают этой стадии, считаются ценными информационными продуктами для компании, поскольку способствуют улучшению процесса принятия решений и повышению ценности бизнеса. Бизнес-аналитики играют ключевую роль на этом этапе, часто выступая экспертами, которые определяют и контролируют, как именно продукт используется в бизнес-операциях.

Мониторинг моделей

В рамках развертывания аналитического решения определяются критерии оценки, которые становятся пороговыми значениями для мониторинга. Отслеживаются такие показатели, как точность, дрейф модели и дрейф признаков, чтобы понять, требуется ли дообучение системы или ее вывод из эксплуатации. С бизнес-аналитиками консультируются, чтобы определить, как текущая производительность модели влияет на ее ценность для бизнеса.

Срок службы модели

Модели строятся на основе исторических данных и хорошо работают, только пока текущие данные сопоставимы с историческими. Поскольку бизнес-процессы меняются, данные также претерпевают изменения. Модели имеют ограниченный срок службы, напрямую связанный со скоростью изменений в организации. Чем активнее меняется среда (например, новые продукты, услуги, предложения, слияния или поглощения), тем короче срок службы модели. Бизнес-аналитики помогают понять, как новые факторы в бизнесе могут повлиять на срок службы аналитического решения.

Дообучение

Модели можно обновить, просто дообучив их на основе новых данных. Цикл дообучения позволяет не только обновить модель, но и улучшить качество ее предсказаний, а также продлить срок службы. Иногда такой процесс

невозможен из-за большого количества изменений, произошедших в организации. В таких случаях следует рассмотреть создание новой версии системы, иначе модель может быть выведена из эксплуатации.

Резюме

В этой главе мы рассмотрели этапы, характерные для аналитического проекта, и выделили те, на которых привлекают бизнес-аналитика. Эти специалисты ценны благодаря своим экспертным знаниям в предметной области и пониманию того, как решения повлияют на бизнес. Они играют ключевую роль в формулировании бизнес-задач, выборе данных для их решения и построения моделей, а также в оценке того, насколько эффективно готовое решение справляется с исходной задачей. В главе 3 мы рассмотрим основные языки, используемые в аналитических проектах: R и Python.