**"Организация команды для работы с данными"**

**РЕШЕНИЕ**

1. **Бизнес-анализ (Business Understanding)**

*1.1 Цель проекта (Business objectives)*

Участники: Аналитики, Владелец продукта, Заказчики

Аналитик знакомится с Владельцем продукта выясняет его истинные цели подробно или предлагает ему типовые решения.

Основные вопросы:

Организационная структура: кто участвует в проекте со стороны Владельца, кто выделяет деньги под проект, кто принимает ключевые решения, кто будет основным пользователем?

**Результат -** создание справочника участника проектов, имеющихся ресурсов.

Какова бизнес-цель проекта?

**Результат -** Удержание клиентов, снабжение только востребованными товарами, чтобы избегнуть затаривания складов.

Существуют ли какие-то уже разработанные решения? Если существуют, то какие и чем именно текущее решение не устраивает?

**Результат –** выбор основного и запасного вариантов достижения целей.

*1.2 Текущая ситуация (Assessing current solution)*

Оценка, достаточности ресурсов для проекта.

**Результат –** ответы на вопросы**:**

Есть ли доступное оборудование или его необходимо закупать?

Где и как хранятся данные, будет ли предоставлен доступ в эти системы, нужно ли дополнительно докупать/собирать внешние данные?

Сможет ли Владелец предоставить для консультаций на данный проект?

Нужно описать вероятные риски проекта, а также определить план действий по их уменьшению.

**Результат –**  определение рисков. (Нарушение сроков. Финансовые риски (потеря инвесторов, изменение условий взаимодействия с банками). Недостаточность количества и качества данных, которые не позволят получить эффективную модель. Данные качественные, но полученные результаты не имеют коммерческого интереса.)

**Результат –**  Глоссарий используемой в рамках проекта терминологии.

**Результат –**  Предварительная оценка ROI.

*1.3 Решаемые задачи с точки зрения аналитики (Data Mining goals)*

**Результат –** ответы на вопросы**:**

Какую метрику мы будем использовать для оценки результата моделирования?

Каков критерий успешности модели?

*1.4 План проекта (Project Plan)*

**Результат –** план проекта. План должен содержать оценку всех шести фаз внедрения.

**2. Анализ данных (Data Understanding)**

Цель шага – описательная аналитика.

Участники: Аналитики, BI специалисты, Data Engineer, Специалисты по БД

*2.1 Сбор данных (Data collection)*

**Результат –** Разделение данных на собственные (1st party data), сторонние данные (3rd party), «потенциальные» данные (для получения которых необходимо организовать сбор).

Необходимо проанализировать все источники, доступ к которым предоставляет заказчик. Если собственных данных недостаточно, возможно, стоит закупить сторонние или организовать сбор новых данных.

*2.2 Описание данных (Data description)*

**Результат –**  описать данные во всех источниках (таблица, ключ, количество строк, количество столбцов, объем на диске). Если объем слишком велик для используемого ПО, создаем сэмпл данных. Считаем ключевые статистики по атрибутам (минимум, максимум, разброс, кардинальность и т.д.).

*2.3 Исследование данных (Data exploration)*

**Результат –** гипотезы относительно того, как имеющиеся данные помогут решить задачу. Список атрибутов, которые потенциально полезны.

*2.4 Качество данных (Data quality)*

**Результат –**  оценка качества данных. Пропущенные значения. Ошибки данных (опечатки). Неконсистентная кодировка значений («M» и «male» в разных системах)

**3. Подготовка данных (Data Preparation)**

Участники: DS специалисты, Data Engineer, Специалисты по БД

*3.1 Отбор данных (Data Selection)*

**Результат –** ответы на вопросы**:**

Какова потенциальная релевантность атрибута решаемой задаче?

Достаточно ли качественный атрибут для использования в модели?

Стоит ли включать коррелирующие друг с другом атрибуты?

Есть ли ограничения на использование атрибутов (персональная информация)?

Например, политика компании может запрещать использование атрибутов с персональной информацией в качестве предикторов.

*3.2 Очистка данных (Data Cleaning)*

**Результат –** Пропущенные значения => заполнить, либо удалить из рассмотрения. Ошибки в данных => исправить вручную либо удалить из рассмотрения. Несоответствующая кодировка => привести к единой кодировке

**Результат –** 3 списка атрибутов – качественные атрибуты, исправленные атрибуты и забракованные.

*3.3 Генерация данных (Constructing new data)*

**Результат –**  генерация признаков (feature engineering)

агрегация атрибутов (расчет sum, avg, min, max, var и т.д.),

генерация кейсов (например, oversampling или алгоритм SMOTE),

конвертация типов данных для использования в разных моделях,

нормализация атрибутов (feature scaling),

заполнение пропущенных данных (missing data imputation).

*3.4 Интеграция данных (Integrating data)*

**Результат –** «горизонтальное» соединение (Merge) и «вертикальное» объединение (Append), а также агрегация данных.

3.5 Форматирование данных (Formatting Data)

**Результат –**  привести данные к формату, пригодному для моделирования.

**4. Моделирование (Modeling)**

Участники: DS специалисты

Цель — обучение моделей.

*4.1 Выбор алгоритмов (Selecting the modeling technique)*

**Результат –** Выбор модели.

**Результат –** ответы на вопросы**:**

Достаточно ли данных?

Сможет ли модель обработать пропуски данных?

Сможет ли модель работать с имеющимися типами данных или необходима конвертация?

*4.2 Планирование тестирования (Generating a test design)*

**Результат –**  на чем мы обучать, а на чем тестировать модель. Разделение выборки на части.

*4.3 Обучение моделей (Building the models)*

**Результат –**  несколько обученных моделей.

Показывает ли модель какие-то интересные закономерности?

Какова скорость обучения/применения модели?

Были ли проблемы с качеством данных?

*4.4 Оценка результатов (Assessing the model)*

**Результат –**  список моделей, отсортированный по объективному и/или субъективному критерию.

технический анализ качества модели,

оценка готовности модели к внедрению,

достигаются ли заданные критерии качества,

оценить результаты с точки зрения достижения бизнес-целей с аналитиками заказчика.

результат моделирования понятен (модель, атрибуты, точность)

результат моделирования логичен

инфраструктура готова к внедрению модели

**5. Оценка результата (Evaluation)**

Участники: Аналитики, BI специалисты

Цель - оценка результатов проекта.

*5.1 Оценка результатов моделирования (Evaluating the results)*

**Результат –**  оцениваем результатов с точки зрения достижения бизнес-целей.

*5.2 Разбор (Review the process)*

**Результат –** ответы на вопросы**:**

Можно ли было какие-то шаги сделать более эффективными?

Какие были допущены ошибки и как их избежать в будущем?

Были ли не сработавшие гипотезы? Если да, стоит ли их повторять?

Были ли неожиданности при реализации шагов? Как их предусмотреть в будущем?

*5.3 Принятие решения (Determining the next steps)*

**Результат –** отбираем модели для внедрения.

**6. Внедрение (Deployment)**

Участники: DS специалисты, Разработчики, Data Engineer

Цель - внедрение модели (если проект предполагает этап внедрения). Физическое добавление функционала или инициирование изменений в бизнес-процессах компании.

*6.1 Планирование развертывания (Planning Deployment)*

**Результат –** что и в каком виде будет внедряться, подготовка технического плана внедрения

*6.2 Настройка мониторинга модели (Planning Monitoring)*

**Результат –** ответы на вопросы**:**

Какие показатели качества модели будут отслеживаться?

Как понимаем, что модель устарела?

Если модель устарела, достаточно ли будет ее переобучить или нужно организовывать новый проект?

*6.3 Отчет по результатам моделирования (Final Report)*

**Результат –**  отчет о результатах моделирования, в который добавляются результаты по каждому шагу, начиная от первичного анализа данных и заканчивая внедрением модели, рекомендации по дальнейшему развитию модели.