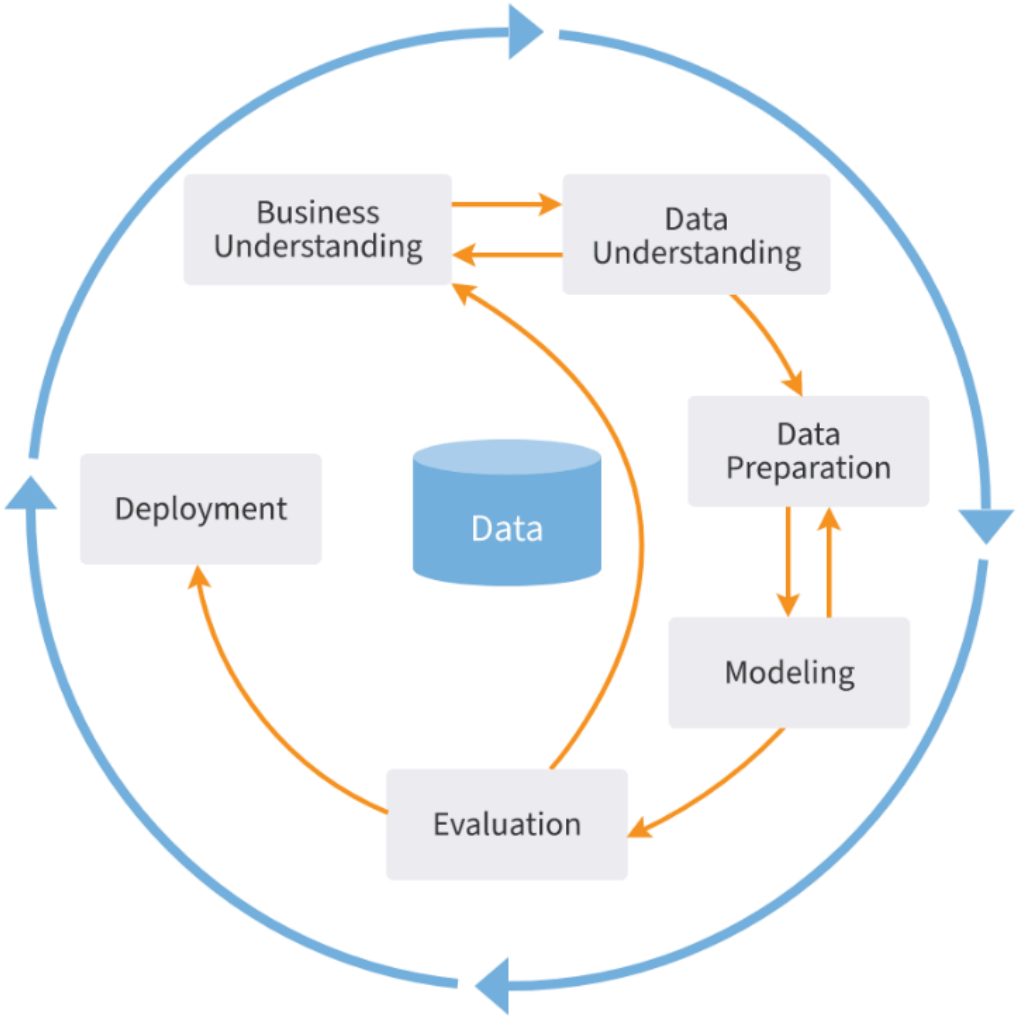
Процесс разработки сервисов машинного обучения

Данная методология разработана на основе стандарта CRISP-DM.

Процесс разработки сервисов машинного обучения итеративный и должен состоять из 6 частей:

1. **Бизнес-анализ (Business Understanding)**
2. **Анализ данных (Data Understanding)**
3. **Подготовка данных (Data Preparation)**
4. **Моделирование (Modeling)**
5. **Оценка результата (Evaluation)**
6. **Внедрение (Deployment)**



1. Бизнес-анализ (Business Understanding)

В первую очередь необходимо определиться с целями и скоупом проекта.
Для этого нужно найти ответы на следующие вопросы:

- *Организационная структура: кто участвует в проекте со стороны заказчика сервиса, кто будет основным пользователем?*
- *Собираем контакты, создаем рабочие чаты.*
- *Какова бизнес-цель проекта? Например, уменьшение оттока клиентов.*
- *Существуют ли какие-то уже разработанные решения? Если существуют, то какие и чем именно текущее решение не устраивает?*

1.1 Текущая ситуация (Assessing current solution)

Оцениваем, хватает ли ресурсов для проекта.

- Есть ли доступное железо 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 заказчик выделить своих экспертов для консультаций на данный проект?

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

Типичные риски следующие.

- Не уложиться в сроки.
- Малое количество или плохое качество данных, которые не позволят получить эффективную модель.
- Данные качественные, но закономерности в принципе отсутствуют и, как следствие, полученные результаты не
- интересны заказчику.

1.2 Решаемые задачи с точки зрения аналитики (Data Mining goals)

Выполняем постановку в технических терминах. Для этого нужно ответить на следующие вопросы:

- Какую метрику мы будем использовать для оценки результата моделирования (а выбрать есть из чего: Accuracy, RMSE,
- AUC, Precision, Recall, F-мера, R2, Lift, Logloss и т.д.)?
- Каков критерий успешности модели (например, считаем AUC равный 0.65 — минимальным порогом, 0.75 — оптимальным)?
- Если объективный критерий качества использовать не будем, то как будут оцениваться результаты?

1.3 План проекта (Project Plan)

Как только получены ответы на все основные вопросы и ясна цель проекта, время составить план проекта. План должен
содержать оценку всех шести фаз внедрения.

2. Анализ данных (Data Understanding)

Цель шага – понять слабые и сильные стороны предоставленных данных, определить их достаточность, предложить идеи, как
их использовать, и лучше понять процессы заказчика. Для этого мы строим графики, делаем выборки и рассчитываем
статистики.

2.1 Сбор данных (Data collection)
Выгружаем необходимые данные (или срез данных если их объем слишком велик) из источников.
Версионируем данные средствами DVC.

2.2 Исследование данных (Data exploration)
Исследуем данные, чтобы сформулировать гипотезы относительно того, как эти данные помогут решить задачу. Проверяем
качество данных.

Ориентировочный список для проверки данных:

1. Загрузить репрезентативную выборку из набора данных
2. Провести предварительный анализ всей выборки.
    - определить тип данных в каждом столбце
    - Категориальные данные (Номинальные, Порядковые)
    - Числовые данные (дискретные, непрерывные, интервальные, отношения)
    - При необходимости преобразовать данные к нужным типам
    - Проверить на выбросы, отсутствующие значения, невалидные значения(например в системе случился сбой и в поле с именемпопала длина просмотра).

(по результатам предварительного анализа сделать визуализацию, обычно это табличка с характеристиками или какой то из
профайлеров)

3. На основе предыдущего анализа выполнить очистку данных (обработать выбросы, отсутствующие значения, удалить невалидные значения)
4. Удалить из рассмотрения неинформативные данные. (лишние идентификаторы, служебные поля, поля с очень малым количеством значений)
5. Провести статистический анализ оставшихся данных
    a. рассчитать ключевые статистики для каждого типа данных
    b. построить распределения (тип графика выбрать в зависимости от данных, часто полезно построить гистограмму, но иногда лучше воспользоваться линейным графиком или посмотреть распределение во времени с помощью scaterplot)
6. Провести корреляционный анализ
    a. Для количественных данных нормализовать данные и построить матрицу корреляции Пирсона

    b. сделать выводы на основе матрицы (найти утечки данных, найти важные признаки линейно влияющие на целевой показатель, определить гипотезы по конструированию признаков)
    c. Для количественных и порядковых данных - построить матрицу корреляции Спирмена сделать выводы аналогично предыдущему анализу, только учесть тип данных
    d. Для всех данных построить матрицу корреляции Пфика
    e. Сделать выводы на основе анализа, сделать оценку между всеми типами корреляции(на пересекающихся данных),
    f. попытаться найти объяснения различиям.
    g. Сделать выводы о наличии или отсутствии нелинейных связей.

7. Провести обработку данных, на основе выводов полученных в прошлых шагах.
    a. провести дополнительную очистку
    b. выполнить нужный тип энкодинга(если требуется)
    c. Сконструировать новые признаки.

8. Построить Графики взаимодействия полученных данных с целевым показателем
    a. Для количественных данных линейные графики на нормализованных данных
    b. для категориальных данных с малым количеством категорий построить ScaterPlot во времени
    c. для категориальных данных с большим количеством показателей построить heatmap во времени(обычно строят, только для ризнаков которые показывают высокие коэффициенты корреляции и потенциально интересны, пример такого графика спектрограмма в анализе звука)
9. Сделать выводы на основе проведенного анализа и учитывая особенности планируемой архитектуры модели.
    a. Какие данные и почему нельзя использовать в модели
    b. какие данные можно использовать без преобразования
    c. какие данные можно использовать выполнив преобразование
    d. какие новые признаки нужно использовать и почему
    e. есть ли смысл использовать один набор признаков или построить разные модели на подмножестве признаков и почему.
    f. Выделить итоговый список необходимых данных
10. Описать ожидания от модели на проанализированных данных

3. Подготовка данных (Data Preparation)

Цель этапа – подготовить обучающую выборку для использования в моделировании.

3.1 Отбор данных (Data Selection)

Для начала нужно отобрать данные, которые мы будем использовать для обучения модели.
Отбираются как атрибуты, так и кейсы.
Например, если мы делаем продуктовые рекомендации посетителям сайта, мы ограничиваемся анализом только
зарегистрированных пользователей или берем всех пользователей.

При выборе данных аналитик отвечает на следующие вопросы.

- Какова потенциальная релевантность атрибута решаемой задаче?
- Так, электронная почта или номер телефона клиента как предикторы для прогнозирования явно бесполезны. А вот домен почты (mail.ru, gmail.com) или код оператора в теории уже могут обладать предсказательной способностью.
- Достаточно ли качественный атрибут для использования в модели? Если видим, что большая часть значений атрибута пуста, то атрибут, скорее всего, бесполезен.
- Стоит ли включать коррелирующие друг с другом атрибуты?
- Есть ли ограничения на использование атрибутов? Например, политика компании может запрещать использование атрибутов с персональной информацией в качестве предикторов.

3.2 Очистка данных (Data Cleaning)

Пропущенные значения => нужно либо их заполнить, либо удалить из рассмотрения
Ошибки в данных => попробовать исправить вручную либо удалить из рассмотрения
Несоответствующая кодировка => привести к единой кодировке

3.3 Генерация данных (Constructing new data)

К генерации данных можно отнести:

- агрегацию атрибутов (расчет sum, avg, min, max, var и т.д.),
- генерацию кейсов (например, oversampling или алгоритм SMOTE),
- конвертацию типов данных для использования в разных моделях (например, SVM традиционно работает с
- интервальными данными, а CHAID с номинальными),
- нормализацию атрибутов (feature scaling),
- заполнение пропущенных данных (missing data imputation).
- аугментация данных.

3.4 Версионирование данных (Data versioning)

- выбираем тип разделения данных
- разбиваем данные на train, test, val
- загружаем данные в хранилище используя пайплайн DVC

4. Моделирование (Modeling)

На четвертом шаге выполняется обучение моделей. Как правило, оно выполняется итерационно – мы пробуем различные
модели, сравниваем их качество, делаем перебор гиперпараметров и выбираем лучшую комбинацию.

Моделирование состоит из этапов:

- Выбор алгоритмов (Selecting the modeling technique)
- Планирование тестирования (Generating a test design)
- Обучение моделей (Building the models)
- Оценка результатов (Assessing the model)

После того, как был сформирован пул моделей, нужно их еще раз детально проанализировать и выбрать модели-победители. На
выходе необходимо иметь список моделей, отсортированный по объективному и/или субъективному критерию.

Задачи шага:

- провести технический анализ качества модели (ROC, Gain, Lift и т.д.),
- оценить, готова ли модель к внедрению в КХД (или куда нужно),
- достигаются ли заданные критерии качества,
- оценить результаты с точки зрения достижения бизнес-целей. Это можно обсудить с аналитиками заказчика.

Если критерий успеха не достигнут, то можно либо улучшать текущую модель, либо пробовать новую.

Прежде чем переходить к внедрению нужно убедиться, что:

- результат моделирования понятен (модель, атрибуты, точность)
- результат моделирования логичен
- Например, мы прогнозируем отток клиентов и получили ROC AUC, равный 95%. Слишком хороший результат – повод
- проверить модель еще раз.
- мы попробовали все доступные модели
- инфраструктура готова к внедрению модели

5. Оценка результата (Evaluation)

Результатом предыдущего шага является построенная математическая модель (model), а также найденные закономерности
(findings). На пятом шаге мы оцениваем результаты проекта.

5.1 Оценка результатов моделирования (Evaluating the results)

Если на предыдущем этапе мы оценивали результаты моделирования с технической точки зрения, то здесь мы оцениваем
результаты с точки зрения достижения бизнес-целей.

Адресуем следующие вопросы:

- Формулировка результата в бизнес-терминах.
- В целом насколько хорошо полученные результаты решают бизнес-задачу?
- Найдена ли какая-то новая ценная информация, которую стоит выделить отдельно?

5.2 Разбор полетов (Review the process)

Необходимо проанализировать ход проекта и сформулировать его сильные и слабые стороны. Для этого нужно пройтись по всем
шагам:

- 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)

Далее нужно либо внедрять модель, если она устраивает заказчика, либо, если виден потенциал для улучшения, попытаться
еще ее улучшить.
Если на данном этапе у нас несколько удовлетворяющих моделей, то отбираем те, которые будем дальше внедрять.

6. Внедрение (Deployment)

На данном шаге осуществляется внедрение модели (если проект предполагает этап внедрения). Причем под внедрением может
пониматься как физическое добавление функционала, так и инициирование изменений в бизнес-процессах компании.

6.1 Планирование развертывания (Planning Deployment)

- Важно зафиксировать, что именно и в каком виде мы будем внедрять, а также подготовить технический план внедрения
- Продумать, как с внедряемой моделью будут работать пользователи
- Определить принцип мониторинга решения. Если нужно, подготовиться к опытно-промышленной эксплуатации.

6.2 Настройка мониторинга модели (Planning Monitoring)

Основные вопросы мониторинга:

- Какие показатели качества модели будут отслеживаться?
- Как понимаем, что модель устарела?
- Если модель устарела, достаточно ли будет ее переобучить или нужно организовывать новый проект?
- Определяемся с параметрами A\B теста

6.3 Документация (documentation)

Предполагается что документирование происходит на всех этапах разработки модели(сервиса машинного обучения) однако на
данном этапе документация финализируется, обогащается информацией о технических нюансах развертывания сервиса.