CRISP-DM

Методология анализа данных

Константин Коточигов, CleverDATA

О Методологии CRISP-DM

CRISP-DM = CRoss Industry Standard Process for Data Mining

Стандарт, описывающий общие процессы и подходы к аналитике данных, используемые в промышленных data-mining проектах независимо от конкретной задачи или индустрии.

История создания

- Разработан в 1996 году
- Авторство: нынешние DaimlerChrysler, SPSS и Teradata
- Дорабатывалась при участии 200 компаний различных индустрий
- IBM SPSS Modeler

Отличительные особенности

- больший фокус на бизнес-целях
- документирование каждого шага

О Методологии CRISP-DM

What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]

2014 poll 2007 poll

CRISP-DM (86)	43% 42%
My own (55)	27.5% 19%
SEMMA (17)	8.5% 13%
Other, not domain-specific (16)	8% 4%
KDD Process (15)	7.5% 7.3%
My organizations' (7)	3.5% 5.3%
A domain-specific methodology (4)	2% 4.7%
None (0)	0% 4.7%

О Методологии CRISP-DM

Основные этапы DM-проекта

1. Business Understanding (Бизнес Постановка)
2. Data Understanding (Анализ данных)
3. Data Preparation (подготовка данных)
4. Modeling (Моделирование)
5. Evaluation (Оценка результата)
6. Deployment (Внедрение)

Основной Workflow

Этапы DM-проекта и включаемые в них задачи

Business Understanding/ Бизнес-анализ	Data Understanding/ Анализ данных	Data Preparation/ Подготовка данных	Modeling/ Моделирование	Evaluation/ Оценка решения	Deployment/ Внедрение
Determine Business Objectives/ **Определение бизнес-целей**	Collect Initial Data/ **Сбор данных**	Select Data/ **Выборка данных**	Select Modeling Techniques/ **Выбор алгоритмов**	Evaluate Results/ **Оценка результатов**	Plan Deployment/ **Внедрение**
Assess Situation/ **Оценка текущей ситуации**	Describe Data/ **Описание данных**	Clean Data/ **Очистка данных**	Generate Test Design/ **Подготовка плана тестирования**	Review Process/ **Оценка процесса**	Plan Monitoring and Maintenance/ **Планирование мониторинга и поддержки**
Determine Data Mining Goals/ **Определение целей аналитики**	Explore Data/ **Изучение данных**	Construct Data/ **Генерация данных**	Build Model/ **Обучение моделей**	Determine Next Steps/ **Определение следующих шагов**	Produce Final Report/ **Подготовка отчета**
Produkt Project Plan/ **Подготовка плана проекта**	Verify Data Quality/ **Проверка качества данных**	Integrate Data/ **Интеграция данных**	Assess Model/ **Оценка качества моделей**		Review Project/ **Ревью проекта**
		Format Data/ **Форматирование данных**			

Несколько оговорок

О методологии

- Не является универсальным рецептом!
- Можно менять последовательность шагов (в рамках этапа)
- Можно исключать какие-то шаги (если они не нужны)
- Не обязательно всегда подходить так же формально (agile на дворе)
- Не нужно документировать каждый шаг (только важные)
- Подходит как для консалтинга, так и для внутренних проектов

- Следует рассматривать просто как типовой шаблон проекта!

1. Business Understanding (Бизнес Постановка)

1. Business Understanding (Бизнес Постановка)

1.1 Business objectives (бизнес-цели проекта)

Основные вопросы

- Организационная структура
    - кто участвует в проекте со стороны заказчика
    - кто выделяет деньги под проект
    - кто принимает ключевые решения
    - кто будет основным пользователем?
    - собираем контакты
- Какова бизнес-цель проекта?
  Например, уменьшение оттока клиентов
  Увеличение конверсии рекламных рассылок
- Существуют ли какие-то уже разработанные решения? Если существуют, то какие и чем именно текущее решение не устраивает?

1. Business Understanding (Бизнес Постановка)

1.2 Assessing current solution (Текущая ситуация)

Определяем

- 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?

1. Business Understanding (Бизнес Постановка)

1.2 Assessing current solution (Текущая ситуация)

- Описание возможных рисков
  - Временные (не уложиться в заявленные сроки)
  - Финансовые (например, спонсор теряет заинтересованность в проекте)
  - 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
- Согласование единой терминоголии, составление глоссария
- Cost-Benefit анализ и оценка рентабельности проекта (ROI)

1. Business Understanding (Бизнес Постановка)

1.3 Data Mining goals (аналитическая постановка)

- Метрика для оценки результатов моделирования. А выбрать есть из чего:
  - Accuracy,
  - RMSE,
  - AUC,
  - Precision,
  - Recall,
  - F-мера,
  - R2,
  - Lift,
  - Logloss
  - и прочее…
- Критерий успешности моделирования
  (например, считаем ROC AUC равный 0.65 — минимальным порогом, 0.75 — оптимальным)?
- Если объективный критерий качества использовать не будем, то как будут оцениваться результаты?

1. Business Understanding (Бизнес Постановка)

1.4 Project Plan (составление плана проекта)

Tasks	Responsible	Start	End	Days	Status
Prerequisites					
Set kick-off meeting	Alice B.	6/2	6/2	0	Complete
Agree on objectives	Frank C.	3-Sep	7-Sep	4	Complete
Initiation					
Detailed Reqs.	Jacob S.	7-Sep	9-Sep	3	Complete
Hardware Reqs.	Jacob S.	9-Sep	11-Sep	2	Overdue
Final Resource Plan	Jacob S.	11-Sep	15-Sep	4	In progress
Staffing	Alice B.	15-Sep	17-Sep	3	In progress
Development					
Technical Reqs.	Frank C.	17-Sep	21-Sep	4	Not started
DB Development	Shari W.	21-Sep	23-Sep	3	Not started
API Development	Shari W.	23-Sep	25-Sep	2	Not started
UI Client	Alice B.	25-Sep	29-Sep	4	Not started
Testing	Kennedy C.	29-Sep	2-Oct	3	Not started
Dev. Complete	Jacob S.	2-Oct	5-Oct	3	Not started
Operations					
Hardware Config.	Alice B.	5-Oct	7-Oct	2	Not started
System Testing	Kennedy C.	7-Oct	9-Oct	2	Not started
Launch		9-Oct	9-Oct		

2. Data Understanding (Анализ данных)

2. Data Understanding (Анализ данных)

2.1 Data collection (Сбор данных)

Задача: понять какие данные есть и какие можно получить

- Классификация типов данных
    - Собственные (1st party data),
    - Сторонние данные (3rd party),
    - «Потенциальные» данные (для получения которых необходимо организовать сбор).
- Получить доступ к источникам

    В крупных компаниях могут сотни источников данных. Корпоративное хранилище данных (КХД)

- Всегда проверять наличие данных

2. Data Understanding (Анализ данных)

2.2 Data description (Описание данных)

Задача: описать все доступные данные

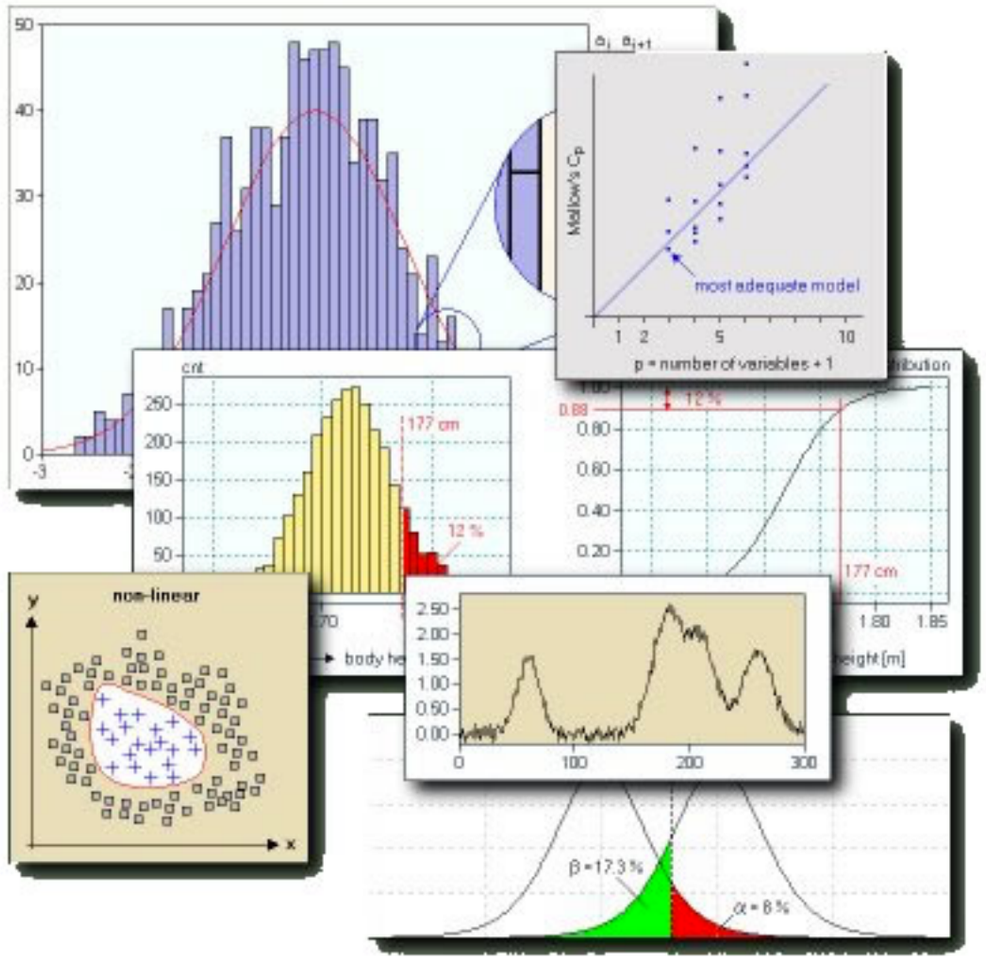
- Описать данные во всех источниках (таблица, ключ, количество строк, количество столбцов, объем на диске)
- Если объем слишком велик для используемого ПО, создаем сэмпл данных.
- Считаем ключевые статистики по атрибутам (минимум, максимум, разброс, кардинальность и т.д.).

2. Data Understanding (Анализ данных)

2.3 Data exploration (Исследование данных)

Делаем Exploratory Data Analysis (EDA)

- Гистограммы
- Распределения
- Коррелограммы
- Feature Importance



Почему это важно:

- Самим понимать данные
- Поиск “случайных” закономерностей
- Поиск возможных несоответствий
- Сам по себе результат

Результаты в отчет

2. Data Understanding (Анализ данных)

2.4 Data quality (Качество данных)

Задача: оценить качество данных и определить возможность их использования

- Пропущенные значения
- Ошибки данных (опечатки)
- Неконсистентная кодировка значений (например «M» и «male» в разных системах)

3. Data Preparation (подготовка данных)

3. Data Preparation (подготовка данных)

3.1 Data selection (Отбор данных)

Типовые критерии отбора атрибутов для моделирования

- Релевантность атрибута
  электронная почта или номер телефона => явно бесполезны
  домен(mail.ru, gmail.com) или код оператора => можно рассмотреть

- Качество атрибута
  Если видим, что большая часть значений атрибута пуста, то атрибут, скорее всего, бесполезен.

- Корреляция с другими атрибутами

- Значимость атрибута (feature importance)

- Возможные ограничения
  Запрет на использование п*ерсональной информации*

3. Data Preparation (подготовка данных)

3.2 Data Cleaning (Очистка данных)

- Пропущенные значения => либо заполнить (data imputation), либо удалить

- Ошибки в данных => исправить вручную, либо удалить

- Несоответствующая кодировка значений => привести к единой кодировке

3. Data Preparation (подготовка данных)

3.3 Constructing new data (Генерация данных)

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

3. Data Preparation (подготовка данных)

3.3 Constructing new data (Генерация данных)

- Агрегация атрибутов (расчет sum, avg, min, max, var и т.д.),
- Генерация кейсов (например, oversampling или алгоритм SMOTE)

3. Data Preparation (подготовка данных)

3.3 Constructing new data (Генерация данных)

- Агрегация атрибутов (расчет sum, avg, min, max, var и т.д.),
- Генерация кейсов (например, oversampling или алгоритм SMOTE)
- Конвертация типов данных
  например, SVM традиционно работает с интервальными данными, а CHAID с номинальными

3. Data Preparation (подготовка данных)

3.3 Constructing new data (Генерация данных)

Атрибут	Пример	Значения	Пример значния
Номинальный	Пол	М/Ж/na	М
Ординальный	Образование	среднее/высшее/ ученая степень	среднее
Непрерывный	Среднее время на сайте	R	23.5s

1. Конвертация Непрерывные => Номинальные
    - Binning/Bucketing transformation
2. Конвертация Номинальные => Непрерывные
    - Dummy values transformation
    - Weight-Of-Evidence transformation
    - Smoothed Likelihood transformation

3. Data Preparation (подготовка данных)

3.3 Constructing new data (Генерация данных)

- Агрегация атрибутов (расчет sum, avg, min, max, var и т.д.),
- Генерация кейсов (например, oversampling или алгоритм SMOTE)
- Конвертация типов данных *например, SVM традиционно работает с интервальными данными, а CHAID с номинальными*
- Нормализация и трансформация атрибутов (feature scaling)

3. Data Preparation (подготовка данных)

3.3 Constructing new data (Генерация данных)

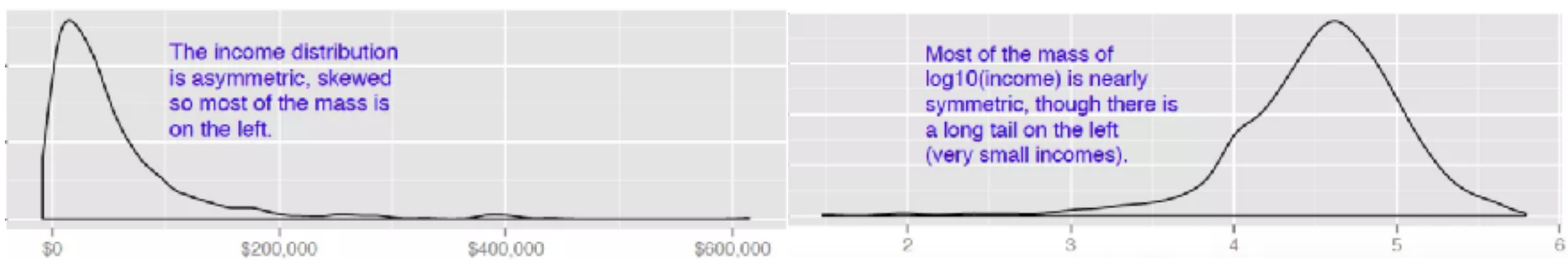
Зачем: влияние шкалы на большинство алгоритмов kNN, PCA

1. Min-Max normalization $$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

2. Mean normalization $$x' = \frac{x - \text{mean}(x)}{\max(x) - \min(x)}$$

3. Standardization $$x' = \frac{x - \bar{x}}{\sigma}$$

4. Logarithm transformation x’ = log(x)



3. Data Preparation (подготовка данных)

3.3 Constructing new data (Генерация данных)

- Агрегация атрибутов (расчет sum, avg, min, max, var и т.д.),
- Генерация кейсов (например, oversampling или алгоритм SMOTE)
- Конвертация типов данных *например, SVM традиционно работает с интервальными данными, а CHAID с номинальными*
- Различные нормализации атрибутов (feature scaling)
- Заполнение пропущенных данных (missing data imputation)

3. Data Preparation (подготовка данных)

3.3 Constructing new data (Генерация данных)

Различные подходы к заполнению пропусков

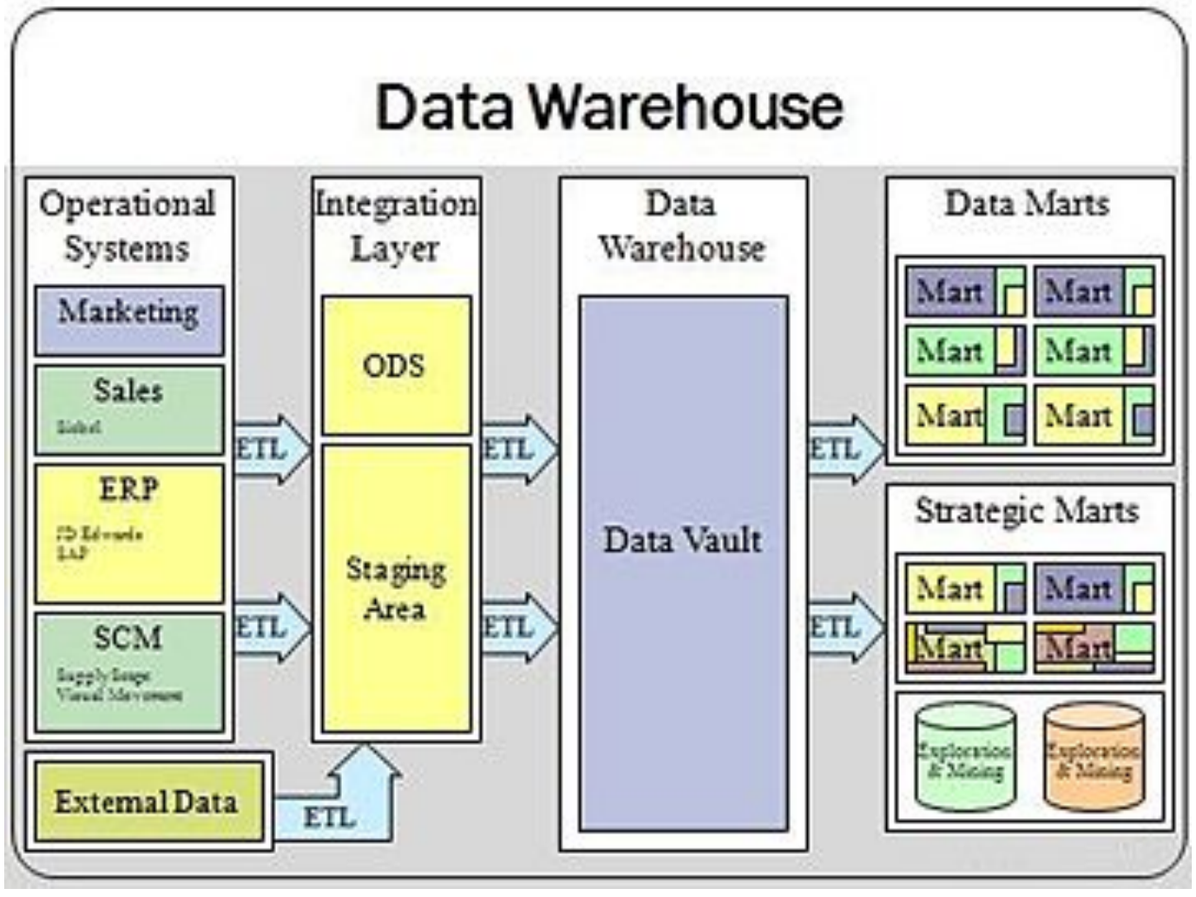
- Random Imputation
- Mean / Median Imputation
- Mode Imputation
- Regression Imputation
- Multiple Imputation

3. Data Preparation (подготовка данных)

3.4 Integrating data (Интеграция данных)

Задача: объединение данных из нескольких разнородных источников

- горизонтальное (join)
- вертикальное (union)



3. Data Preparation (подготовка данных)

3.5 Formatting Data (Форматирование данных)

(Опционально) Привести данные к формату, пригодному для моделирования

Например, сортировка временного ряда по временным отметкам

Нормализация/денормализация данных для алгоритмов Market Basket Analysis или ассоциативных правил.

Order Number	Category	Product	Product Price
SO61269	Helmets	Sport-100	53.99
SO61269	Jerseys	Long-Sleeve Logo Jersey	49.99
SO61270	Fenders	Fender Set - Mountain	21.98
SO61271	Tires and Tubes	LL Road Tire	21.49
SO61271	Tires and Tubes	Patch kit	564.99
SO61272	Tires and Tubes	Mountain Tire Tube	4.99
SO61272	Tires and Tubes	Patch kit	564.99
SO61273	Bottles and Cages	Water Bottle	4.99
SO61274	Caps	Cycling Cap	8.99
SO61274	Shorts	Women's Mountain Shorts	69.99
SO61275	Helmets	Sport-100	53.99
SO61276	Jerseys	Short-Sleeve Classic Jersey	539.99
SO61276	Caps	Cycling Cap	8.99
SO61277	Mountain Bikes	Mountain-500	539.99
SO61277	Jerseys	Short-Sleeve Classic Jersey	539.99
SO61277	Caps	Cycling Cap	8.99

4. Modeling (Моделирование)

4. Modeling (Моделирование)

4.1 Selecting the modeling technique (Выбор алгоритмов)

На что обращаем внимание при выборе

- Достаточно ли данных, поскольку сложные модели как правило требуют большей выборки?

- Сможет ли модель обработать пропуски данных (какие-то реализации алгоритмов умеют работать с пропусками, какие-то нет)?

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

4. Modeling (Моделирование)

4.2 Generating a test design (Планирование тестирования)

Планирование процесса обучения моделей

- 2 выборки: Train/Test (примерно 80/20)
- 3 выборки: Train/Validation/Test в примерной пропорции 60/20/20
- Различные варианты кросс-валидации
- Оптимизация гиперпараметров – сколько будет итераций по каждому алгоритму, grid-search или random-search

4. Modeling (Моделирование)

4.3 Building the models (Обучение моделей)

Запускаем цикл обучения и после каждой итерации фиксируем результат.
На выходе получаем несколько обученных моделей.

Помимо основной метрики качества

- Показывает ли модель какие-то интересные закономерности?
  Например, что точность предсказания на 99% объясняется всего одним атрибутом.

- Какова скорость обучения/применения модели?
  Если модель обучается 2 дня, возможно, стоит поискать более эффективный алгоритм или уменьшить обучающую выборку.

- Были ли проблемы с качеством данных?
  Например, в тестовую выборку попали кейсы с пропущенными значениями, и из-за этого не вся выборка проскорилась.

4. Modeling (Моделирование)

4.4 Assessing the model (Оценка результатов)

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.

5. Evaluation (Оценка решения)

5. Evaluation (Оценка результата)

5.1 Evaluating the results (Оценка результатов моделирования)

- Формулировка результата в бизнес-терминах

  Бизнесу гораздо легче общаться в терминах $ и ROI, чем в абстрактных Lift или R2

- В целом насколько хорошо полученные результаты решают бизнес-задачу?

- Найдена ли какая-то новая ценная информация, которую стоит выделить отдельно?

  Пример с целевой аудиторией

5. Evaluation (Оценка результата)

5.2 Review the process (Разбор полетов)

Темы для обсуждения

- Можно ли было какие-то шаги сделать более эффективными?
  “*месяц ушел на согласование доступов…*”

- Какие были допущены ошибки и как их избежать в будущем?
  “*недооценили сложность выгрузки данных из источников и в результате не уложились в сроки…*”

- Были ли не сработавшие гипотезы? Если да, стоит ли их повторять?
  “давайте здесь попробуем тип лёрнинг… всё становится лучше с дип лёрнингом”

- Были ли неожиданности при реализации шагов? Как их предусмотреть в будущем?
  Заказчик: «Ok. А мы думали, что обучающая выборка для разработки модели не нужна…»

5. Evaluation (Оценка результата)

5.3 Determining the next steps (Принятие решения)

Решаем => “внедрять модель” | “улучшать дальше”

6. Deployment (Внедрение)

6. Deployment (Внедрение)

6.1 Planning Deployment (Планирование развертывания)

- Важно зафиксировать, что именно и в каком виде мы будем внедрять, а также подготовить технический план внедрения (пароли, явки и прочее)

- Продумать, как с внедряемой моделью будут работать пользователи
  Например, на экране сотрудника колл-центра показываем склонность клиента к подключению дополнительных услуг.

- Определить принцип мониторинга решения. Если нужно, подготовиться к опытно-промышленной эксплуатации.
  Например, договариваемся об использовании модели в течение года и тюнинге модели раз в 3 месяца.

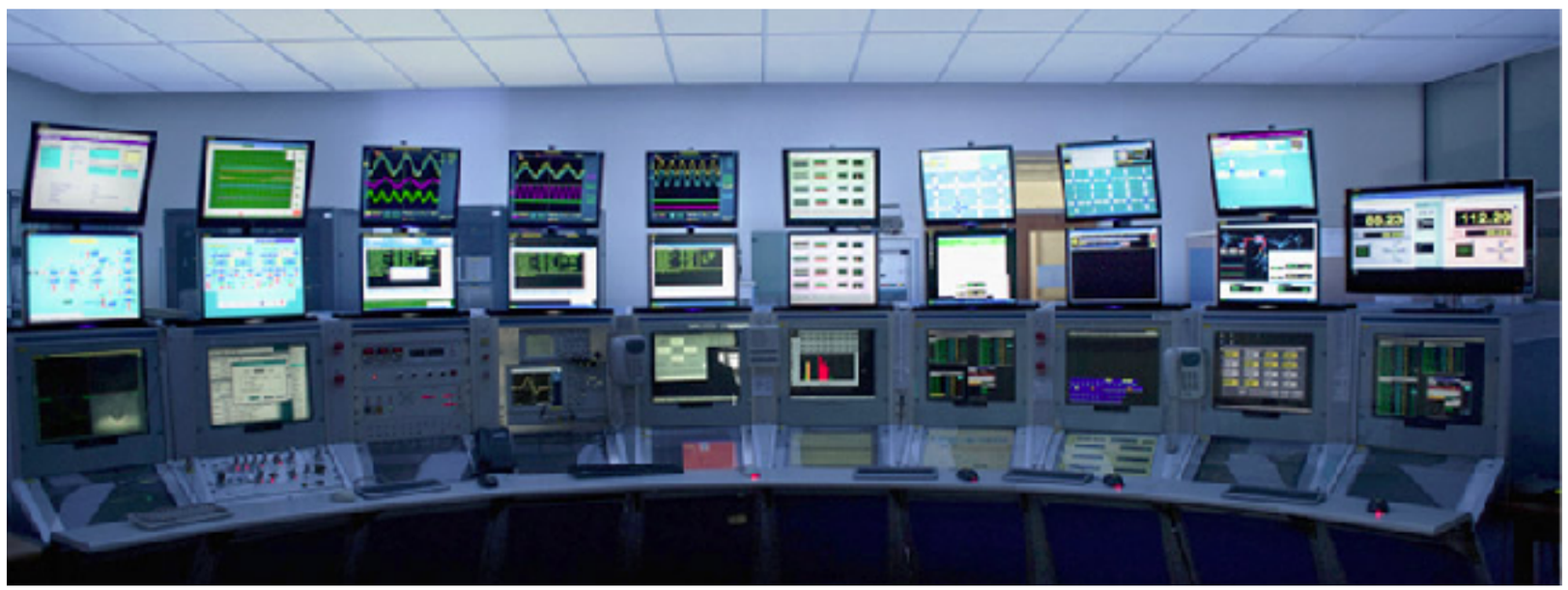
6. Deployment (Внедрение)

6.2 Planning Monitoring (Настройка мониторинга модели)

- Какие показатели качества модели будут отслеживаться?
  В своих банковских проектах мы часто используем популярный в банках показатель population stability index PSI.

- Критерий устаревания модели
  Например, если PSI больше 0.15, либо просто договариваемся о регулярном пересчете раз в 3 месяца.

- Достаточно ли переобучить уставшую модель или нужно организовывать новый проект?
  При существенных изменениях в данных как правило нужен полный цикл переобучения



6. Deployment (Внедрение)

6.3 Final Report (Отчет по результатам моделирования)

- Написанный отчет презентуется заказчику и всем заинтересованным лицам. В отсутствие ТЗ этот отчет является главным документом проекта.
- В отчете фиксируются основные результаты по каждому шагу
- Рекомендации по дальнейшему развитию модели
- Провести встречу со всеми задействованными на проекте сотрудниками и собрать их мнение о проекте

Несколько простых советов по взаимодействию с заказчиком

- Поддерживайте постоянную коммуникацию с заказчиком
  Периодические статусы с описанием прогресса

- С самого начала формируйте правильные ожидания заказчика
  Не стоит обещать “с три короба”, если нет уверенности в реализуемости

- Всегда практикуйте подход MVP
  Сначала простое решение, затем постепенное усложнение. Не пишите сразу в прод

Небольшая Практика

Задание

Примените шаблон CRISP-DM к своим дипломным работам.

По каждому этапу выпишите:
- примерные сроки
- состав работ

Время: 15 мин

Business Understanding/ Бизнес-анализ	Data Understanding/ Анализ данных	Data Preparation/ Подготовка данных	Modeling/ Моделирование	Evaluation/ Оценка решения	Deployment/ Внедрение
Determine Business Objectives/ **Определение бизнес-целей**	Collect Initial Data/ **Сбор данных**	Select Data/ **Выборка данных**	Select Modeling Techniques/ **Выбор алгоритмов**	Evaluate Results/ **Оценка результатов**	Plan Deployment/ **Внедрение**
Assess Situation/ **Оценка текущей ситуации**	Describe Data/ **Описание данных**	Clean Data/ **Очистка данных**	Generate Test Design/ **Подготовка плана тестирования**	Review Process/ **Оценка процесса**	Plan Monitoring and Maintenance/ **Планирование мониторинга и поддержки**
Determine Data Mining Goals/ **Определение целей аналитики**	Explore Data/ **Изучение данных**	Construct Data/ **Генерация данных**	Build Model/ **Обучение моделей**	Determine Next Steps/ **Определение следующих шагов**	Produce Final Report/ **Подготовка отчета**
Produkt Project Plan/ **Подготовка плана проекта**	Verify Data Quality/ **Проверка качества данных**	Integrate Data/ **Интеграция данных**	Assess Model/ **Оценка качества моделей**		Review Project/ **Ревью проекта**
		Format Data/ **Форматирование данных**			

Вопросы