Анализ данных в организации

NewProLab, осень 2020 г.
Олег Хомюк

Олег Хомюк

Lamoda, R&D Director

oleg.khomyuk@gmail.com
telegram: @khomyuk
https://www.linkedin.com/in/olegkhomyuk

Yandex, Consultant Plus, Ezhome

1. Зачем бизнесу анализ данных?

Зачем бизнесу анализ данных

Основные цели бизнеса

- **рост**
  (увеличение выручки, рыночной доли, аудитории и т.д.)

- **оптимизация**
  (сокращение издержек, улучшение качества продуктов / сервиса, повышение эффективности процессов)

Зачем бизнесу анализ данных

Монетизация данных – процесс извлечения/повышения прибыли за счет применения практик анализа данных.

- повышение эффективности существующих собственных бизнес-процессов организации или процессов другой (внешней) организации

- создание принципиально новых продуктов, основанных на данных, а также продажа данных и их производных

Принятие решений - это основополагающий процесс и одна из главных функций управления различными структурами, в том числе и **бизнесом**.

Можно влиять на достижение бизнесом своих целей с помощью более эффективного процесса принятия решений!

Виды принятия решений

Gut-feeling	• Creative: fast-paced, lack of information
Judgement	• Intuitive: incomplete outcome certainty, low quality data
Information	• Rational: able to predict outcomes and choose best options
Data-driven	• Programmed: automated intelligence

Описательная аналитика

Что происходит сейчас?

Реализуется с помощью:

- Описания данных
- Анализа случайных наборов и объектов
- Визуализации данных

Диагностическая аналитика

В чем причина происходящего?

Реализуется с помощью:

- Разведочного анализа
- Статистического анализа

Используются:

- Визуализация распределений, диаграммы, гистограммы
- Статистики, корреляционный анализ
- Проверка статистических гипотез (в том числе множественная)

Предиктивная аналитика

Что произойдет в будущем?

Реализуется с помощью:

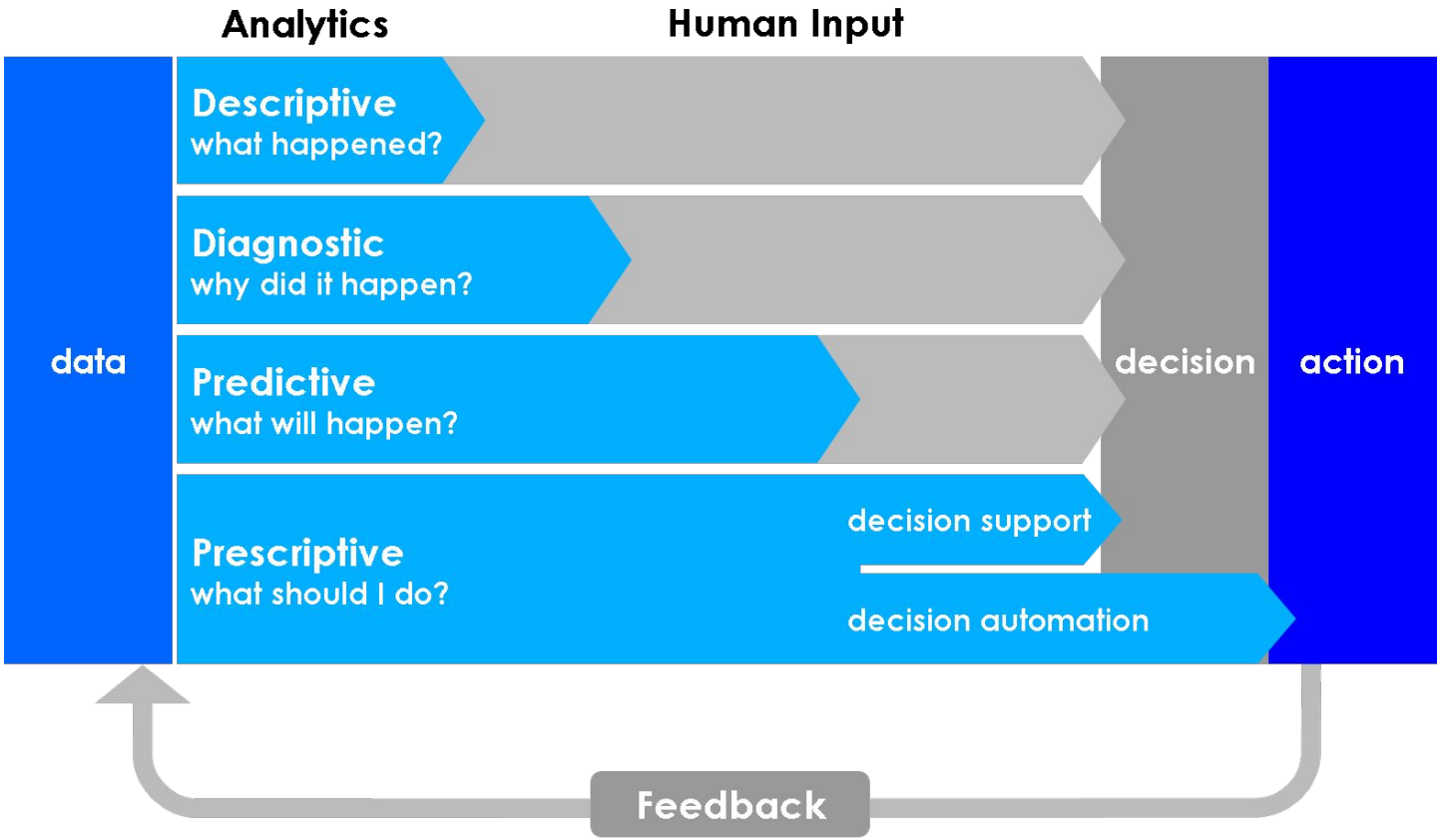
- Классификации, регрессии
- Кластеризации
- Прогнозирования временных рядов
- Методов выявления аномалий

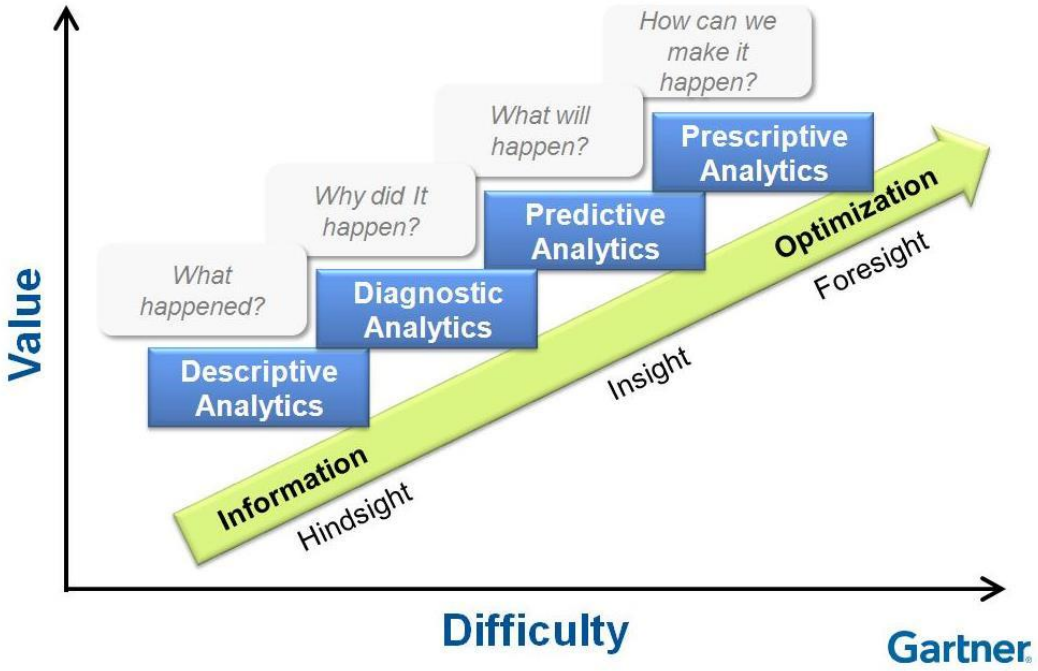
Прескриптивная аналитика

Что мы должны предпринять для достижения цели?

Реализуется с помощью:

- Рекомендательных систем
- Систем поддержки принятия решений
- Алгоритмов оптимизации
- Решений по автоматизации процессов


Analytics
Human Input
data
Descriptive
what happened?
Diagnostic
why did it happen?
Predictive
what will happen?
Prescriptive
what should I do?
decision support
decision automation
decision
action
Feedback



Предписывающая аналитика имеет наибольшую ценность для бизнеса.

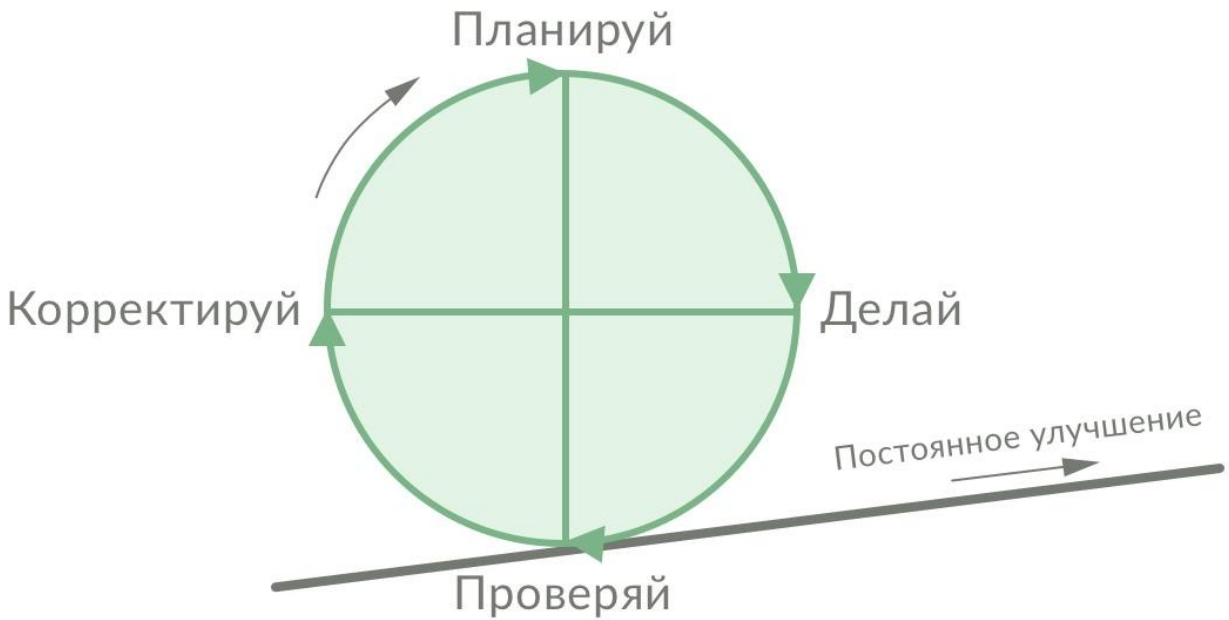
Жизненный цикл DS проектов

NewProLab, осень 2020 г.
Олег Хомюк

CRISP-DM

CRoss Industry Standard Process for Data Mining

- Бизнес-анализ (Business understanding)
- Анализ данных (Data understanding)
- Подготовка данных (Data preparation)
- Моделирование (Modeling)
- Оценка результата (Evaluation)
- Внедрение (Deployment)


Планируй
Делай
Проверяй
Корректируй
Постоянное улучшение

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**			

1. Бизнес-анализ / Business understanding

- Бизнес-цель проекта
  (заказчик, бюджет, бизнес-цель, чем не устраивает текущее решение)

- Аудит текущей ситуации
  (ресурсы - железо, инфраструктура, доступность данных, эксперты по предметной области, анализ текущего решения, риски)

- Цели по аналитике
  (метрики качества, критерии приемки / успешности)

- План проекта
  (оценка всех этапов, сроки, роли, команда, ответственные)

В чем сложность этапа постановки задачи

Необходимо:

- собрать полную информацию о бизнес задаче
- корректно конвертировать ее в математическую постановку

Ошибки и неточности на этом этапе

- могут весьма драматическим образом сказаться на результате
- к сожалению, не редкость.

Трудности перевода:

В реальности существует колоссальный разрыв между тем, что нужно бизнесу, и тем, что привыкли делать аналитики, data scientist-ы и математики.

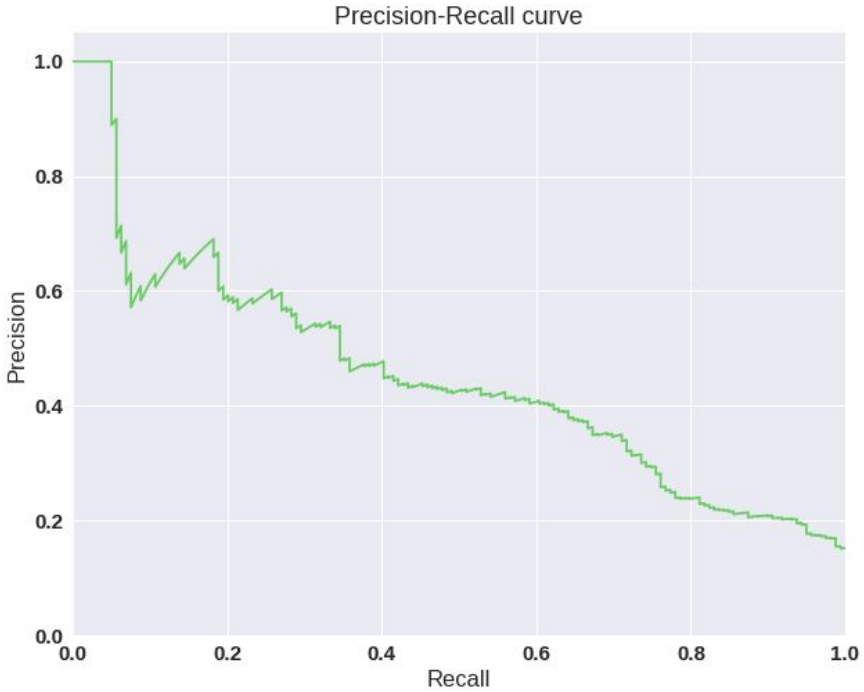
В чем сложность этапа постановки задачи

Бизнес-задача:

- Сформулированная задача, позволяющая достигать цели компании
- Требует экспертных знаний в предметной области
- Во многих случаях успех измеряется в деньгах

Математическая постановка:

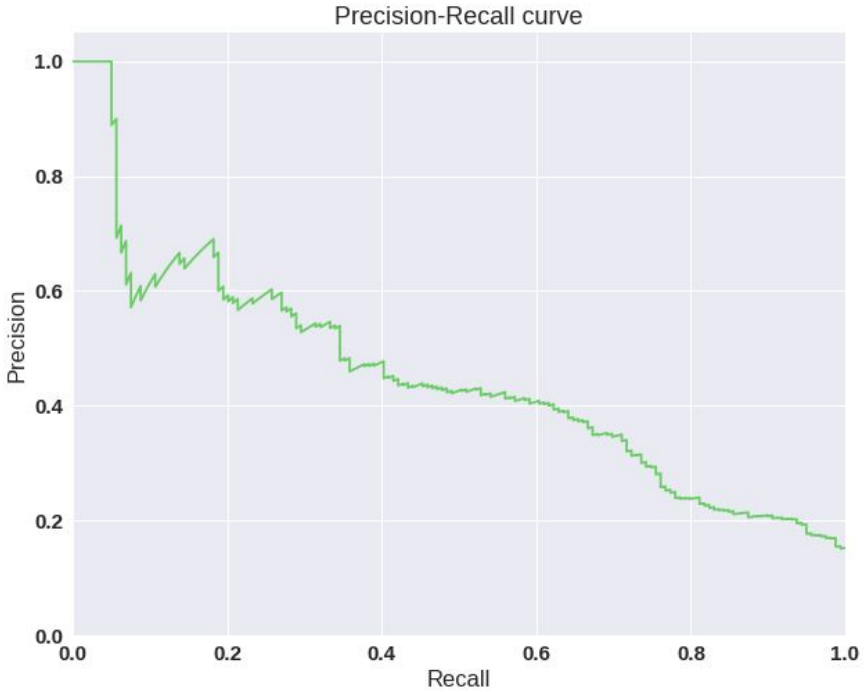
- Постановка в терминах анализа данных
- Требует экспертизы в математике и машинном обучении
- Успех измеряется численно (точность, полнота)



Что можно сказать о выборке по данной кривой?

Можно ли определить по ней баланс классов? (какую долю составляют объекты Positive класса)

Что будет, если мы захотим максимизировать Precision? A Recall?

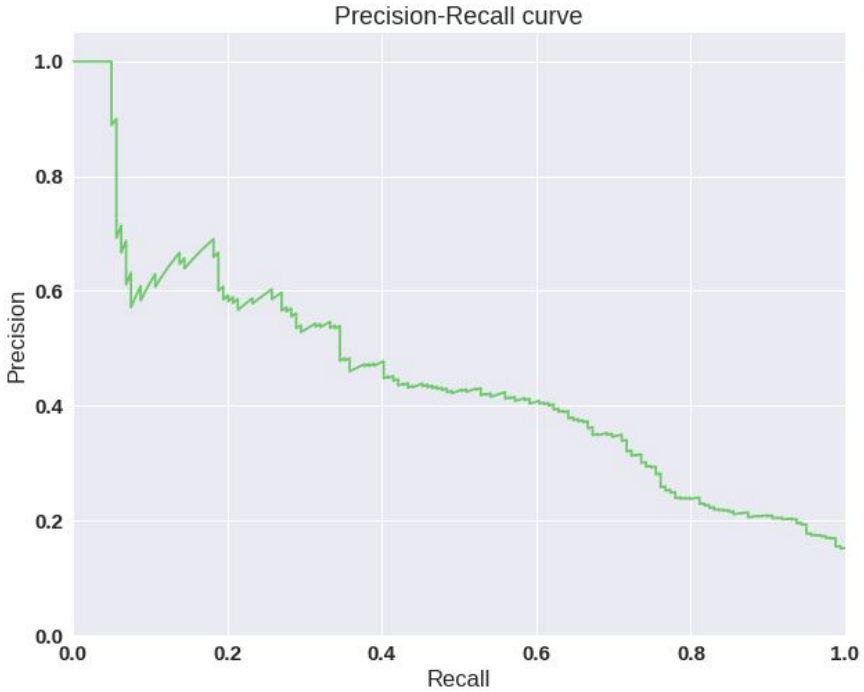


Trade-off между точностью и полнотой.

Для разных задач разное может быть важно.

Что важнее в кейсе обнаружения заболевания в результатах анализов

Precision VS Recall



Trade-off между точностью и полнотой.

Для разных задач разное может быть важно.

- Детекция заболеваний
- Фильтрация взрослого контента
- Таргетинг рассылок
- FaceID / TouchID

Что же делать

Работать над постановкой задачи в формате кросс-функциональной команды и делиться экспертизой!

Чек-лист по постановке задачи

Вводные:

- Какой процесс хотим оптимизировать? Как он работает?
- Где мы видим точки роста / уязвимости? Что хотим улучшить? Какие есть идеи?

Проработка:

- Потенциал в плане экономического эффекта
- Как и где будет использоваться модель?
- Как оценить экономический эффект в случае внедрения? Как будем понимать, что проект успешен?

Постановка задачи. Кейс

На входе:

- Нужно сделать модель, прогнозирующую продажи товаров на следующую неделю

Какую метрику взять?

- MAE, MSE, RMSE, MAPE, sMAPE

Разные последствия для бизнеса от:

- Недопрогноза
- Перепрогноза

А стоит ли вообще браться за этот проект?

Перед тем, как приступать к следующим этапам надо оценить экономический потенциал проекта

Кейс по оттоку

Что хотим оптимизировать?

Хотим оптимизировать стоимость для нас денег, заработанных на пользователе

Кейс по оттоку

Экономический эффект на одного пользователя

$$\mathbf{MQ} * Z * ARPU - COST$$

MQ - качество модели (доля правильно угаданных отточников)
Z - успешность удержания
ARPU - средняя выручка на пользователя
COST - стоимость удержания одного пользователя

Кейс по оттоку

Экономический эффект

$$TP * Z * ARPU - (TP + FP) * COST$$

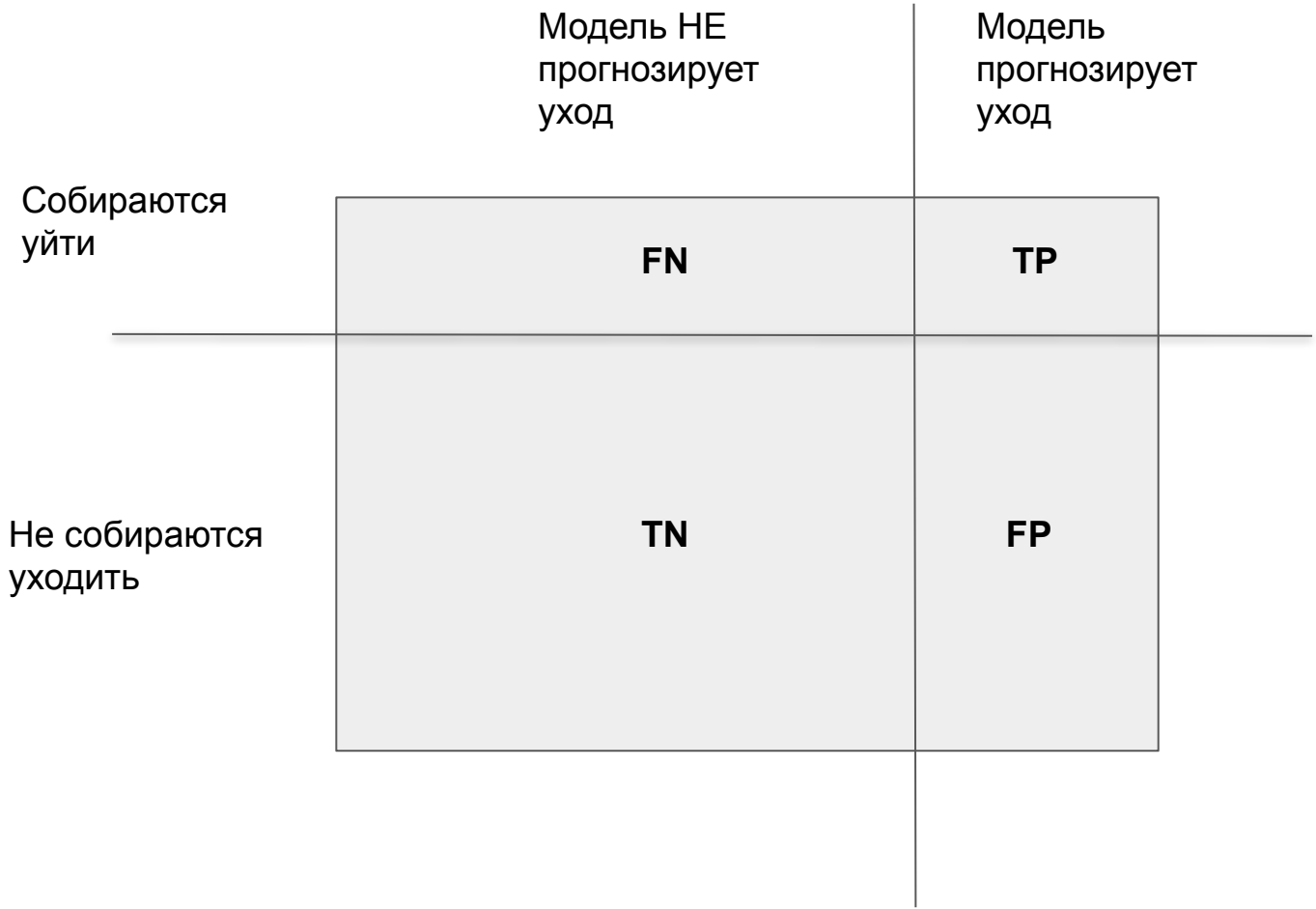
TP - число верно классифицированных пользователей, уходящих в отток
FP - “ложная тревога”
Z - успешность удержания
ARPU - средняя выручка на пользователя
COST - затраты на удержание одного пользователя

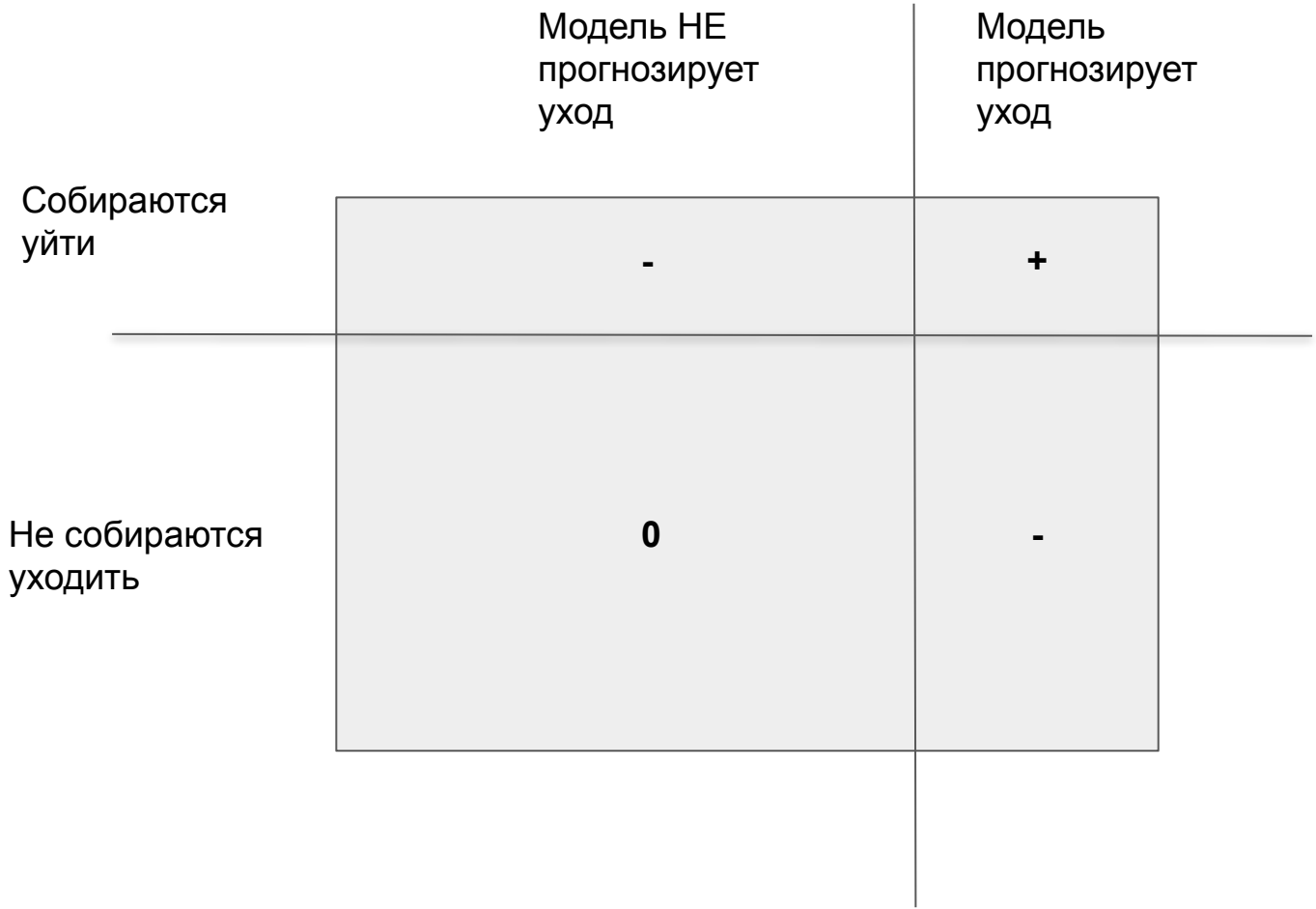
TP * Z * ARPU - (TP + FP) * COST

TP = N_users * Churn_rate * Recall
TP / (TP + FP) = Precision

N * Churn_rate * Recall * (Z * ARPU - COST / Precision)

N - число клиентов
Churn_rate - коэффициент оттока
Precision / Recall - точность / полнота модели классификации
Z - успешность удержания
ARPU - средняя выручка на пользователя
COST - затраты на удержание одного пользователя


Модель НЕ прогнозирует уход
Модель прогнозирует уход
Собираются уйти
FN
TP
Не собираются уходить
TN
FP


Модель НЕ прогнозирует уход
Модель прогнозирует уход
Собираются уйти
-
+
Не собираются уходить
0
-

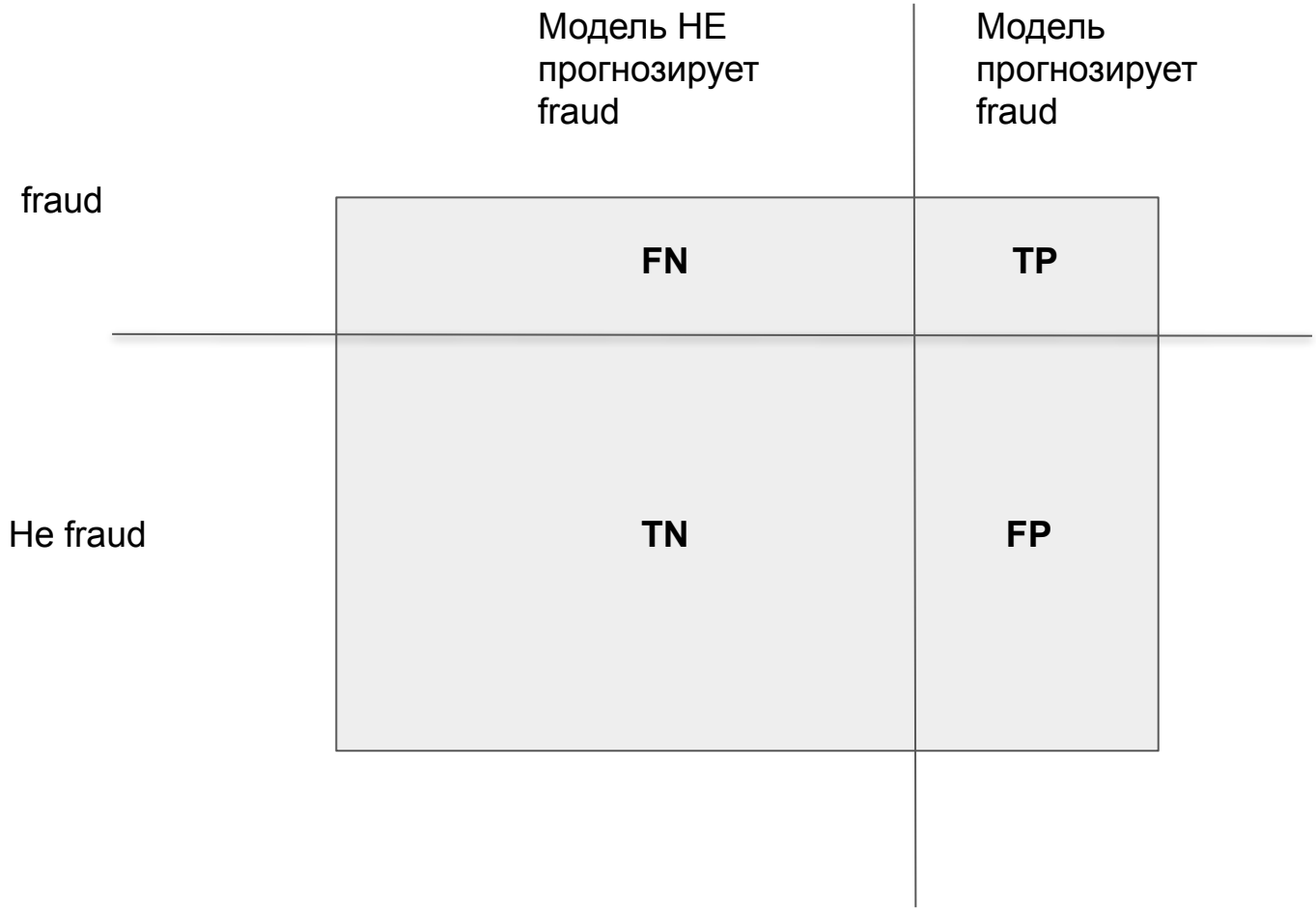
Кейс Fraud detection в банковских транзакциях

Дано

- 100 тыс. транзакций в сутки на сумму 100 млн рублей
- Примерно 1% (1000) транзакций - мошеннические, на сумму 10 млн рублей

Вы

- Делаете fraud detector, который автоматизировано блокирует транзакцию
- Разблокировка - по звонку оператора


Модель НЕ прогнозирует fraud
Модель прогнозирует fraud
fraud
FN
TP
Не fraud
TN
FP

	Расшифровка	С моделью	Без модели
TP	Верно обнаружили мошенническую транзакцию	- Потратили деньги на прозвон	- Терпим убытки от мошенничества
FP	Неверно посчитали фродом нормальную транзакцию	- Потратили деньги на прозвон - Потенциально потеряли в лояльности клиента	Нет эффекта
FN	Не обнаружили мошенническую транзакцию	- Терпим убытки от мошенничества	- Терпим убытки от мошенничества
TN	Верно посчитали транзакцию нормальной	Нет эффекта	Нет эффекта
#	ПРОЕКТ	- Траты на проект по обнаружению фрода	Нет эффекта

Экономика проектов DS

- **NPV** (Net Present Value)
  Чистая приведенная стоимость

- **ROI** (Return of Investment)
  Рентабельность инвестиций

Более продвинутые метрики, которые измеряют стоимость поступления денежных средств от проекта по сравнению с затратами.

Экономика проектов DS

$$NPV = \sum_{t=0}^{N} \frac{CFt}{(1+\mathrm{i})^{\mathrm{t}}}$$

- *t* - год, начиная с нулевого
- *CFt* - кэшфлоу (доходы минус расходы)
- *i* - ставка дисконтирования (стоимость денег)

Год	Денежный поток	Приведённая стоимость
T=0	$\frac{-100\,000}{(1+0.10)^0}$	- $ 100 000
T=1	$\frac{30\,000-5000}{(1+0.10)^1}$	$ 22 727
T=2	$\frac{30\,000-5000}{(1+0.10)^2}$	$ 20 661
T=3	$\frac{30\,000-5000}{(1+0.10)^3}$	$ 18 783
T=4	$\frac{30\,000-5000}{(1+0.10)^4}$	$ 17 075
T=5	$\frac{30\,000-5000}{(1+0.10)^5}$	$ 15 523
T=6	$\frac{30\,000-5000}{(1+0.10)^6}$	$ 14 112

Экономика проектов DS

$$ROI = \frac{NPV}{Inv} \cdot 100\%$$

- *NPV* - чистая приведенная стоимость
- *Inv* - объем инвестиций

2. Анализ данных / Data understanding

- Сбор данных
  (собственные / сторонние / потенциальные)

- Описание данных
  (ключи, объемы, доступность, возможные значения, статистики)

- Исследование данных
  (основные статистики, гипотезы, какие данные помогут решить задачу)

- Качество данных
  (пропущенные значения, опечатки / ошибки, противоречия)

Оценка доступных данных

- Какие данные доступны?
- Есть ли историчность? За какой период? (для выявления сезонности нужно >2 года)
- Есть ли возможность использовать данные совместно (ключи)
- Есть ли нужный для задачи сигнал в данных
- Будет ли модель потом работать на live данных в production

3. Подготовка данных / Data preparation

- Отбор данных
  (отбор релевантных данных, полезных для решения задачи)

- Очистка данных
  (удаление / обработка пропусков, ошибок, кодировки, шумов)

- Генерация новых данных
  (построение новых признаков из имеющихся данных)

- Интеграция данных и форматирование
  (объединение данных из разных источников)

4. Моделирование / Modeling

- Выбор алгоритмов
  (сложные / простые, учет специфики задачи)

- Планирование тестирования
  (кросс-валидация, train/test/validation, подбор гипер-параметров)

- Обучение моделей
  (непосредственное написание программного кода для обучения и валидации и его запуск)

- Оценка результатов обучения
  (выбрать лучшие модели, провести анализ качества, принять решение о готовности к внедрению)

5. Оценка результата / Evaluation

- Оценка результатов моделирования
  (насколько хорошо модель решает бизнес-задачу)

- Ретроспектива по проекту
  (разбор полетов, возникшие проблемы, можно ли было что-нибудь сделать лучше / быстрее / эффективнее?)

- Определение следующих шагов
  (внедряем или нет, если да, то какую модель и куда. Надо ли строить новый сервис?)

Отличие модели от сервиса

- Оффлайн моделям могут быть доступны любые данные, которые вы подготовите, даже те, что сложно получать в реальном времени

- Качество работы сервиса естественнее измерять в бизнес показателях, моделей - в ML метриках

- Сервис реализует действие, которое рекомендует модель - например, сервис автоматических рассылок

Мониторинг качества решения

За чем надо следить?

- Изменилось ли качество модели?
- Изменилось ли распределение во входящих данных?
- Триггеры для поддержки качества (нужно отличать случайные изменения качества и “протухание”)

Автоматизация:

- Расчет триггеров
- Регулярное обновление моделей (расписание / триггеры)

6. Внедрение / Deployment

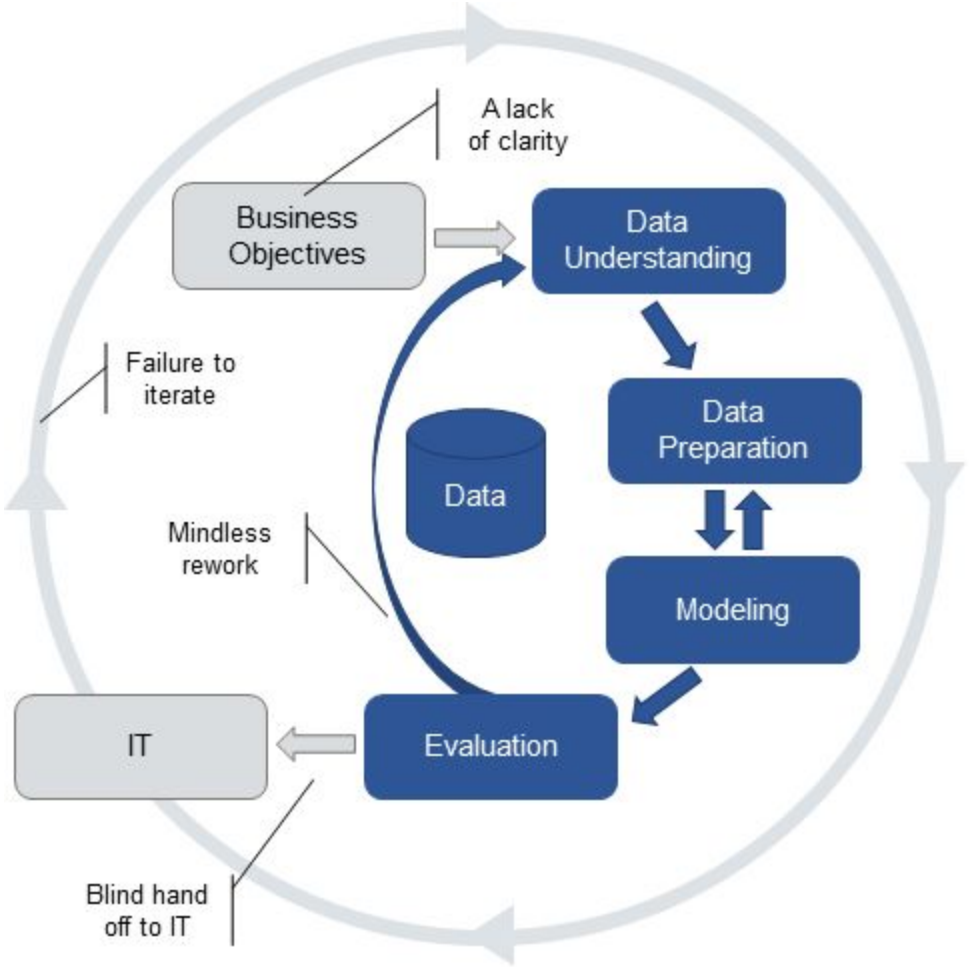
- Развертывание
  (определение вида конечного решения / сервиса, внедрение)

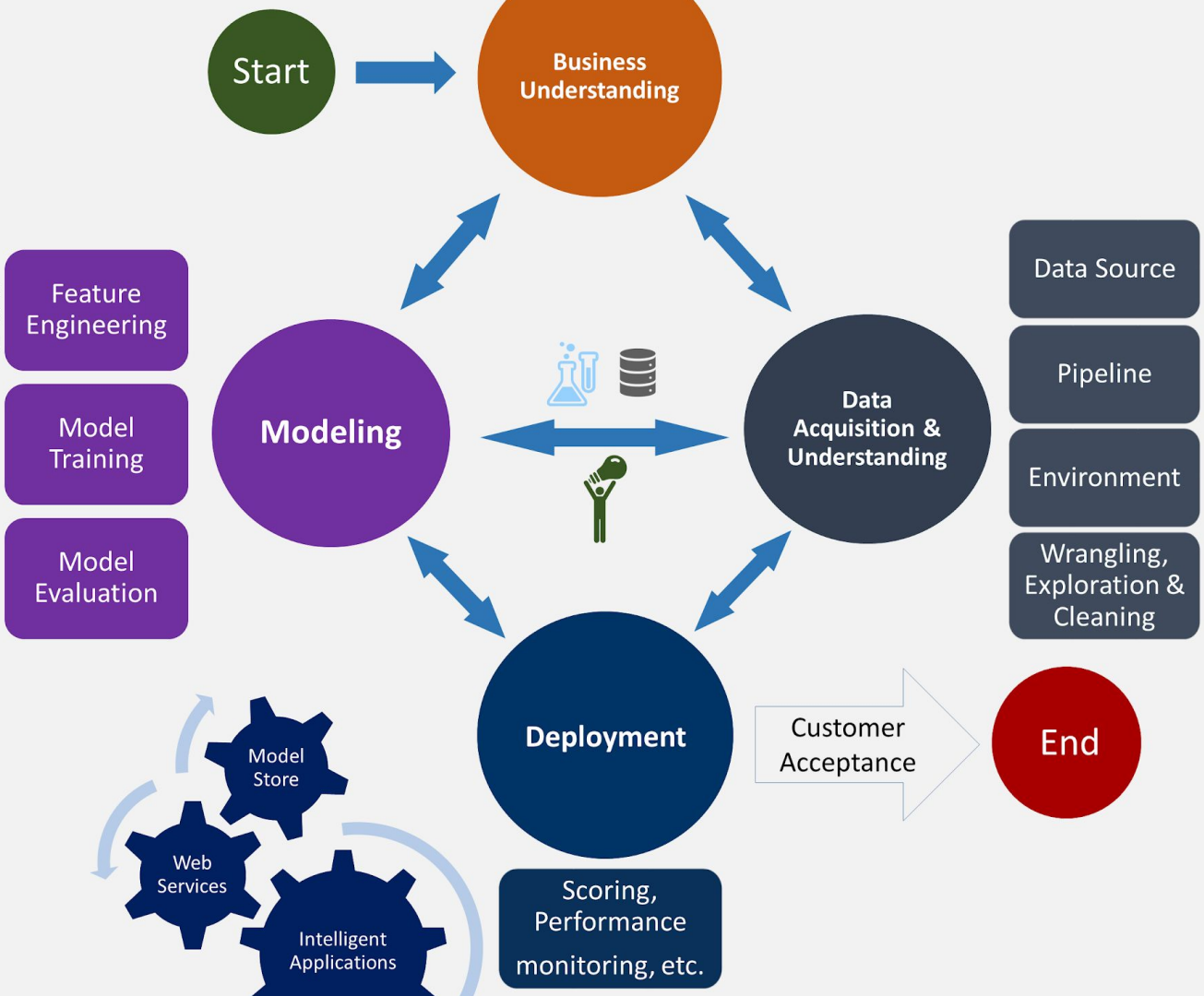
- Настройка мониторинга модели
  (мониторинг качества модели, протухание, частота переобучения)

- Подготовка отчета
  (отчет по проекту)

- Ревью проекта
  (финальный отчет по результатам внедрения)

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**			


A lack of clarity
Business Objectives
Data Understanding
Failure to iterate
Data Preparation
Data
Mindless rework
Modeling
IT
Evaluation
Blind hand off to IT


Start
Business Understanding
Feature Engineering
Model Training
Model Evaluation
Modeling
Data Acquisition & Understanding
Data Source
Pipeline
Environment
Wrangling, Exploration & Cleaning
Deployment
Customer Acceptance
End
Model Store
Web Services
Intelligent Applications
Scoring, Performance monitoring, etc.

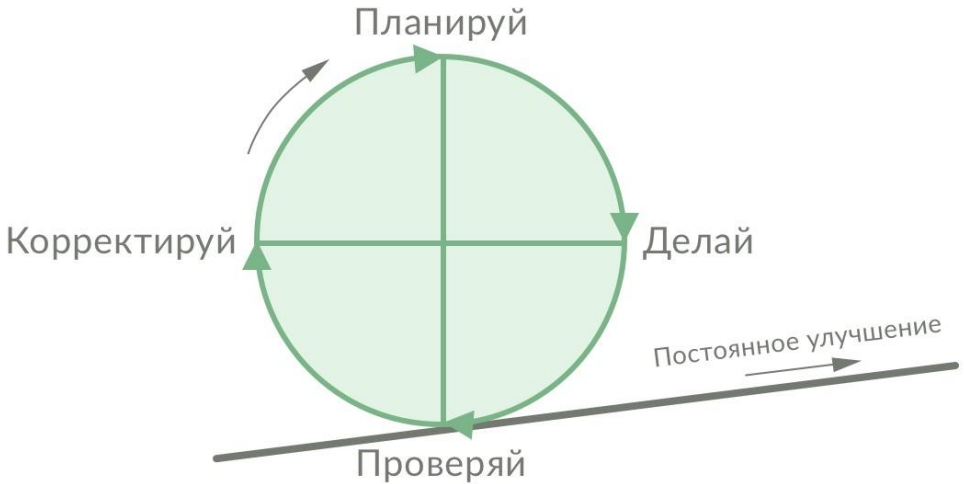
Какие компетенции могут понадобиться

- Product / Project Manager
- Бизнес аналитик
- Data Scientist
- Data Engineer / Software Developer
- Server administrator / DevOps

Кроме этого:

- Эксперты в предметной области
- Команды сервисов и IT-систем, с которыми необходима интеграция

Дальнейшая поддержка решения



Спасибо за внимание!