Анализ данных в бизнесе

Олег Хомюк

NewProLab, Специалист по большим данным
Весна 2018г.

Хомюк Олег

Yandex, Consultant+, Ezhome, Lamoda

oleg.khomyuk@gmail.com
Telegram: @khomyuk
Skype: oleg.khomyuk

https://www.facebook.com/oleg.khomyuk
https://www.linkedin.com/in/olegkhomyuk

1. Зачем бизнесу анализ данных?

Зачем бизнесу анализ данных

Основные цели бизнеса

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

Основные цели бизнеса

Зачем бизнесу анализ данных

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

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

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**.

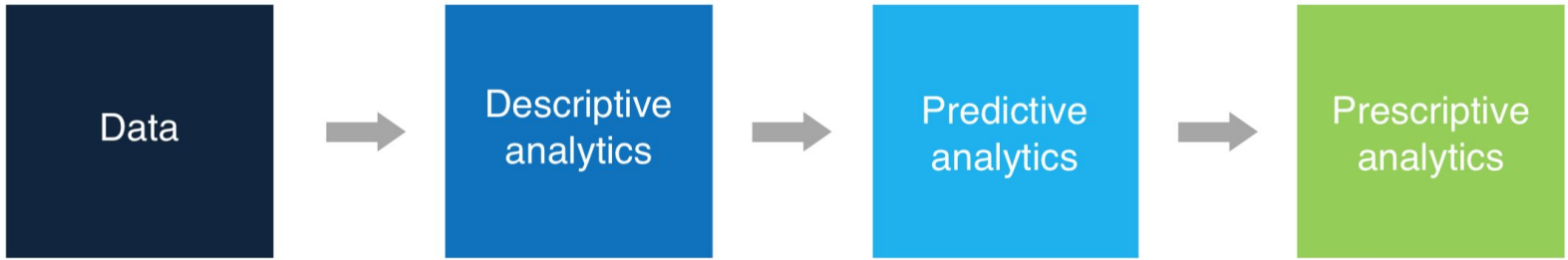
Можно влиять на достижение бизнесом своих целей с помощью более эффективного процесса принятия решений!

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


Rational decision-making process
1. Identify the problem
2. Establish decision criteria
3. Weigh decision criteria
4. Generate alternatives
5. Evaluate the alternatives
6. Choose the best alternative
7. Implement the decision
8. Evaluate the decision

От данных к достижению целей



Описательная аналитика

Что происходит сейчас?

Описательная аналитика

Что происходит сейчас?

Реализуется с помощью:

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

Диагностическая аналитика

В чем причина происходящего?

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?

Предиктивная аналитика

Что произойдет в будущем?

Реализуется с помощью:

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

Прескриптивная аналитика

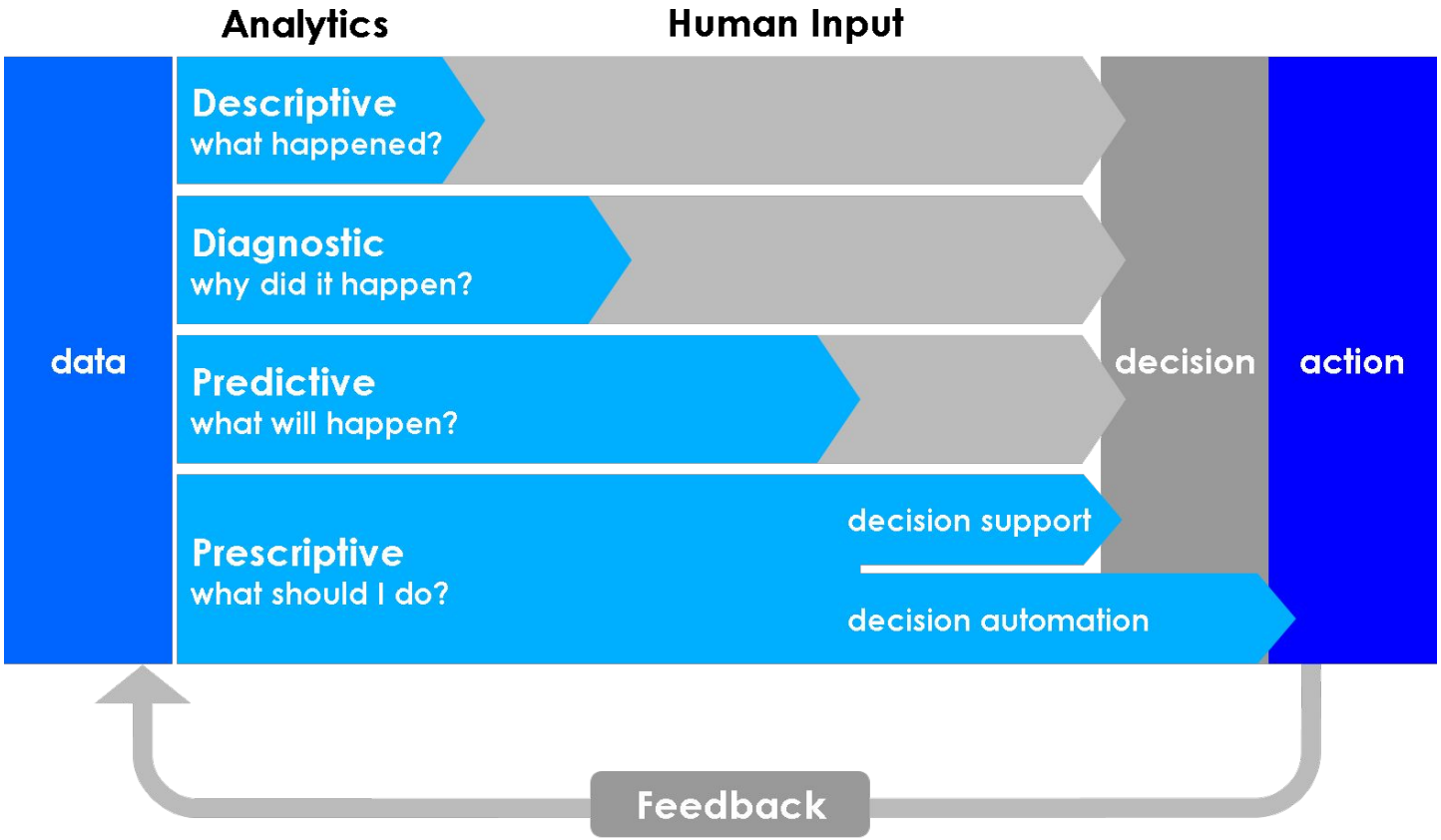
Что мы должны предпринять для достижения цели?

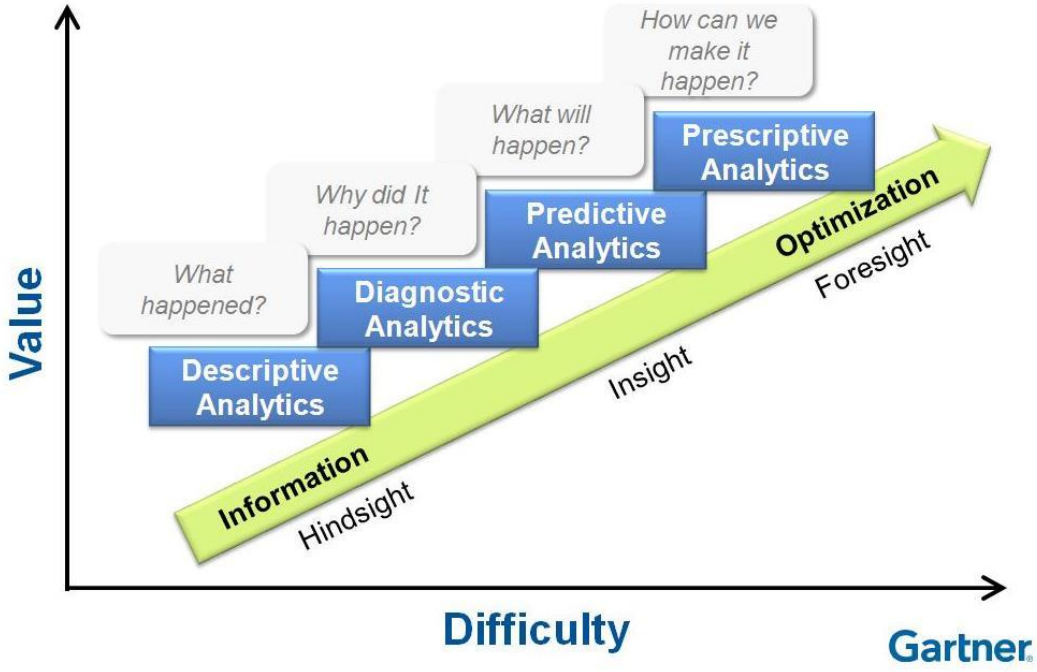
Прескриптивная аналитика

Что мы должны предпринять для достижения цели?

Реализуется с помощью:

- Рекомендательных систем
- Систем поддержки принятия решений
- Систем скоринга возможных сценариев
- 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.

2. Жизненный цикл DS проектов

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)

CRISP-DM

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, организационная структура, 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)

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)

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
  (объединение данных из разных источников)

- Форматирование данных

4. Моделирование / Modeling

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

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

- Обучение моделей
  (построение новых признаков из имеющихся данных)

- 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)

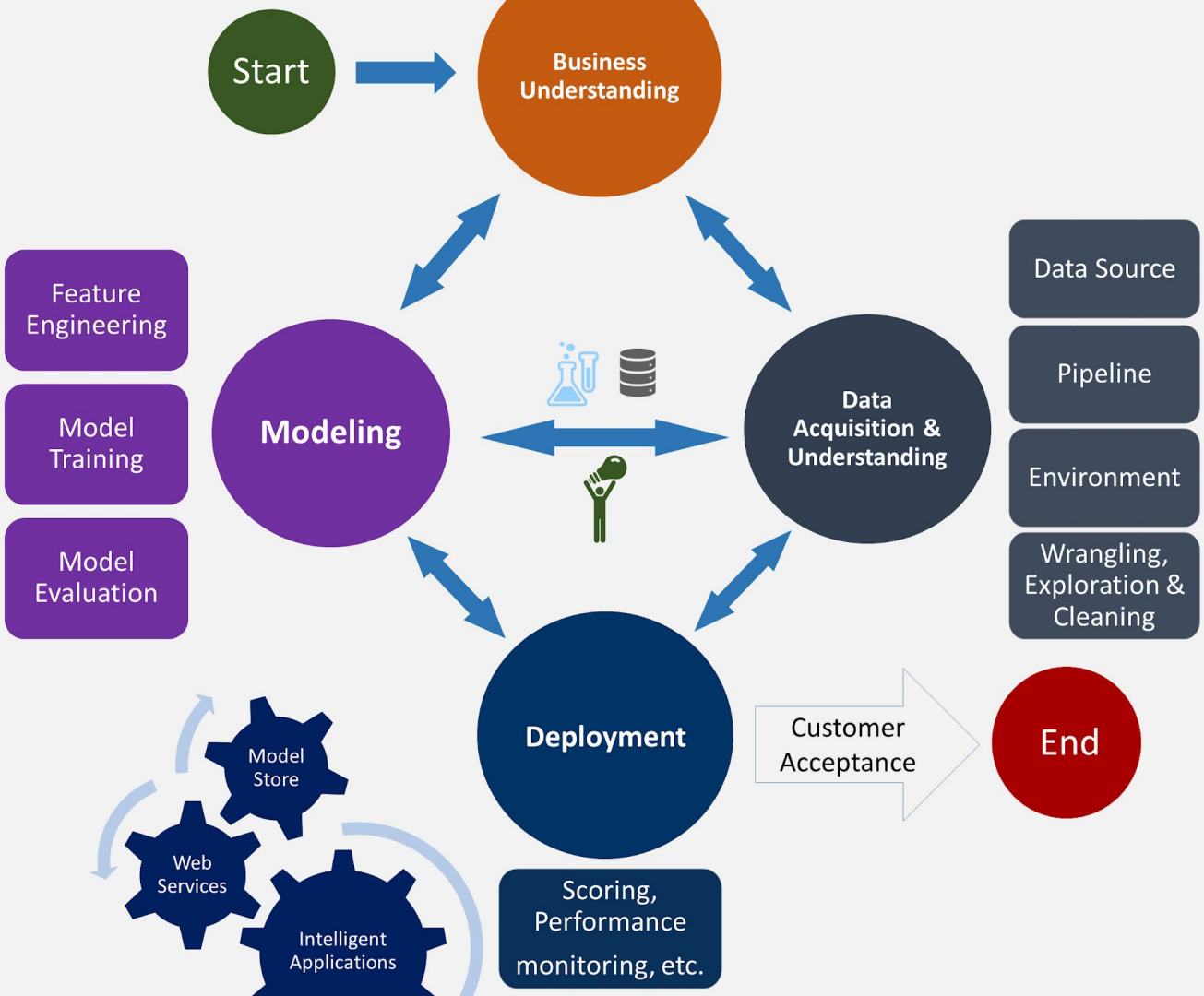
6. Внедрение / Deployment

- Планирование развертывания
  (вид конечного решения / сервиса)

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

- Отчет и презентация
  (финальный отчет по результатам моделирования)

3. Команда


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.

Команда

Внутренние специалисты

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

Команда

Внутренние специалисты

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

Внешние:

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

4. Что может пойти не так?

4. Что может пойти не так?

ВСЁ :(

Постановка задачи

Необходимо:

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

Постановка задачи

Необходимо:

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

Ошибки и неточности на этом этапе

- могут быть фатальными
- 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.

Постановка задачи

Бизнес-задача:

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

Постановка 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)

Постановка задачи. Кейс 1

На входе:

- Сделайте сегментацию пользователей для email рассылок

Постановка задачи. Кейс 1

На входе:

- Выделите сегмент премиальных пользователей для рассылок

Постановка задачи. Кейс 1

На входе:

- Мы хотим увеличить прибыль с помощью email рассылок со спецпредложениями для премиальных пользователей, надо выделить их из потока.

Постановка задачи. Кейс 1

После нескольких встреч:

В рамках цели на среднесрочное увеличение выручки мы хотим запустить специальное предложение со скидками на товары категории премиум.

Задача - необходимо максимизировать полезность от данного предложения в плане эффекта на P&L. В идеале понять, кому данное предложение необходимо отправить, учитывая риски, связанные с упущенной выгодой от продажи по меньшей цене и возможного негативного эффекта от увеличения числа рассылок.

Постановка задачи. Кейс 2

На входе:

- Нужно сделать модель, прогнозирующую продажи товаров на следующую неделю

Какую метрику взять?

Постановка задачи. Кейс 2

На входе:

- Нужно сделать модель, прогнозирующую продажи товаров на следующую неделю

Какую метрику взять?

- MAE, MSE, RMSE, MAPE, sMAPE

Постановка задачи. Кейс 2

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

Постановка задачи. Кейс 3

На входе:

- Нужно сделать модель, составляющую расписания и маршруты для курьеров собственной доставки

Спасибо за внимание!