

# Предиктивная аналитика

Эмели Драль  
Data Mining In Action  
весна 2018г.

Предиктивная  
аналитика

Виды анализа данных

Работа над проектом

# 1. Виды анализа данных

# Виды анализа

## От данных к результатам



## Виды анализа

# Исторические данные

## Виды анализа

# Исторические данные

Что происходило раньше?

# Виды анализа

## Исторические данные

Что происходило раньше?

Важно наладить процессы:

- Сбора данных
- Хранения данных
- Обработки данных

## Виды анализа

# Описательная аналитика

## Виды анализа

# Описательная аналитика

Что происходит сейчас?

# Виды анализа

## Описательная аналитика

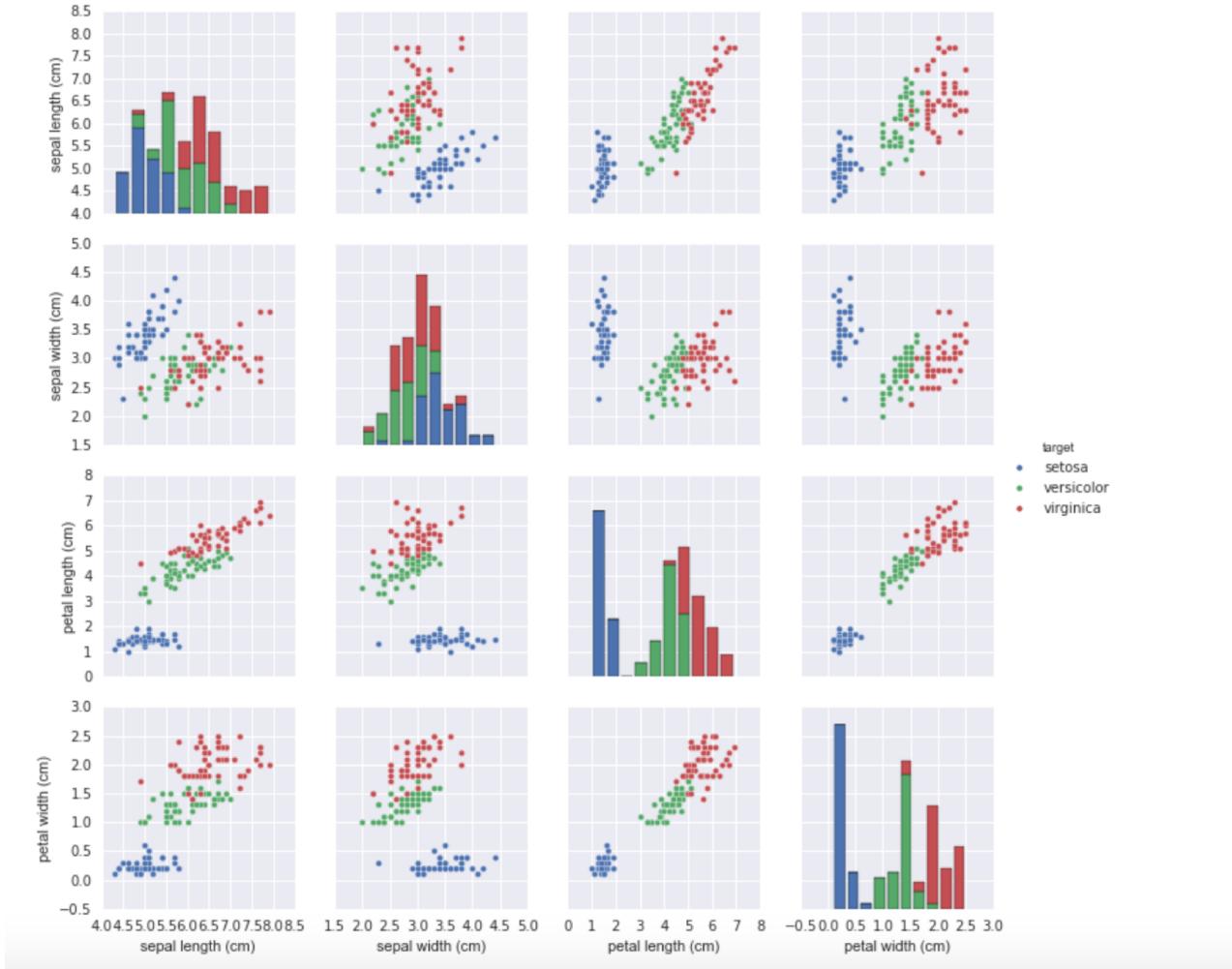
Что происходит сейчас?

Реализуется с помощью:

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

# Виды анализа

# Описательная аналитика



Виды  
анализа

# Диагностическая аналитика

## Виды анализа

# Диагностическая аналитика

В чем причина происходящего?

## Виды анализа

# Диагностическая аналитика

В чем причина происходящего?

Реализуется с помощью:

- Разведочного анализа
- Статистического анализа

## Виды анализа

### Диагностическая аналитика

- Диаграммы, гистограммы
- Статистики и распределения
- Корреляционный анализ
- Проверка статистических гипотез
- Множественная проверка гипотез

# Виды анализа

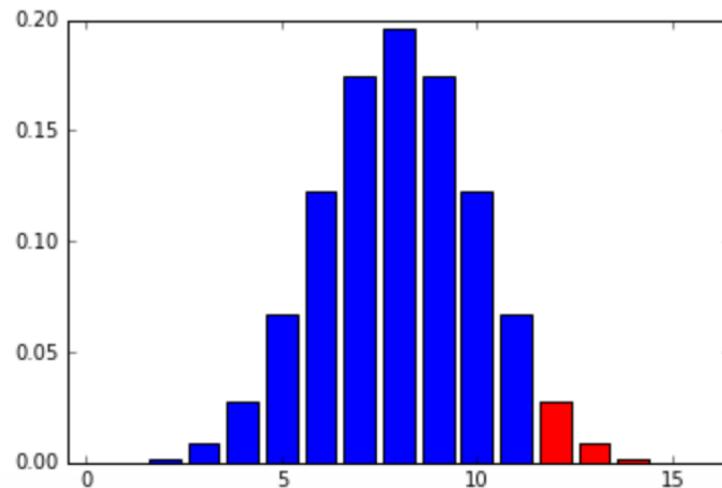
# Диагностическая аналитика

## Односторонняя альтернатива

гипотеза  $H_1$ : Джеймс Бонд предпочитает взболтанный мартини.

При такой альтернативе более вероятны большие значения статистики; при расчёте достигаемого уровня значимости будем суммировать высоту столбиков в правом хвосте распределения.

```
pylab.bar(x, F_H0.pmf(x), align = 'center')
pylab.bar(np.linspace(12,16,5), F_H0.pmf(np.linspace(12,16,5)), align = 'center', color='red')
xlim(-0.5, 16.5)
pylab.show()
```



## Виды анализа

# Предсказательная аналитика

## Виды анализа

# Предсказательная аналитика

Что произойдет в будущем?

# Виды анализа

## Предсказательная аналитика

Что произойдет в будущем?

Реализуется с помощью:

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

## Виды анализа

# Предсказательная аналитика



## Виды анализа

# Предписывающая аналитика

## Виды анализа

# Предписывающая аналитика

Что мы должны предпринять для достижения цели?

## Виды анализа

### Предписывающая аналитика

Что мы должны предпринять для достижения цели?

Реализуется с помощью:

- Рекомендательных систем
- Систем поддержки принятия решений
- Систем скоринга возможных сценариев
- Решений по автоматизация процессов

## 2. Работа над проектом

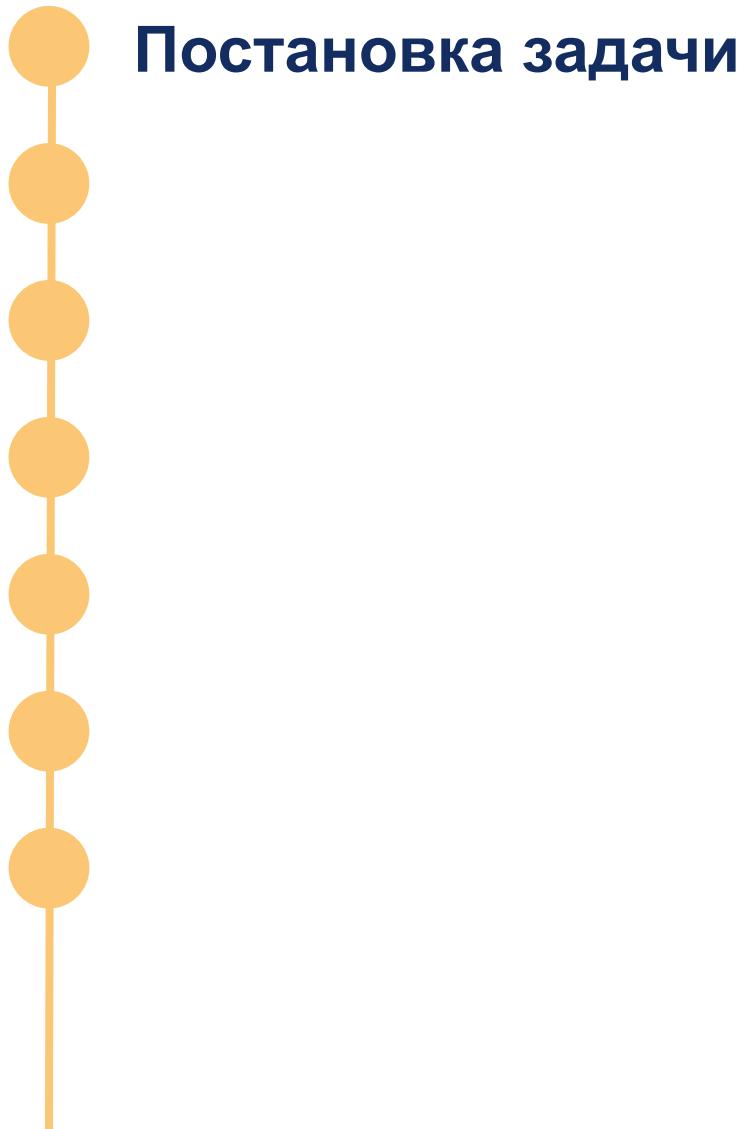
# Работа над проектом

## Этапы проекта

- 
- Постановка задачи
  - Определение метрик и критериев успеха
  - Оценка доступных данных
  - Обучение предсказательной модели
  - Тестирование модели (эксперимент)
  - Тестирование качества работы сервиса
  - Поддержка качества, дообучение модели

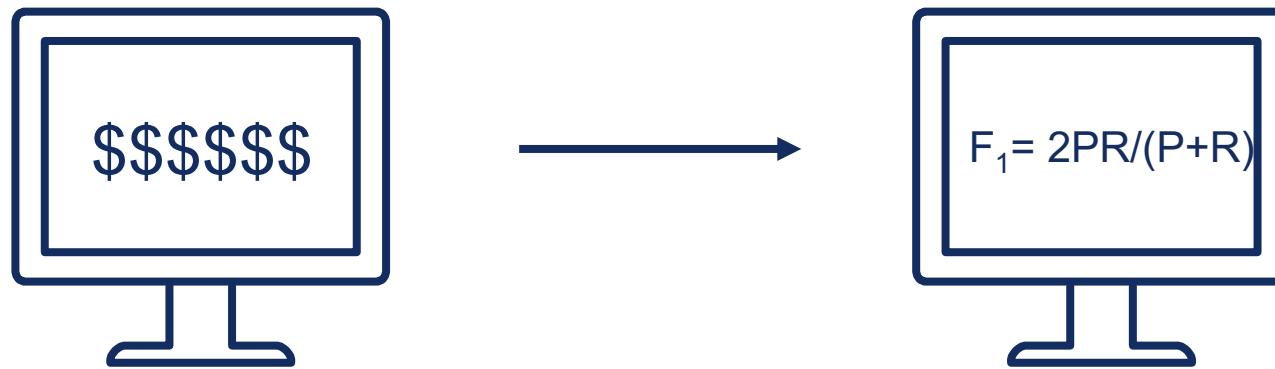
# Работа над проектом

## Этапы проекта



# Постановка задачи

## Постановка задачи



Бизнес-задача

Математическая  
постановка

# Постановка задачи

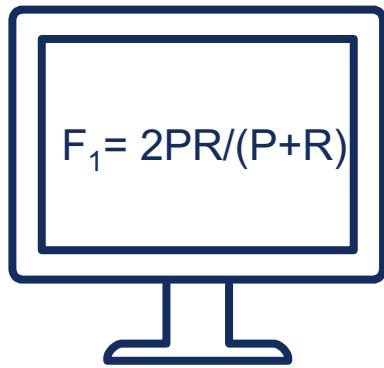
## Бизнес-задача



- Сформулированная бизнес-цель
- Требует экспертных знаний в предметной области
- Обычно хорошо измеряется в деньгах

# Постановка задачи

## Математическая постановка



- Формальная постановка задачи в терминах анализа данных
- Требует экспертных знаний в математике
- Обычно хорошо измеряется в числах (точность, полнота, аккуратность)

## Постановка задачи

# Пример: что такое отток?

Отказ пользователя от некоторого продукта или услуги



# Определение оттока

Постановка  
задачи

## Определение оттока

- Разрыв договора
- Отсутствие платных транзакций более 10/90 дней
- Отсутствие на сервисе более 14/28 дней

## Постановка задачи

## Постановка задачи

Какая доля пользователей попадает под это определение?

Определение	Доля
Разрыв договора	0,1%
Отсутствие платных транзакций более 10 дней	22%
Отсутствие платных транзакций более 90 дней	90%
Отсутствие в сети более 14 дней	16%
Отсутствие в сети более 28 дней	9%

## Постановка задачи

# Какая доля пользователей потом возвращается?

Определение	Доля
Разрыв договора	0%
Отсутствие платных транзакций более 10 дней	80%
Отсутствие платных транзакций более 90 дней	12%
Отсутствие в сети более 14 дней	90%
Отсутствие в сети более 28 дней	78%

# Постановка задачи

## Формализация задачи

Отток	?
Модель	?
Горизонт прогнозирования	?
Методика оценки	?
Дизайн эксперимента	?
Требования к модели	?

# Постановка задачи

## Формализация задачи

Отток	разрыв договора подключения к сервису
Модель	бинарная классификация
Горизонт прогнозирования	2 недели
Методика оценки	точность модели в топ-5% (precision@5)
Дизайн эксперимента	А/Б тест на 10% сегменте случайных пользователей
Требования к модели	вероятностная модель

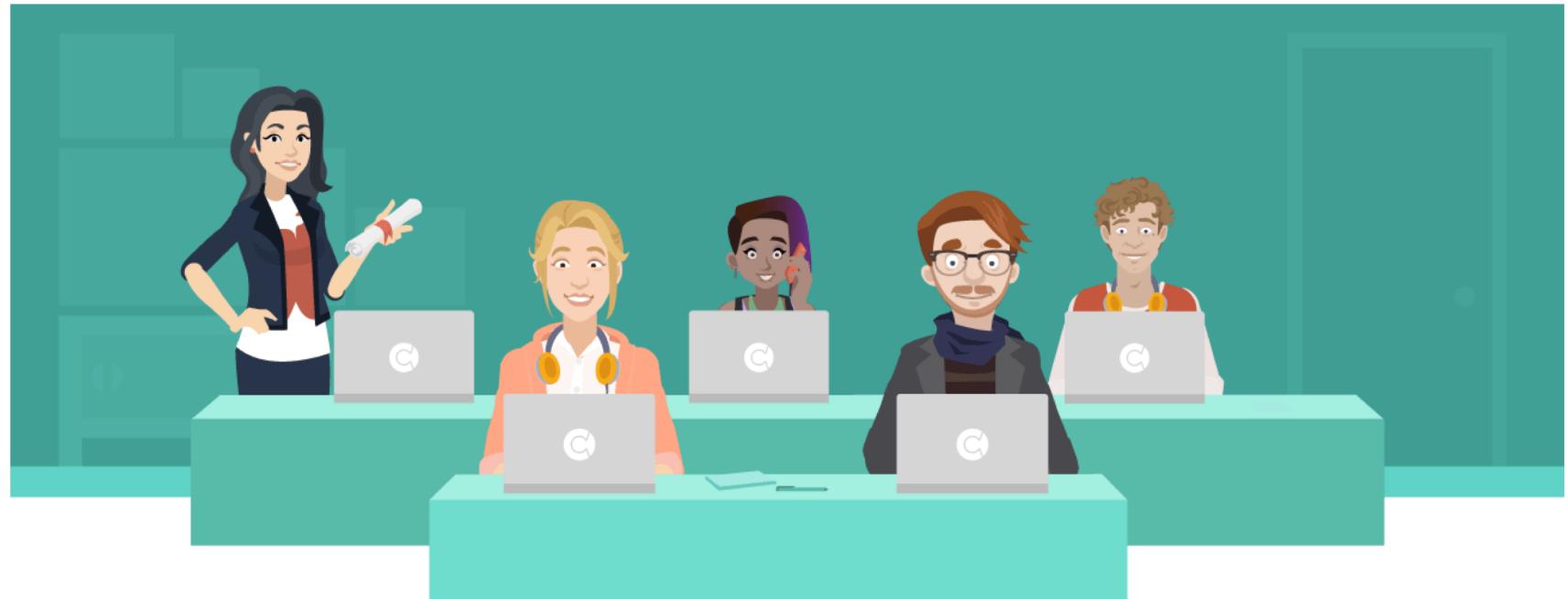
## Постановка задачи

### Бизнес-процесс

- Нуждается ли бизнес процесс в оптимизации?

# Постановка задачи

# Бизнес-процесс



## Постановка задачи

### Бизнес-процесс

- Обоснован ли проект экономически?

# Бизнес-процесс

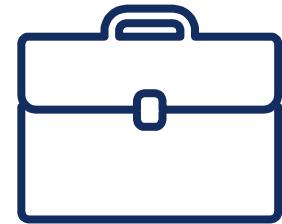
## Постановка задачи

**Are you making more money than you are spending?**

Knowing when to stop improving a once profitable model is also an important lesson to learn.

<http://www.cbronline.com/news/big-data/analytics/five-questions-every-business-must-ask-starting-machine-learning-project/>

# Постановка задачи



Постановка должна соответствовать бизнес-цели



Нужна отраслевая экспертиза



Нужна экспертиза в машинном обучении



Важно делиться экспертизой и работать над постановкой командно

# Работа над проектом

## Этапы проекта

- 
- Постановка задачи
  - Определение метрик и критериев успеха
  - Изучение требований
  - Создание прототипа
  - Тестирование и отладка
  - Реализация функций
  - Финальный тест

# Метрики и критерии успеха

# Фиксация метрик, критериев успеха



# Метрики и критерии успеха

## Фиксация метрик, критериев успеха

Метрики оценки ошибки прогноза:

- MAE
- MSE
- RMSE
- MAPE
- SMAPE
- и пр.

# Метрики и критерии успеха

## Фиксация метрик, критериев успеха

### MAPE

- Недопрогнозирование:

actual = 100

forecast = 90

mape = 10%

- Перепрогнозирование

actual = 100

forecast = 110

mape = 10%

# Метрики и критерии успеха

## Фиксация метрик, критериев успеха

### MAPE

- Недопрогнозирование:

actual = 100

forecast = 90

mape = 10%

- Перепрогнозирование

actual = 100

forecast = 110

mape = 10%

### SMAPE

- Недопрогнозирование:

actual = 100

forecast = 90

mape = 10,5%

- Перепрогнозирование

actual = 100

forecast = 110

mape = 9,5%

# Метрики и критерии успеха

## Специфичные метрики

- Return rate
- Churn rate
- X-day retention
- Rolling retention

# Метрики и критерии успеха

## Специфичные метрики

### Return rate

$$RR = \frac{\text{current number of customers from the original set}}{\text{number of customers at the original set}} * 100$$

### Churn rate

$$CR = \frac{\text{number of churned customers}}{\text{total number of customers}} * 100$$

# Работа над проектом

## Этапы проекта

- 
- Постановка задачи
  - Определение метрик и критериев успеха
  - Оценка доступных данных**
  - 
  - 
  - 
  -

# Оценка доступных данных

# Оценка доступности данных

Schicht/ Datum	Sorte	Einheiten	D/02.02.03
<b>FLG</b>			
Stoffverhältnis DIP / Etik.			
V - Sieb			
V - Poperoller			
Arbeitsbreite			
Stoffauflauf			
Auslaufverhältnis			
Druck			
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
FU-Stoff			
Schüttel			
Schüttel			
Duofo			
Obersieben			
Scimmer / Entw.			
1. Zg			
2. Zg			
Druck Leiste			
Einlaufwalze Duofo			
Vakuumeinstell			
1.Vakufoil			
2. Vakufoil / Näh			
Doppelvakuf			
Scimmer			
1.Formationszone			
2. Zone (Trockeng)			
Trennsauger			
Flachsäger			
SSW			
PU Haltezone			
PU PreZone			
Pressenpartie / Linie			
1.Presse			
2.Presse			
3.Presse			
Pressmantelstellung			
B			
Schicht/ Datum	Sorte	Einheiten	Newspaper heatset
V - Sieb	FLG	g/m <sup>2</sup>	
V - Poperoller		%	46,5
Arbeitsbreite		m/min.	49
Stoffauflauf		m/min.	859
Auslaufverhältnis		m	400
Druck			3,075
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
Pumpendre			
Schüttelbock			
Schüttelbock Hub			
Duofo			
Scimmer			
Oberse			
Druck			
Einlaufwalze			
Vakuum			
2. Vak			
1. Form			
2. Zon			
Press			
Korre			
C			
Schicht/ Datum	Sorte	Einheiten	
V - Sieb	FLG	g/m <sup>2</sup>	
V - Poperoller		%	49
Arbeitsbreite		m/min.	859
Stoffauflauf		m/min.	400
Auslaufverhältnis		m	3,075
Druck			
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
Pumpendre			
Schüttelbock			
Schüttelbock Hub			
Duofo			
Scimmer			
Oberse			
Druck			
Einlaufwalze			
Vakuum			
2. Vak			
1. Form			
2. Zon			
Press			
Korre			
D			
Schicht/ Datum	Sorte	Einheiten	R
Stoffverhältnis DIP / Etik.	FLG	g/m <sup>2</sup>	05.01.03
V - Sieb		%	Newspap.
V - Poperoller		m/min.	46,5
Arbeitsbreite		m/min.	49
Stoffauflauf		m	859
Auslaufverhältnis			400
Druck			3,075
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
Pumpendre			
Schüttelbock			
Schüttelbock Hub			
Duofo			
Scimmer			
Oberse			
Druck			
Einlaufwalze			
Vakuum			
2. Vak			
1. Form			
2. Zon			
Press			
Korre			
E			
Schicht/ Datum	Sorte	Einheiten	
V - Sieb	FLG	g/m <sup>2</sup>	
V - Poperoller		%	
Arbeitsbreite		m/min.	
Stoffauflauf		m	
Auslaufverhältnis			
Druck			
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
Pumpendre			
Schüttelbock			
Schüttelbock Hub			
Duofo			
Scimmer			
Oberse			
Druck			
Einlaufwalze			
Vakuum			
2. Vak			
1. Form			
2. Zon			
Press			
Korre			
F			
Schicht/ Datum	Sorte	Einheiten	
V - Sieb	FLG	g/m <sup>2</sup>	
V - Poperoller		%	
Arbeitsbreite		m/min.	
Stoffauflauf		m	
Auslaufverhältnis			
Druck			
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
Pumpendre			
Schüttelbock			
Schüttelbock Hub			
Duofo			
Scimmer			
Oberse			
Druck			
Einlaufwalze			
Vakuum			
2. Vak			
1. Form			
2. Zon			
Press			
Korre			
G			
Schicht/ Datum	Sorte	Einheiten	
V - Sieb	FLG	g/m <sup>2</sup>	
V - Poperoller		%	
Arbeitsbreite		m/min.	
Stoffauflauf		m	
Auslaufverhältnis			
Druck			
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
Pumpendre			
Schüttelbock			
Schüttelbock Hub			
Duofo			
Scimmer			
Oberse			
Druck			
Einlaufwalze			
Vakuum			
2. Vak			
1. Form			
2. Zon			
Press			
Korre			
H			
Schicht/ Datum	Sorte	Einheiten	
V - Sieb	FLG	g/m <sup>2</sup>	
V - Poperoller		%	
Arbeitsbreite		m/min.	
Stoffauflauf		m	
Auslaufverhältnis			
Druck			
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
Pumpendre			
Schüttelbock			
Schüttelbock Hub			
Duofo			
Scimmer			
Oberse			
Druck			
Einlaufwalze			
Vakuum			
2. Vak			
1. Form			
2. Zon			
Press			
Korre			
I			
Schicht/ Datum	Sorte	Einheiten	
V - Sieb	FLG	g/m <sup>2</sup>	
V - Poperoller		%	
Arbeitsbreite		m/min.	
Stoffauflauf		m	
Auslaufverhältnis			
Druck			
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
Pumpendre			
Schüttelbock			
Schüttelbock Hub			
Duofo			
Scimmer			
Oberse			
Druck			
Einlaufwalze			
Vakuum			
2. Vak			
1. Form			
2. Zon			
Press			
Korre			
J			
Schicht/ Datum	Sorte	Einheiten	
V - Sieb	FLG	g/m <sup>2</sup>	
V - Poperoller		%	
Arbeitsbreite		m/min.	
Stoffauflauf		m	
Auslaufverhältnis			
Druck			
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
Pumpendre			
Schüttelbock			
Schüttelbock Hub			
Duofo			
Scimmer			
Oberse			
Druck			
Einlaufwalze			
Vakuum			
2. Vak			
1. Form			
2. Zon			
Press			
Korre			
K			
Schicht/ Datum	Sorte	Einheiten	
V - Sieb	FLG	g/m <sup>2</sup>	
V - Poperoller		%	
Arbeitsbreite		m/min.	
Stoffauflauf		m	
Auslaufverhältnis			
Druck			
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
Pumpendre			
Schüttelbock			
Schüttelbock Hub			
Duofo			
Scimmer			
Oberse			
Druck			
Einlaufwalze			
Vakuum			
2. Vak			
1. Form			
2. Zon			
Press			
Korre			
L			
Schicht/ Datum	Sorte	Einheiten	
V - Sieb	FLG	g/m <sup>2</sup>	
V - Poperoller		%	
Arbeitsbreite		m/min.	
Stoffauflauf		m	
Auslaufverhältnis			
Druck			
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
Pumpendre			
Schüttelbock			
Schüttelbock Hub			
Duofo			
Scimmer			
Oberse			
Druck			
Einlaufwalze			
Vakuum			
2. Vak			
1. Form			
2. Zon			
Press			
Korre			
M			
Schicht/ Datum	Sorte	Einheiten	
V - Sieb	FLG	g/m <sup>2</sup>	
V - Poperoller		%	
Arbeitsbreite		m/min.	
Stoffauflauf		m	
Auslaufverhältnis			
Druck			
PD Innendruck			
Druckwasge / Spülung			
Lippeneöffnung			
Vorderwand			
Pumpendre			
Schüttelbock			
Schüttelbock Hub			
Duofo			
Scimmer			
Oberse			
Druck			
Einlaufwalze			
Vakuum			
2. Vak			
1. Form			
2. Zon			
Press			
Korre			
N			
Schicht/ Datum	Sorte	Einheiten	
V - Sieb	FLG	g/m <sup>2</sup>	
V - Poperoller		%	
Arbeitsbreite		m/min.	
Stoffauflauf		m	
Auslaufverh			

# Оценка доступных данных

## Данные

- Какие данные доступны?
- За какой исторический период?
- Как объединять данные ?
- Есть ли в данных сигнал?
- Как данные следует обработать?
- Как рассчитать признаки на основе данных?

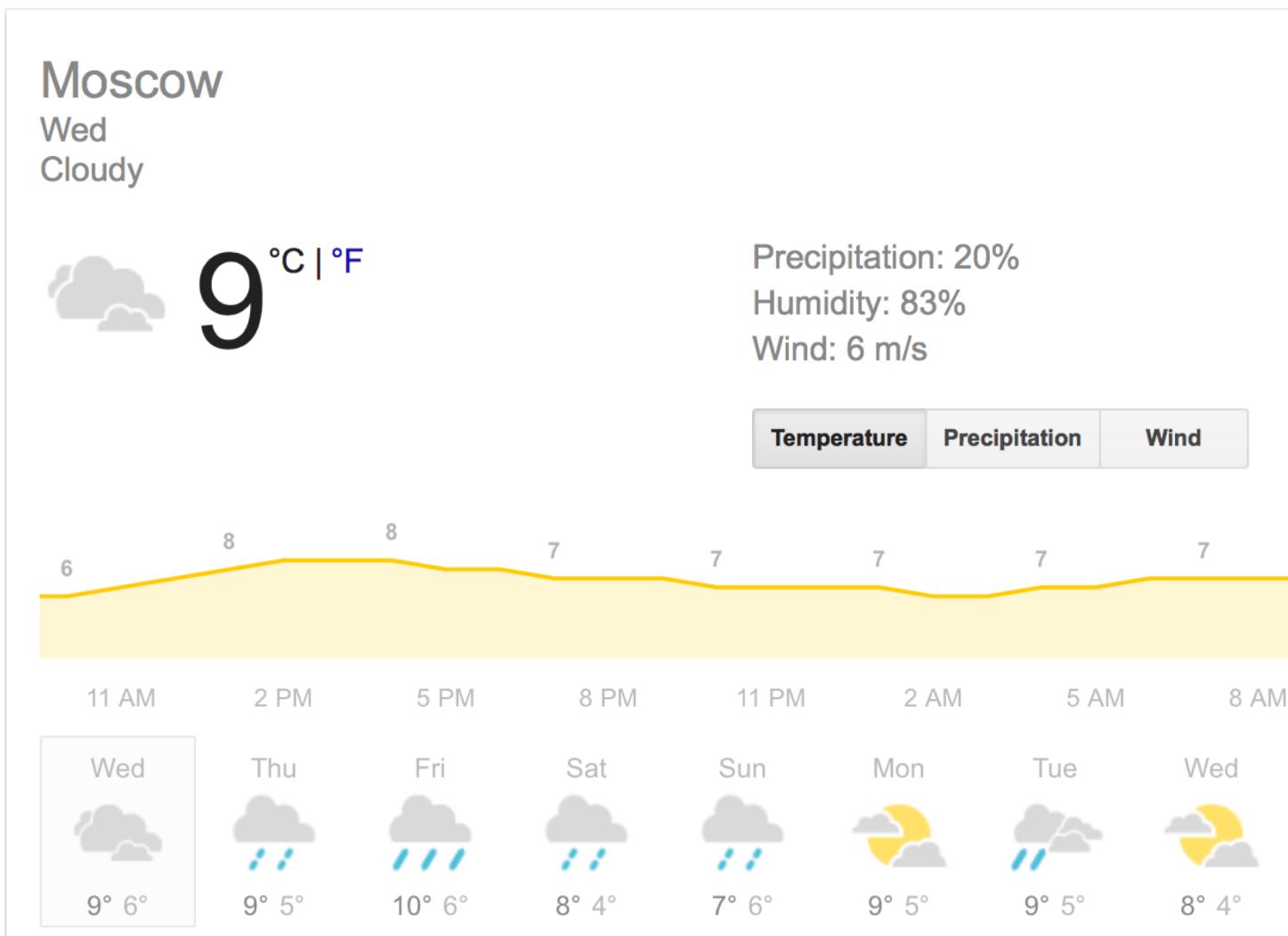
## Оценка доступных данных

## Данные

- Будет ли модель, построенная по историческим данным, работать в production?

# Оценка доступных данных

## Данные



# Работа над проектом

## Этапы проекта

- 
- Постановка задачи
  - Определение метрик и критериев успеха
  - Оценка доступных данных
  - Обучение предсказательной модели**
  - 
  - 
  -

# Обучение модели

## Обучение предсказательной модели

- Обучение на данных, доступных НЕ только за исторический период
- Контроль обучения на данных из будущего
- Контроль переобучения

# Обучение модели

## Кросс-валидация

- По объектам
- По времени

# Обучение модели

## Подбор параметров

- Используем кросс-валидацию
- Сразу фиксируем hold-out dataset
- Их может быть несколько
- Используем их для финальной проверки решения

# Обучение модели

## Подготовка датасета

- Данные разного типа
  - числовые
  - номинальные
  - порядковые
- Временные ряды

# Обучение модели

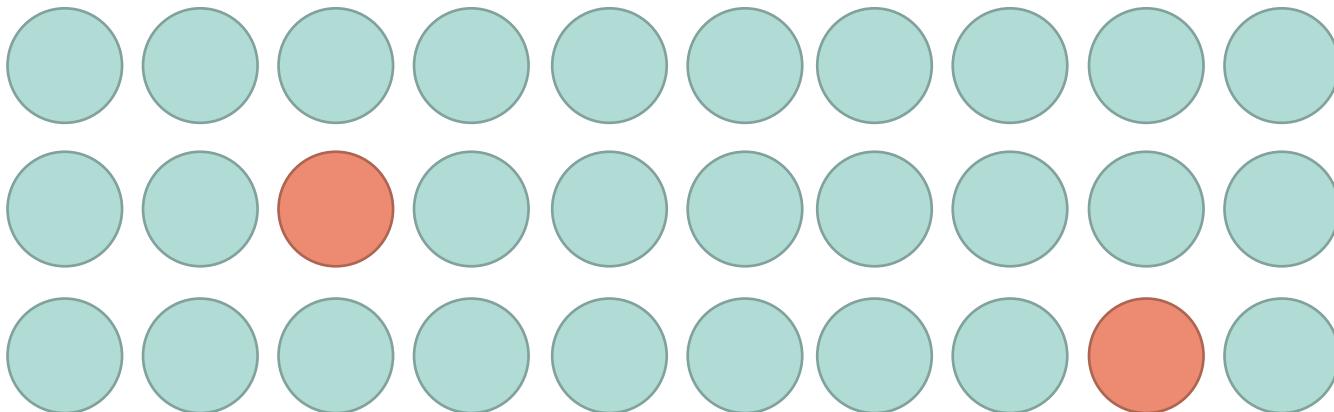
## Несбалансированная выборка

- Доля целевого класса может быть намного меньше доли нецелевого класса (0.1% vs 99.9%)
- Несбалансированность выборки может негативно сказаться на качестве модели
- Важно заметить это в процессе построения модели!

# Обучение модели

## Несбалансированная выборка

- Доля целевого класса может быть намного меньше доли нецелевого класса (0.1% vs 99.9%)
- Несбалансированность выборки может негативно сказаться на качестве модели
- Важно заметить это в процессе построения модели!



# Обучение модели

## Reweighting

Задать веса для объектов таким образом, чтобы:

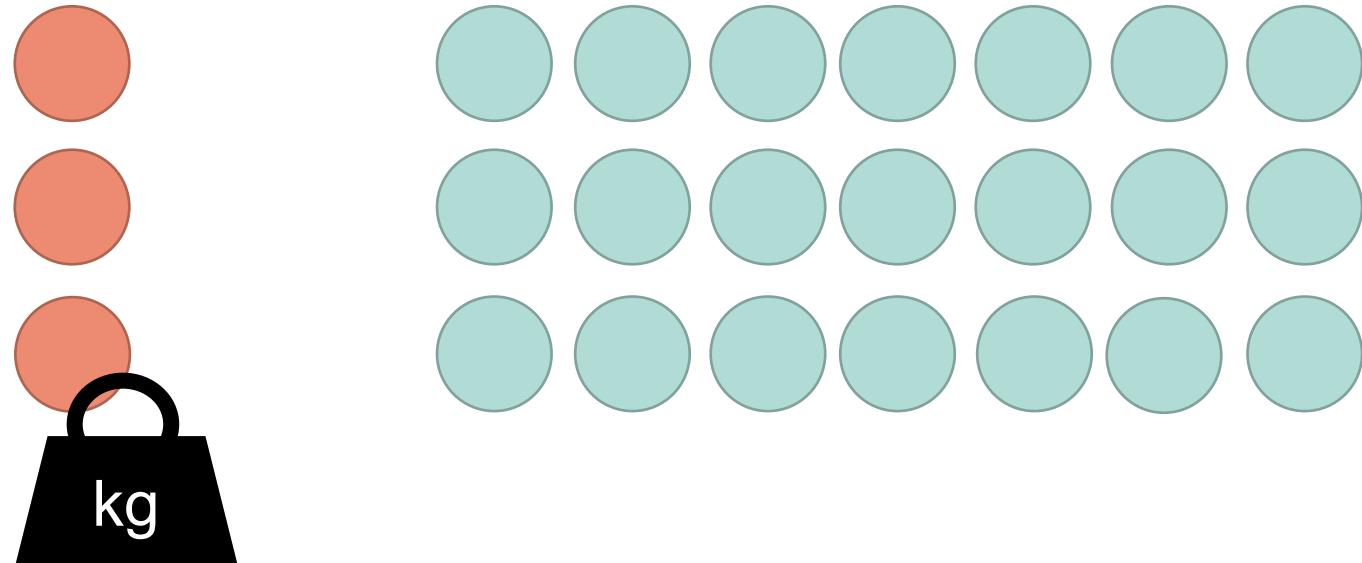
- Скомпенсировать количество объектов меньшего класса их важностью
- Задать стоимость ошибки классификации

# Обучение модели

## Reweighting

Задать веса для объектов таким образом, чтобы:

- Скомпенсировать количество объектов меньшего класса их важностью
- Задать стоимость ошибки классификации



# Обучение модели

## Oversampling

Сгенерировать больше объектов меньшего класса:

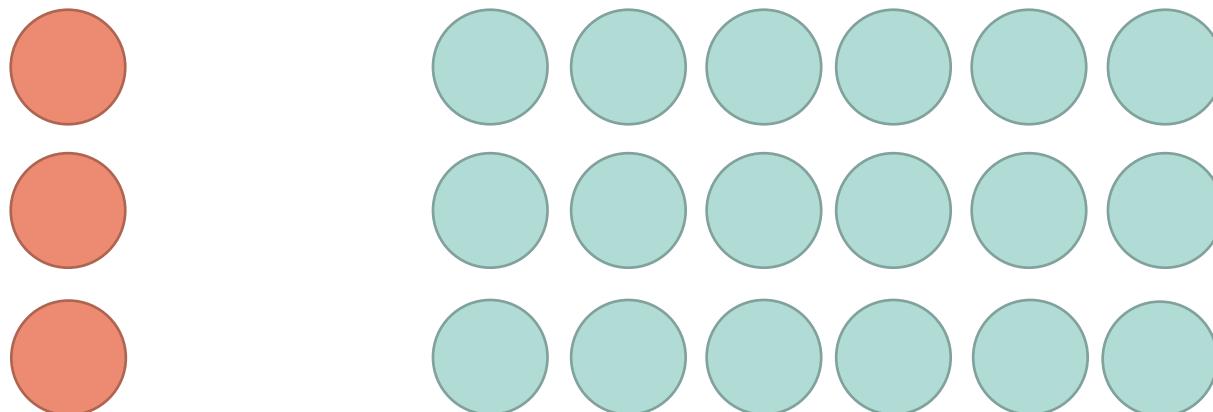
- Дублирование объектов
- Генерация новых объектов путем изменения некоторых признаков существующих объектов
- Генерация на основе нескольких объектов

# Обучение модели

## Oversampling

Сгенерировать больше объектов меньшего класса:

- Дублирование объектов
- Генерация новых объектов путем изменения некоторых признаков существующих объектов
- Генерация на основе нескольких объектов

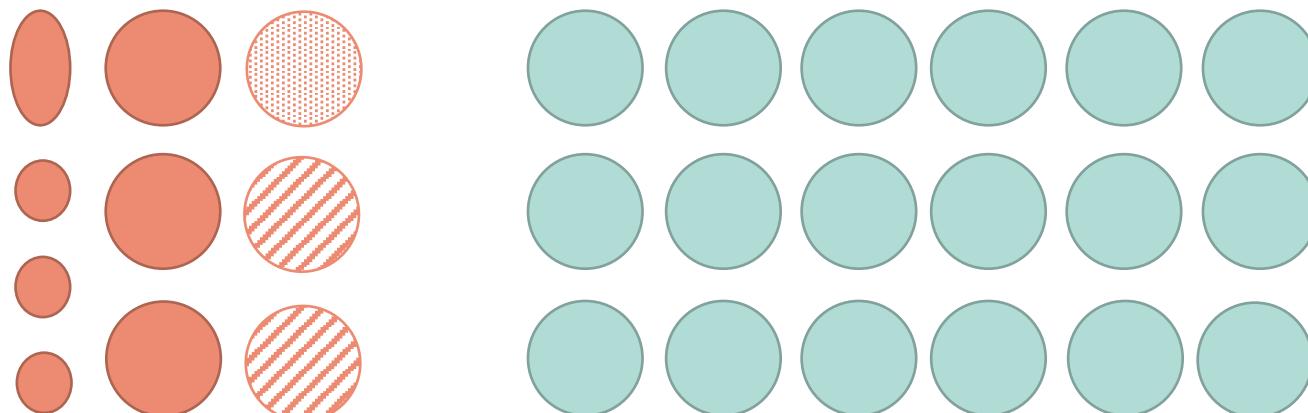


# Обучение модели

## Oversampling

Сгенерировать больше объектов меньшего класса:

- Дублирование объектов
- Генерация новых объектов путем изменения некоторых признаков существующих объектов
- Генерация на основе нескольких объектов



# Обучение модели

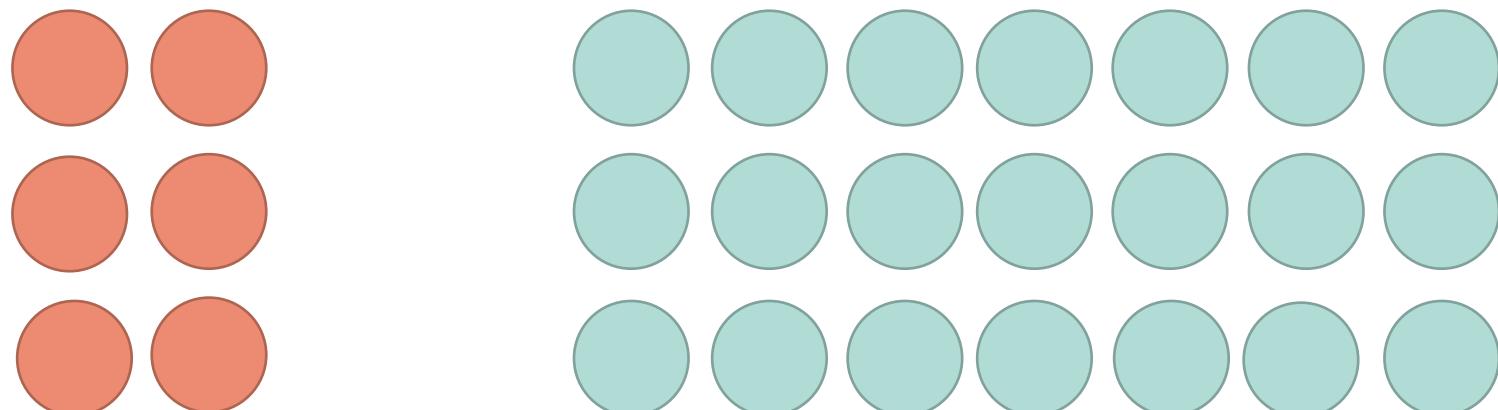
## Undersampling

- Исключить из обучения объекты преобладающего класса:
  - Удаление из выборки случайных объектов преобладающего класса
  - Удаление из выборки групп схожих объектов из преобладающего класса

# Обучение модели

## Undersampling

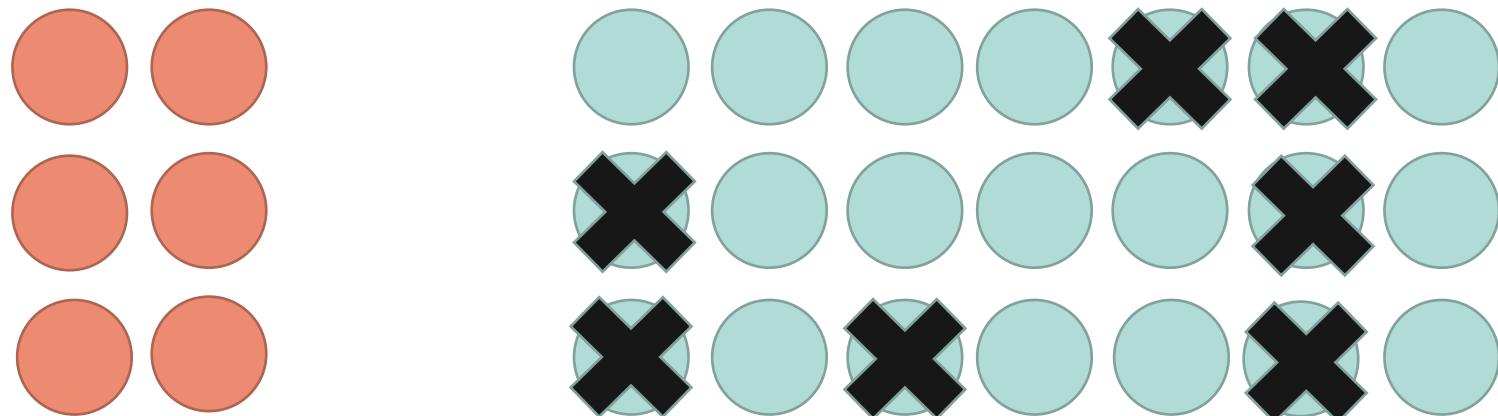
- Исключить из обучения объекты преобладающего класса:
  - Удаление из выборки случайных объектов преобладающего класса
  - Удаление из выборки групп схожих объектов из преобладающего класса



# Обучение модели

## Undersampling

- Исключить из обучения объекты преобладающего класса:
  - Удаление из выборки случайных объектов преобладающего класса
  - Удаление из выборки групп схожих объектов из преобладающего класса



# Работа над проектом

## Этапы проекта

- 
- Постановка задачи
  - Определение метрик и критериев успеха
  - Оценка доступных данных
  - Обучение предсказательной модели
  - Тестирование модели (эксперимент)**
  - 
  -

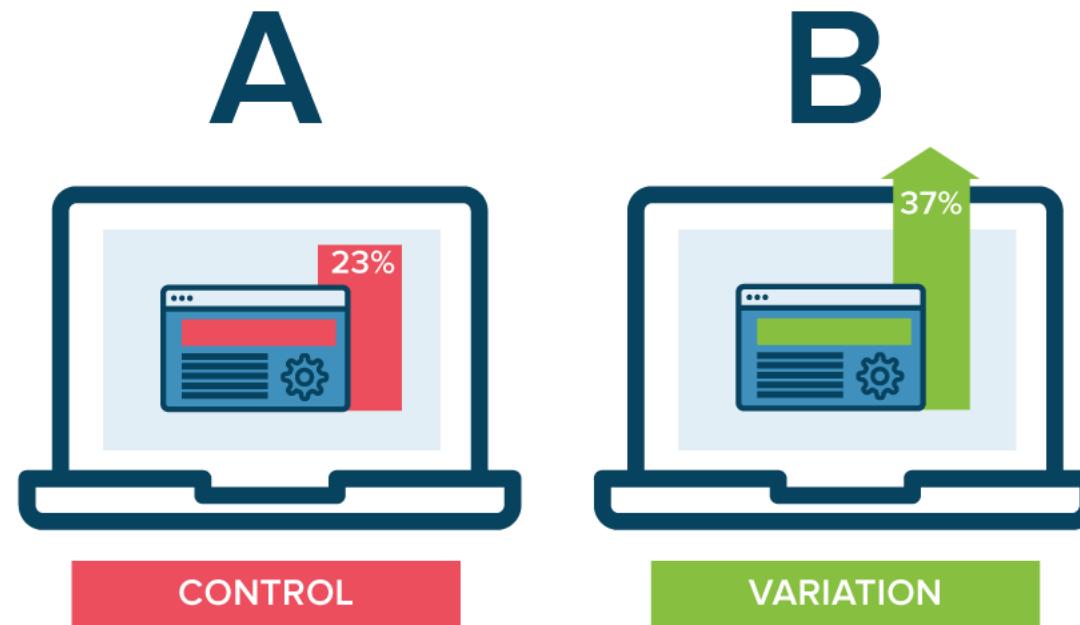
## Тестирование модели

### Дизайн эксперимента

- Позволяет ли эксперимент достоверно оценить качество модели в production?

## Тестирование модели

# Дизайн эксперимента



Тестирование  
модели

# Coca-Cola



# Работа над проектом

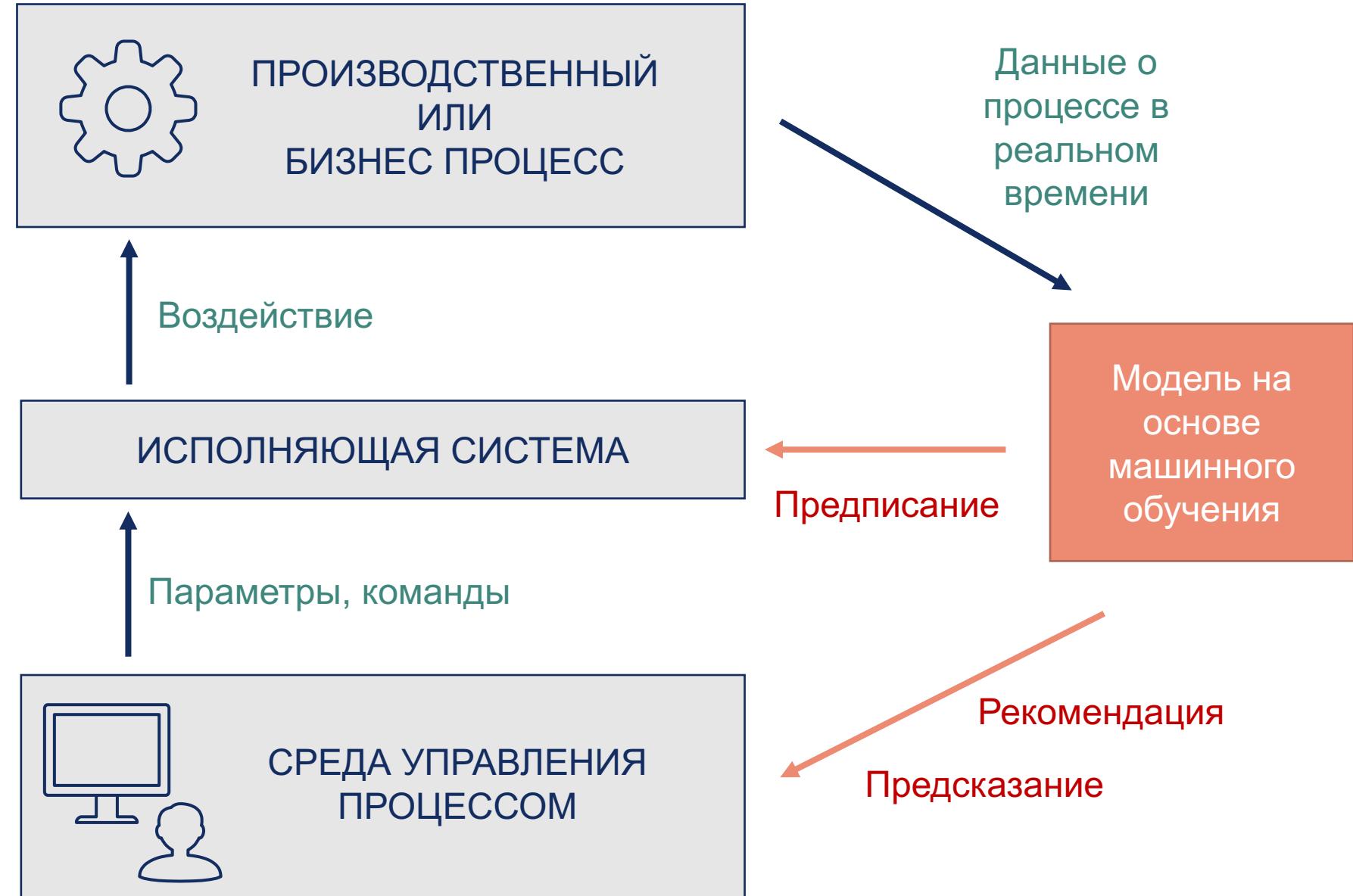
## Этапы проекта

- 
- Постановка задачи
  - Определение метрик и критериев успеха
  - Оценка доступных данных
  - Обучение предсказательной модели
  - Тестирование модели (эксперимент)
  - Тестирование качества работы сервиса**

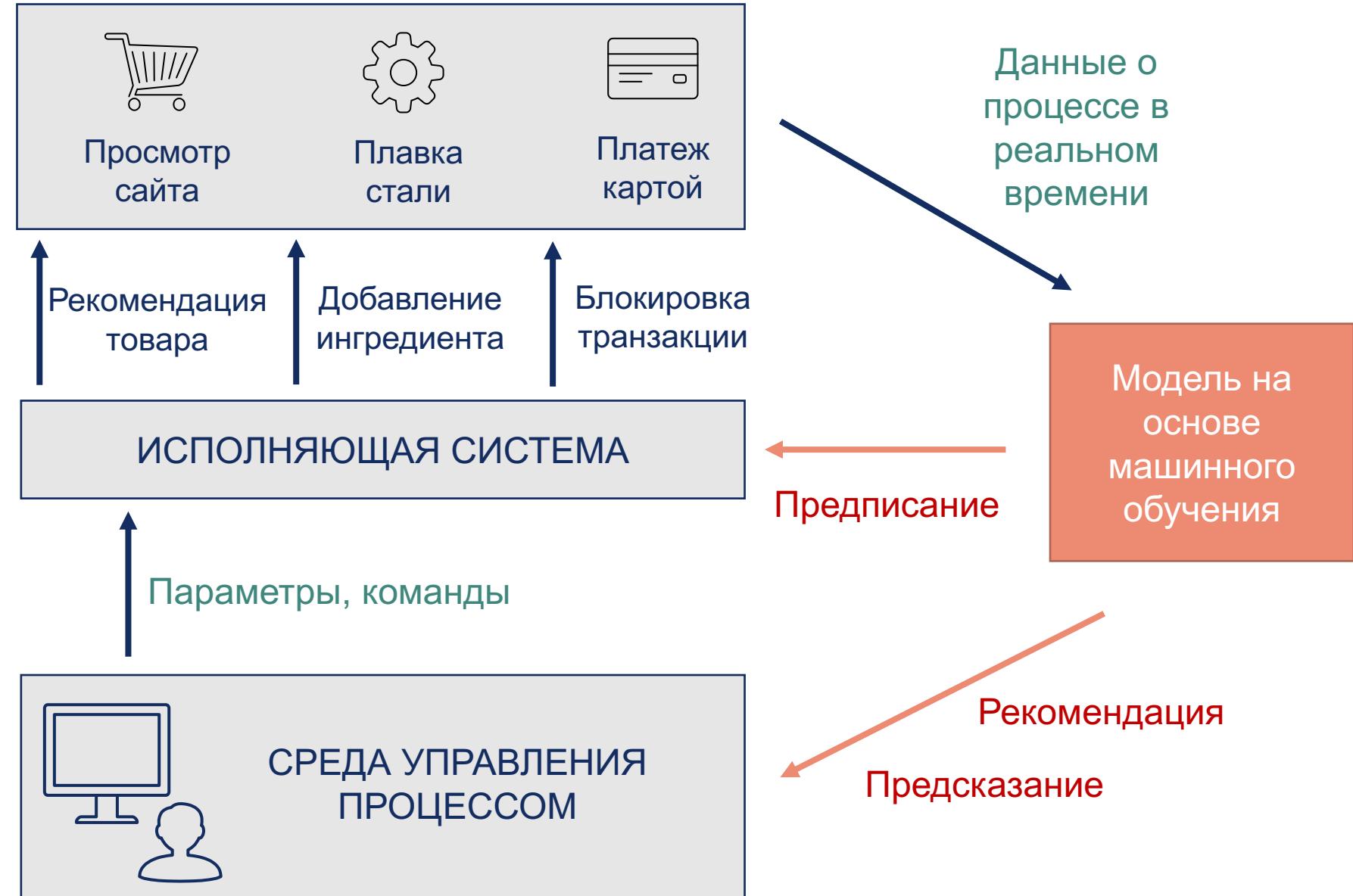
# Как выглядит сервис



# Как выглядит сервис



# Как выглядит сервис



# Качество сервиса

## Мониторинг качества

- Изменились ли данные?

# Качество сервиса

## Мониторинг качества

- Изменились ли данные?
- Изменилось ли качество модели?

# Качество сервиса

## Мониторинг качества

- Изменились ли данные?
- Изменилось ли качество модели?
- Хорошо оценивать модель с разных сторон с помощью набора метрик

# Работа над проектом

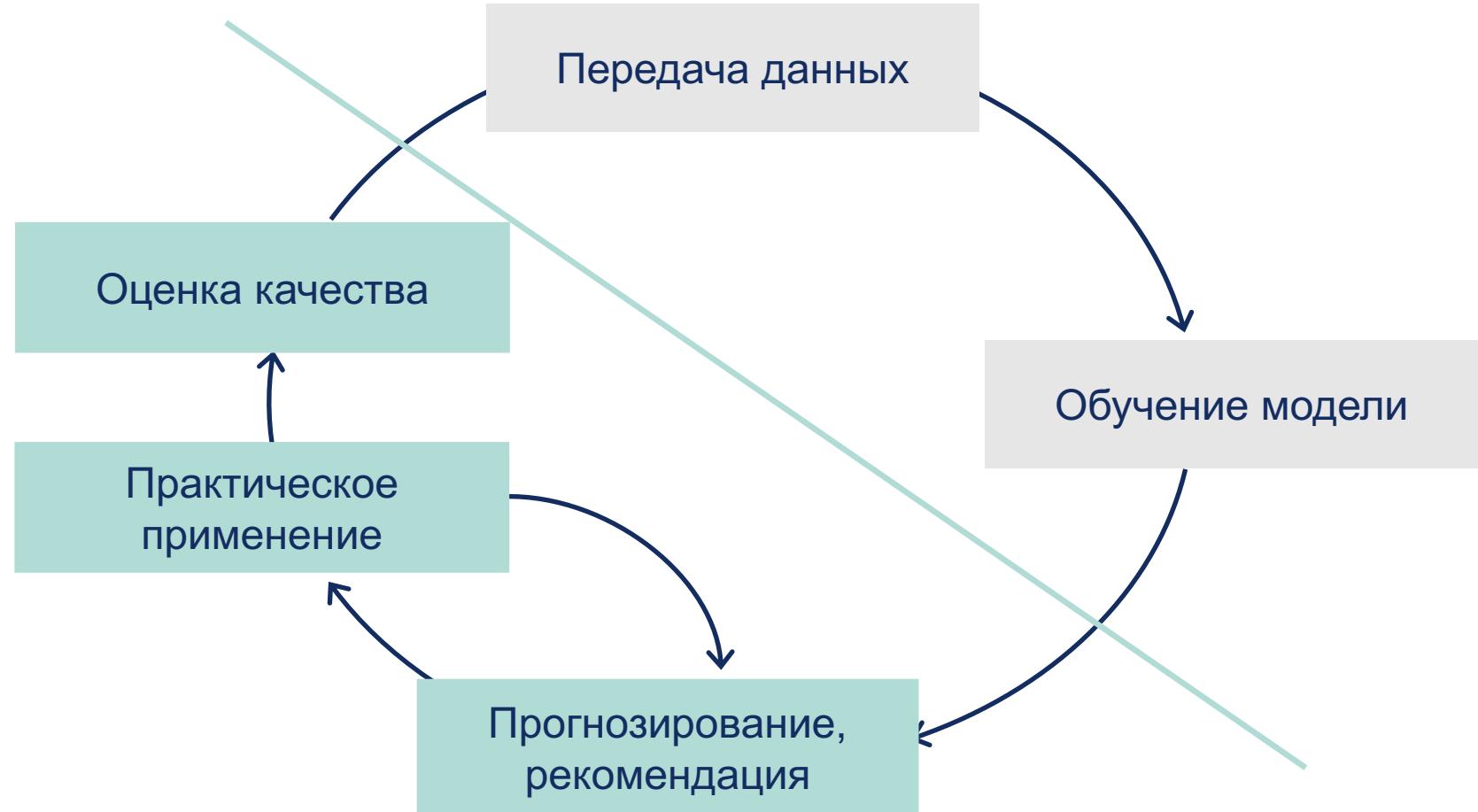
## Этапы проекта

- 
- Постановка задачи
  - Определение метрик и критериев успеха
  - Оценка доступных данных
  - Обучение предсказательной модели
  - Тестирование модели (эксперимент)
  - Тестирование качества работы сервиса
  - Поддержка качества, дообучение модели

# Жизненный цикл модели

near  
real-time

offline



# Регулярное обновление модели

## Перспективы улучшения модели

- Как меняется качество модели во времени?

## Регулярное обновление модели

## Перспективы улучшения модели

- Как меняется качество модели во времени?
- Как быстро она «протухает»?

# Регулярное обновление модели

## Перспективы улучшения модели

- Как меняется качество модели во времени?
- Как быстро она «протухает»?
- Сколько времени занимает переобучение модели?

# Регулярное обновление модели

## Перспективы улучшения модели

- Как меняется качество модели во времени?
- Как быстро она «протухает»?
- Сколько времени занимает переобучение модели?
- Сколько времени требуется на переключение с одной модели на другую?

## Регулярное обновление модели

## Перспективы улучшения модели

- На каких группах объектов модель ошибается?
- Является ли инвестиция в дальнейшее улучшение модели экономически оправданной?

# Работа над проектом

## Этапы проекта

- 
- Постановка задачи
  - Определение метрик и критериев успеха
  - Оценка доступных данных
  - Обучение предсказательной модели
  - Тестирование модели (эксперимент)
  - Тестирование качества работы сервиса
  - Поддержка качества, дообучение модели

# Что может пойти не так?

Работа над  
проектом

# Работа над проектом

## ВСЁ

- Постановка задачи
- Определение метрик и критериев успеха
- Оценка доступных данных
- Обучение предсказательной модели
- Тестирование модели (эксперимент)
- Тестирование качества работы сервиса
- Поддержка качества, дообучение модели

Предиктивная  
аналитика

Виды анализа данных

Работа над проектом

# Спасибо!

Эмели Драль