НЕТОЛОГИЯ
групп

Блок

FEATURE ENGINEERING

ЕГОР САЧКО

Lead Data Scientist

Сбербанк

egsachko@gmail.com

sachkoe

ЦЕЛИ БЛОКА

ЦЕЛИ БЛОКА

В КОНЦЕ БЛОКА ВЫ СМОЖЕТЕ

1. Обрабатывать и очищать данные
2. Осуществлять feature engineering
3. Оценивать значимость признаков
4. Feature selection
5. Уменьшать размерность пространства

СТРУКТУРА БЛОКА

СТРУКТУРА БЛОКА

7
занятий

3
домашних задания

...
самостоятельное изучение материалов

ROADMAP БЛОКА

ROADMAP БЛОКА

1 Потенциальные проблемы с данными

2 Анализ и подготовка данных

3 Feature engineering

4 Feature selection

5 Уменьшение размерности и методы декомпозиции

МАТЕРИАЛЫ ПО БЛОКУ

МАТЕРИАЛЫ ПО БЛОКУ

1. **“Learning scikit-learn: Machine Learning in Python”**
Raul Garreta,
Guillermo Moncecchi,
2013,
Packt

2. **“Hands-On Machine Learning with Scikit-Learn and Tensorflow: Concepts, Tools and Techniques to Build Intelligent Systems”**
Geron, A., 2017, O’Reilly Media

3. **“Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists”**
Zheng, A., Casari, A.,
2018,
O’Reilly Media

4. blog.kaggle.com/ - No Free Hunch

Занятие 1

ПРОБЛЕМЫ КАЧЕСТВА И РАЗМЕРНОСТИ ДАННЫХ

ЦЕЛИ ЗАНЯТИЯ

ЦЕЛИ ЗАНЯТИЯ

В КОНЦЕ ЗАНЯТИЯ ВЫ СМОЖЕТЕ

1. Описывать **основные проблемы с данными**
2. Определять **валидность и правильность данных**
3. Работать с **пропущенным и значениями**
4. Обрабатывать **категориальные переменные**
5. Использовать **pandas и numpy для очистки данных**

ЧТО БУДЕМ ОБСУЖДАТЬ

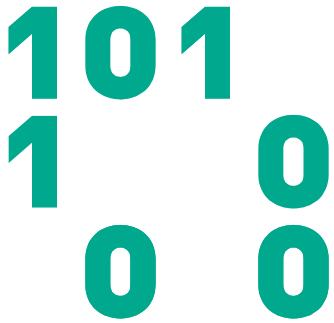
ПЛАН ЗАНЯТИЯ

1. Данные, фичи, пространства данных
2. Сопутствующие проблемы с данными
3. Обработка ошибочных и пропущенных значений
4. Обработка категориальных переменных
5. Pandas и numpy для обработки данных

Часть 1

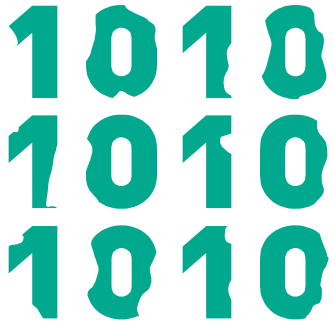
ДАННЫЕ И СОПУТСТВУЮЩИЕ ПРОБЛЕМЫ

ДАННЫЕ И СОПУТСТВУЮЩИЕ ПРОБЛЕМЫ

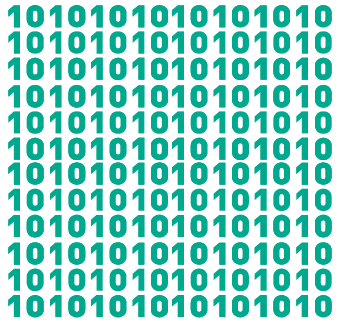


Недостаточное количество данных

Нерепрезентативные данные

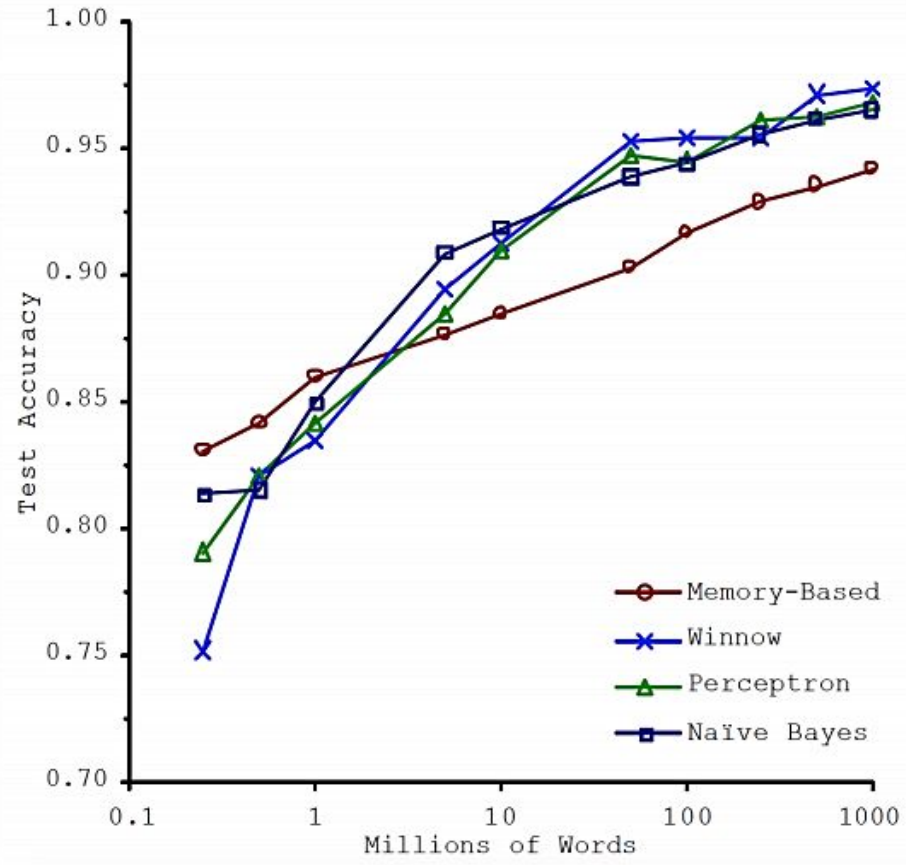


Данные **плохого качества** *аномалии, выбросы*



Большая размерность пространства данных

Чем больше данных, тем лучше



“Scaling to Very Very Large Corpora for Natural Language Disambiguation”

Michele Banko and Eric Brill
2001, Microsoft Research

The Literary Digest poll 1936

The Literary Digest

NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

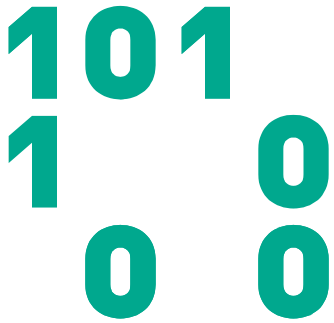
Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

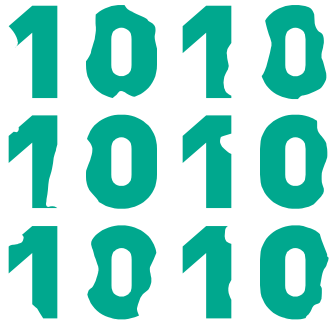
"We never make any claims before election but we respectfully refer you to the

ДАННЫЕ И СОПУТСТВУЮЩИЕ ПРОБЛЕМЫ

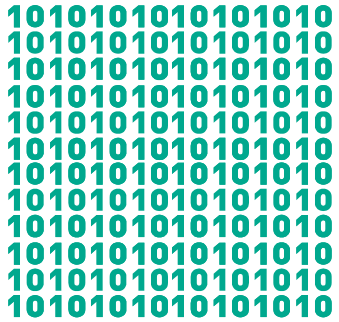


Недостаточное количество данных

Нерепрезентативные данные



Данные **плохого качества** *аномалии, выбросы*



Большая размерность пространства данных

Типы входных признаков

1

Количественные признаки

2

Бинарные признаки

3

Категориальные признаки

- Номинальные *нельзя отранжировать*
- Порядковые *можно отранжировать*

Датасет

- Датасет из бюро переписи населения США
- Каждая строка описывает географическую единицу (район) Калифорнии
- Стоимость жилья данные ~ о районе

github.com/ageron/handson-ml/tree/master/ datasets/housing

Датасет

- Какие атрибуты?
- Какие типы у атрибутов?
- Есть ли NULL значения?
- Какие масштабы значений у атрибутов?
- Есть в данных аномалии/ограничения?
- Визуализация данных

Практика

АНАЛИЗ ДАТАСЕТА HOUSING

Часть 2

ПЕРВИЧНЫЙ АНАЛИЗ ДАННЫХ

Training & Test

A

Разбиваем на Training и Test сеты как можно раньше

B

Data snooping bias

C

Training set — выбор, тренировка и тюнинг моделей

D

Testing set — оценка финальной модели

Балансировка данных

Перекос данных

- 90 % данных — класс A, 10 % данных — класс B
- Модель всегда отвечает A — accuracy 90 %

Как бороться? Часть методов

- Oversampling and undersampling
- Синтетические данные
- Другие метрики *AUC, F1-score*
- Другие способы machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset

Масштабирование и нормализация

Масштабирование

- Standard $x' = \frac{x - \bar{x}}{\bar{o}}$
- Min-Max $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$

Нормализация

L1, L2, ...

Трансформация данных

Feature
SELECTION

Feature
ENGINEERING

Заключение

1

Подготовка данных ≈ тренировка моделей

2

Поиск аномалий и способы их решений — только training set

3

Полученные решения применяются к данным в обучающую модель

- Test set
- Новые данные

Практика

АНАЛИЗ БАНКОВСКИХ ТРАНЗАКЦИЙ

Часть 3

ОБРАБОТКА КАТЕГОРИАЛЬНЫХ ПЕРЕМЕННЫХ

Виды трансформации категориальных переменных

1

One hot encoding

2

Counts
«Каунты»

3

Weights of evidence
WOE

One Hot Encoding

Company Name	Categorical value	Price
VW	1	20.000
Acura	2	10.011
Honda	3	50.000
Honda	3	10.000

VW	Acura	Honda	Price
1	0	0	20.000
0	1	0	10.011
0	0	1	50.000
0	0	1	10.000

Counts *Likelihood encoding*

- Для каждой категории считаем среднюю долю целевого события
- С математической точки зрения — условная вероятность целевого события при известной категории соответствующей фичи
- Получается в какой-то степени смещенный результат
- В идеале делать KFold, исключая возможность переобучения

Weights of Evidence *WOE*

$$W\ eighto\int Evidence = \ln\left(\frac{DistributionGood_i}{DistributionBad_i}\right)$$

$$IV = \sum(DistributionGood_i - DistributionBad_i) \times ln\left(\frac{DistributionGood_i}{DistributionBad_i}\right)$$

$$IV = \sum(DistributionGoodi - DistributionBad_i) \times WOE_i$$

DistributionGood (p_good) — отношение количества хороших в категории к числу всех хороших

DistributionBad (p_bad) — отношение количества плохих в категории к числу всех плохих

Information value

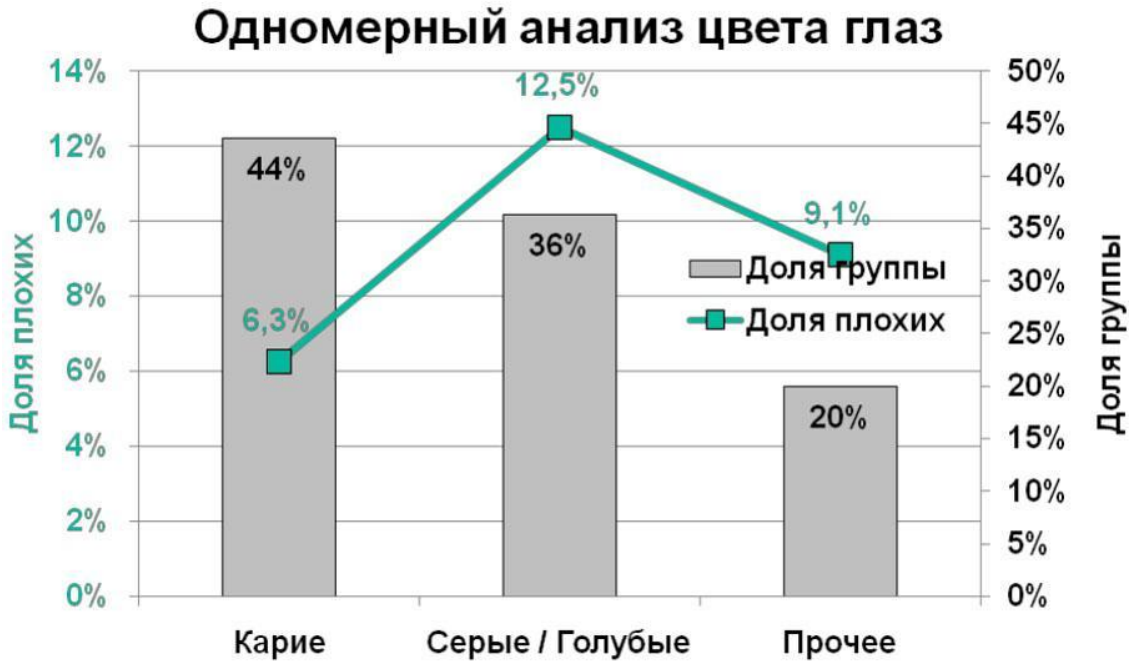
$$W\ eighto\int Evidence = \ln\left(\frac{DistributionGood_i}{DistributionBad_i}\right)$$

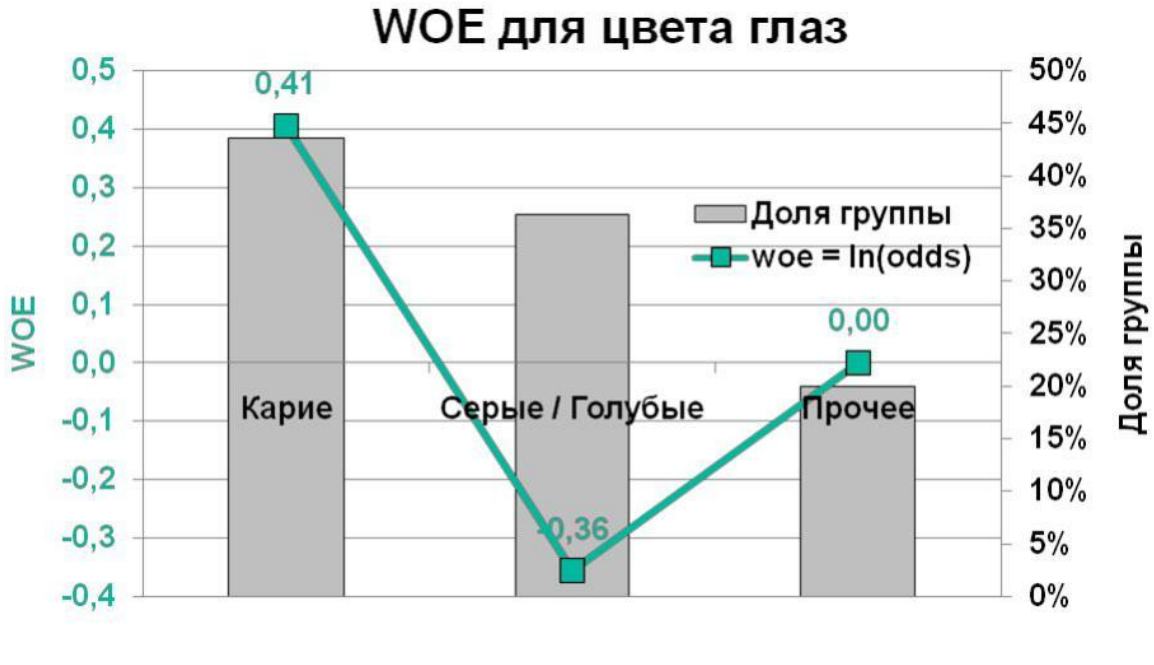
$$IV = \sum(DistributionGood_i - DistributionBad_i) \times ln\left(\frac{DistributionGood_i}{DistributionBad_i}\right)$$

$$IV = \sum(DistributionGoodi - DistributionBad_i) \times WOE_i$$

- Мера прогностической силы переменной
- Оценка качества группировки переменной
- Оценка информативности переменной

Weights of Evidence *WOE*





www.machinelearning.ru/wiki/images/b/b2/MSU-BI-Filipenkov-2010-10-01.pdf

Практика

РАБОТА С ДАННЫМИ ТАКСИ

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

ИТОГИ

1. Какие существуют **основные проблемы с данными**
2. Как осуществить **первичный анализ данных**
3. Как обрабатывать **пропущенные значения**
4. Как преобразовывать **категориальные признаки**

СПАСИБО ЗА ВНИМАНИЕ