



НЕТОЛОГИЯ
групп

Блок

FEATURE ENGINEERING



Константин Гусев

Старший аналитик-моделист
Vi.zone (кибер-безопасность)

Ex-аналитик McKinsey & Co.



ks.gusev@physics.msu.ru



ksgusev



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

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

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

4 [blog.kaggle.com/-
No Free Hunch](https://blog.kaggle.com/-/No-Free-Hunch)



Занятие 1

ПРОБЛЕМЫ КАЧЕСТВА И РАЗМЕРНОСТИ ДАННЫХ



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

1 0 1
1 0
 0 0

Недостаточное
количество
данных

1 0 1 0
1 0 1 0
1 0 1 0

Нерепрезентативные
данные

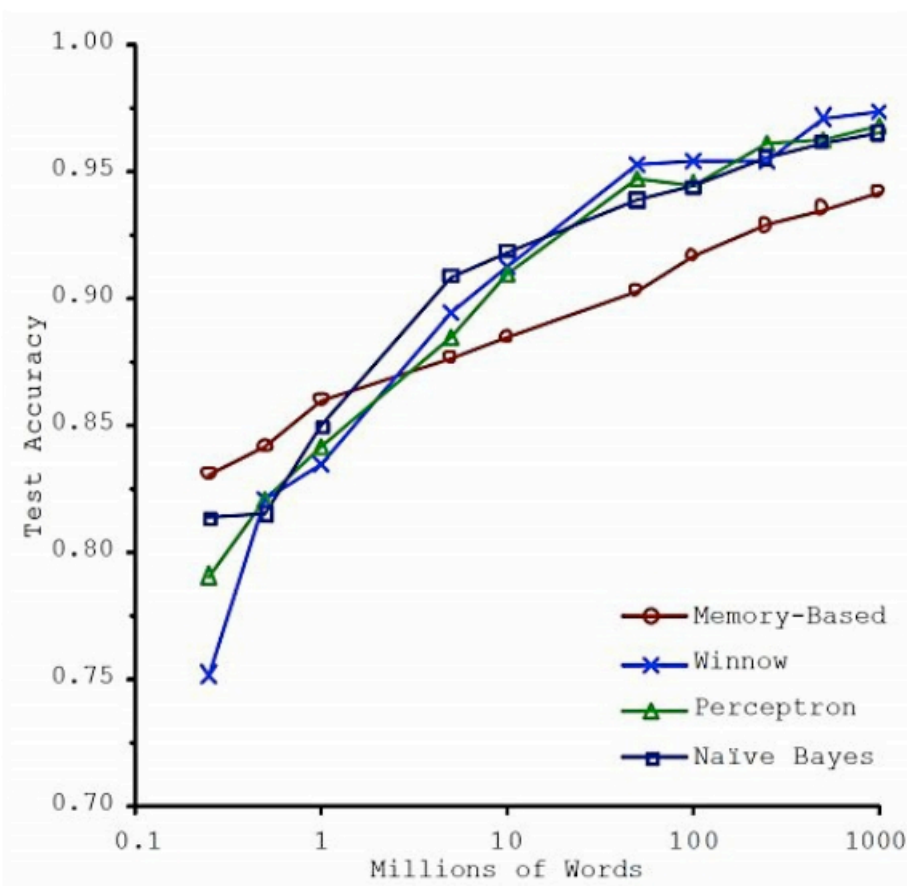
1 0 1 0
1 0 1 0
1 0 1 0

Данные
плохого качества
аномалии,
выбросы

1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010

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

Republican 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 opinion of one of the most quoted citizens

ДАННЫЕ И СОПУТСТВУЮЩИЕ ПРОБЛЕМЫ

1 0 1
1 0
 0 0

Недостаточное
количество
данных

1 0 1 0
1 0 1 0
1 0 1 0

Нерепрезентативные
данные

1 0 1 0
1 0 1 0
1 0 1 0

Данные
плохого качества
аномалии,
выбросы

1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010

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 DATASETА

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

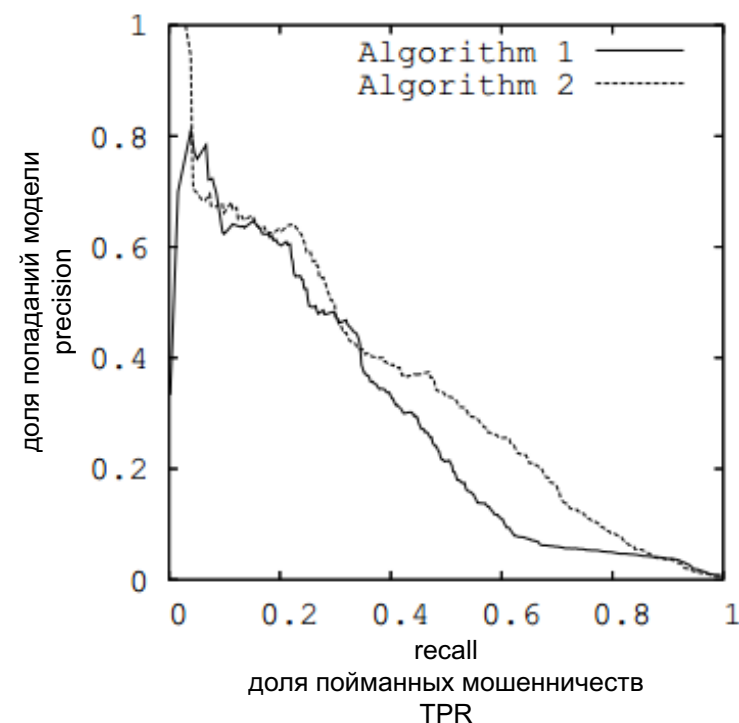
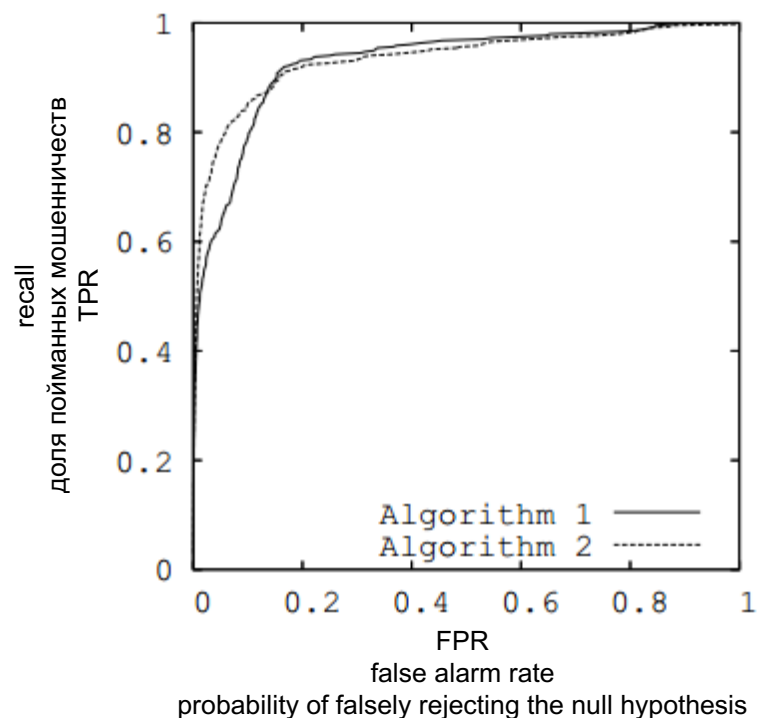
Перекося данных

- 90 % данных — класс А,
10 % данных — класс В
- Модель всегда отвечает
А — 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](https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset)

Построение ROC и ROC-PR прямых



Процедура построения ROC:

1. Сначала все объекты сортируются по оценке классификатора (от большего)
2. Начинаем построение из точки (0,0)
3. Начинаем идти от большей оценки к меньшей:
 - Если текущий объект имеет класс «1», то у алгоритма увеличивается TPR, ROC-кривая сдвигается вверх на $1/I_+$ (I_+ — число объектов положительного класса)
 - Если у текущего объекта класс «0», то алгоритм допускает на одну ошибку больше, чем предыдущий, ROC-кривая сдвигается вправо на $1/I_-$ (I_- — число объектов отрицательного класса)

Процедура построения ROC-PR:

1. Сначала все объекты сортируются по оценке классификатора (от большего)
2. Рассчитываются кумулятивные precision и recall (от большего)
3. Полученные точки наносятся на график

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

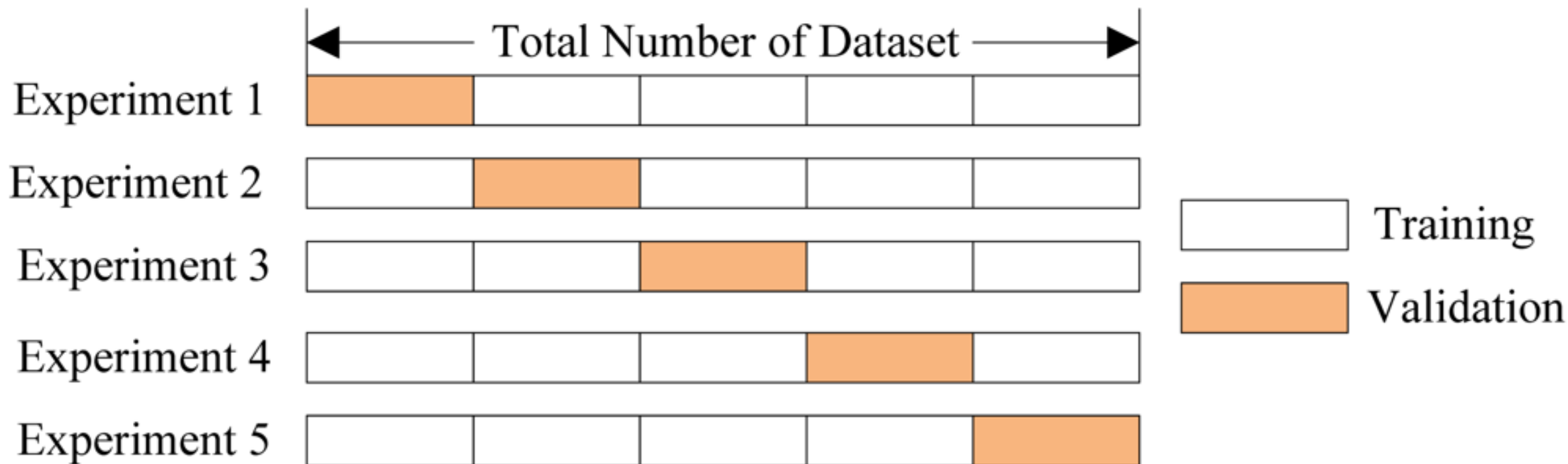
One Hot Encoding

Company Name	Categorical value	Price		VW	Acura	Honda	Price
VW	1	20.000		1	0	0	20.000
Acura	2	10.011		0	1	0	10.011
Honda	3	50.000		0	0	1	50.000
Honda	3	10.000		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

K-fold кросс-валидация



Weights of Evidence WOE

$$Weight of Evidence = \ln\left(\frac{DistributionGood_i}{DistributionBad_i}\right)$$

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

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

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

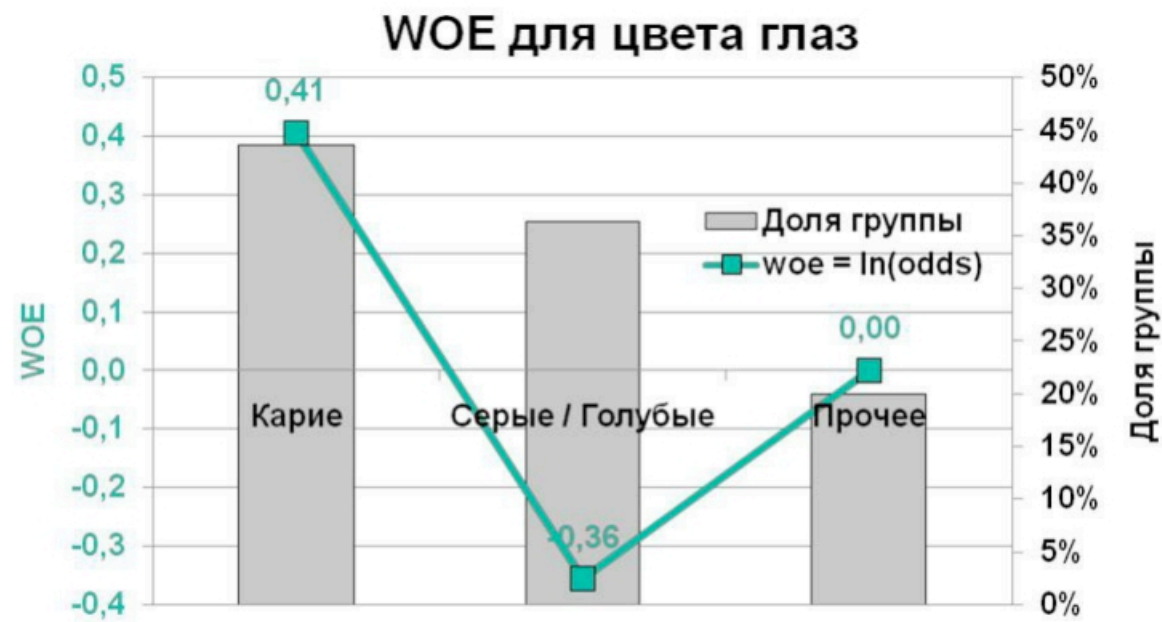
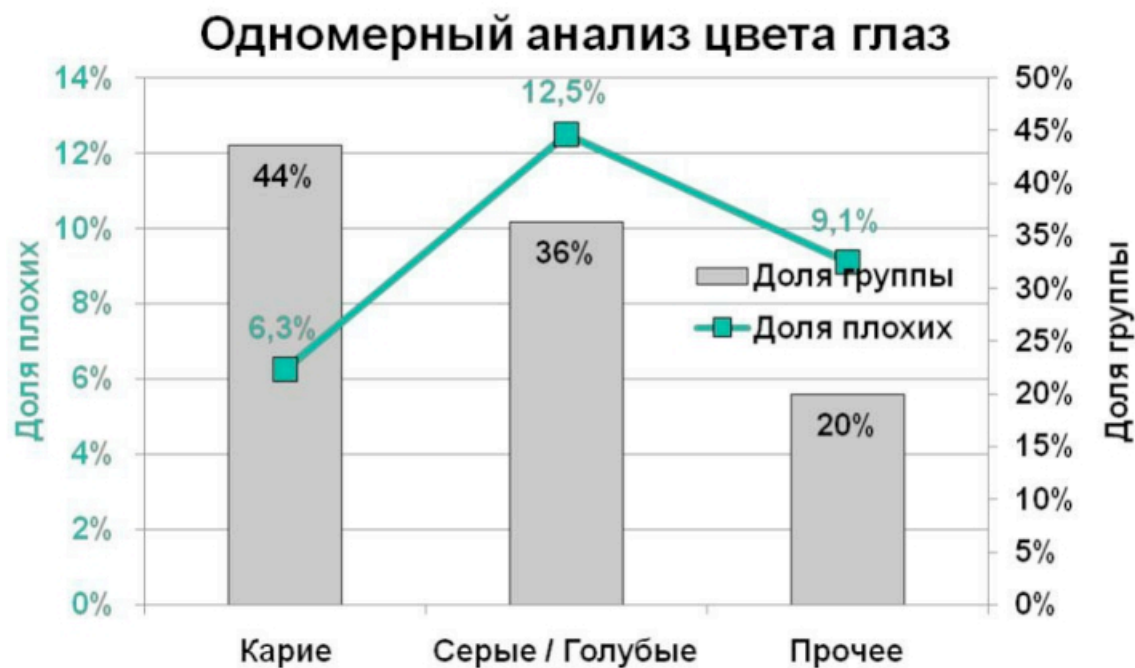
$$Weight of Evidence = \ln\left(\frac{DistributionGood_i}{DistributionBad_i}\right)$$

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

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

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

Weights of Evidence WOE





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

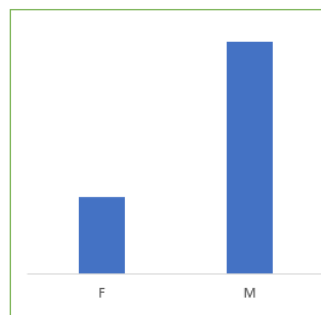
Как
преобразовывать
**категориальные
признаки**



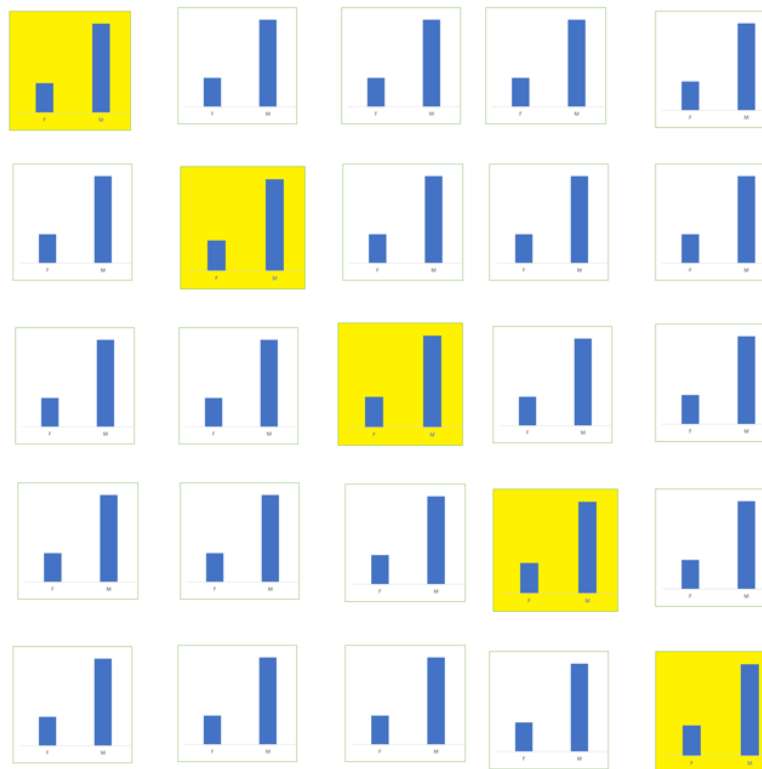
СПАСИБО ЗА ВНИМАНИЕ

Stratified K-fold **кросс-валидация**

Stratified K-Fold
Cross Validation
(K=5)



Class Distributions



Fold 1

Fold 2

Fold 3

Fold 4

Fold 5