



НЕТОЛОГИЯ  
групп

Блок

# FEATURE ENGINEERING



# Константин Гусев

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

Ex-аналитик McKinsey & Co.



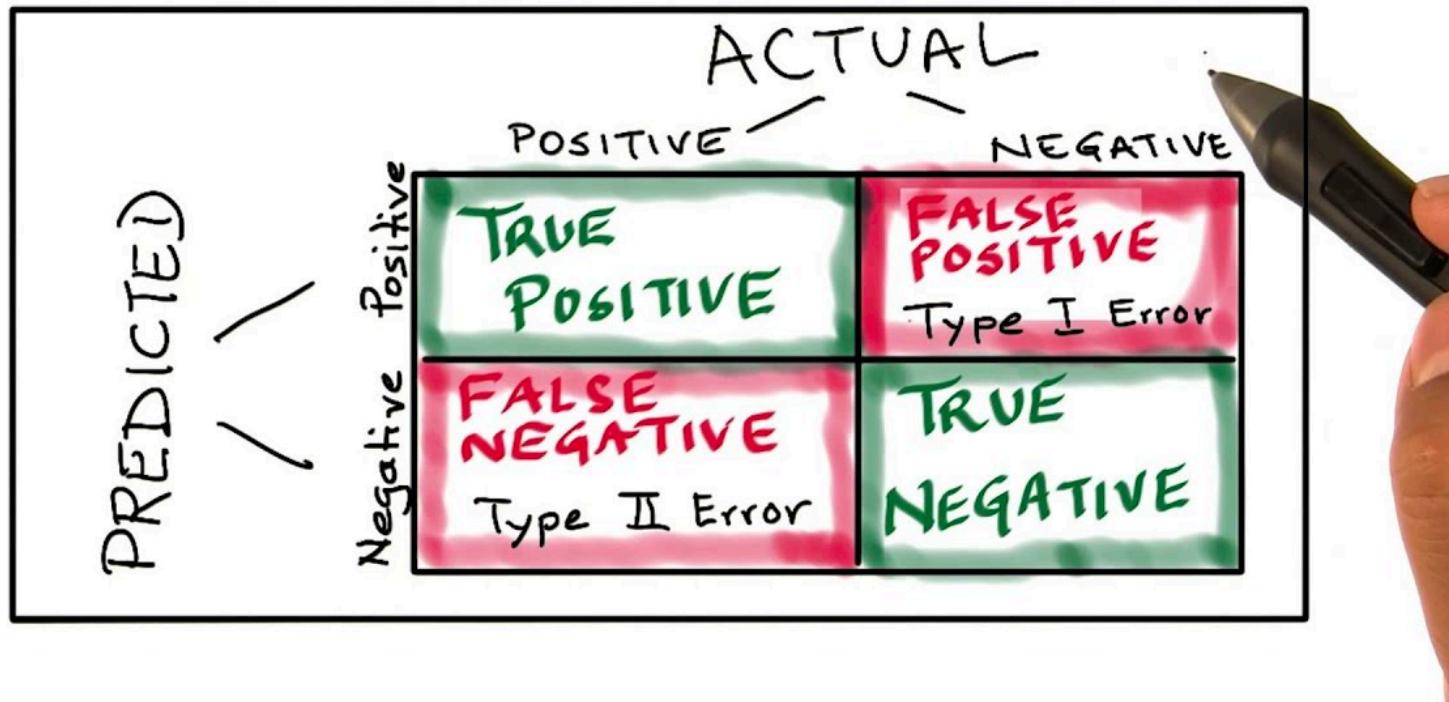
[ks.gusev@physics.msu.ru](mailto:ks.gusev@physics.msu.ru)



ksgusev

## ВОПРОСЫ С ПРОШЛОГО СЕМИНАРА

# Матрица ошибок



- TP – попадание
- TN – корректный отказ
- FP – ложная тревога (Ошибка 1 рода)
- FN – пропуск (Ошибка 2 рода)

## Метрики

- Полнота (recall) – Доля пойманых мошенничеств

$$= \text{TP}/(\text{TP}+\text{FN}) = \text{Попаданий}/(\text{Попаданий} + \text{Пропусков})$$

- Точность (precision) – Доля попаданий модели

$$= \text{TP}/(\text{TP}+\text{FP}) = \text{Попаданий}/(\text{Попаданий} + \text{Ложных тревог})$$

$$F_{\beta} = \frac{(1 + \beta^2) \cdot (\text{precision} \cdot \text{recall})}{(\beta^2 \cdot \text{precision} + \text{recall})}$$

- TP – попадание
- TN – корректный отказ
- FP – ложная тревога (Ошибка 1 рода)
- FN – пропуск (Ошибка 2 рода)

## ВОПРОСЫ С ПРОШЛОГО СЕМИНАРА

---

### F beta

$$\begin{aligned} F_1 &= \frac{1}{\frac{1}{2} \frac{1}{\text{precision}} + \frac{1}{2} \frac{1}{\text{recall}}} \\ &= 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \end{aligned}$$

$$\begin{aligned} F_\beta &= \frac{1}{\frac{1}{\beta+1} \frac{1}{\text{precision}} + \frac{\beta}{\beta+1} \frac{1}{\text{recall}}} \\ &= (1 + \beta) \frac{\text{precision} \cdot \text{recall}}{\beta \cdot \text{precision} + \text{recall}} \end{aligned}$$

## ВОПРОСЫ С ПРОШЛОГО СЕМИНАРА

# Нам необходимо выбрать 100 релевантных документов из 1 миллиона документов

- Алгоритм 1 возвращает 100 документов, 90 из которых релевантны. Таким образом,

$$TPR = \frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0.9$$

$$FPR = \frac{FP}{FP + TN} = \frac{10}{10 + 999890} = 0.00001$$

- Алгоритм 2 возвращает 2000 документов, 90 из которых релевантны. Таким образом,

$$TPR = \frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0.9$$

$$FPR = \frac{FP}{FP + TN} = \frac{1910}{1910 + 997990} = 0.00191$$

- Алгоритм 1

$$precision = \frac{TP}{TP + FP} = 90 / (90 + 10) = 0.9$$

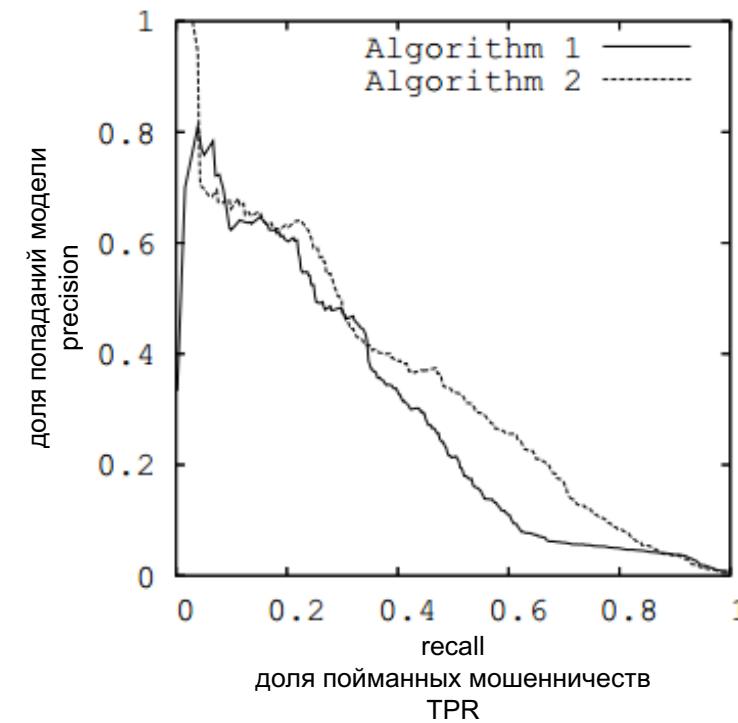
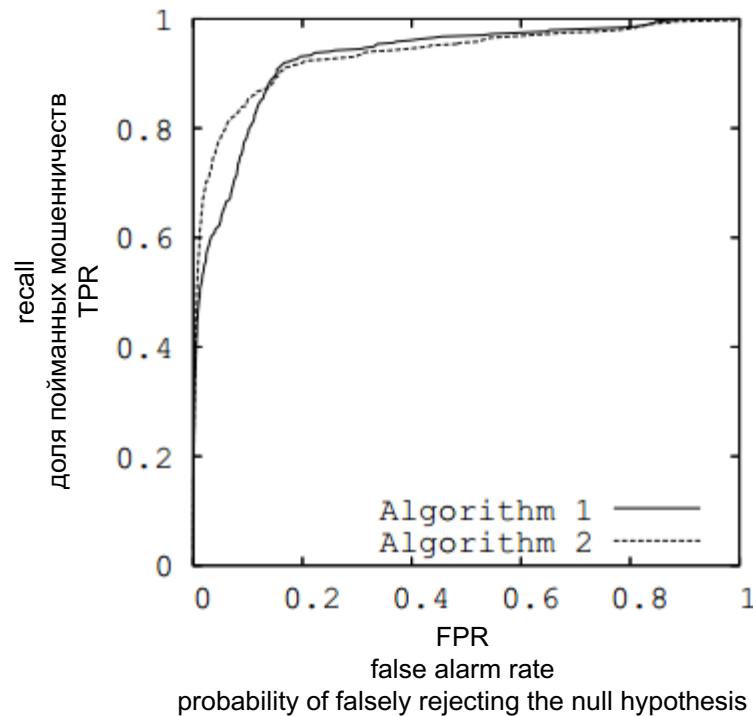
$$recall = \frac{TP}{TP + FN} = 90 / (90 + 10) = 0.9$$

- Алгоритм 2

$$precision = \frac{TP}{TP + FP} = \frac{90}{90 + 1910} = 0.045$$

$$recall = \frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0.9$$

# Построение ROC и ROC-PR прямых



Процедура построения ROC:

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

Процедура построения ROC-PR:

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

# ЦЕЛИ ЗАНЯТИЯ

## ЦЕЛИ ЗАНЯТИЯ

---

# В КОНЦЕ ЗАНЯТИЯ ВЫ СМОЖЕТЕ

1

Осуществлять  
поиск  
подмножества  
признаков

2

Использовать  
дilemму  
Смещения-  
Дисперсии

3

Осознанно  
применять  
линейную  
регрессию

4

Правильно  
использовать  
P-value

5

Использовать  
sklearn  
для Feature  
selection

---

ЧТО БУДЕМ ОБСУЖДАТЬ

## ПЛАН ЗАНЯТИЯ

---

1

Обзор домашнего  
задания

2

Первичный анализ  
данных

3

Оценка значимости  
переменных

4

Сокращение  
размерности  
пространства данных

Часть 1-2

Обзор домашнего задания  
Первичный анализ данных

---

# Практика АНАЛИЗ БАНКОВСКИХ ТРАНЗАКЦИЙ

---

Часть 3

# Осознанное применение линейной регрессии

### Датасет

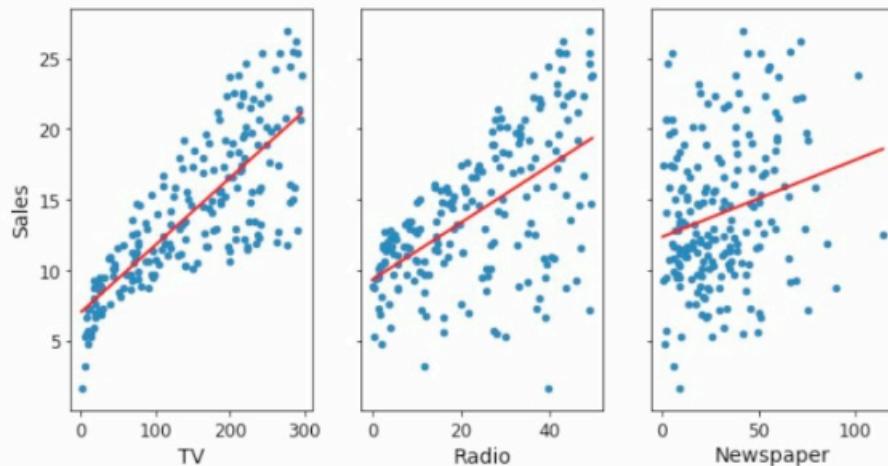
- Продажи продукта ~ рекламные бюджеты на разные медиа
- Медиа: ТВ, радио и газеты
- URL: <http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv>

## ОЦЕНКА ЗНАЧИМОСТИ ПЕРЕМЕННЫХ

---

### Цель

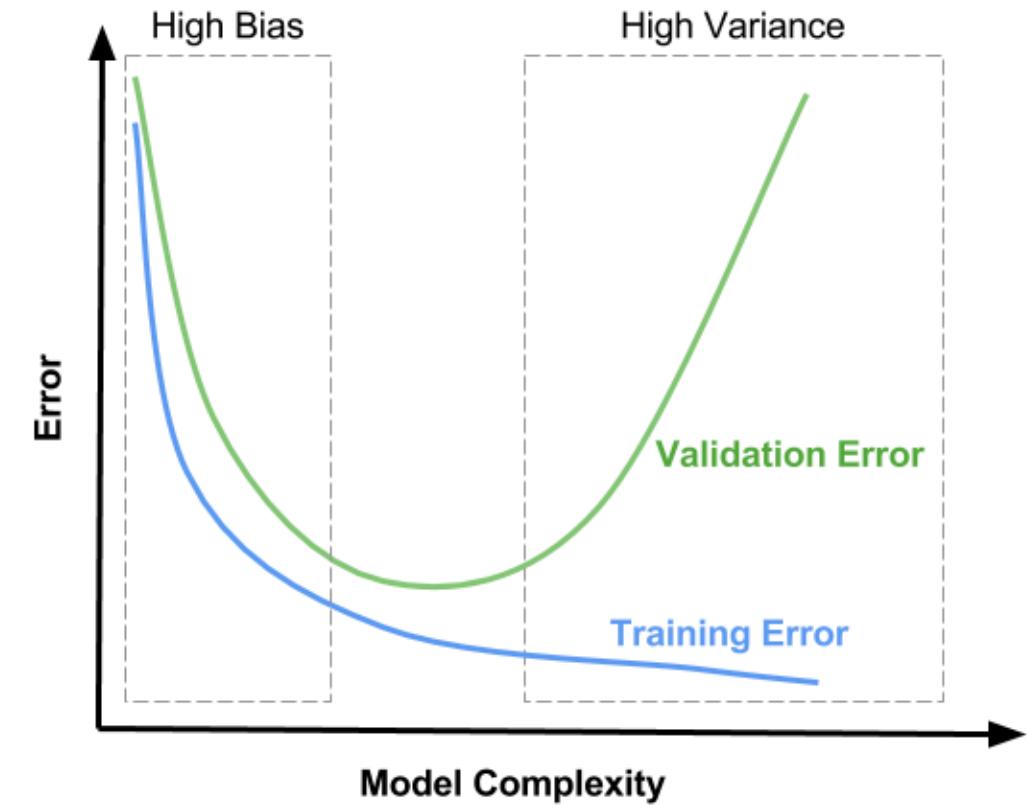
Создать маркетинговый план на следующий год на основе данных из датасета, так, чтобы продажи продукта были высокими.



## ОЦЕНКА ЗНАЧИМОСТИ ПЕРЕМЕННЫХ

# Bias-Variance Tradeoff (Дilemma Смещения-Дисперсии)

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"><li>• High training error</li><li>• Training error close to test error</li><li>• High bias</li></ul>	<ul style="list-style-type: none"><li>• Training error slightly lower than test error</li></ul>	<ul style="list-style-type: none"><li>• Very low training error</li><li>• Training error much lower than test error</li><li>• High variance</li></ul>
Regression illustration			
Classification illustration			



## ОЦЕНКА ЗНАЧИМОСТИ ПЕРЕМЕННЫХ

---

# На какие вопросы пытаемся ответить?

- Есть ли связь между рекламным бюджетом и продажами?
- Насколько сильна связь между бюджетом и продажами?  
Можем ли мы предсказывать продажи на основе бюджета?
- Какие медиа способствуют продажам?
- Насколько точно мы можем предсказывать будущие продажи?
- Линейная ли зависимость между бюджетом и продажами?
- Есть ли эффект взаимодействия (synergy/interaction effect)  
между медийными бюджетами?

# Линейная регрессия

$$sales = \beta_0 + \beta_1 * TV + \beta_2 * Radio + \beta_3 * Newspaper$$

- Предположим, что медиийные бюджеты не зависят друг от друга
- Определим  $\beta_0, \beta_1, \beta_2, \beta_3$

## Линейная регрессия - Проблемы

1. Признаки должны быть независимыми
2. Разный масштаб признаков
3. Функция зависит от параметров линейно
4. Ошибки не зависимы и имеют нормальное распределение

# Проверка гипотезы

- Нулевая и альтернативная гипотезы
  - $H_0$ : между  $x_i$  и  $y$  нет зависимости
  - $H_A$ : между  $x_i$  и  $y$  есть зависимость
- Для проверки гипотезы используется t-test

# T-Statistics & P-value

$$t = \frac{\hat{\beta}_i - 0}{SE(\hat{\beta}_i)}$$

- Если между  $x_i$  и  $y$  нет зависимости, то  $t$  соответствует  $t$ -распределению с  $n-2$  степенями свободы
- p-value - вероятность того, что при известном распределении наблюдаемое значение  $\geq |t|$  (при условии, что  $\beta_i = 0$ )
- Если p-value достаточно маленький ( $< 1\%$ ), то мы можем отклонить  $H_0$

## ОЦЕНКА ЗНАЧИМОСТИ ПЕРЕМЕННЫХ

---

# О чём нам говорит p-value?

Существует ли взаимосвязь между пристрастием к кровавым компьютерным играм и агрессивностью в реальной жизни? Для этого были случайным образом сформированы две группы школьников по 100 человек в каждой (1 группа – фанаты стрелялок, вторая группа – не играющие в компьютерные игры).

В качестве показателя агрессивности выступает число драк со сверстниками.

В нашем воображаемом исследовании оказалось, что группа школьников-игроманов действительно заметно чаще конфликтует с товарищами.

P-value – это вероятность получить такие или более выраженные различия при условии, что в генеральной совокупности никаких различий на самом деле нет. Пусть P-value = 0.04.

1. Компьютерные игры – причина агрессивного поведения с вероятностью 96%.
2. Вероятность того, что агрессивность и компьютерные игры не связаны, равна 0.04.
3. Если бы мы получили р-уровень значимости больше, чем 0.05, это означало бы, что агрессивность и компьютерные игры никак не связаны между собой.
4. Вероятность случайно получить такие различия равняется 0.04.
5. Все утверждения неверны.

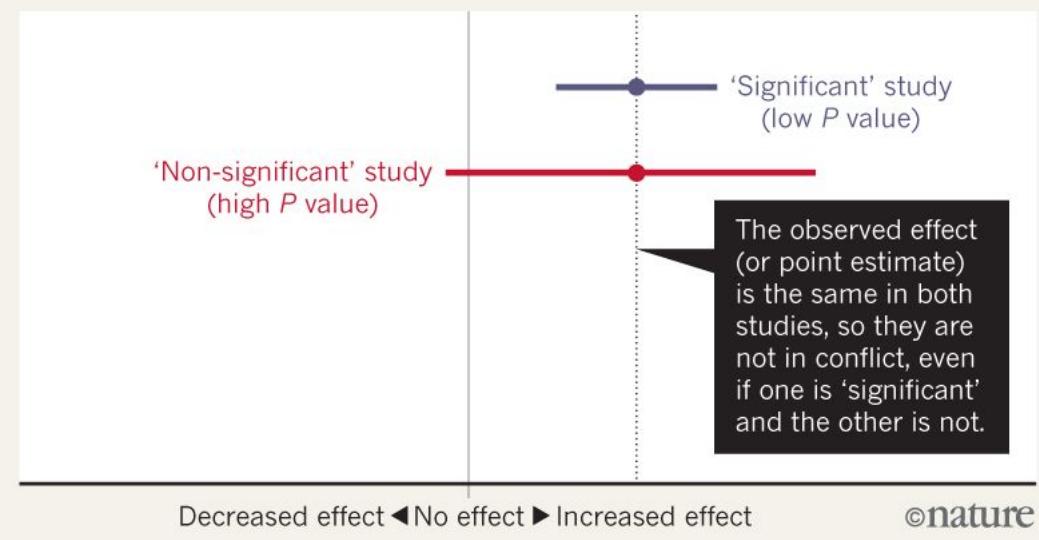
# Разберем все ответы по порядку:

1. Первое утверждение – пример ошибки корреляции: факт значимой взаимосвязи двух переменных ничего не говорит нам о причинах и следствиях. Может быть, это более агрессивные люди предпочитают проводить время за компьютерными играми, а вовсе не компьютерные игры делают людей агрессивнее.
2. Это уже более интересное утверждение. Все дело в том, что мы изначально принимаем за данное, что никаких различий на самом деле нет. И, держа это в уме как факт, рассчитываем значение p-value. Поэтому правильная интерпретация: «Если предположить, что агрессивность и компьютерные игры никак не связаны, то вероятность получить такие или еще более выраженные различия составила 0.04».
3. А что делать, если мы получили незначимые различия? Значит ли это, что никакой связи между исследуемыми переменными нет? Нет, это означает лишь то, что различия, может быть, и есть, но наши результаты не позволили их обнаружить.
4. Это напрямую связано с самим определением p-value. 0.04 – это вероятность получить такие или еще более экстремальные различия. Оценить вероятность получить **именно такие** различия, как в нашем эксперименте, в принципе невозможно!

# Проверка гипотез и статистическая значимость

### BEWARE FALSE CONCLUSIONS

Studies currently dubbed ‘statistically significant’ and ‘statistically non-significant’ need not be contradictory, and such designations might cause genuine effects to be dismissed.



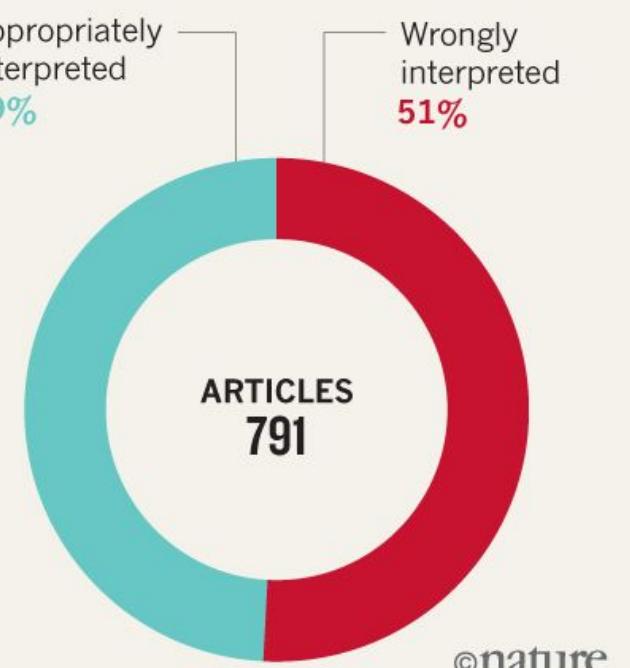
### WRONG INTERPRETATIONS

An analysis of 791 articles across 5 journals\* found that around half mistakenly assume non-significance means no effect.

\*Data taken from: P. Schatz et al. *Arch. Clin. Neuropsychol.* **20**, 1053–1059 (2005); F. Fidler et al. *Conserv. Biol.* **20**, 1539–1544 (2006); R. Hoekstra et al. *Psychon. Bull. Rev.* **13**, 1033–1037 (2006); F. Bernardi et al. *Eur. Sociol. Rev.* **33**, 1–15 (2017).

Appropriately interpreted  
49%

Wrongly interpreted  
51%



## ОЦЕНКА ЗНАЧИМОСТИ ПЕРЕМЕННЫХ

# Бюджеты и продажи

- 4 независимые гипотезы:

- $H_0: \beta_i = 0$

- $H_A: \beta_i \neq 0$

- Недостаток t-statistics: оценка важности каждого атрибута производится независимо от других

	coef	std err	t	P> t	[0.025	0.975]
Intercept	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

## ОЦЕНКА ЗНАЧИМОСТИ ПЕРЕМЕННЫХ

---

### R<sup>2</sup>

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}, \quad TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

- TSS - Total Sum of Squares
- R<sup>2</sup> - показывает, какой процент вариативности (variance) объяснен моделью
- R<sup>2</sup> ∈ [0, 1] - относительная величина, чем ближе к 1, тем лучше

Объясненная дисперсия  
Общая дисперсия

Проблема - с добавлением  
переменных R<sup>2</sup> растет

—

# Практика АНАЛИЗ РЕКЛАМНЫХ БЮДЖЕТОВ

## ВОПРОСЫ С ПРОШЛОГО СЕМИНАРА

В задачах где есть юзер\_айди, одна строка должна соответствовать одному юзер\_айди?

Primary key  
unique

Row_id	User_id	Value
1	12	10000
2	12	30000
3	13	300
4	14	250



Primary key  
unique

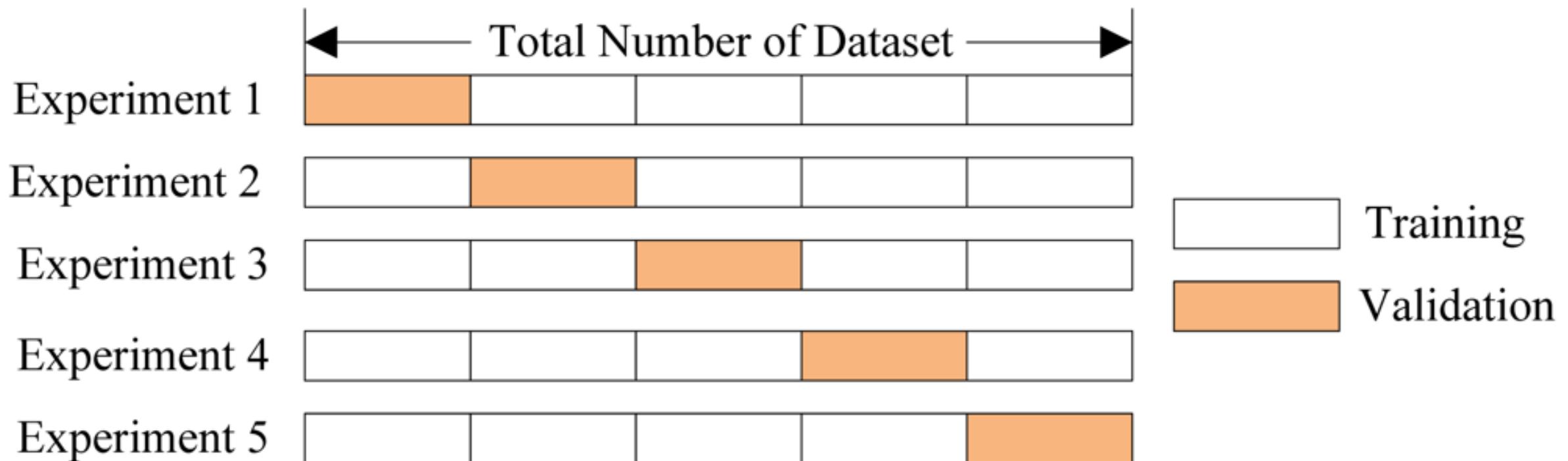
User_id	Value_sum	Value_max
12	40000	30000
13	300	300
14	250	250

Вариант хранения данных в базе – поток транзакций

Вариант хранения описания клиента,  
которое имеет период обновления

## ВАЛИДАЦИЯ

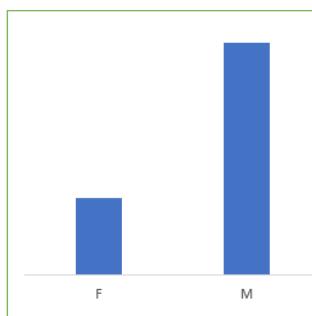
### K-fold кросс-валидация



## ВАЛИДАЦИЯ

# Stratified K-fold кросс-валидация

Stratified K-Fold  
Cross Validation  
(K=5)

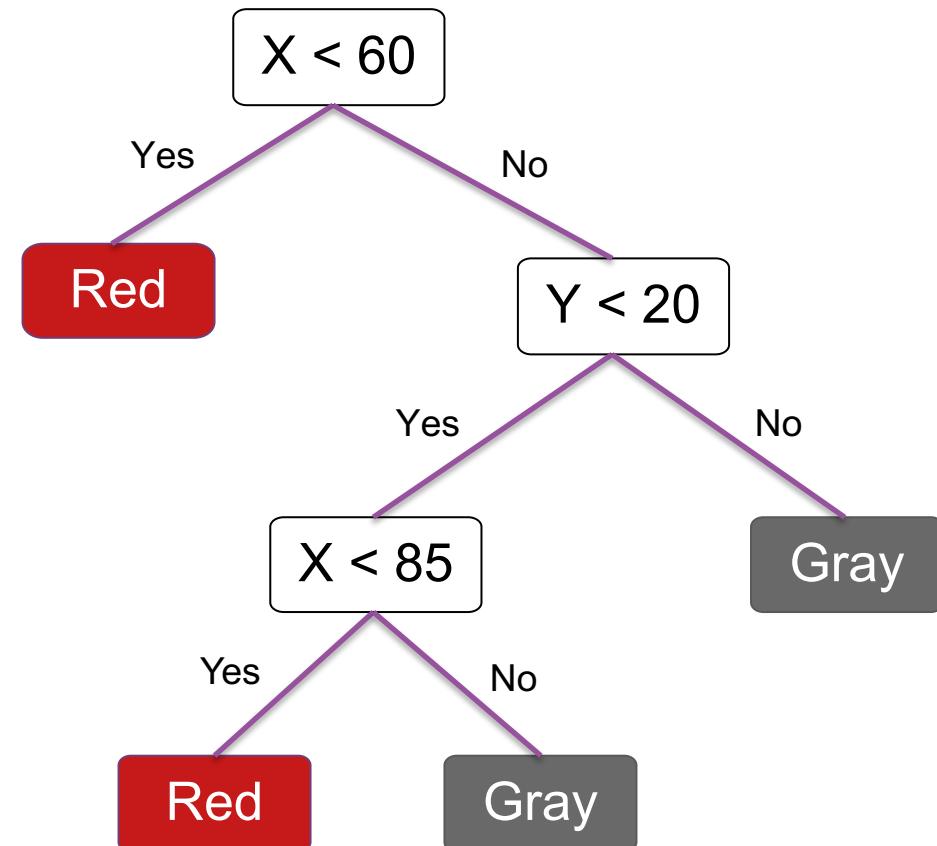
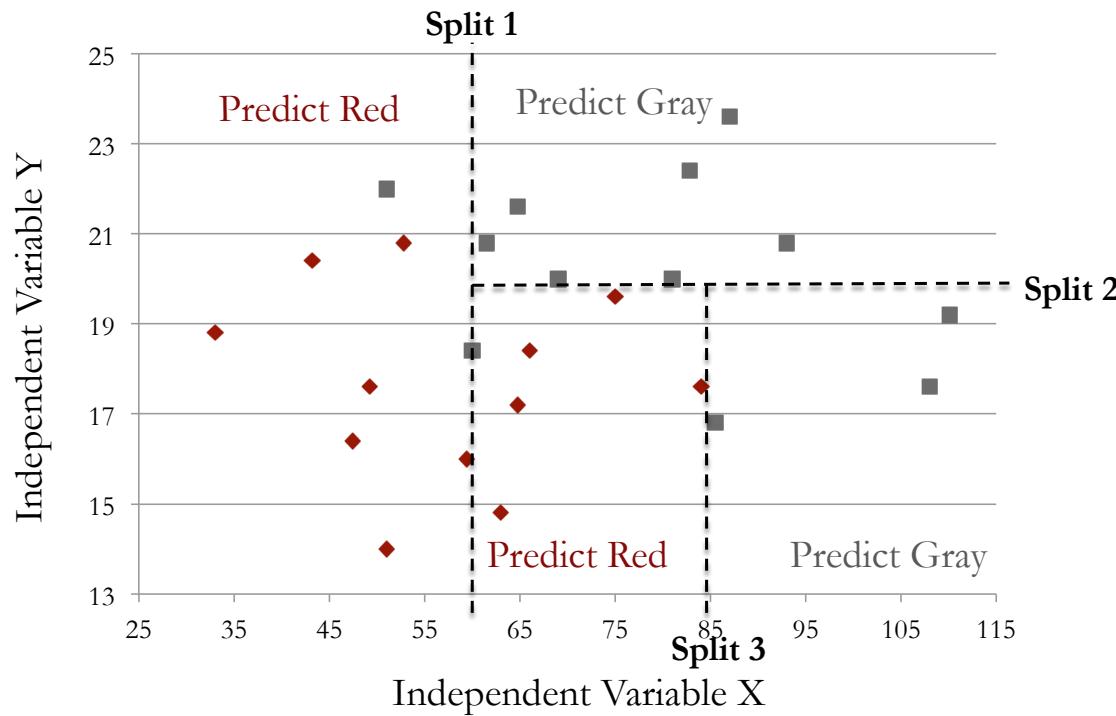


## ТИПЫ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

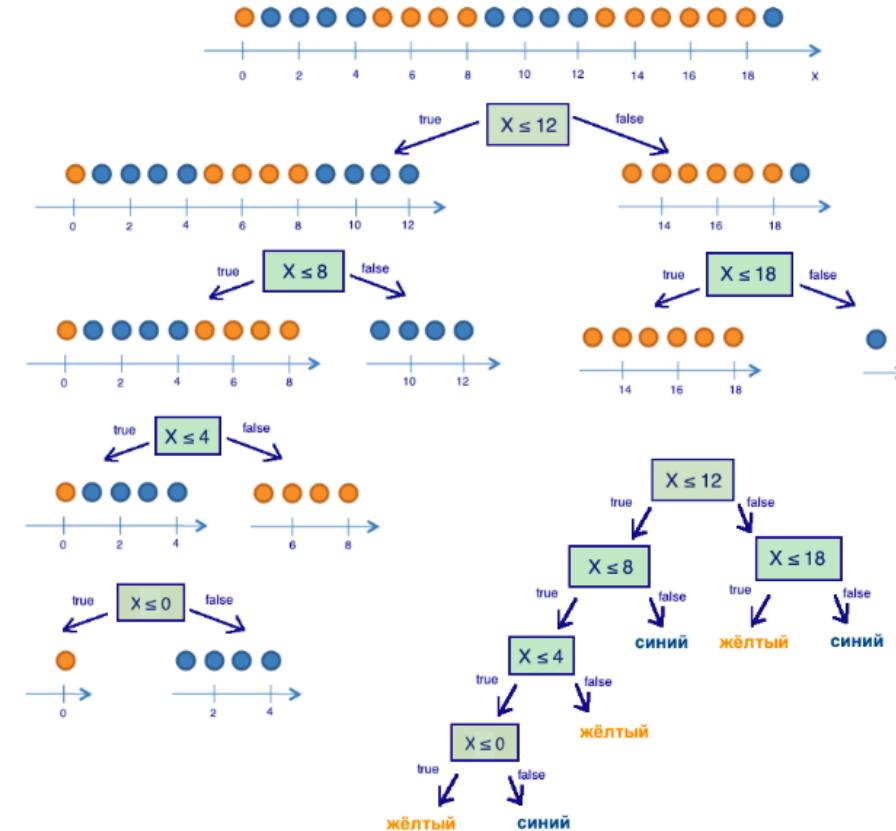
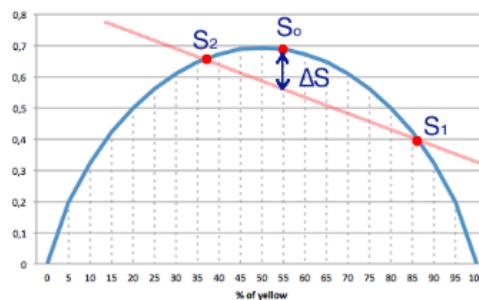
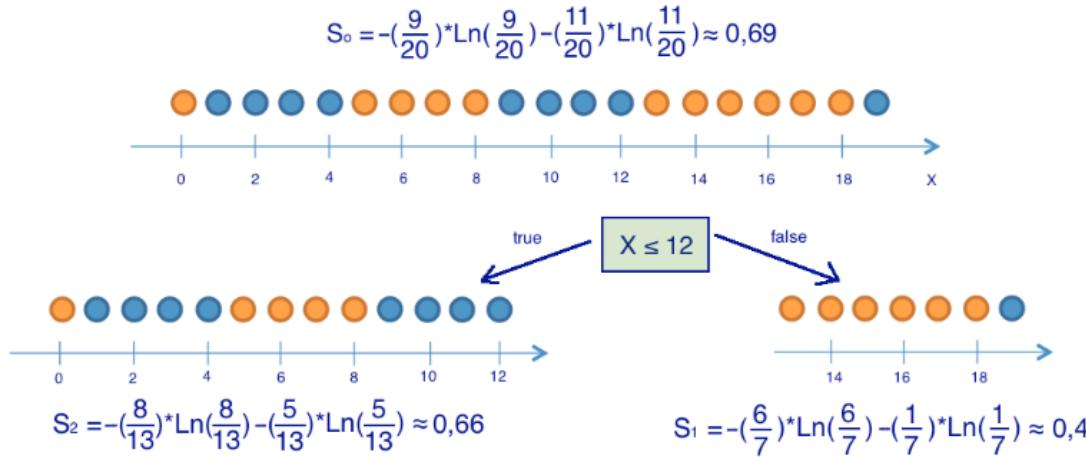
# Классическое обучение



# Решающее дерево (Decision tree)



# Решающее дерево (Decision tree)



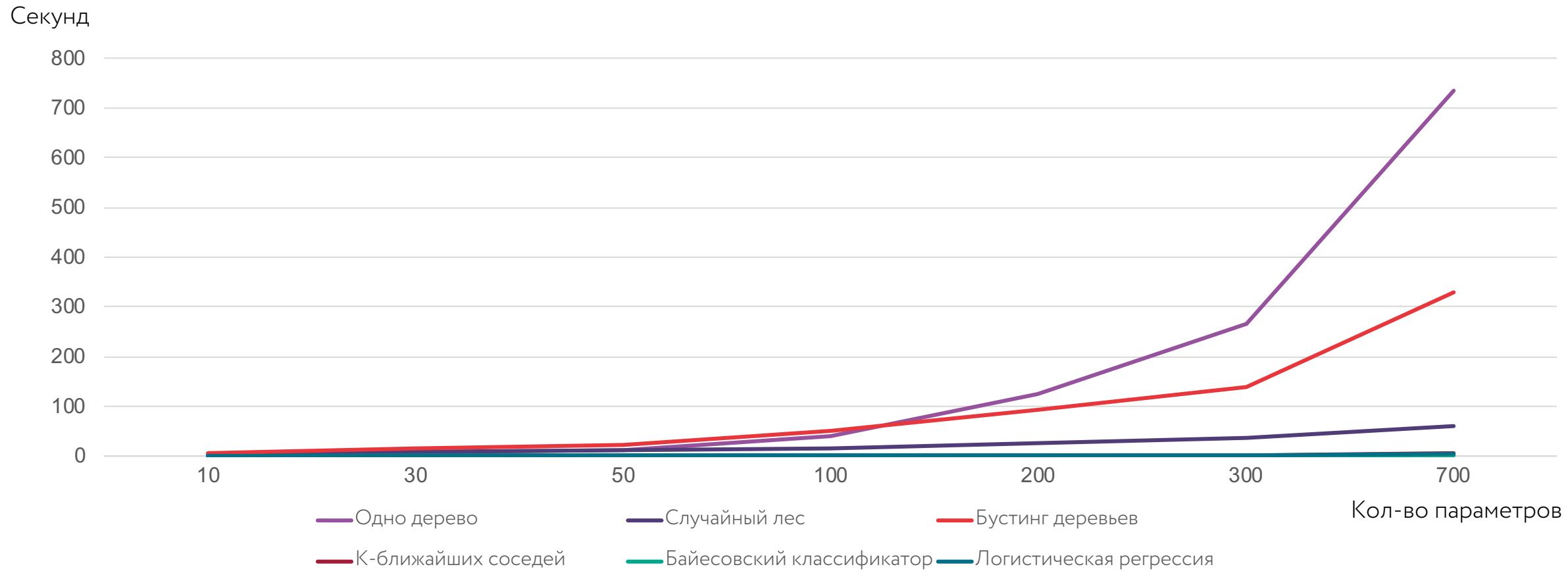
---

Часть 4

**Поиск подмножества признаков**

## СОКРАЩЕНИЕ РАЗМЕРНОСТИ ПРОСТРАНСТВА ДАННЫХ

# Время тренировки различных моделей на тестовой выборке



## Теоретические показатели времени обучения и предсказания для различных типов алгоритмов

Algorithm	Classification/Regression	Training	Prediction
Decision Tree	C+R	$O(n^2p)$	$O(p)$
Random Forest	C+R	$O(n^2pn_{trees})$	$O(pn_{trees})$
Random Forest	R Breiman implementation	$O(n^2pn_{trees})$	$O(pn_{trees})$
Random Forest	C Breiman implementation	$O(n^2\sqrt{p}n_{trees})$	$O(pn_{trees})$
Extremely Random Trees	C+R	$O(npn_{trees})$	$O(npn_{trees})$
Gradient Boosting ( $n_{trees}$ )	C+R	$O(npn_{trees})$	$O(pn_{trees})$
Linear Regression	R	$O(p^2n + p^3)$	$O(p)$
SVM (Kernel)	C+R	$O(n^2p + n^3)$	$O(n_{sv}p)$
k-Nearest Neighbours (naive)	C+R	—	$O(np)$
Nearest centroid	C	$O(np)$	$O(p)$
Neural Network	C+R	?	$O(pn_{l_1} + n_{l_1}n_{l_2} + \dots)$
Naive Bayes	C	$O(np)$	$O(p)$

Где  $n$  – количество строк,  $p$  – количество параметров,  $ntree$  – количество деревьев,  $nsv$  – количество опорных векторов

## СОКРАЩЕНИЕ РАЗМЕРНОСТИ ПРОСТРАНСТВА ДАННЫХ

---

# Причины по которым необходимо отбирать признаки перед тренировкой модели:

- Это позволяет алгоритму обучаться быстрее
- Это уменьшает сложность модели и облегчает интерпретацию
- Это повышает точность модели, если выбрано правильное подмножество признаков
- Это уменьшает переобучение

## Подходы к отбору признаков

1. Экспертный/теоретический
2. ‘Статистический’
  - Методы фильтрации
  - Оберточные методы
  - Встроенные методы

# Методы фильтрации

Основаны на статистических методах, как правило рассматривают каждый признак отдельно, проверяя его влияние на целевую переменную и ранжируя по нему

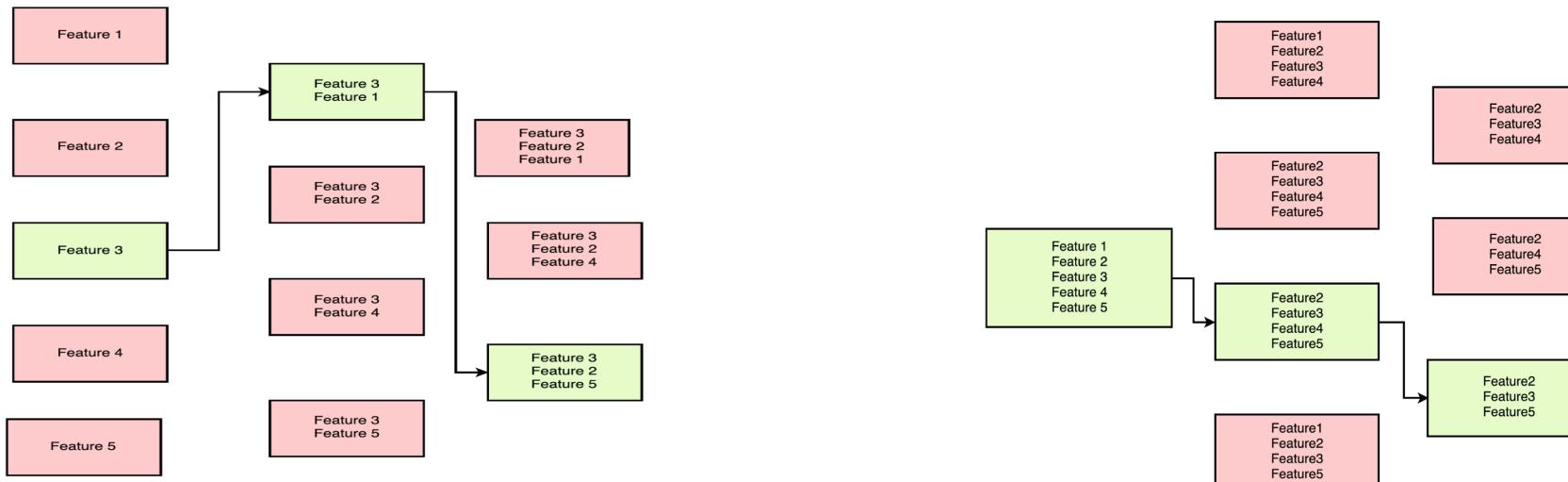
- Correlation – определение насколько переменные линейно зависимы друг от друга
- LDA (Linear Discriminant Analysis) – метод поиска комбинации параметров, которая лучше всегда разделяет целевой класс
- ANOVA – дисперсионный анализ (Analysis of Variance)
- Chi-Square (Хи-квадрат) – метод позволяет оценить статистическую значимость различий двух или нескольких относительных значений переменных
- F Test
- Mutual Information (MIC) – тест на основе информационного критерия
- Variance Threshold – отбрасывание значений, дисперсия которых меньше определенного значения
- mRmR - минимальная избыточность при максимальной релевантности

## СОКРАЩЕНИЕ РАЗМЕРНОСТИ ПРОСТРАНСТВА ДАННЫХ

# Оберточные методы (wrapper methods)

Генерируются подмножества признаков для обучения модели и выбирается подмножество дающее наиболее высокий результат

- **Метод прямого отбора (Forward Selection)** – к пустому множеству признаков мы добавляем на каждой итерации признак, который лучше всего увеличивает показатели модели. До тех пор пока, добавление не перестанет приносить результаты
- **Рекурсивное исключение признаков (Recursive Feature elimination)** - модель обучается на всех данных. Для каждого признака вычисляется его значимость. На каждом следующем шаге отбрасывается наиболее малозначимый признак. И операция повторяется до достижения необходимого числа признаков или пока исключение признаков не перестанет приносить результаты

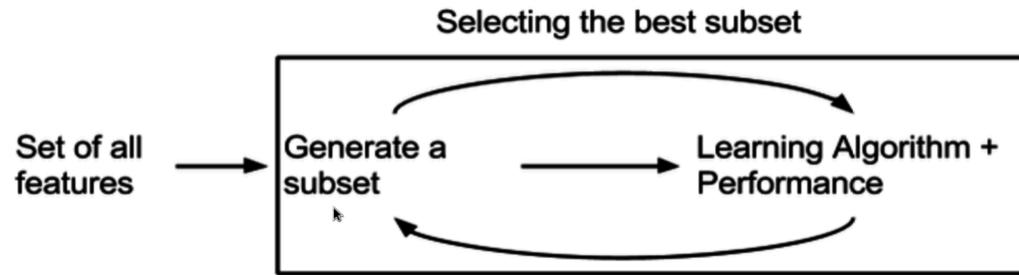


Источник: <https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/>

<https://machinelearningmastery.com/an-introduction-to-feature-selection/>

<https://towardsdatascience.com/why-how-and-when-to-apply-feature-selection-e9c69adfabf2>

## Встроенные методы (embedded methods)



- **Lasso регуляризация** - метод при котором накладывается ограничение в виде суммы абсолютных величин весов в функции ошибок.
- **Ridge регуляризация** - метод при котором накладывается ограничение в виде суммы квадратов весов в функции ошибок
- **Методы в моделях основанных на деревьях** – для построения узлов дерева на каждой итерации выбирается параметр, который лучше всего разделяет обучающую выборку по определенному критерию (Энтропии, Критерий Джини)

## СОКРАЩЕНИЕ РАЗМЕРНОСТИ ПРОСТРАНСТВА ДАННЫХ

### Регуляризация

Это мера расхождения между данными и моделью оценки

- Линейная регрессия

$$RSS = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j * x_{ij})^2$$

- Ridge

$$RSS = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j * x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2$$

- Lasso

$$RSS = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j * x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j|$$

## “Не классические” методы отбора:

- Генетические алгоритмы
- Перебор всех возможных вариаций
- Mixed selection – работает как метод прямого отбора, но на каждом шаге может сделать шаг назад

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

**1** Как оценивать  
значимость переменных

**2** Как устроена  
линейная регрессия

**3** Какие существуют типы  
регуляризации

**4** Как осуществить  
отбор признаков

**СПАСИБО ЗА ВНИМАНИЕ**