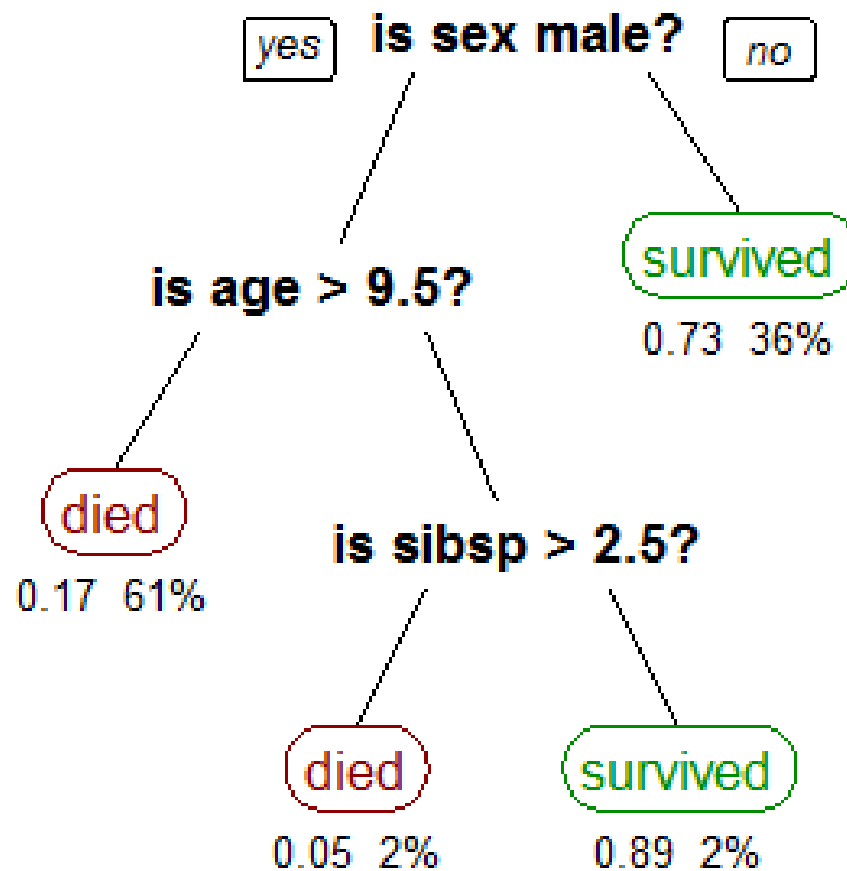


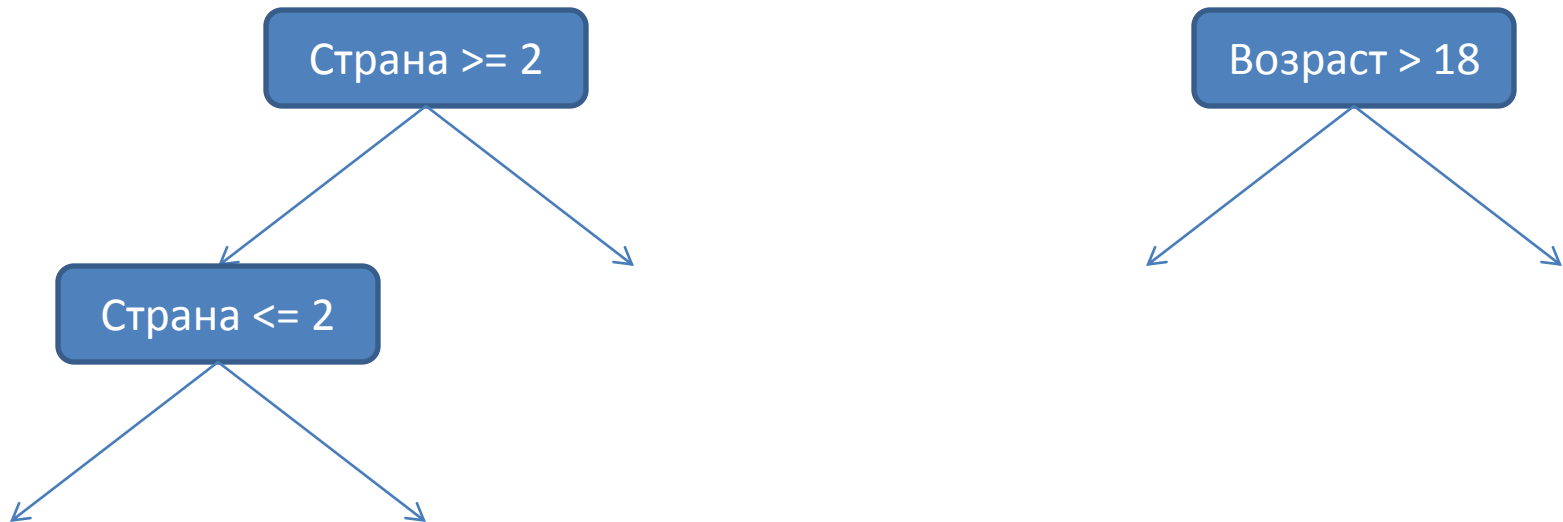
Data Mining in Action

Decision trees



Проблема №1

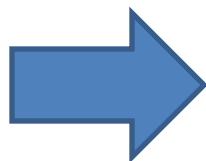
- Нужен линейный порядок на каждой из фичей



Решение

- OneHotEncoding

Страна
2
2
1
3



Страна=1	Страна=2	Страна=3
0	1	0
0	1	0
1	0	0
0	0	1

“Страна=2” ≥ 0.5

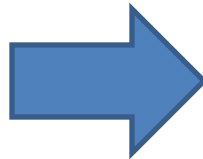


Все равно 1 сплит на 1 категорию!

Хорошее решение

- Посчитать статистику по другой колонке

Страна	Доход
2	2,000
2	20,000
1	10,000
3	12,000
2	2,000



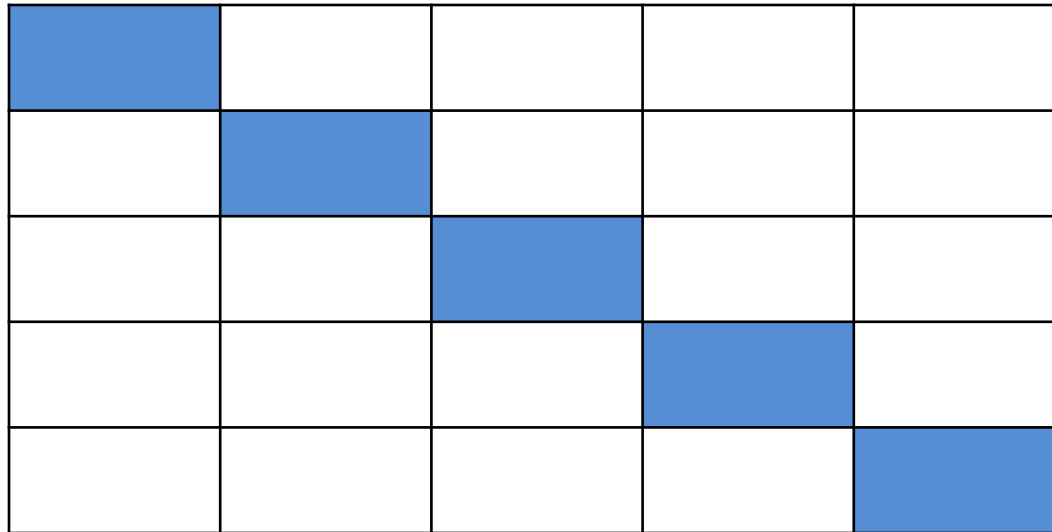
Страна	Доход	Средний доход по стране
2	2,000	8,000
2	20,000	8,000
1	10,000	10,000
3	12,000	12,000
2	2,000	8,000

Ср. ВВП > 8,000



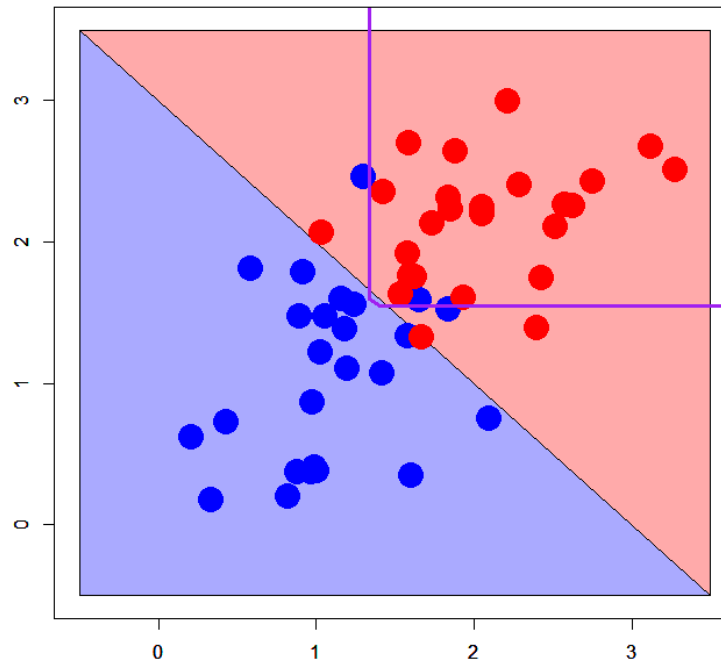
Опасное решение

- Брать статистику по целевому признаку
- Ведет к переобучению
- Выход – K-fold предсказания



Проблема №2

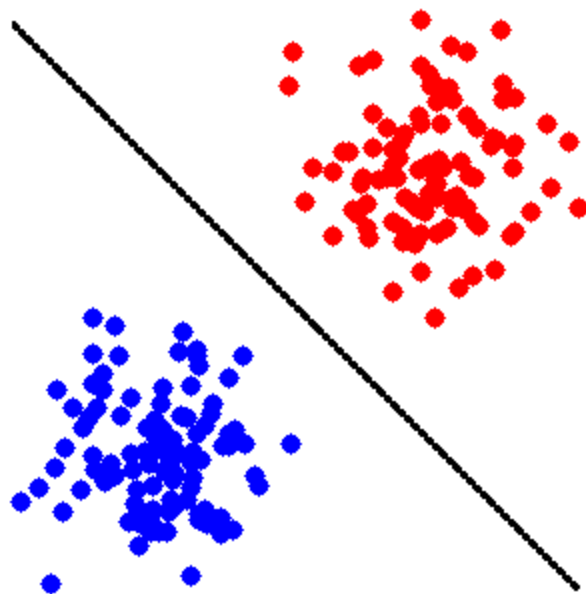
- Плохое восстановление линейных (арифметических) зависимостей



Решение

- Добавить комбинации признаков: сумма, разность, произведение, частное
- Бинарные признаки: конъюнкции, XOR
- Разность скоррелированных признаков, сумма антикоррелированных

Линейные методы



Проблема №1

- Неупорядоченные категориальные признаки

Страна
2
2
1
3

Решение

- OneHotEncoding

Страна	2	2	1	3
--------	---	---	---	---



Страна=1	Страна=2	Страна=3
0	1	0
0	1	0
1	0	0
0	0	1

$$a(x) = 0.5 * \text{«Страна=1»} + 0.3 * \text{«Страна=2»} - 0.4 * \text{«Страна=3»}$$

Проблема №2

- Чувствительность к большому разбросу значений

```
all['delta_imp_аport_var13_1y3'].value_counts()
```

0.000000e+00	147710
-1.000000e+00	3328
1.000000e+10	747
-3.333333e-01	3
-5.000000e-01	2
4.000000e+00	2
-1.000000e-01	2
-8.250875e-01	1
-6.000000e-01	1
1.000000e+00	1

Решение

- Если это значение имеет смысл NaN, то заменим его на ноль или среднее по колонке
- Добавим еще один бинарный признак isNaN

Проблема №3

- Неоднородность признакового пространства

Признак1	Признак2	Признак3	Признак4	Признак5
1	100000	1	0.345	54
2	-55500	0	-0.123	76
3	3000	0	0.864	12
4	150000	1	0.0023	34
5	500	1	1.1867	87

Решение 1

- По мотивам соревнования Springleaf: заменим все признаки на среднее по целевому

Признак1	Признак2	Признак3	Признак4	Признак5
0.1	0.3	0.35	0.04	0.1
0.2	0.1	0.65	0.03	0.2
0.3	0.4	0.65	0.12	0.9
0.4	0.5	0.35	0.05	0.76
0.5	0.6	0.35	0.43	0.2

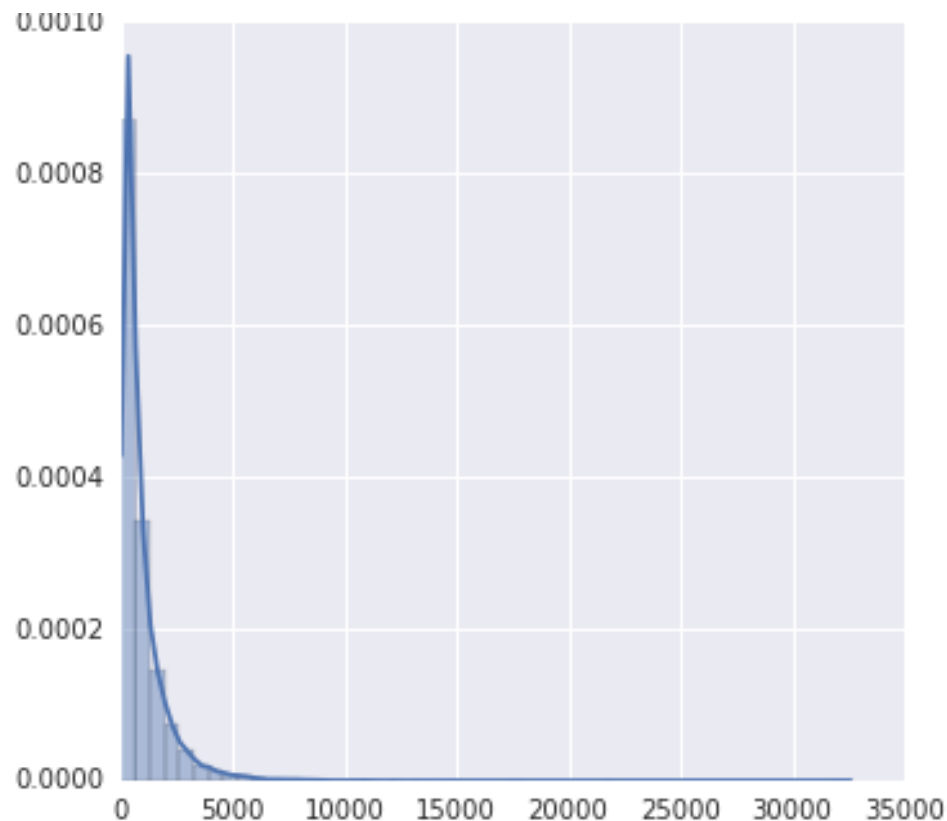
Решение 2

- Заменим все значения на их ранки!

Признак1	Признак2	Признак3	Признак4	Признак5
0	3	1	2	2
1	0	0	0	3
2	2	0	3	0
3	4	1	1	1
4	1	1	4	4

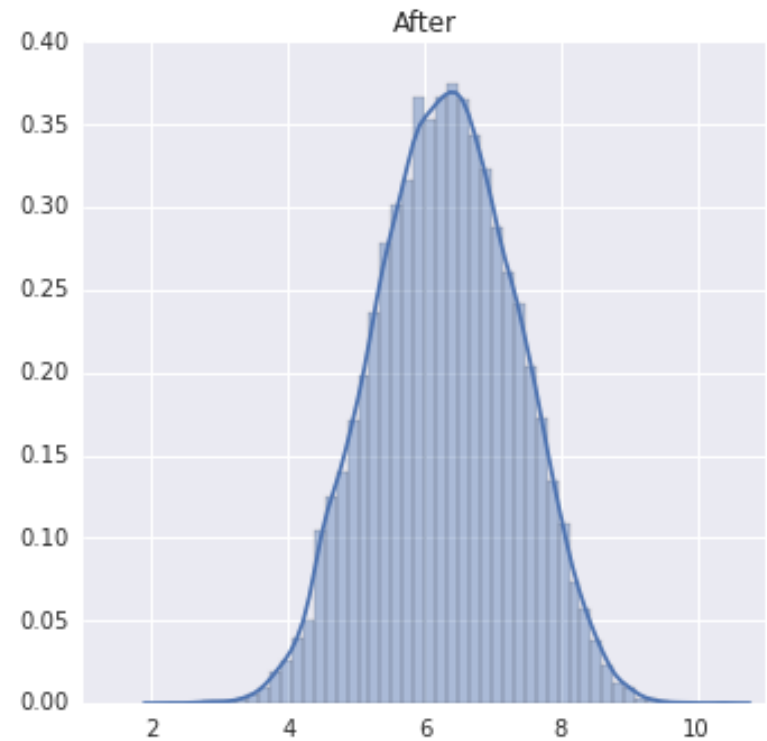
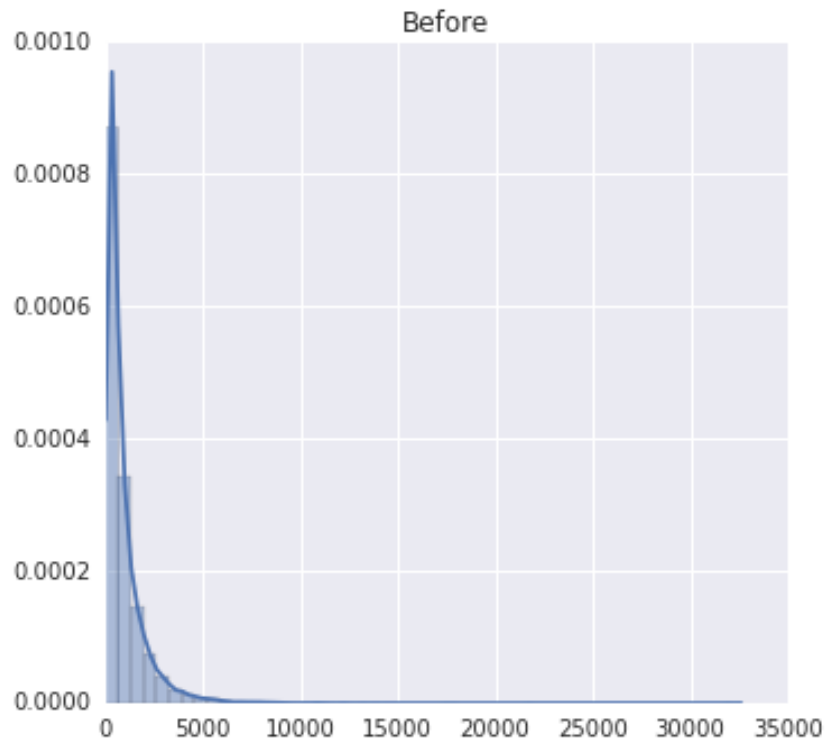
Проблема №4

- Распределения с длинными хвостами



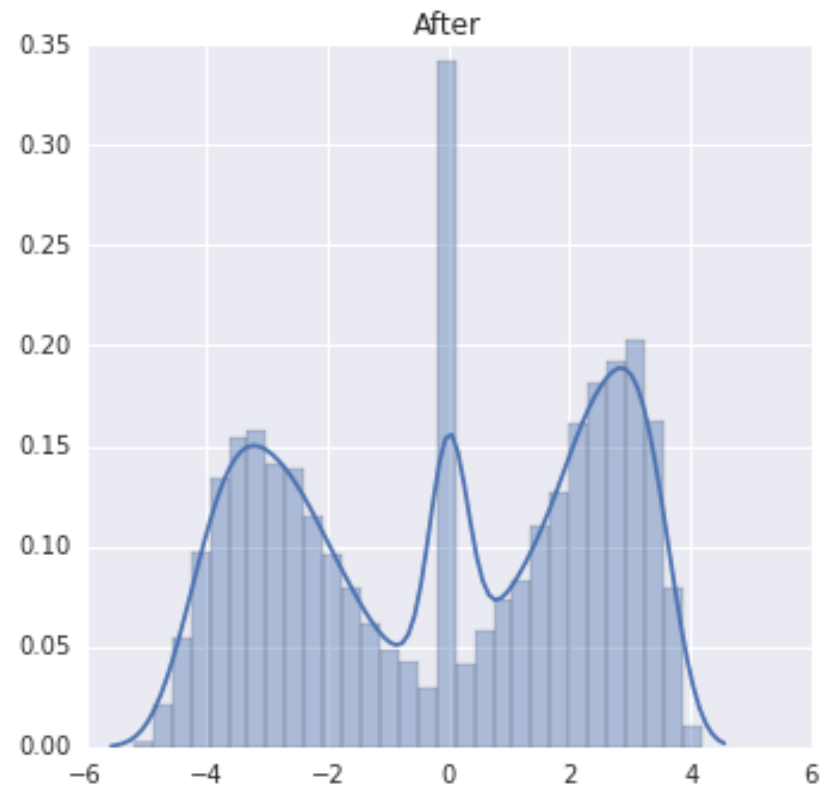
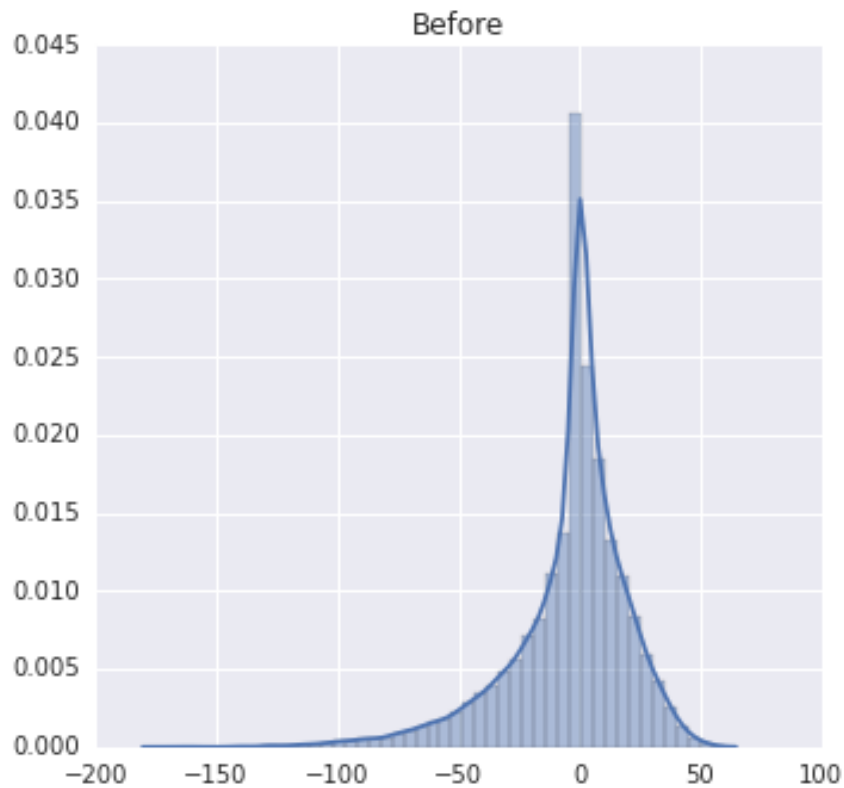
Решение

- $\text{sign}(x) * \log(1 + |x|)$
- $\text{sign}(x) * \sqrt{|x|}$



Решение

- Зачем $\text{sign}(x)$?



Проверим на данных

Logistic Regression, данные - Santander

Данные	ROC AUC
train	0.6036
scale(train)	0.7909
scale(train_rank)	0.8206
scale(train_rank) + преобразование распределений	0.8215