

Машинное обучение

Лекция 16

Заключение

Ковалев Евгений

НИУ ВШЭ, 2020

Куда двигаться дальше?

Open Data Science



- <https://ods.ai/>
- <https://www.youtube.com/channel/UCeq6ZIlvC9SVsfhfKnSvM9w>
- <https://www.youtube.com/channel/UCM9ECBAZtILeEr-m3ldZ7Tw>
- Международное сообщество – почти 50 тысяч человек
- Конференции – датафесты: <https://datafest.ru/video/>
- Курсы: ML (<https://mlcourse.ai/>), DL (<https://dlcourse.ai/>), совместное прохождение курсов Stanford University, Carnegie Mellon University
- Соревнования: <https://ods.ai/competitions>, канал #kaggle_crackers
- Вакансии, полезные ссылки, разбор статей, общение

Курсы

- ML и анализ данных: <https://www.coursera.org/specializations/machine-learning-data-analysis>
- Конечно, курсы от ODS (предыдущий слайд)
- Обзор курсов: <https://www.learndatasci.com/best-machine-learning-courses/>
- Отличный курс по DL: <https://www.coursera.org/specializations/deep-learning>
- Жесткая прокачка: ШАД/MADE Mail.ru/Ozon Masters (сравнение: <https://youtu.be/orygeynBakI>)

Курсы

When you complete an online course on machine learning and realise it won't be enough for you to get a job in the field



Соревнования

- Платформы:
 - <https://www.kaggle.com/>
 - <https://www.topcoder.com/challenges>
 - <https://www.drivendata.org/competitions/>
 - <https://zindi.africa/competitions>
- ML-тренировки:
<https://www.youtube.com/channel/UCeq6ZIlvC9SVsfhfKnSvM9w>
- Как правильно «фармить» Kaggle:
<https://habr.com/ru/company/ods/blog/426227/>
- Kaggle – используй платформу на 100%:
<https://youtu.be/bmpkXlykXjk>

Работа


- Каналы в ODS: #_jobs, #_jobs_hr, #ods_resume_mastering
- Сайты:
 - <https://www.linkedin.com>
 - <https://hh.ru/>
 - <https://your.gms.tech/all>
 - <https://www.glassdoor.com/index.htm>
 - <https://www.upwork.com/>
 - <https://angel.co/>

Еще интересное

- Lex Fridman: <https://www.youtube.com/user/lexfridman>
- Каналы в Telegram:
 - Denis Sexy IT: <https://t.me/denissexy>; https://youtu.be/hZ1OgQL9_Cw
 - Жалкие низкочастотники: https://t.me/pathetic_low_freq
 - Futuris: <https://t.me/Futuris>
 - Уже написали: <https://t.me/kontsarenko>
 - Стать специалистом по машинному обучению: <https://t.me/toBeAnMLspecialist>
- Habr:
 - Тэг «Машинное обучение»: https://habr.com/ru/hub/machine_learning/
 - Блог ODS: <https://habr.com/ru/company/ods/>
 - [https://habr.com/ru/search/?target_type=hubs&order_by=relevance&q=машинное обучение](https://habr.com/ru/search/?target_type=hubs&order_by=relevance&q=машинное_обучение)


Kaggle: Fraud detection

Обо мне




Evgeny Kovalev
ML Developer at SberCloud
Moscow, Russia
Joined 4 years ago · last seen in the past day
[GitHub](#) [in](#)

Followers 25
Following 29


Competitions
Expert


[Home](#) [Competitions \(47\)](#) [Datasets \(84\)](#) [Notebooks \(13\)](#) [Discussion \(71\)](#) [...](#) [Edit Profile](#)


Competitions
Expert




Current Rank
774
of 149,277

Highest Rank
737

0

3

1

Deepfake Detec...
6 months ago
Top 4%

72nd
of 2265


IEEE-CIS Fraud ...
a year ago
Top 2%

76th
of 6381


Jigsaw Uninten...
a year ago
Top 4%


106th
of 3165


Datasets
Contributor



Unranked

0

0

0

efficientnet_pyt...
8 months ago

3
votes


efficientnet_pyt...
8 months ago

2
votes


gensim_embedd...
a year ago


2
votes


Notebooks
Contributor



Unranked

0

0

0

LSTM (cp clip n...
a year ago

4
votes


[texts-classifica...
6 months ago

4
votes


[texts-classifica...
6 months ago


3
votes


Discussion
Contributor



Unranked

2

4

19

Shakeup memes
a year ago

27
votes

106th place solu...
a year ago

13
votes

My submissions...
2 years ago

13
votes

<https://www.kaggle.com/blackitten13>

Kaggle

- <https://www.kaggle.com/>
- Целое комьюнити
- Соревнования с денежными призами, где участвуют сильные DS со всего мира
- Рейтинг и звания участников
- Участие в соревновании = мини-проект: репозиторий на GitHub, блог-посты, статьи, строчки в резюме
- Отличное место для получения опыта
- Датасеты, ноутбуки, курсы, вакансии, форум

Fraud detection



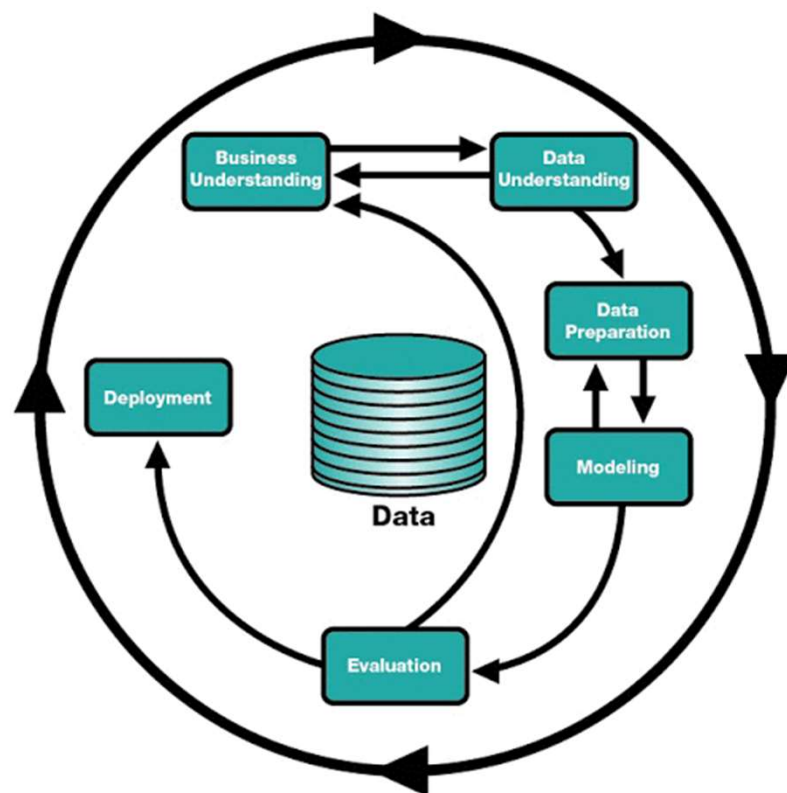
Fraud detection

- Методы обнаружения мошенничества: косвенные признаки, информаторы
- Долго и неэффективно
- Данных очень много – большое число мошеннических транзакций может пройти незамеченным

Fraud detection: Data Science

- Обработка большого объема транзакций
- Анализ данных с помощью статистических методов
- Формирование профиля пользователя
- Выявление паттернов пользователей/транзакций (кластеризация)
- Анализ временных рядов
- Обнаружение аномальных активностей

Машинное обучение










CRISP-DM (межотраслевой стандартный процесс для исследования данных)

Обучающая выборка

- Наблюдения: транзакции
- Целевая переменная (target): нормальная транзакция или мошенническая? (0 или 1)
- Транзакции описываются какими-то признаками
 - Информация о карте (тип карты, банк выпуска, страна, ...)
 - Информация о транзакции (сумма, время совершения, тип устройства, ...)
 - Информация о пользователях (адрес, email, дата заключения договора, ...)

Соревнование

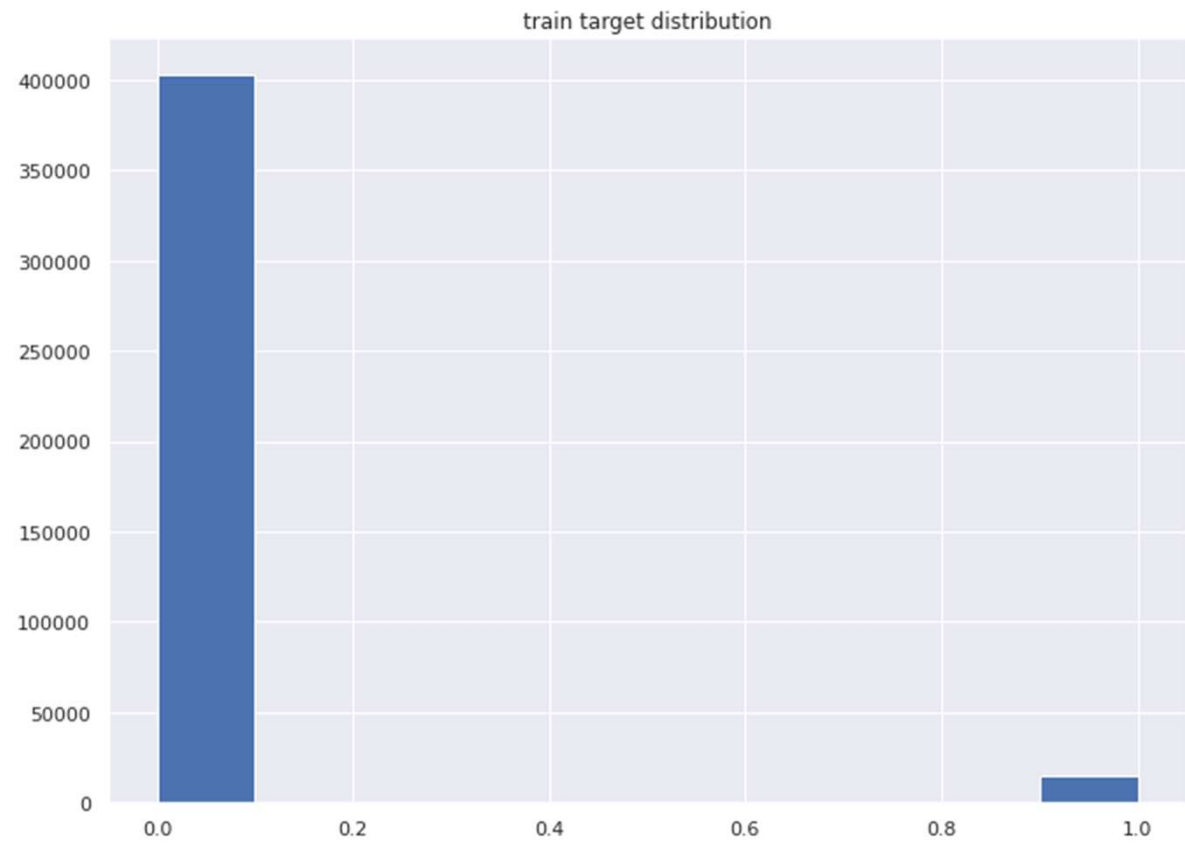
- <https://www.kaggle.com/c/ieee-fraud-detection>
- Призовой фонд: \$20,000
- 6381 команда
- Наше место: 76 (top-2%, серебряная медаль)

75	▼ 25	Thierry		0.932069	53	1y
76	▲ 23	[ods.ai] No Fosters for Fr...	    	0.931944	124	1y
77	▲ 39	Kaz-U		0.931688	170	1y

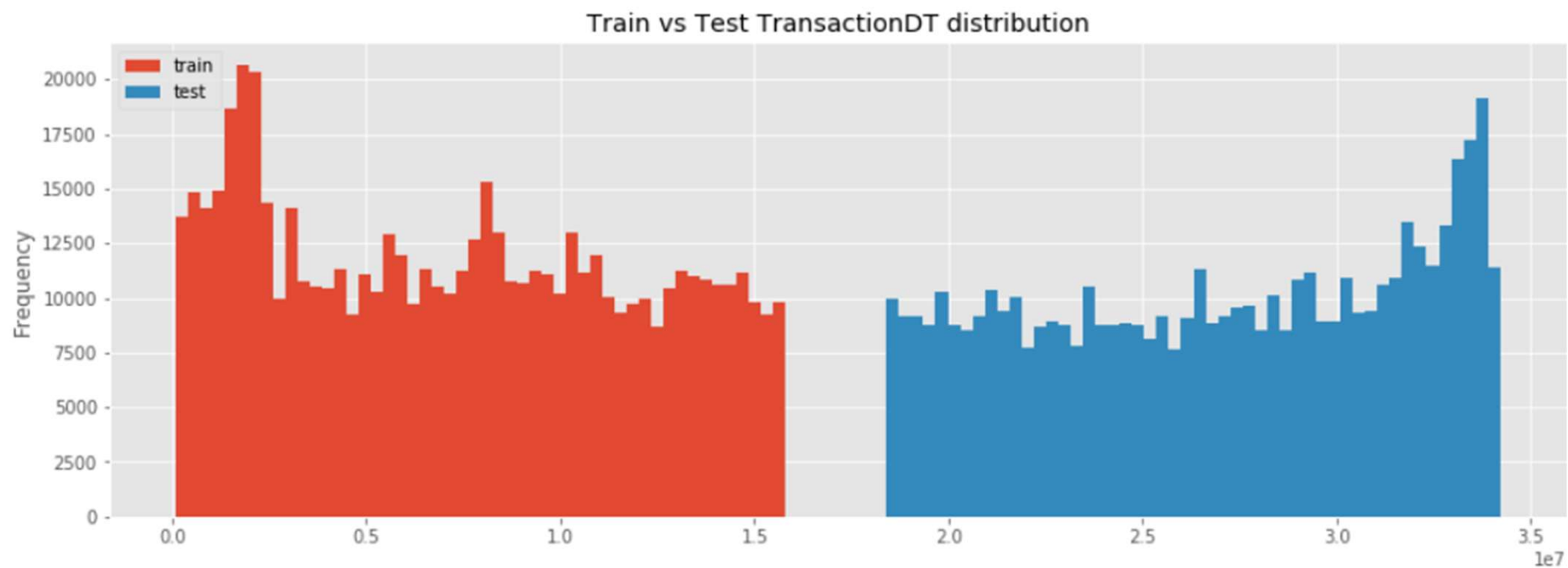
Exploratory Data Analysis (EDA)

- Как выглядят данные?
- Как выборка разбита на обучение и тест (проверку)?
- Как распределены значения признаков?
- Как связаны признаки и целевая переменная?
- (анонимизированные признаки) Что означают те или иные признаки?
- Как сделать предобработку данных?
- Какие паттерны есть в данных?

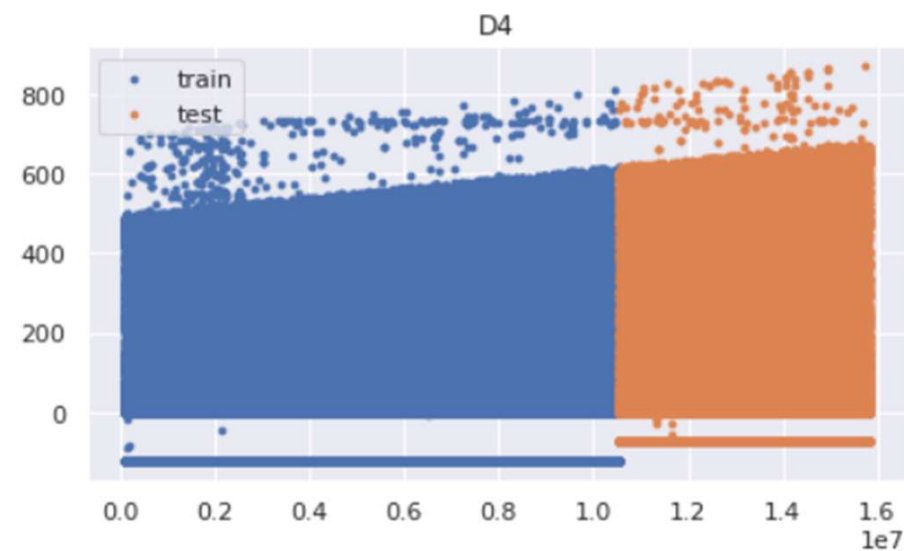
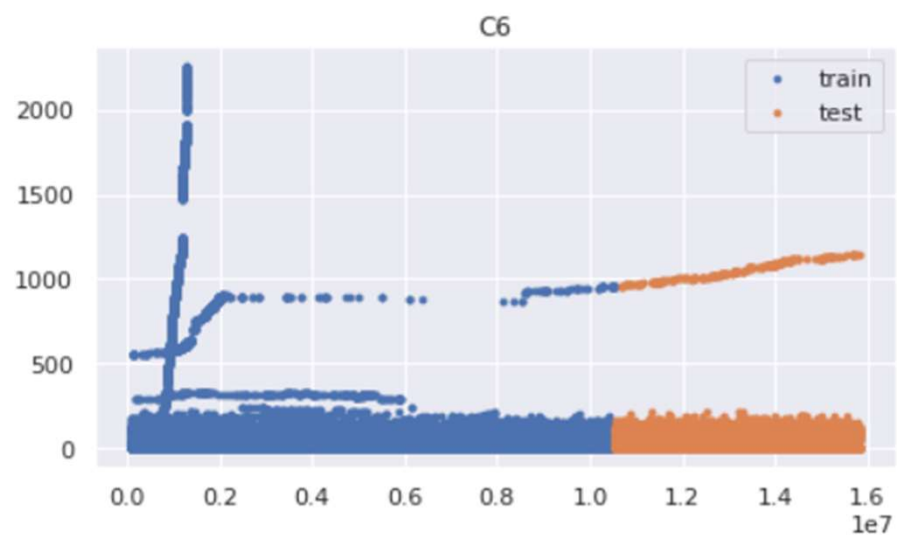
Баланс данных



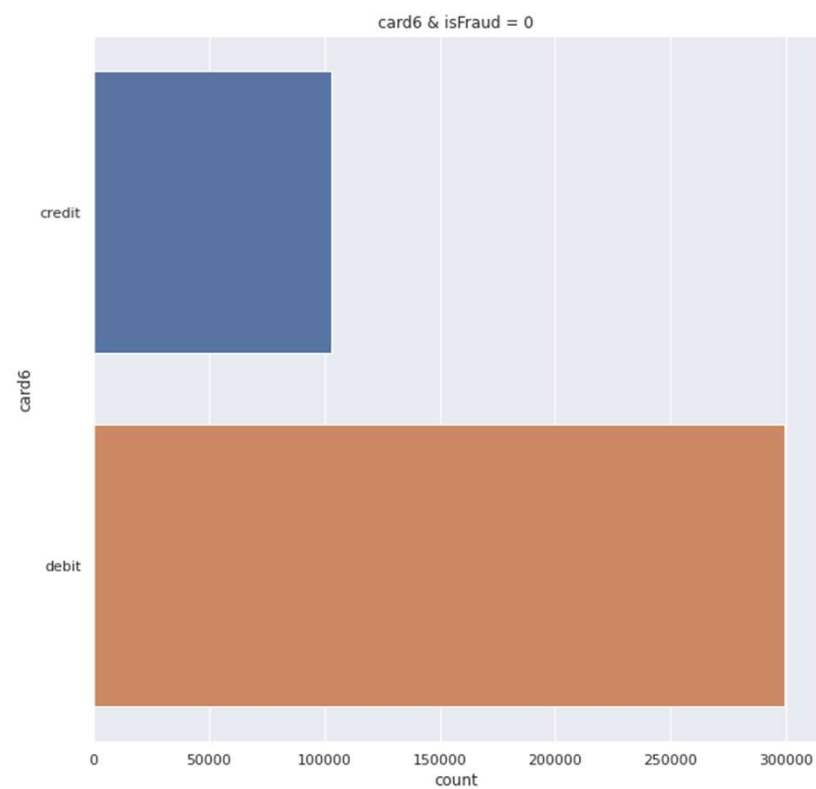
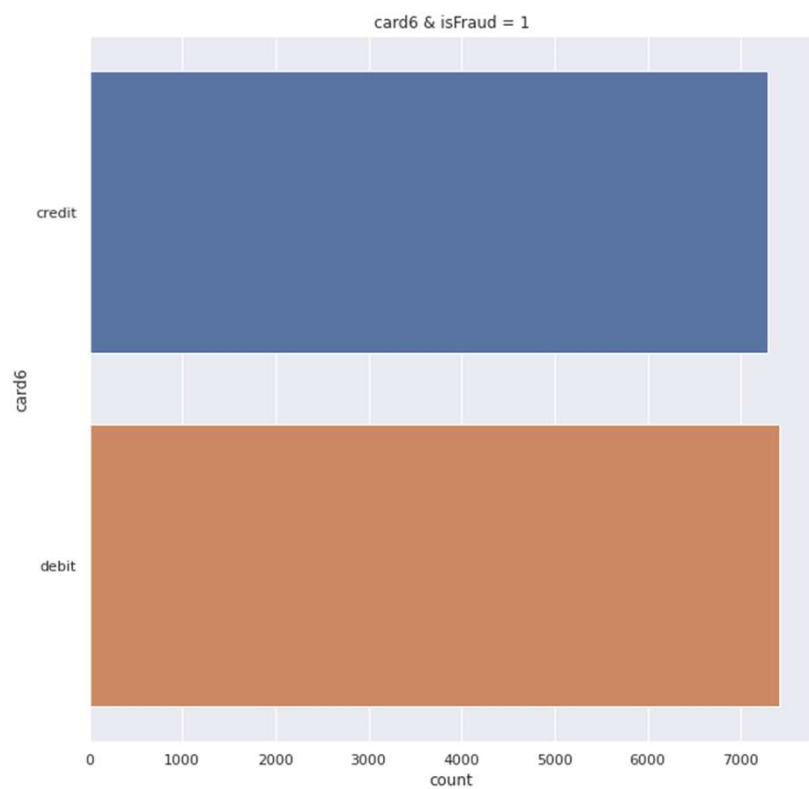
Обучающая и тестовая выборки



Распределение значений признаков



Связь признаков и целевой переменной



Деанонимизация признаков

D9	
0	NaN
1	0.000000
2	0.041666
3	0.083333
4	0.125000
5	0.166666
6	0.208333
7	0.250000
8	0.291666
9	0.333333
10	0.375000
11	0.416666
12	0.458333
13	0.500000
14	0.541666
15	0.583333
16	0.625000
17	0.666666
18	0.708333
19	0.750000
20	0.791666
21	0.833333
22	0.875000
23	0.916666
24	0.958333

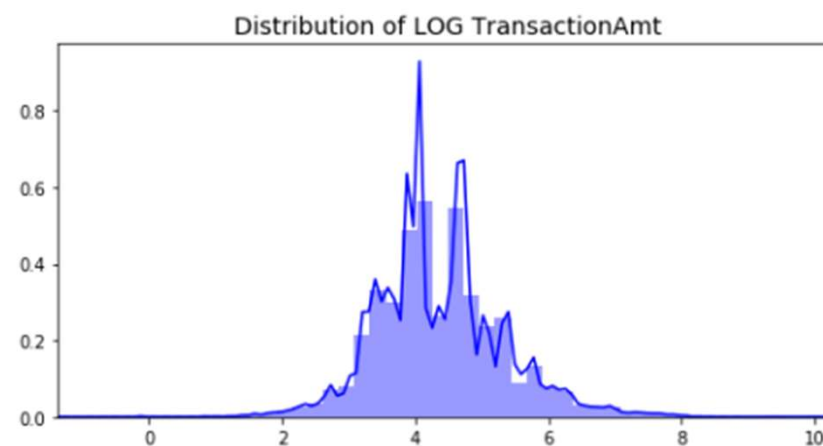
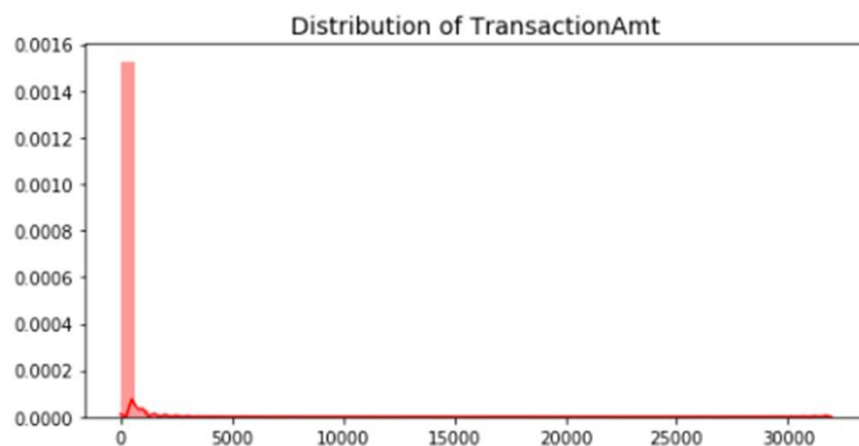
Деанонимизация признаков

D9	
0	NaN
1	0.000000
2	0.041666
3	0.083333
4	0.125000
5	0.166666
6	0.208333
7	0.250000
8	0.291666
9	0.333333
10	0.375000
11	0.416666
12	0.458333
13	0.500000
14	0.541666
15	0.583333
16	0.625000
17	0.666666
18	0.708333
19	0.750000
20	0.791666
21	0.833333
22	0.875000
23	0.916666
24	0.958333



D9	
0	NaN
1	0.000000
2	0.999984
3	1.999992
4	3.000000
5	3.999984
6	4.999992
7	6.000000
8	6.999984
9	7.999992
10	9.000000
11	9.999984
12	10.999992
13	12.000000
14	12.999983
15	13.999992
16	15.000000
17	15.999983
18	16.999992
19	18.000000
20	18.999983
21	19.999992
22	21.000000
23	21.999983
24	22.999992

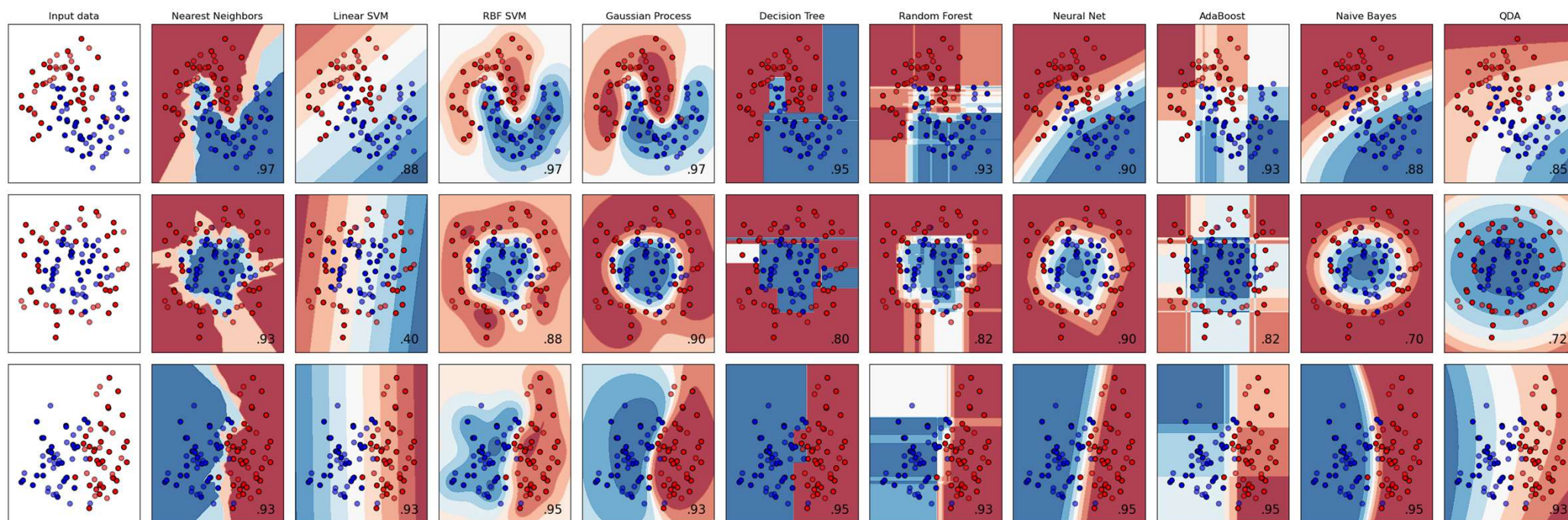
Предобработка данных



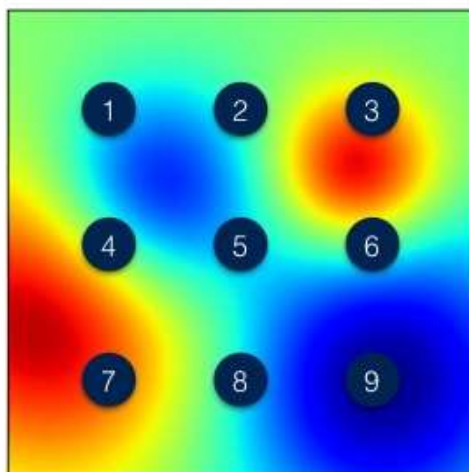
Обучение

- Какую модель выбрать?
- Как подобрать гиперпараметры?
- Как проверить, что модель хорошо работает на новых данных?
- Как проанализировать результат?
- Как улучшить результат?

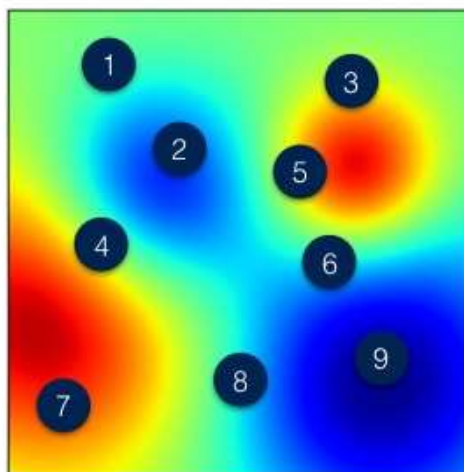
Модели машинного обучения



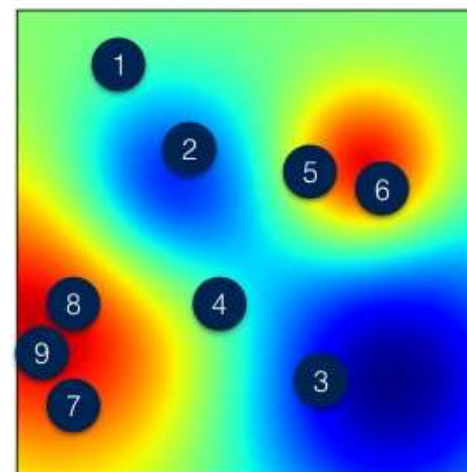
Настройка модели (обучение)



Grid Search



Random Search



Adaptive Selection

Валидация

70% train data

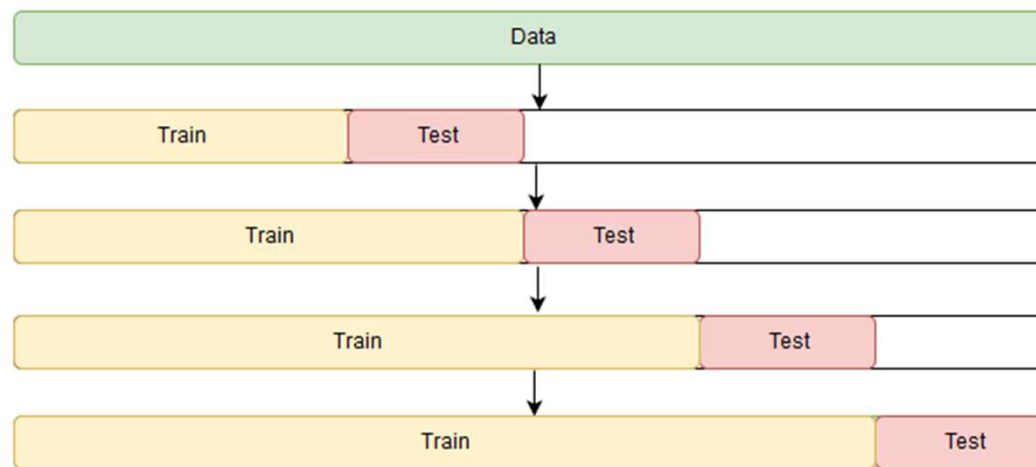


Обучение

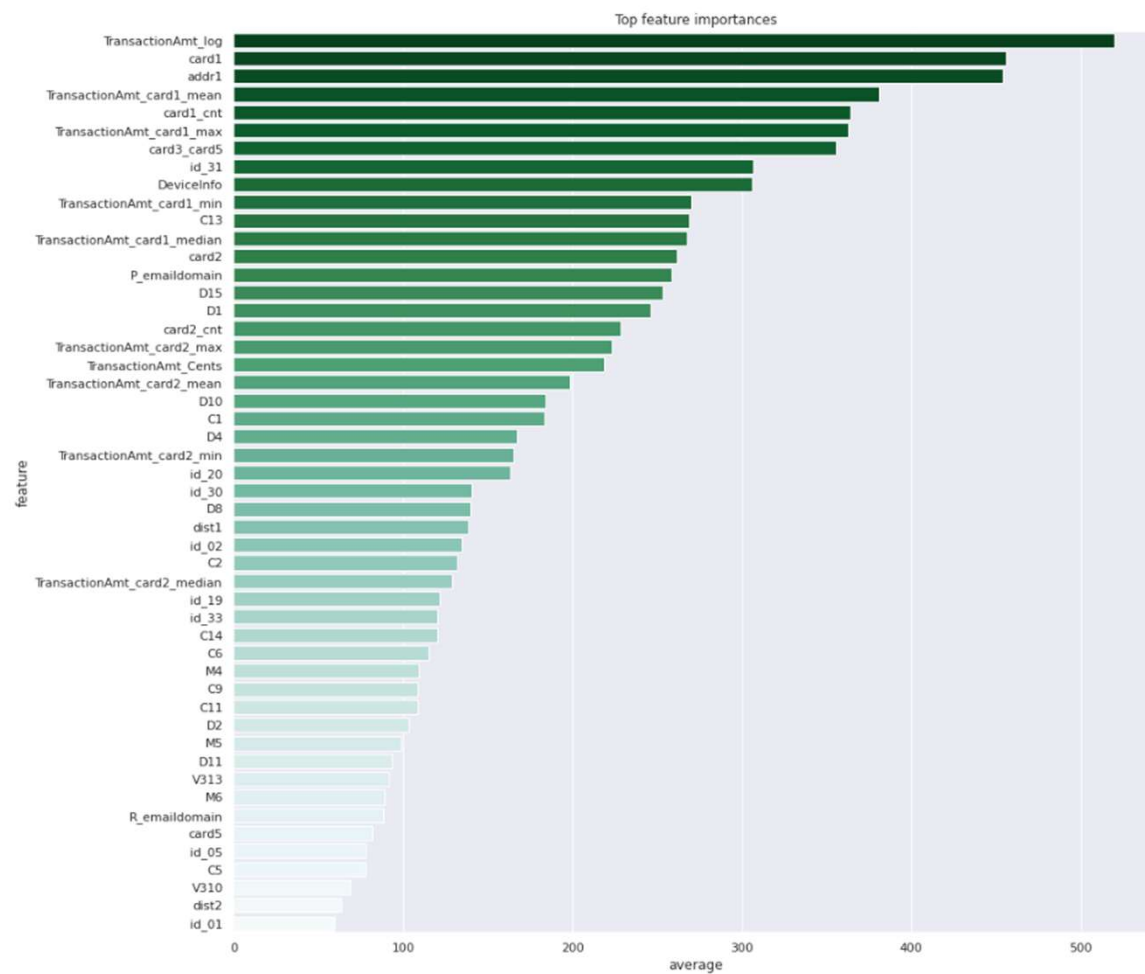
30% train data



Валидация



Анализ результатов



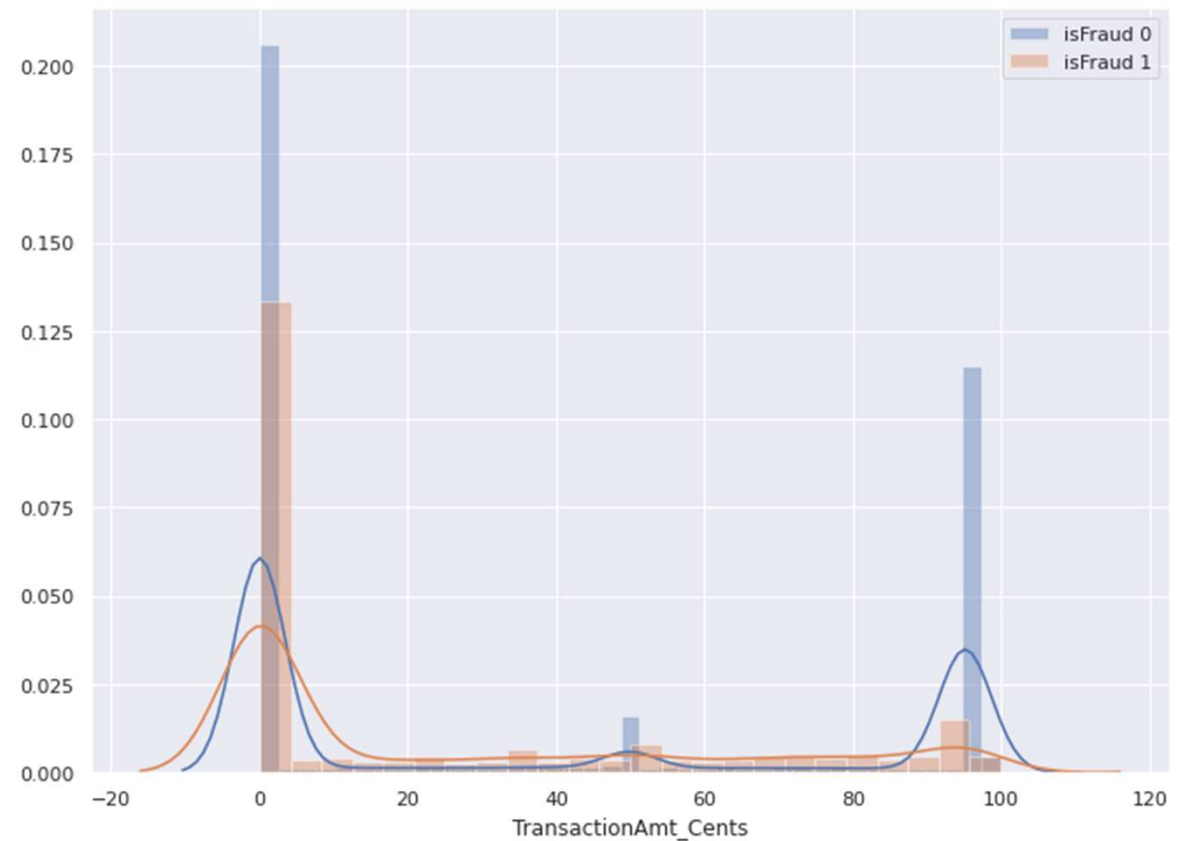
Улучшение результатов: feature engineering

Заметим, что отнюдь не все суммы транзакций (в долларах) - целочисленные:

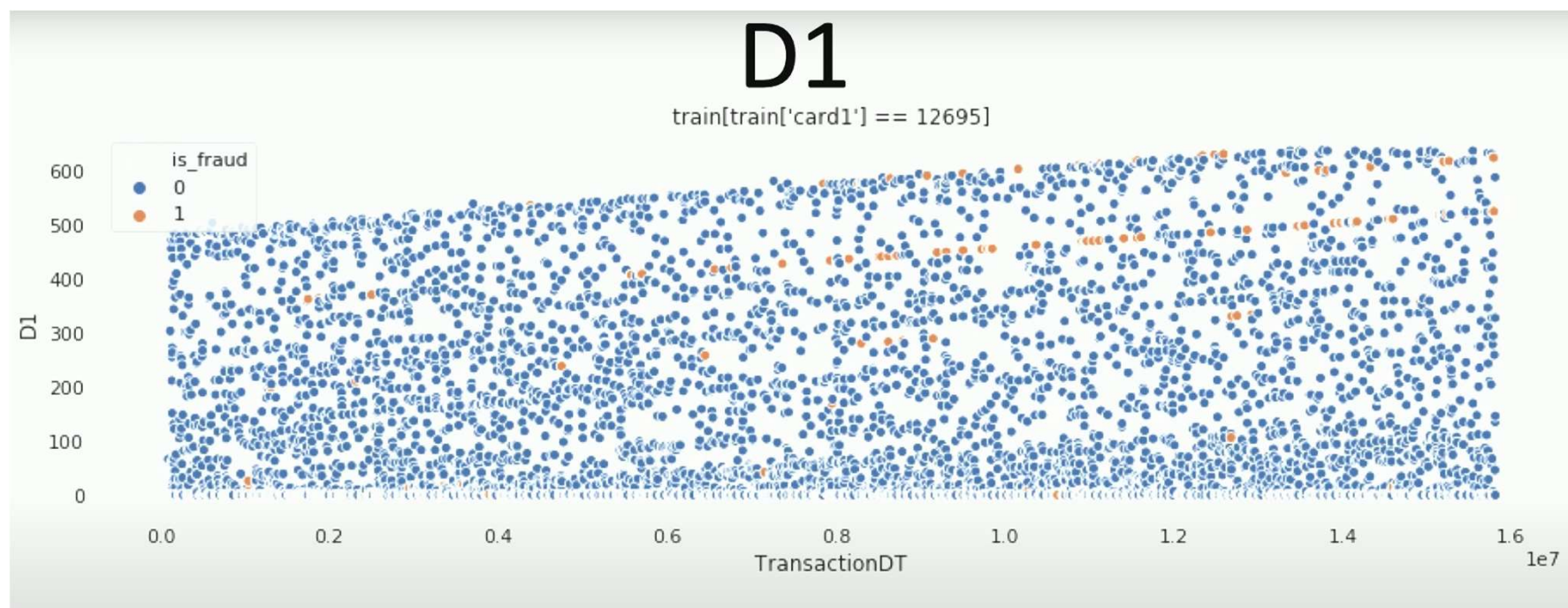
```
[ ] df_train['TransactionAmt'].value_counts()[:15]
```

59.000000	20509
117.000000	19544
100.000000	16685
107.949997	15869
57.950001	15615
50.000000	14270
49.000000	10725
226.000000	7966
39.000000	7490
29.000000	7177
150.000000	6700
47.950001	6448
25.000000	6166
35.950001	5888
171.000000	5245
34.000000	5209
200.000000	5183
30.950001	5082
77.000000	4974
25.950001	4905
75.000000	4109
209.949997	4099
335.000000	3900
67.949997	3533
159.949997	3176
97.000000	3163
92.000000	2887
30.000000	2842
250.000000	2781
58.950001	2455
40.000000	2412
108.500000	2316
108.949997	2288
15.000000	2200

Name: TransactionAmt, dtype: int64



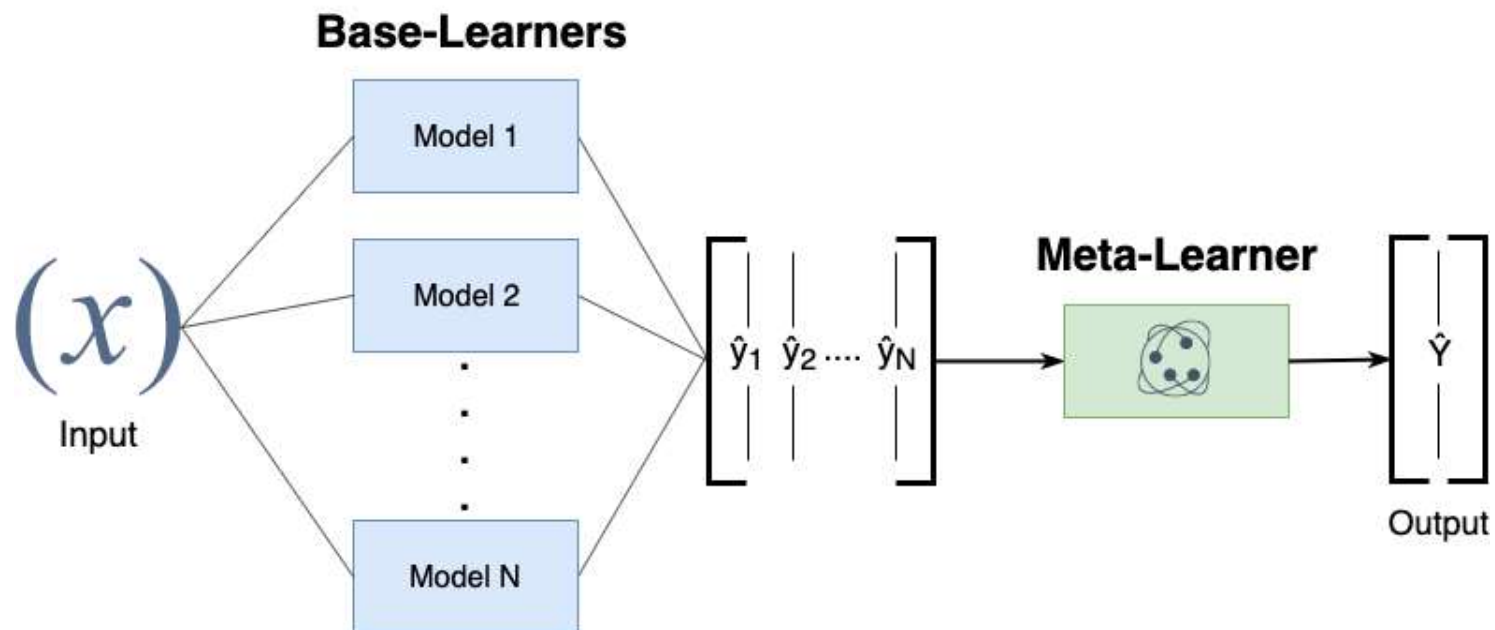
Улучшение результатов: magic features



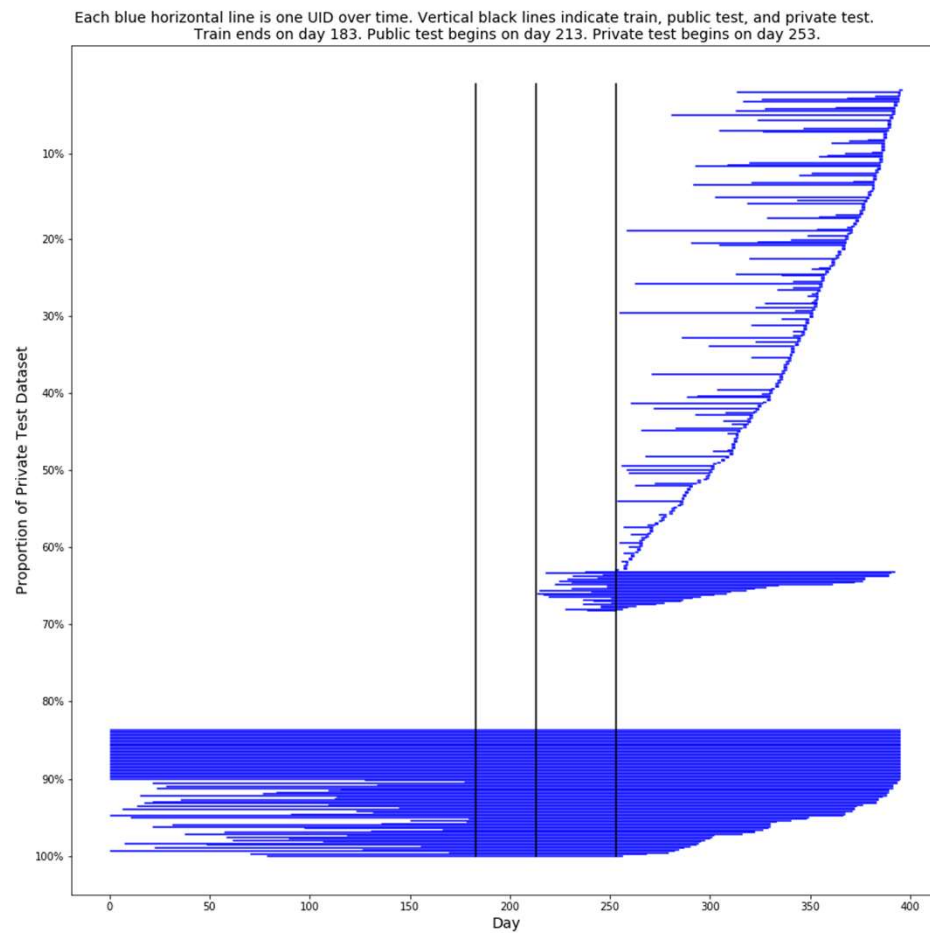
```
df.ProxyUserID1 = (to_day(df.TransactionDT) - df.D1).astype(str) + "_" + df.card1.astype(str)
```

<https://www.youtube.com/watch?v=jtE3FhMUJDw>

Улучшение результатов: model stacking



Постпроцессинг: определение мошенников



	TransactionID	isFraud	TransactionAmt	card1	addr1	D1n	day	D3n	dist1	P_emaildomain	UID
1694	2988694	1	240.0	15775	251.0	-81.0	1.0	0.0	NaN	yahoo.com	2988694.0
10046	2997046	1	260.0	15775	251.0	-81.0	3.0	1.0	NaN	yahoo.com	2988694.0
34029	3021029	1	250.0	15775	251.0	-81.0	9.0	3.0	NaN	yahoo.com	2988694.0
36812	3023812	1	315.0	15775	251.0	-81.0	10.0	9.0	NaN	yahoo.com	2988694.0
40459	3027459	1	390.0	15775	251.0	-81.0	11.0	10.0	NaN	yahoo.com	2988694.0
43926	3030926	1	475.0	15775	251.0	-81.0	12.0	11.0	NaN	yahoo.com	2988694.0
43941	3030941	1	445.0	15775	251.0	-81.0	12.0	12.0	NaN	yahoo.com	2988694.0
44717	3031717	1	445.0	15775	251.0	-81.0	12.0	12.0	NaN	yahoo.com	2988694.0
44727	3031727	1	445.0	15775	251.0	-81.0	12.0	12.0	12.0	NaN	2988694.0
58485	3045485	1	295.0	15775	251.0	-81.0	15.0	12.0	NaN	yahoo.com	2988694.0

Data Science: соревнования vs бизнес

Соревнования	Бизнес
Данные чистые и хорошо подготовленные	Данные нужно собирать, размечать, чистить
Итог: предсказания модели + код	Итог: развертывание модели на сервере
Метрика качества дана	Метрику качества нужно выбрать
Нужна высокая точность по метрике	Гибкость в плане точности
Результат интерпретировать не нужно	Интерпретация может быть крайне важна
Слабые ограничения на сложность модели	Часто есть разнообразные ограничения