Предсказание пульсаров на основе датасета HTRU2

...

Артём Давыдов
Группа 205

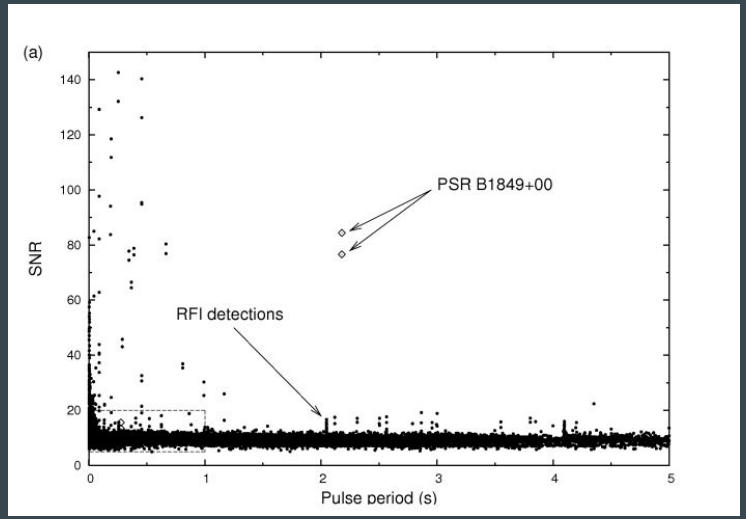
Постановка задачи

Одним из объектов исследования современной астрофизики являются пульсары - космические источники электромагнитного излучения, быстро вращающиеся вокруг своей оси. Сигналы от пульсаров приходят на Землю в виде периодических импульсов, которые отслеживаются с помощью телескопов.
Однако такой метод поиска затруднен ложными срабатываниями телескопов на другие сигналы, которые не относятся к пульсарам.

Процесс поиска пульсаров

Процесс поиска является очень долгим. Его можно разбить на 2 этапа:

1. Сбор данных с радиотелескопа и первичная обработка, в результате которой каждый пульсар изображается точкой на графике (пример-справа)
2. Далее для отсеивания ложных срабатываний делается, как правило, визуальная проверка.



Для ускорения второго этапа может быть использовано машинное обучение.

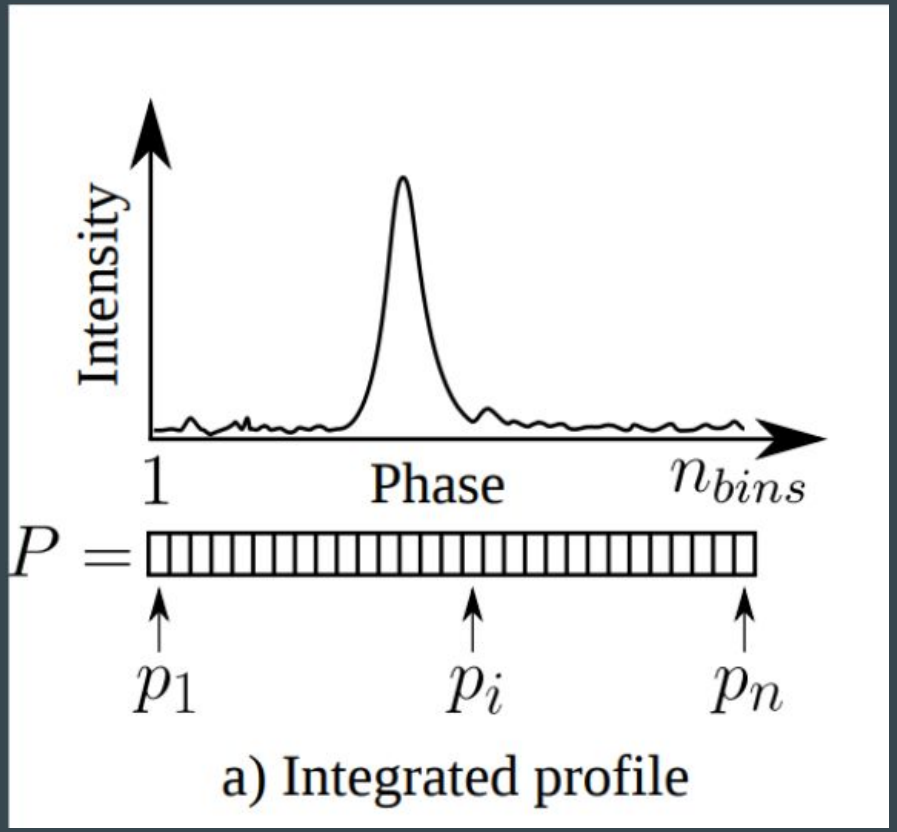
Набор данных (DataSet)

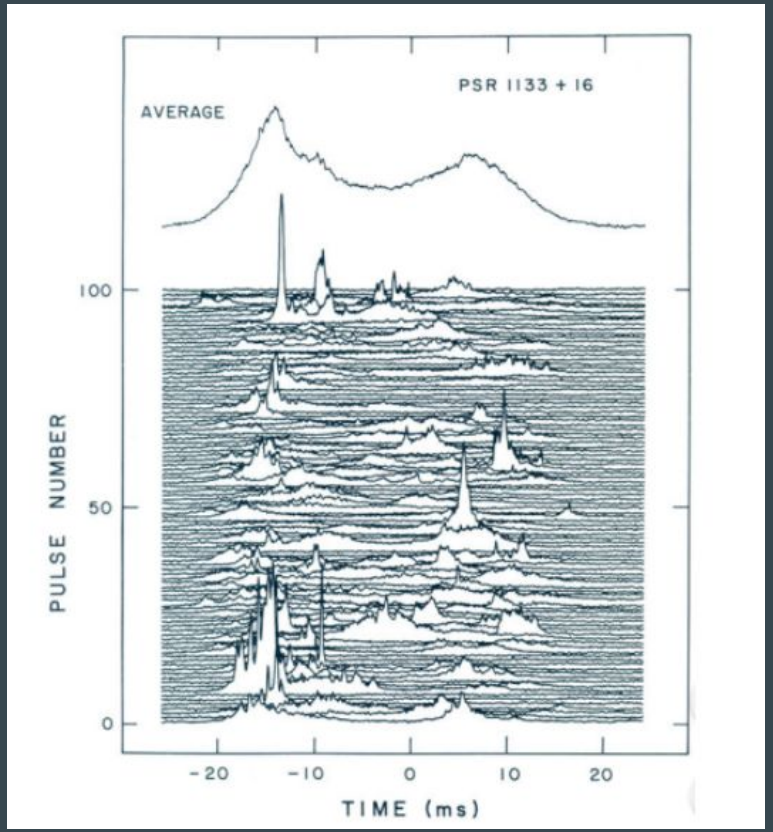
Набор данных называется HTRU2, он представляет собой 17898 различных сигналов, которые были зарегистрированы телескопами, из них:

1. 16259 ложных срабатываний
2. 1639 истинных срабатываний

Каждому пульсару соответствует строка из 8 чисел, каждое из которых - это интегральные характеристики двух кривых - интегрированного профиля и DM-SNR кривой.

Кривые с данными





DM и SNR

- **по мере дисперсии DM**. При всех недостатках этого метода он является основным. Для подавляющего большинства пульсаров расстояние определено только таким образом.

  Измерение времени запаздывания сигналов на разных частотах позволяет определить меру дисперсии для данного пульсара:

$$\Delta t = \frac{2\pi e^2}{m_e c}\left(\frac{1}{\omega_1^2} - \frac{1}{\omega_2^2}\right) DM,$$

  где m_e — масса электрона, e — его заряд, c — скорость света, $\omega_{1,2}$ — измеряемые частоты. Поскольку

$$DM = \int_0^L n_e dl = \bar{n_e} L,$$

$$\mathrm{SNR} = \frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}},$$

Что за данные мы имеем?

1. Mean(среднее значение) M(x)

$$M(X) = \int_{-\infty}^{+\infty} x f_X(x) dx$$

2. Standard deviation (стандартное отклонение) **σ**(x)

$$\sigma(X) = \sqrt{\int_{-\infty}^{+\infty} \left(x - \int_{-\infty}^{+\infty} x f_X(x) dx \right)^2 f_X(x) dx}$$

3. Excess kurtosis (коэффициент эксцесса) Ek(x)

$$Ek(X) = \frac{\int_{-\infty}^{\infty} \left(x - \int_{-\infty}^{\infty} x f_X(x) dx \right)^4 f_X(x) dx}{\left[\int_{-\infty}^{\infty} \left(x - \int_{-\infty}^{\infty} x f_X(x) dx \right)^2 f_X(x) dx \right]^2} - 3$$

4. Skewness (коэффициент асимметрии) As(x)

$$As(X) = \frac{\int_{-\infty}^{\infty} \left(x - \int_{-\infty}^{\infty} x f_X(x) dx \right)^3 f_X(x) dx}{\left[\int_{-\infty}^{\infty} \left(x - \int_{-\infty}^{\infty} x f_X(x) dx \right)^2 f_X(x) dx \right]^{\frac{3}{2}}}$$

Методы решения поставленной задачи

Для решения задачи будет использоваться обычная полносвязная нейронная сеть (Dense-слои), количество нейронов в которых подбирается (примерно) так, чтобы не было переобучения(не слишком много). На выходе должен быть 1 нейрон, т.к. рассматриваемая задача - задача бинарной классификации.

Layer (type)	Output Shape	Param #
dense_3 (Dense)	multiple	144
dropout_2 (Dropout)	multiple	0
dense_4 (Dense)	multiple	544
dropout_3 (Dropout)	multiple	0
dense_5 (Dense)	multiple	33

Total params: 721
Trainable params: 721
Non-trainable params: 0

Возможные проблемы

При решении задачи с помощью нейронных сетей могут возникнуть следующие проблемы:

1. Данные сильно несбалансированы: “не пульсаров” гораздо больше, чем пульсаров. Поэтому нейронная сеть может научиться говорить, что всегда “не пульсар”
2. Слишком мало самих данных: всего 17898 пульсаров, а на каждый пульсар приходится только 8 характеристик.

Методы решения озвученных проблем

Чтобы решить указанные выше проблемы были использованы следующие методы:

1. Для несбалансированных данных, согласно статье [1], используется метод “пересчета веса класса”, при этом вес класса “пульсар” становится больше веса класса “не пульсар”. И сети становится невыгодно выдавать на выходе все ответы “не пульсар”
2. Чтобы увеличить количество данных, согласно статье [2], используется метод “Feature engineering”, который на основе имеющихся данных, с помощью стандартных математических операций, составляет новые данные. Это также помогает обнаружить скрытые закономерности в данных.

Обучение и проверка работы сети

Исходные данные были разбиты на 3 группы:

1. 80% - на обучение сети
2. 10% - на проверку
3. 10% - на тест

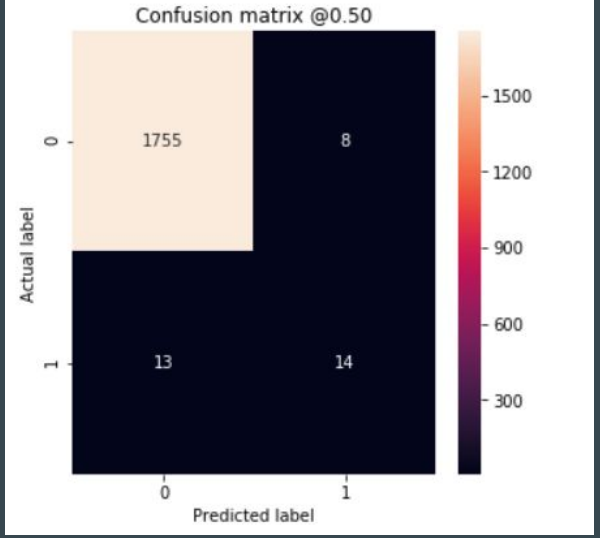
Работа была проведена с 3-мя типами данных:

1. Обычные данные из HTRU2 без применения методов “Feature engineering”
2. Центированные данные из HTRU2 без применения методов “Feature engineering”
3. Данные с применением метода “Feature engineering”

Сравнение результатов с пересчетом веса класса и без

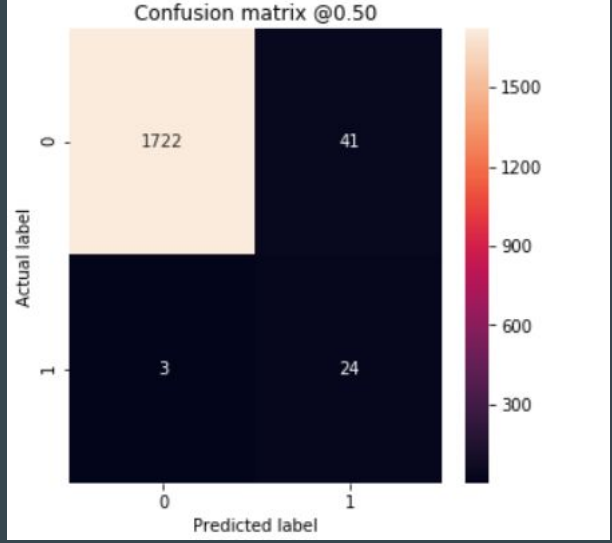
Данные без пересчета веса класса

```
loss :  0.05196929413346605
accuracy :  0.98826814

всего 27 "пульсаров"
всего 1763 "не пульсаров"
не пульсар предсказанный как не пульсар:  1755
не пульсар предсказанный как пульсар : 8
пульсар предсказанный как не пульсар : 13
пульсар предсказанный как пульсар :  14
количество правильных ответов 1769 из 1790
```

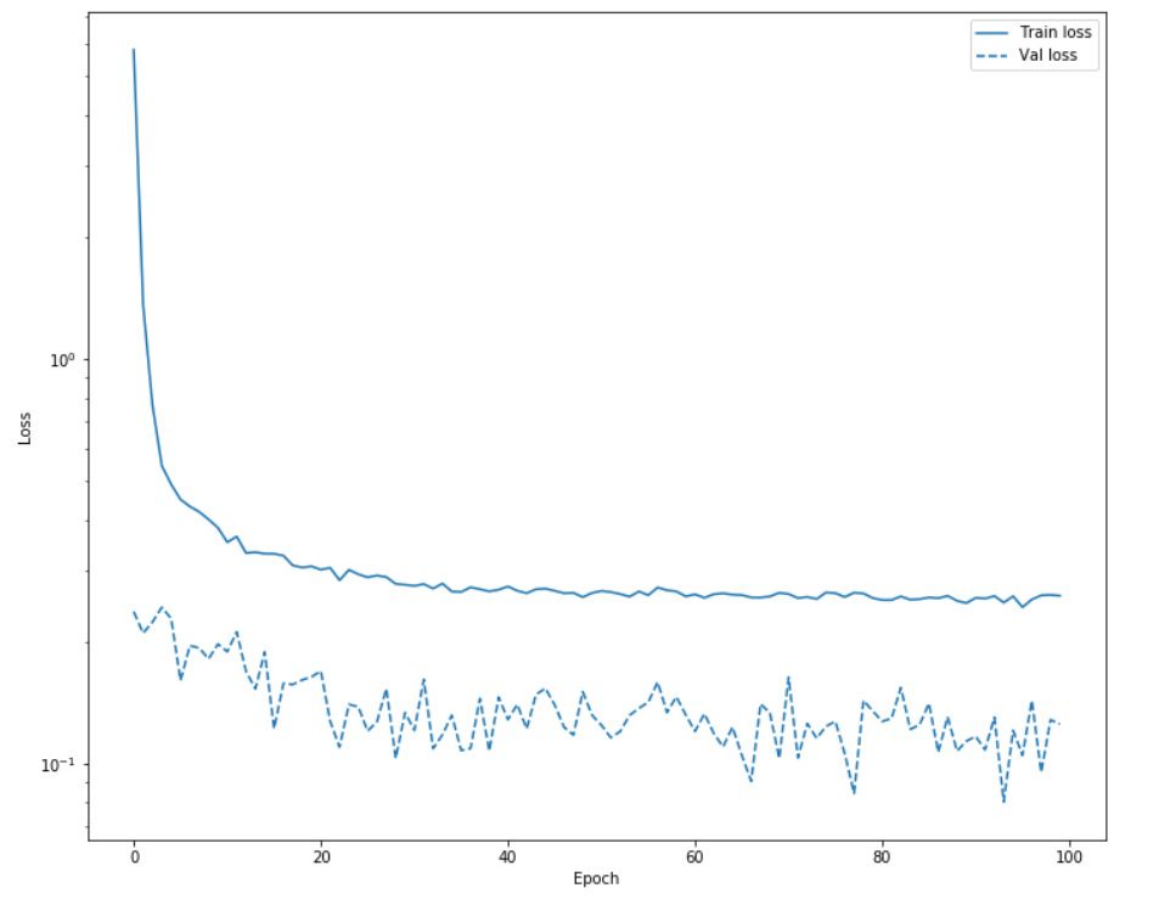


Данные с пересчетом веса класса

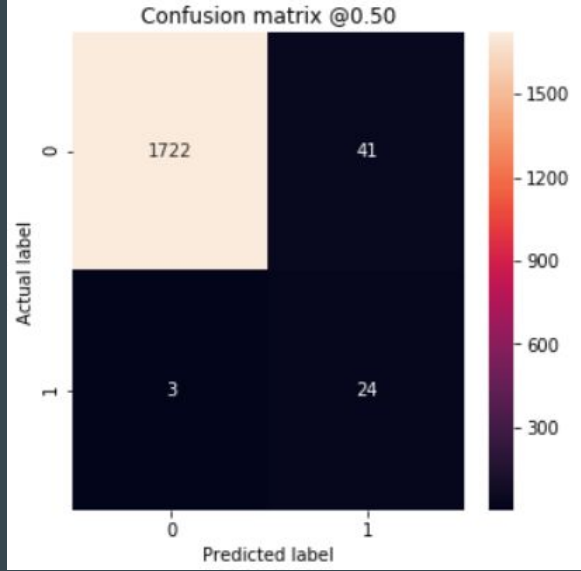
```
loss :  0.20917624711324384
accuracy :  0.975419

всего 27 "пульсаров"
всего 1763 "не пульсаров"
не пульсар предсказанный как не пульсар:  1722
не пульсар предсказанный как пульсар : 41
пульсар предсказанный как не пульсар : 3
пульсар предсказанный как пульсар :  24
количество правильных ответов 1746 из 1790
```



Обучение на “сырых” данных




```
loss :  0.20917624711324384
accuracy :  0.975419

всего 27 "пульсаров"
всего 1763 "не пульсаров"
не пульсар предсказанный как не пульсар:  1722
не пульсар предсказанный как пульсар : 41
пульсар предсказанный как не пульсар : 3
пульсар предсказанный как пульсар :  24
количество правильных ответов 1746 из 1790
```



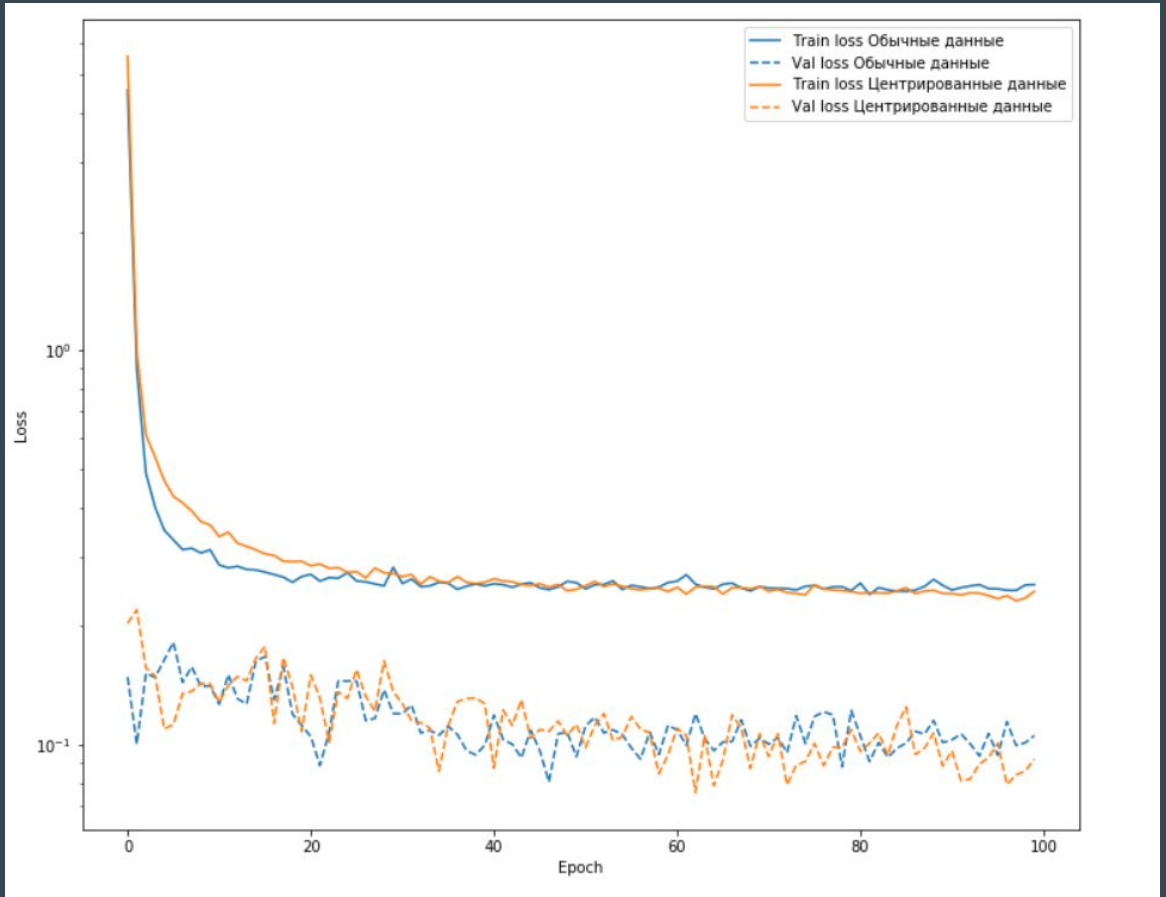
Применение нормализации

Нормализация данных выполняется в соответствии с формулой

$$z = \frac{x - \mu}{\sigma}$$

Необходимость нормализации обусловлена тем, что изначальные данные имеют сильный разброс в значениях: отличия на 2-3 порядка, что может негативно сказаться точности работы сети. Для этого большинство колонок были приведены к распределению со средним значением 0 и стандартным отклонением 1.

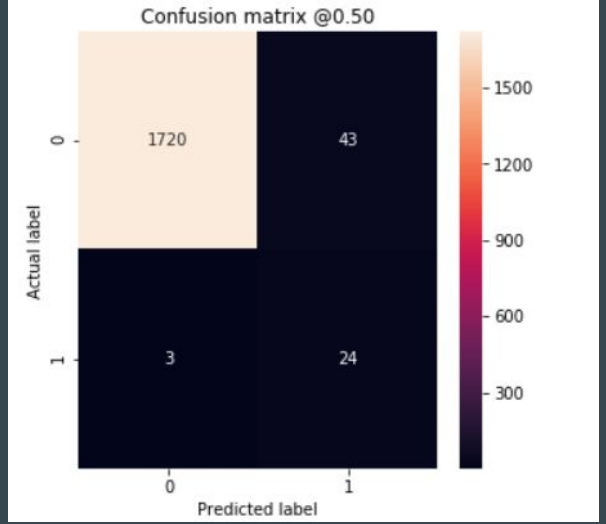
После применения центровки можем увидеть промежуточные итоги



Сравнение Confusion matrix для двух случаев

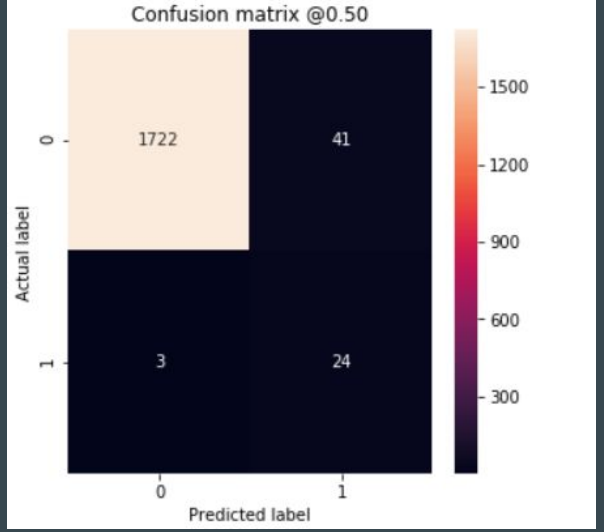
Нормализованные данные

```
loss :  0.2177614570329975
accuracy :  0.9743017

всего 27 "пульсаров"
всего 1763 "не пульсаров"
не пульсар предсказанный как не пульсар:  1720
не пульсар предсказанный как пульсар : 43
пульсар предсказанный как не пульсар : 3
пульсар предсказанный как пульсар :  24
количество правильных ответов 1744 из 1790
```



Ненормализованные данные

```
loss :  0.20917624711324384
accuracy :  0.975419

всего 27 "пульсаров"
всего 1763 "не пульсаров"
не пульсар предсказанный как не пульсар:  1722
не пульсар предсказанный как пульсар : 41
пульсар предсказанный как не пульсар : 3
пульсар предсказанный как пульсар :  24
количество правильных ответов 1746 из 1790
```



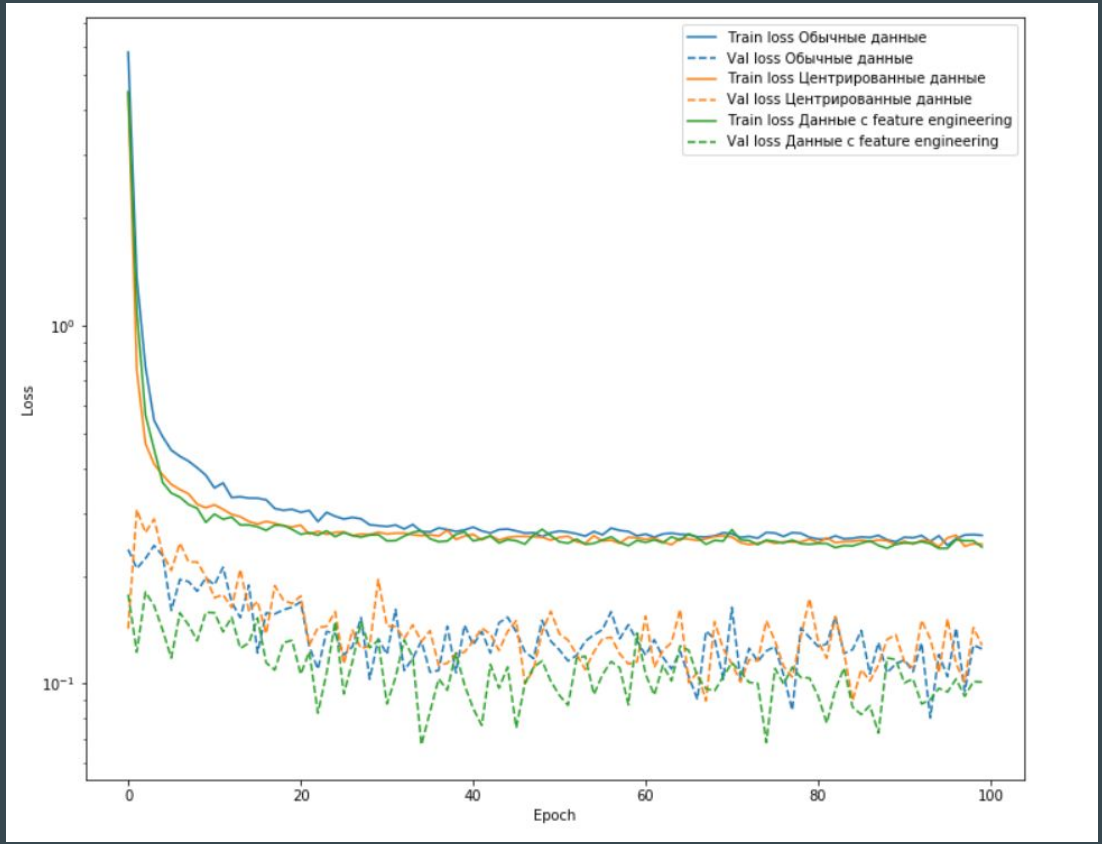
Применение метода feature engineering

Item_ID	Item_Weight	Item_Price	Price_per_Weight
FDA15	9.3	249.81	26.86
DRC01	5.9	48.27	8.15
FDN15	17.5	141.62	8.09
FDX07	19.2	182.10	9.48

В нашем случае

number	MeanIP	StdIP	ExckurtIP	SkewIP	MeanDMSNR	StdDMSNR	ExckurtDMSNR	SkewDMSNR	new_signals.MEAN(signals.StdDMSNR)
0	140.562500	55.683782	-0.234571	-0.699648	3.199833	19.110426	7.975532	74.242225	19.110426
1	102.507812	58.882430	0.465318	-0.515088	1.677258	14.860146	10.576487	127.393580	15.252106
2	103.015625	39.341649	0.323328	1.051164	3.121237	21.744669	7.735822	63.171909	18.196718
3	136.750000	57.178449	-0.068415	-0.636238	3.642977	20.959280	6.896499	53.593661	19.568731
4	88.726562	40.672225	0.600866	1.123492	1.178930	11.468720	14.269573	252.567306	11.468720
5	93.570312	46.698114	0.531905	0.416721	1.636288	14.545074	10.621748	131.394004	33.296463
6	119.484375	48.765059	0.031460	-0.112168	0.999164	9.279612	19.206230	479.756567	32.433255
7	130.382812	39.844056	-0.158323	0.389540	1.220736	14.378941	13.539456	198.236457	17.301946
8	107.250000	52.627078	0.452688	0.170347	2.331940	14.486853	9.001004	107.972506	17.726237
9	107.257812	39.496488	0.465882	1.162877	4.079431	24.980418	7.397080	57.784738	21.537289
10	142.078125	45.288073	-0.320328	0.283953	5.376254	29.009897	6.076266	37.831393	29.009897
11	133.257812	44.058244	-0.081060	0.115362	1.632107	12.007806	11.972067	195.543448	25.233969
12	134.960938	49.554327	-0.135304	-0.080470	10.696488	41.342044	3.893934	14.131206	30.485614
13	117.945312	45.506577	0.325438	0.661459	2.836120	23.118350	8.943212	82.475592	26.786372
14	138.179688	51.524484	-0.031852	0.046797	6.330268	31.576347	5.155940	26.143310	21.577881

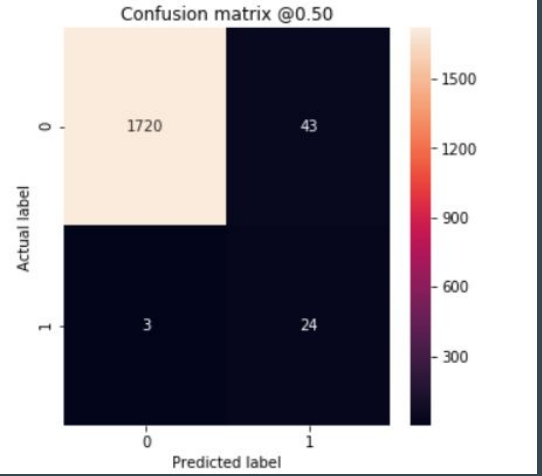
После применения feature engineering можем сравнить работу сети для 3-ёх случаев



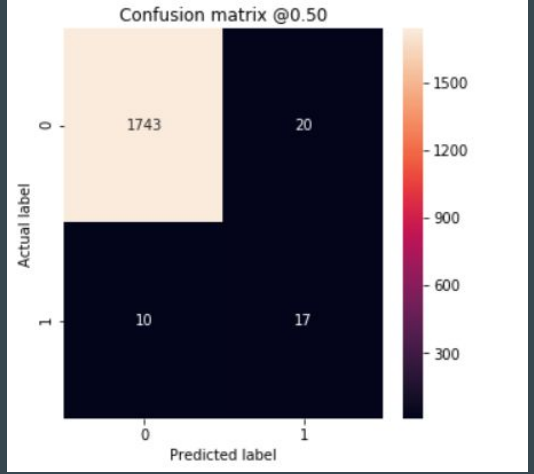
Сравнение Confusion matrix для трёх случаев

Нормализованные данные Данные с feature engineering Ненормализованные данные

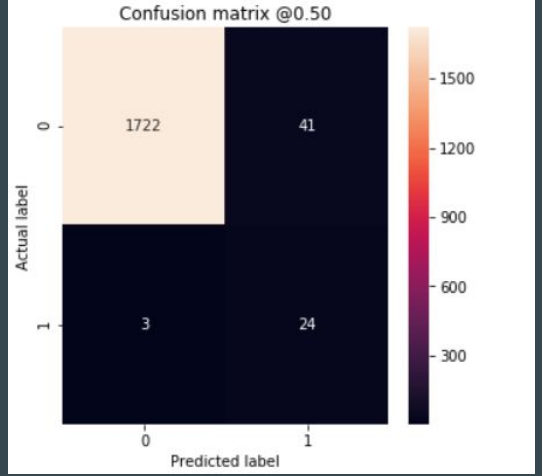
```
loss :  0.2177614570329975
accuracy :  0.9743017

всего 27 "пульсаров"
всего 1763 "не пульсаров"
не пульсар предсказанный как не пульсар:  1720
не пульсар предсказанный как пульсар : 43
пульсар предсказанный как не пульсар : 3
пульсар предсказанный как пульсар :  24
количество правильных ответов 1744 из 1790
```




```
loss :  0.14739838046068585
accuracy :  0.98324025

всего 27 "пульсаров"
всего 1763 "не пульсаров"
не пульсар предсказанный как не пульсар:  1743
не пульсар предсказанный как пульсар : 20
пульсар предсказанный как не пульсар : 10
пульсар предсказанный как пульсар :  17
количество правильных ответов 1760 из 1790
```




```
loss :  0.20917624711324384
accuracy :  0.975419

всего 27 "пульсаров"
всего 1763 "не пульсаров"
не пульсар предсказанный как не пульсар:  1722
не пульсар предсказанный как пульсар : 41
пульсар предсказанный как не пульсар : 3
пульсар предсказанный как пульсар :  24
количество правильных ответов 1746 из 1790
```



Итоги

1. Изменение веса класса дает существенное улучшение результатов.
2. Применение центровки не дает существенных изменений.
3. Применение Feature Engineering даёт ухудшение результатов, поскольку он более пригоден для датасетов с несколькими сущностями, имеющими так называемые дочерние-родительские отношения, например одна таблица с данными о клиентах, которые делают транзакции, другая таблица с теми покупками, которые могут быть приобретены через эти транзакции.

Список использованной литературы

1. [1] - www.tensorflow.org/tutorials/structured_data/imbalanced_data
2. [2] - medium.com/dataexplorations/tool-review-can-featuretools-simplify-the-process-of-feature-engineering-5d165100b0c3