

Отчет по заданию в Kaggle

Гончарова Наталия, группа БПМ-17-2

Декабрь 2020 г.

1 Постановка задачи

Решение задачи с Kaggle с использованием нейронной сети. Ссылка на задачу - <https://www.kaggle.com/nasa/kepler-exoplanet-search-results>

2 Входные переменные

Ознакомимся со структурой файла **kepler-exoplanet-search-results.csv**

rowid	kepid	kepoi_name	kepler_name	koi_disposition	koi_pdisposition	koi_score	koi_fpflag_nt	koi_fpflag_ss	koi_fpflag_co	...	koi_steff_err2	koi_
0	1	10797460	K00752.01	Kepler-227 b	CONFIRMED	CANDIDATE	1.000	0	0	0	...	-81.0
1	2	10797460	K00752.02	Kepler-227 c	CONFIRMED	CANDIDATE	0.969	0	0	0	...	-81.0
2	3	10811496	K00753.01	NaN	FALSE POSITIVE	FALSE POSITIVE	0.000	0	1	0	...	-176.0
3	4	10848459	K00754.01	NaN	FALSE POSITIVE	FALSE POSITIVE	0.000	0	1	0	...	-174.0
4	5	10854555	K00755.01	Kepler-664 b	CONFIRMED	CANDIDATE	1.000	0	0	0	...	-211.0

Рис. 1: Структура файла

3 Решение

Посмотрим на распределение параметров **koi_pdisposition** и **koi-disposi**.

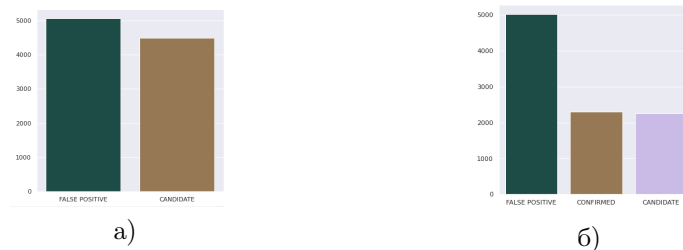


Рис. 2: Распределение параметров

Перекодируем категориальные признаки для дальнейшего анализа. Реализуем функцию, которая принимает на вход DataFrame, кодирует числовыми значениями категориальные признаки и возвращает обновленный DataFrame и сами кодировщики.

	rowid	kepid	kepoi_name	koi_disposition	koi_pdisposition	koi_score	koi_fpflag_nt	koi_fpflag_ss	koi_fpflag_co	koi_fpflag_ec	...	koi_steff_err2	koi_
0	1	10797460	1080	1	0	1.000	0	0	0	0	...	-81.0	
1	2	10797460	1081	1	0	0.969	0	0	0	0	...	-81.0	
2	3	10811496	1082	2	1	0.000	0	1	0	0	...	-176.0	
3	4	10848459	1083	2	1	0.000	0	1	0	0	...	-174.0	
4	5	10854555	1084	1	0	1.000	0	0	0	0	...	-211.0	

Рис. 3: Категориальные признаки

Построим матрицу корреляций:

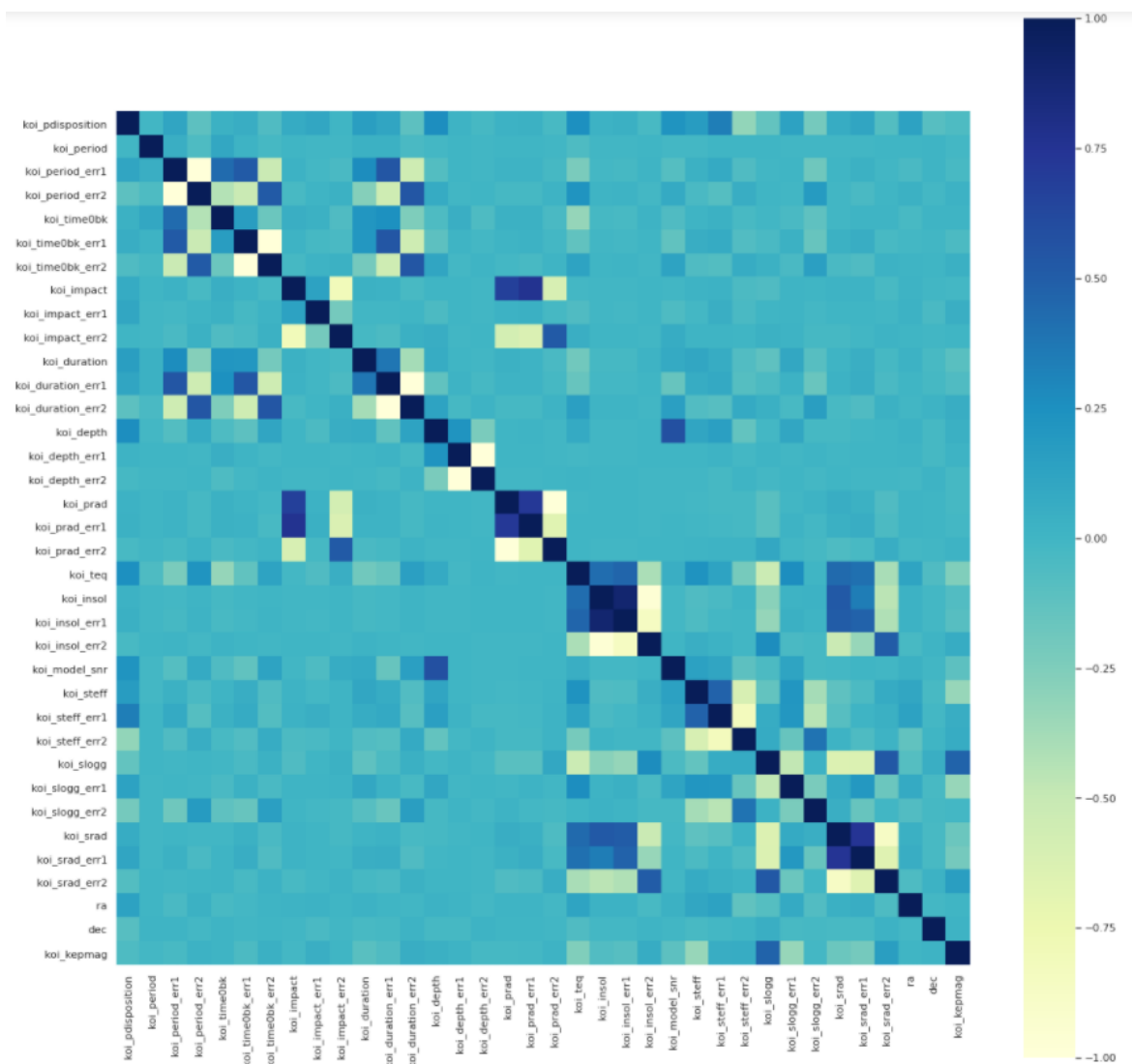


Рис. 4: Матрица корреляций

Посмотрим на распределение величин по признакам в данных:

Задаются следующие характеристики:

1. на первом слое содержится 128 нейронов с relu функцией активации;
2. на втором слое содержится 64 нейронов с relu функцией активации;
3. на третьем слое содержится 64 нейрона с relu функцией активации;
4. на третьем слое содержится 64 нейрона с relu функцией активации;
5. на выходное слое содержится 1 нейрон с сигмоидной функцией;

В качестве метода оптимизации был выбран adam. Было проведено обучение на 100 эпохах с 32 батчами

Посмотрим на графики изменения ошибки и точности по эпохе тренировочной и валидационных выборках.

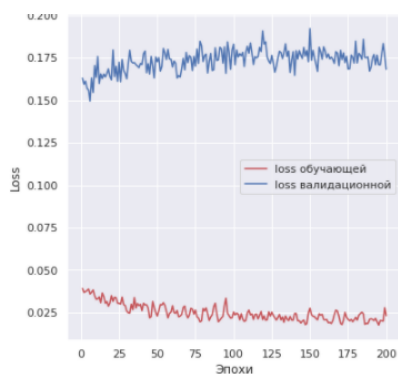


Рис. 6: Графики изменения ошибки

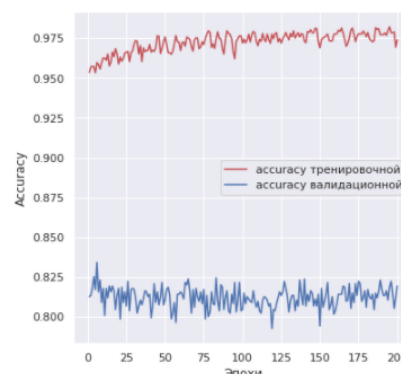


Рис. 7: Графики изменения точности.

Получим пример предсказания:

	precision	recall	f1-score	support
0	0.84	0.80	0.82	679
1	0.81	0.84	0.82	670
accuracy			0.82	1349
macro avg	0.82	0.82	0.82	1349
weighted avg	0.82	0.82	0.82	1349

Рис. 8: Пример предсказания

4 Вывод

В результате данной работы была написана и обучена нейронная сеть с использованием библиотеки keras для предсказывания, а также построены графики обучения.