16 - Kaggle Competition

Data Science Practicum 2021/22, Lesson 16

Marko Tkalčič

Univerza na Primorskem

Table of Contents

Kaggle Competition

- Deadline: 18. January 2022 23:59
- Presentations: 19. January 2022 in-class
- Max 20 submission/day
- Public:Private test sets = 50:50

Public vs. Private Leaderboard

- Leaderboard fitting: public vs private test sets
- Public Leaderboard: 50% of the test set
  - is just an estimate of the final score
- Private leaderboard: other 50% of the test set
  - used to calculate the final score and ranking
  - three submissions selected by the competitors
  - highest scored used to compute the final score
  - may be different than the public leaderboard
- To Read:
  - https://www.quora.com/What-is-the-difference-between-public-and-private-leaderboard-in-Kaggle
  - http://gregpark.io/blog/Kaggle-Psychopathy-Postmortem/
  - https://www.kdnuggets.com/2015/05/data-science-contest-leaderboard-without-reading-data.html

Public vs. Private Leaderboard

1. Kaggle competitions are decided by your model's performance on a test data set. Kaggle has the answers for this data set, but withholds them to compare with your predictions. Your Public score is what you receive back upon each submission (that score is calculated using a statistical evaluation metric, which is always described on the Evaluation page). BUT: Your Public Score is being determined from only a fraction of the test data set – usually between 25-33%. This is the Public Leaderboard, and it shows some relative performance during the competition.
2. When the competition ends, we take your selected submissions (see below) and score your predictions against the REMAINING FRACTION of the test set, or the private portion. You never receive ongoing feedback about your score on this portion, so it is the Private leaderboard. Final competition results are based on the Private leaderboard, and the Winner is the person(s) at the top of the Private Leaderboard. Why? This separation of the test set into public and private portions is what ensures that the most accurate but generalized model is the one that wins the challenge. If you based your model solely on the data which gave you constant feedback, you run the danger of a model that overfits to the specific noise in that data. One of the hard challenges in data science is to avoid overfitting, by leaving your model flexible to out-of-sample data.

Dataset

	id	season	stage	date	home_team_api_id	away_team_api_id	home_team_goal	away_team_goal	B365H	B365D	...	VCH	VCD	VCA	GBH	GBD	GBA	BSH	BSD	BSA	outcome
0	1729	2008/2009	1	2008-08-17 00:00:00	1	13	1	1	1.29	5.5	...	1.28	5.5	12.00	1.30	4.75	10.00	1.29	4.50	11.00	0
1	1730	2008/2009	1	2008-08-16 00:00:00	2	16	1	0	1.20	6.5	...	1.25	6.0	13.00	1.22	5.50	13.00	1.22	5.00	13.00	1
2	1731	2008/2009	1	2008-08-16 00:00:00	3	12	0	1	5.50	3.6	...	5.50	3.8	1.65	5.00	3.40	1.70	4.50	3.40	1.73	-1
3	1732	2008/2009	1	2008-08-16 00:00:00	4	20	2	1	1.91	3.4	...	1.90	3.5	4.35	1.91	3.25	4.00	1.91	3.25	3.80	1
4	1733	2008/2009	1	2008-08-17 00:00:00	5	18	4	2	1.91	3.4	...	1.90	3.5	4.35	1.91	3.25	4.00	1.91	3.30	3.75	1
5	1734	2008/2009	1	2008-08-16 00:00:00	6	19	2	3	2.00	3.3	...	2.05	3.3	4.00	2.00	3.25	3.75	2.00	3.25	3.50	-1
6	1735	2008/2009	1	2008-08-16 00:00:00	7	15	2	1	3.20	3.4	...	3.20	3.4	2.30	3.00	3.25	2.30	2.80	3.25	2.30	1
7	1736	2008/2009	1	2008-08-16 00:00:00	8	11	3	1	1.83	3.5	...	1.85	3.4	4.80	1.83	3.25	4.50	1.80	3.25	4.33	1
8	1737	2008/2009	1	2008-08-16 00:00:00	9	14	2	1	2.60	3.2	...	2.60	3.4	2.80	2.60	3.25	2.60	2.60	3.25	2.50	1
9	1738	2008/2009	1	2008-08-17 00:00:00	10	17	4	0	1.33	5.0	...	1.33	5.0	11.00	1.33	4.75	9.00	1.33	4.20	10.00	1

Submission File

id	outcome
4509	1
4510	-1
4511	-1
4512	0
4513	1
...	...
4708	1

Table of Contents

Assessment

- 50% Weekly Assignments (pass/fail)
- 50% Kaggle Competition
  - Thresholds (90%)
  - Ranking (10%)

Assessment

- 50% Weekly Assignments (pass/fail)
- 50% Kaggle Competition
  - Thresholds (90%)
  - Ranking (10%)

- Thresholds (0-90 points):
  - No submission: Fail Exam
  - < Random : 0
  - > 1 baseline: 30 points
  - > 2 baselines: 50 points
  - > 3 baselines: 85 points
  - > 4 baselines: 90 points
- Ranking (up to 10 points):
  - 1st, 2nd, 3rd etc -> 10,9,8, etc points

Table of Contents

Exercise

- Register/login to Kaggle
- Competition page: https://www.kaggle.com/c/up-data-practicum-ii-202122-football/host/launch-checklist
- Read

Exercise

- Register/login to Kaggle
- Competition page: https://www.kaggle.com/c/up-data-practicum-ii-202122-football/host/launch-checklist
- Read

- Download Data
- Load train, test sets and submission example into dataframes
- Inspect the dataframes

Exercise

- Create four result files:
  - all 0
  - all 1
  - all -1
  - random
- Submit

Exercise

- Create four result files:
  - all 0
  - all 1
  - all -1
  - random
- Submit

```
# Random Baseline
import numpy as np
import pandas as pd
df_test = pd.read_csv("test.csv")
answer = pd.DataFrame(df_test["id"])
answer["outcome"] = np.random.randint(-1,2,size=(len(answer["id"]), 1))
answer.to_csv("random_baseline.csv",index=False)
answer.head()
print(answer)
```

```
       id  outcome
0    4509        0
1    4510       -1
2    4511       -1
3    4512       -1
4    4513       -1
..    ...      ...
185  4704        0
186  4705        0
187  4706        1
188  4707        0
189  4708        1

[190 rows x 2 columns]
```

Table of Contents

References

Part of the material has been taken from the following sources. The usage of the referenced copyrighted work is in line with fair use since it is for nonprofit educational purposes.

- https://towardsdatascience.com/preprocessing-with-sklearn-a-complete-and-comprehensive-guide-670cb98fcfb9
- https://stackoverflow.com/questions/18691084/what-does-1-mean-in-numpy-reshape