Kaggle_Data_Competition_Final_Project

January 29, 2026

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

pd.set_option('display.max_rows', 1000)
```

[2]:
```
df = pd.read_csv("./train_competition_2026.csv")
```

1 Kaggle Data Competition Project

1.0.1 Project Description

Context: Medical data from patients in the ER.

There are multiple timepoints observed per patient. Some of the predictors are constant in time, some are dynamic. The responses y_1 and y_2 are the measurements of indicators of health (e.g. arterial pressure, heart rate) recorded after a 5 minute gap from the predictors. These responses are usually a sign of whether someone's condition is deterioriating to an extreme point.

We want to build a model to predict these values so we can know when someone's condition is at risk 5 minutes before it actually happens. Your task for this project is to build a machine learning model which takes in the predictors and returns predictions for y_1 and y_2. At each of the checkpoints (2/17 at 11:59p, 3/2 at 11:59p, and 3/11 at 11:59p) you will use your model to predict the 2 responses for a test set. Your relative performance (based on MAE) compared to your competitors will dictate part of your score for the project.

1.0.2 Grading Breakdown

Evaluation of your project will be determined by the following breakdown:

1. Relative performance at first checkpoint, Feb 17 (5 pts)

```
- Top 25% of scorers on test set will receive 5/5 points
- Next 50% of scorers will receive 4.5/5
- Bottom 25% of scorers will receive 4/5, conditional on meeting a minimum threshold average l
    the private test set
- Failure to beat an avg MAE of 10 on the private test set will receive 2.5/5
- No submission = 0/5
```

2. Relative performance at second checkpoint, Mar 2 (5 pts)

- Top 25% of scorers on test set will receive 5/5 points
- Next 50% of scorers will receive 4/5
- Bottom 25% of scorers will receive 3/5, conditional on meeting a minimum threshold average
  the private test set
- Failure to improve upon your MAE from the first checkpoint on the private test set will rec
- No submission = 0/5

3. Relative performance at final due date, Mar 11 (10 pts)

- Top 25% of scorers on test set will receive 10/10 points
- Next 50% of scorers will receive 8/10
- Bottom 25% of scorers will receive 6/10

4. Write up of your model development process in a blog (NOT hosted online – submit a .html or .pdf over Canvas). (60 pts)

- Components of the blog:
- Sections separated by checkpoint and what models you had explored at that point /5
- Visualization & EDA -- should at least be in first section (or in multiple sections if you c
- Description of feature engineering you used /10
- Model comparison & validation across candidate models + vizualization of results /20
- Summary of why you think your winning model performed the best /10
- Description of division of work across partners /5

Feel free to include other details in your blog like navigating computational issues, bug fixi

5. Peer evaluation of partners' teamwork, attendance in class, and contributions to model development blogs (20 pts)

Extra. The top 5 teams will have the opportunity to present their models and development process on the final day of class. This presentation can add up 1 extra credit to your overall course grade.

1.0.3 First checkpoint

Your first task should be to perform EDA and construct and tune a baseline model which has an average MAE of 10 or less. The due date for this checkpoint will be Feb 17.

[3]:
```
df.head(10)
```

[3]:

	obs	sub_id	time	num_0	num_1	num_2	cat_0	cat_1	cat_2	\
0	0	0	2068-09-19 23:34:11	1.38	49	7	1	3	1	
1	0	0	2068-09-19 23:35:11	1.38	49	7	1	3	1	
2	0	0	2068-09-19 23:36:11	1.38	49	7	1	3	1	
3	0	0	2068-09-19 23:37:11	1.38	49	7	1	3	1	

```
4    0       0  2068-09-19 23:38:11   1.38     49      7      1      3      1
5    0       0  2068-09-19 23:39:11   1.38     49      7      1      3      1
6    0       0  2068-09-19 23:40:11   1.38     49      7      1      3      1
7    0       0  2068-09-19 23:41:11   1.38     49      7      1      3      1
8    0       0  2068-09-19 23:42:11   1.38     49      7      1      3      1
9    0       0  2068-09-19 23:43:11   1.38     49      7      1      3      1
```

	cat_3	cat_4	t_0	t_1	t_2	t_3	t_4	y_1	y_2
0	0	1	105.5	95.0	67.4	36.6	23.2	33.4	107.4
1	0	1	104.4	95.0	66.4	37.8	22.7	33.4	107.4
2	0	1	104.0	95.0	65.2	37.0	22.1	33.4	107.4
3	0	1	102.8	95.0	63.4	35.9	20.7	33.4	107.4
4	0	1	101.3	95.1	59.1	34.5	18.1	33.4	107.4
5	0	1	99.7	94.8	56.8	33.4	16.5	33.4	107.4
6	0	1	101.2	94.7	60.8	35.4	19.5	33.4	107.4
7	0	1	99.9	94.2	64.3	36.2	20.7	33.4	107.4
8	0	1	107.2	94.8	67.7	38.3	24.2	33.4	107.4
9	0	1	109.0	95.4	68.0	35.5	22.9	33.4	107.4

1.1 Test set data predictors, without y_1, y_2

[4]:
```
ex_test = pd.read_csv("./test_no_outcome.csv")

ex_test.head(50)
```

[4]:

	obs	sub_id	time	num_0	num_1	num_2	cat_0	cat_1	\
0	18	1	2134-04-01 22:23:14	-1.0	38	1	1	1	
1	18	1	2134-04-01 22:24:14	-1.0	38	1	1	1	
2	18	1	2134-04-01 22:25:14	-1.0	38	1	1	1	
3	18	1	2134-04-01 22:26:14	-1.0	38	1	1	1	
4	18	1	2134-04-01 22:27:14	-1.0	38	1	1	1	
5	18	1	2134-04-01 22:28:14	-1.0	38	1	1	1	
6	18	1	2134-04-01 22:29:14	-1.0	38	1	1	1	
7	18	1	2134-04-01 22:30:14	-1.0	38	1	1	1	
8	18	1	2134-04-01 22:31:14	-1.0	38	1	1	1	
9	18	1	2134-04-01 22:32:14	-1.0	38	1	1	1	
10	18	1	2134-04-01 22:33:14	-1.0	38	1	1	1	
11	18	1	2134-04-01 22:34:14	-1.0	38	1	1	1	
12	18	1	2134-04-01 22:35:14	-1.0	38	1	1	1	
13	18	1	2134-04-01 22:36:14	-1.0	38	1	1	1	
14	18	1	2134-04-01 22:37:14	-1.0	38	1	1	1	
15	18	1	2134-04-01 22:38:14	-1.0	38	1	1	1	
16	18	1	2134-04-01 22:39:14	-1.0	38	1	1	1	
17	18	1	2134-04-01 22:40:14	-1.0	38	1	1	1	
18	18	1	2134-04-01 22:41:14	-1.0	38	1	1	1	
19	18	1	2134-04-01 22:42:14	-1.0	38	1	1	1	
20	18	1	2134-04-01 22:43:14	-1.0	38	1	1	1	

21	18	1	2134-04-01 22:44:14	-1.0	38	1	1	1
22	18	1	2134-04-01 22:45:14	-1.0	38	1	1	1
23	18	1	2134-04-01 22:46:14	-1.0	38	1	1	1
24	18	1	2134-04-01 22:47:14	-1.0	38	1	1	1
25	18	1	2134-04-01 22:48:14	-1.0	38	1	1	1
26	18	1	2134-04-01 22:49:14	-1.0	38	1	1	1
27	18	1	2134-04-01 22:50:14	-1.0	38	1	1	1
28	18	1	2134-04-01 22:51:14	-1.0	38	1	1	1
29	18	1	2134-04-01 22:52:14	-1.0	38	1	1	1
30	19	1	2134-04-01 23:58:14	-1.0	38	1	1	1
31	19	1	2134-04-01 23:59:14	-1.0	38	1	1	1
32	19	1	2134-04-02 00:00:14	-1.0	38	1	1	1
33	19	1	2134-04-02 00:01:14	-1.0	38	1	1	1
34	19	1	2134-04-02 00:02:14	-1.0	38	1	1	1
35	19	1	2134-04-02 00:03:14	-1.0	38	1	1	1
36	19	1	2134-04-02 00:04:14	-1.0	38	1	1	1
37	19	1	2134-04-02 00:05:14	-1.0	38	1	1	1
38	19	1	2134-04-02 00:06:14	-1.0	38	1	1	1
39	19	1	2134-04-02 00:07:14	-1.0	38	1	1	1
40	19	1	2134-04-02 00:08:14	-1.0	38	1	1	1
41	19	1	2134-04-02 00:09:14	-1.0	38	1	1	1
42	19	1	2134-04-02 00:10:14	-1.0	38	1	1	1
43	19	1	2134-04-02 00:11:14	-1.0	38	1	1	1
44	19	1	2134-04-02 00:12:14	-1.0	38	1	1	1
45	19	1	2134-04-02 00:13:14	-1.0	38	1	1	1
46	19	1	2134-04-02 00:14:14	-1.0	38	1	1	1
47	19	1	2134-04-02 00:15:14	-1.0	38	1	1	1
48	19	1	2134-04-02 00:16:14	-1.0	38	1	1	1
49	19	1	2134-04-02 00:17:14	-1.0	38	1	1	1

	cat_2	cat_3	cat_4	t_0	t_1	t_2	t_3	t_4
0	0	0	0	105.4	99.8	50.7	61.4	36.8
1	0	0	0	105.4	99.4	49.4	61.1	36.2
2	0	0	0	104.6	99.0	49.7	61.4	36.6
3	0	0	0	104.5	99.6	51.7	61.8	37.2
4	0	0	0	104.6	99.5	52.5	61.9	37.5
5	0	0	0	102.8	98.9	60.8	66.0	42.7
6	0	0	0	103.3	100.0	58.2	63.9	40.1
7	0	0	0	104.5	100.0	63.2	67.0	43.9
8	0	0	0	108.1	99.8	69.7	72.0	49.3
9	0	0	0	107.7	98.8	65.6	69.7	46.4
10	0	0	0	108.5	97.0	61.9	66.6	43.5
11	0	0	0	108.0	96.9	57.3	64.5	41.0
12	0	0	0	106.2	97.1	60.2	65.7	42.6
13	0	0	0	108.0	97.8	56.1	63.9	40.3
14	0	0	0	108.4	98.0	53.1	62.3	38.3
15	0	0	0	108.2	98.1	51.4	61.5	37.2

16	0	0	0	108.0	99.0	49.7	60.4	35.9
17	0	0	0	106.8	98.4	50.4	60.6	36.1
18	0	0	0	106.8	98.8	52.1	61.0	36.8
19	0	0	0	107.0	99.1	53.4	61.2	37.1
20	0	0	0	104.3	99.0	60.3	64.5	41.4
21	0	0	0	104.1	100.0	63.5	65.6	43.1
22	0	0	0	105.0	100.0	57.7	63.0	39.5
23	0	0	0	106.4	100.0	51.1	61.0	36.5
24	0	0	0	105.7	100.0	47.8	60.8	35.8
25	0	0	0	106.1	99.9	46.4	60.4	35.1
26	0	0	0	105.3	100.0	48.9	61.3	36.3
27	0	0	0	105.4	100.0	51.8	62.1	37.6
28	0	0	0	104.1	99.9	56.7	63.3	39.5
29	0	0	0	105.2	100.0	57.2	62.9	39.3
30	0	0	0	103.7	100.0	40.1	53.5	28.6
31	0	0	0	102.5	100.0	43.1	55.1	30.7
32	0	0	0	101.9	100.0	45.4	55.3	31.5
33	0	0	0	102.4	100.0	43.4	54.6	30.4
34	0	0	0	102.5	100.0	44.3	55.1	31.1
35	0	0	0	102.2	100.0	43.1	54.5	30.2
36	0	0	0	101.5	100.0	46.3	56.0	32.3
37	0	0	0	101.8	100.0	43.6	54.6	30.3
38	0	0	0	102.3	100.0	41.5	54.1	29.4
39	0	0	0	102.3	100.0	42.6	54.6	30.1
40	0	0	0	102.2	100.0	41.5	54.0	29.5
41	0	0	0	102.6	100.0	40.7	54.0	29.4
42	0	0	0	102.2	100.0	41.0	53.8	29.3
43	0	0	0	101.4	100.0	40.9	53.9	29.2
44	0	0	0	102.1	100.0	39.8	53.6	28.6
45	0	0	0	102.0	99.8	41.0	54.0	29.5
46	0	0	0	101.8	99.9	41.1	54.1	29.5
47	0	0	0	102.0	99.8	40.2	53.6	28.9
48	0	0	0	101.8	99.6	40.5	53.8	29.3
49	0	0	0	101.4	99.3	40.6	54.0	29.3

1.2 Example Submission Format

[5]:
```
ex_submission = pd.read_csv("./sample_submission.csv")

ex_submission.head(50)
```

[5]:

	obs	y_1	y_2
0	18	42.0	82.0
1	19	42.0	82.0
2	20	42.0	82.0
3	21	42.0	82.0
4	22	42.0	82.0

```
5    23  42.0  82.0
6    24  42.0  82.0
7    25  42.0  82.0
8    26  42.0  82.0
9    27  42.0  82.0
10   28  42.0  82.0
11   29  42.0  82.0
12   30  42.0  82.0
13   31  42.0  82.0
14   32  42.0  82.0
15   33  42.0  82.0
16  127  42.0  82.0
17  128  42.0  82.0
18  129  42.0  82.0
19  146  42.0  82.0
20  147  42.0  82.0
21  148  42.0  82.0
22  149  42.0  82.0
23  150  42.0  82.0
24  151  42.0  82.0
25  152  42.0  82.0
26  153  42.0  82.0
27  154  42.0  82.0
28  195  42.0  82.0
29  196  42.0  82.0
30  197  42.0  82.0
31  198  42.0  82.0
32  199  42.0  82.0
33  200  42.0  82.0
34  201  42.0  82.0
35  202  42.0  82.0
36  203  42.0  82.0
37  204  42.0  82.0
38  205  42.0  82.0
39  206  42.0  82.0
40  207  42.0  82.0
41  208  42.0  82.0
42  209  42.0  82.0
43  210  42.0  82.0
44  211  42.0  82.0
45  212  42.0  82.0
46  213  42.0  82.0
47  214  42.0  82.0
48  215  42.0  82.0
49  216  42.0  82.0
```

[ ]: