Common validation problems

Validation

1. We discussed the concept of validation and overfitting
2. We understood how to choose validation strategy
3. We learned to identify data split made by organizers.

Validation

1. We discussed the concept of validation and overfitting
2. We understood how to choose validation strategy
3. We learned to identify data split made by organizers.
4. Validation problems
   a. Validation stage
   b. Submission stage

Validation stage

Holidays in Russia

January

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

February

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28				

8 holidays

14 weekend

12 working days

Validation stage

Causes of different scores and optimal parameters

1. Too little data
2. Too diverse and inconsistent data

We should do extensive validation

1. Average scores from different KFold splits
2. Tune model on one split, evaluate score on the other

Validation stage: extensive validation

- Liberty Mutual Group:
  Property Inspection Prediction

In both of them, scores of the competitors were very close to each other. And thus participants tried to squeeze more from the data. But do not overfit, so the thorough validation was crucial.

- Santander Customer Satisfaction

Submission stage

We can observe that:

- LB score is consistently higher/lower that validation score
- LB score is not correlated with validation score at all

Now remember that the main rule of making a reliable validation, is to mimic a train tests pre made by organizers.

I won't lie to you, it can be quite hard to identify and mimic the exact train tests here. Because of that, I highly recommend you to start submitting your solutions right after you enter the competition.

Submission stage

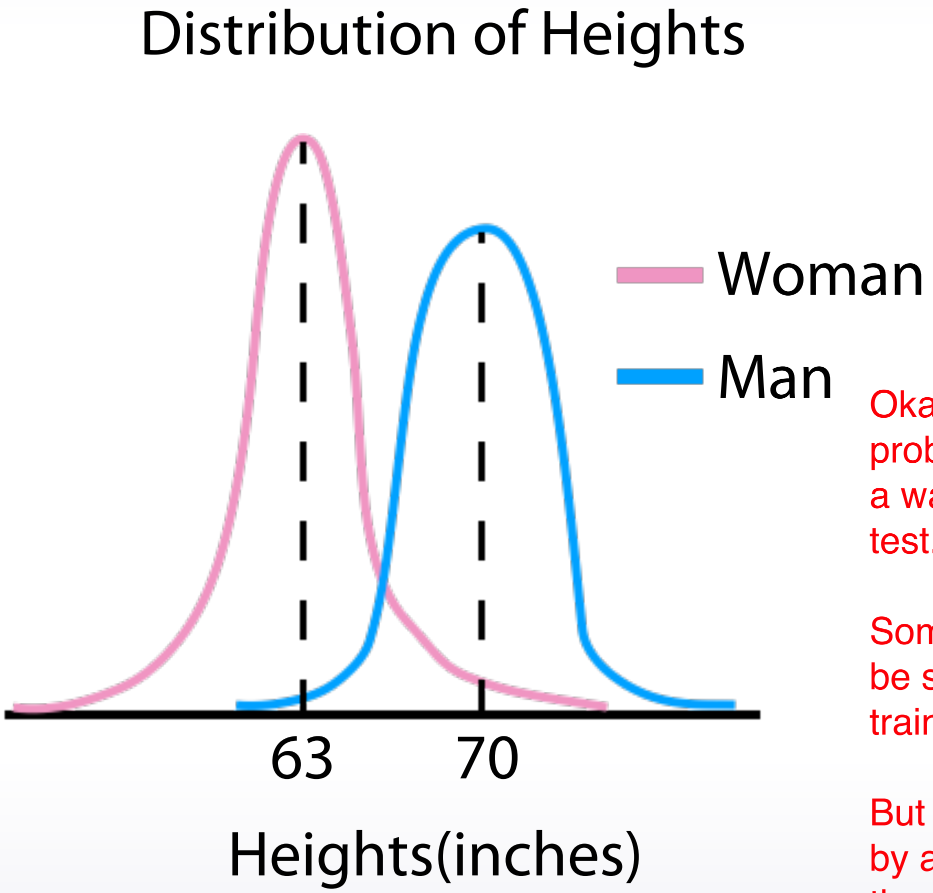
0. We may already have quite different scores in Kfold

Other reasons:

Here it is useful to see a leaderboard as another validation fold. Then, if we already have different scores in KFold, getting a not very similar result on the leaderboard is not suprising. More we can calculate mean and standard deviation of the validation scores and estimate if the leaderboard score is expected. But if this is not the case, then something is definitely wrong.

1. too little data in public leaderboard
2. train and test data are from different distributions

Submission stage: different distributions

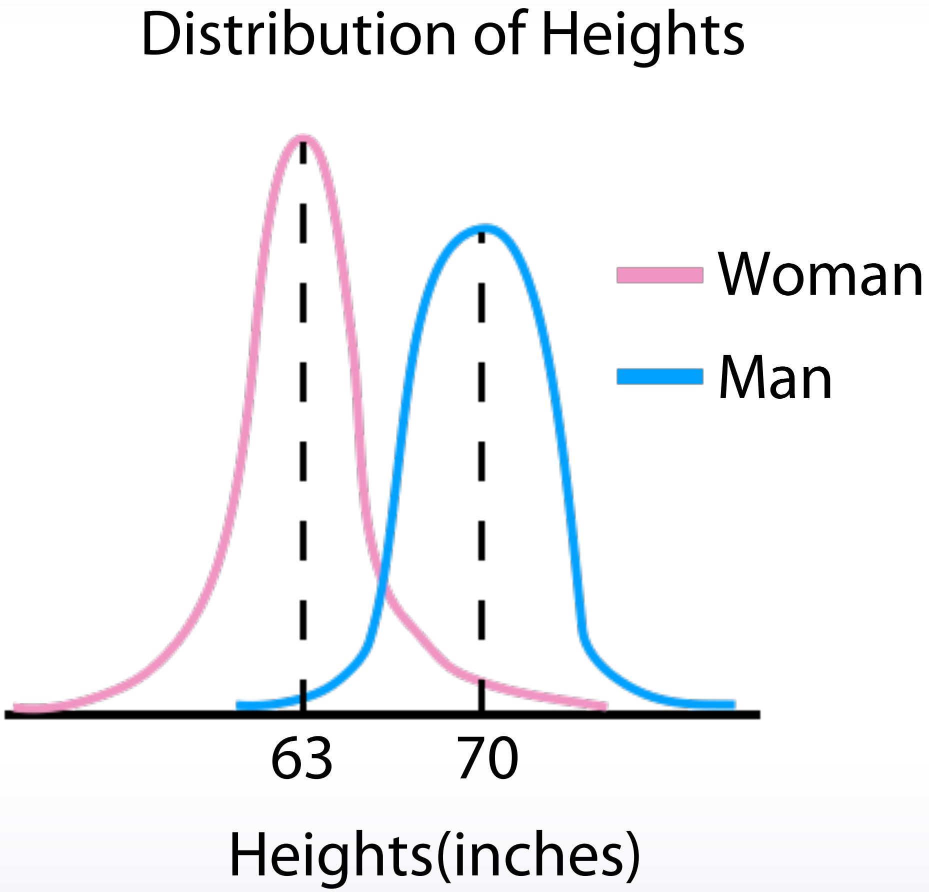


Okay, let's start with a general approach to such problems. At the broadest level, we need to find a way to tackle different distributions in train and test.

Sometimes, these kind of problems could be solved by adjusting your solution during the training procedure.

But sometimes, this problem can be solved only by adjusting your solution through the leaderboard.

Submission stage: different distributions

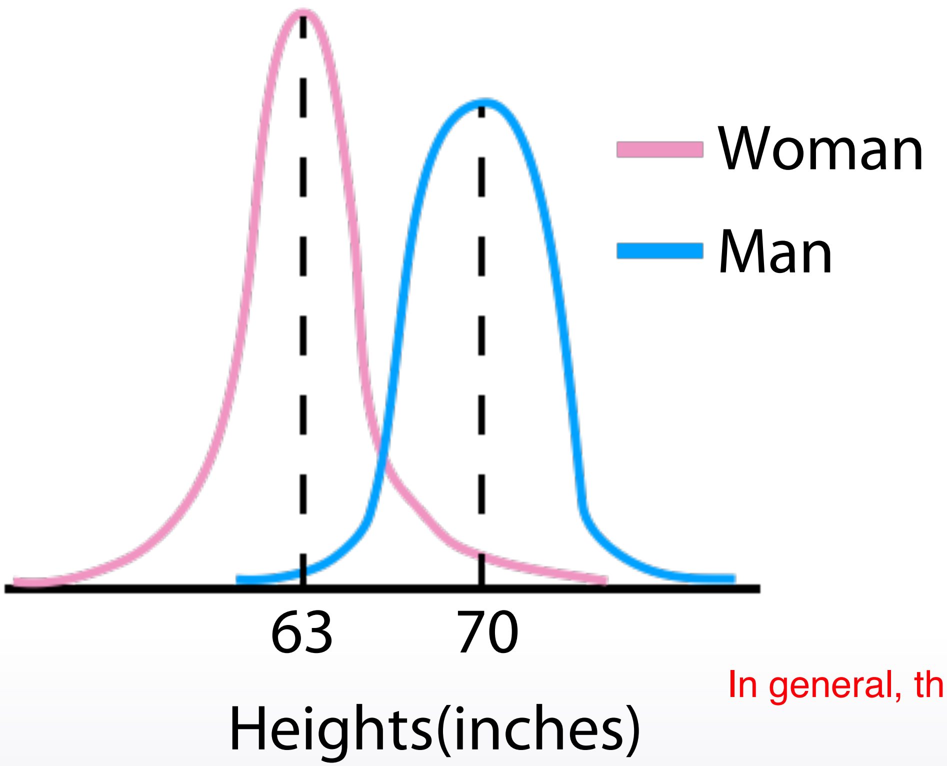


- Mean for train:
  Calculate from the train data

- Mean for test:
  Leaderboard probing

Submission stage: different distributions

Distribution of Heights

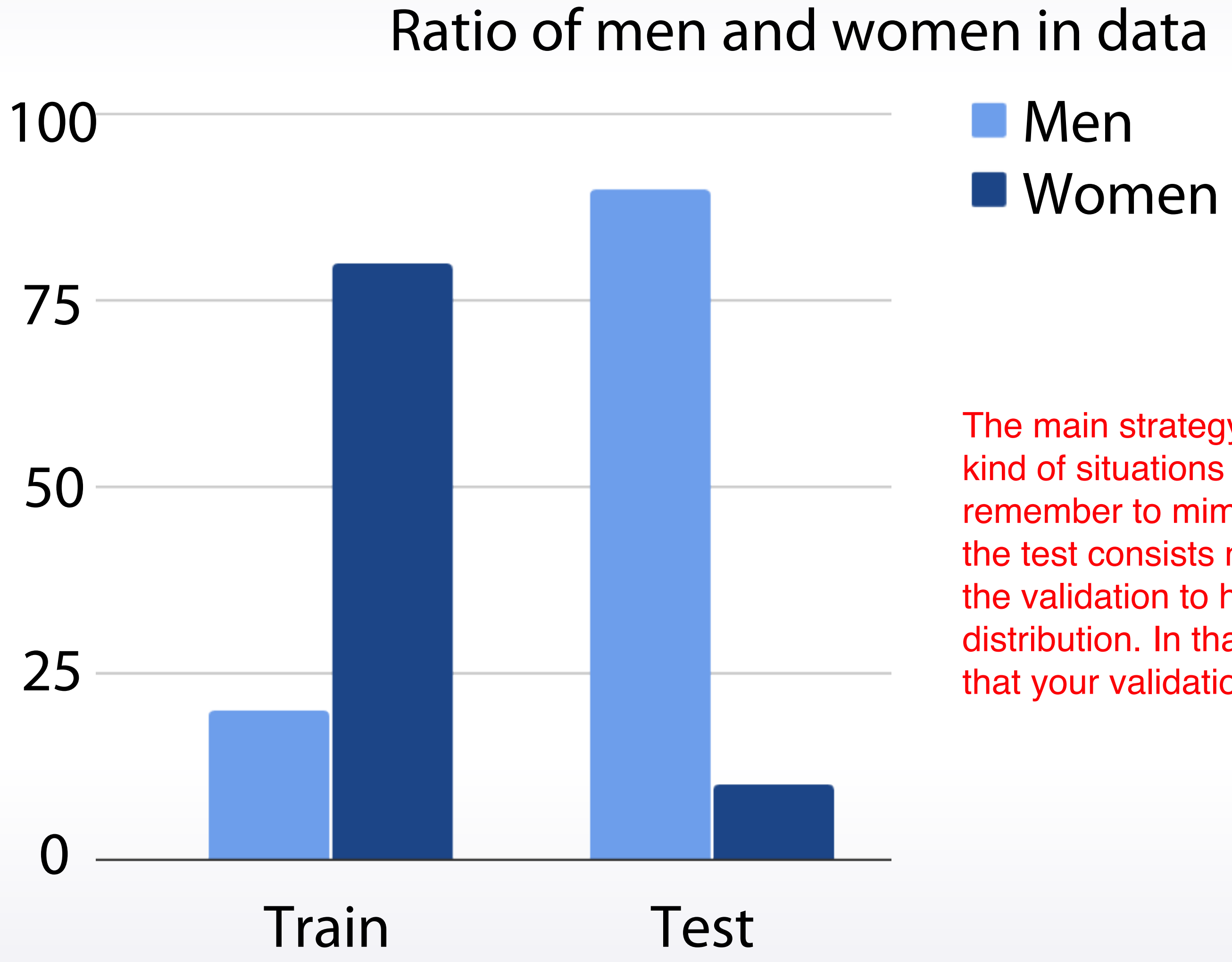


Quora Question Pairs

Quora

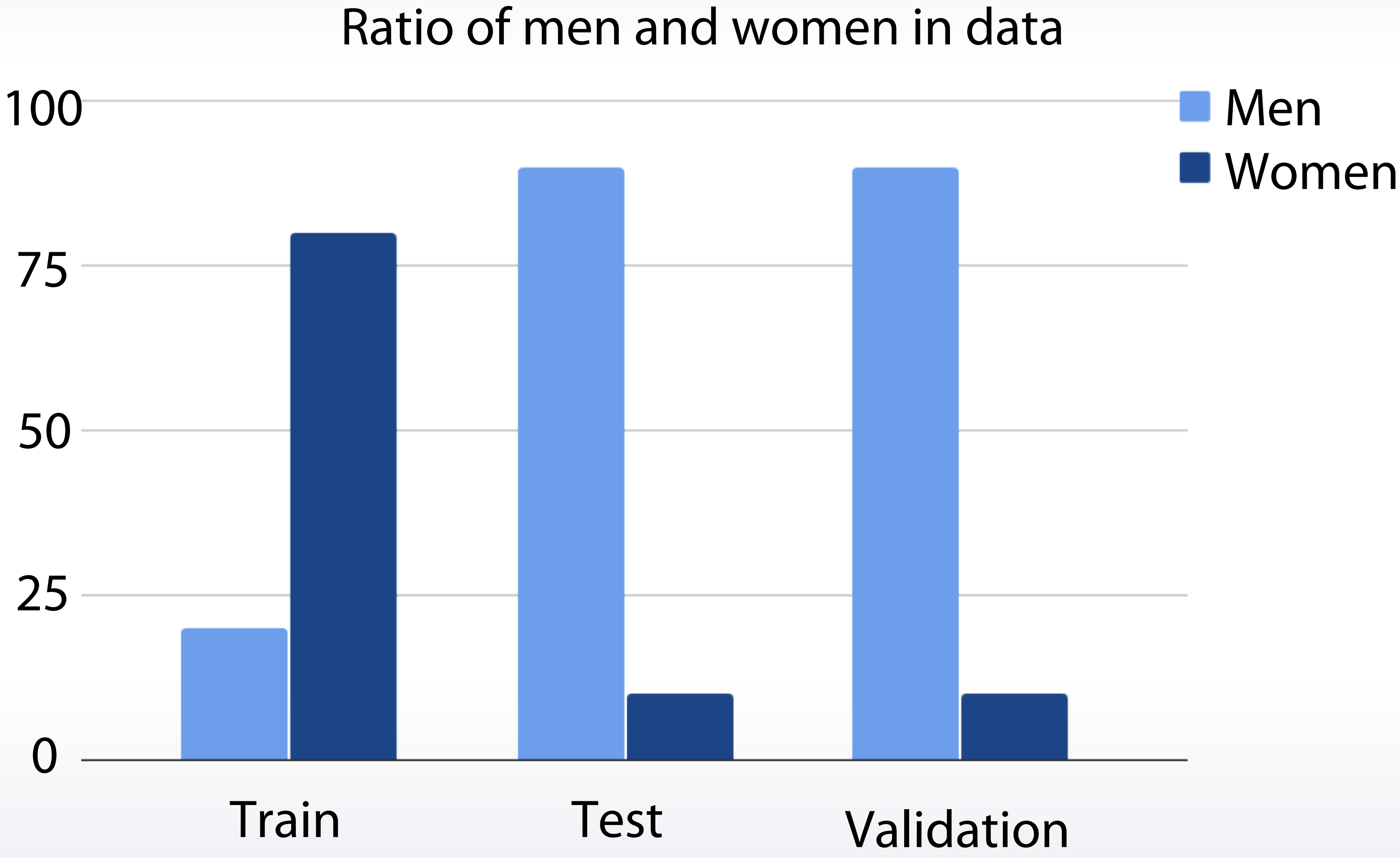
In general, this technique is known as leaderboard probing

Submission stage: different distributions



The main strategy to deal with these kind of situations is simple. Again, remember to mimic the train test split. If the test consists mostly of Men, force the validation to have the same distribution. In that case, you ensure that your validation will be fair.

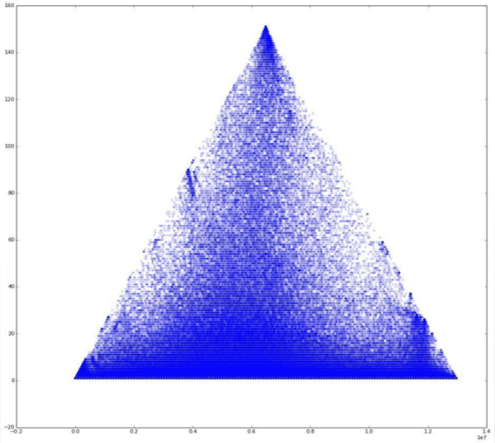
Submission stage: different distributions



Submission stage: different distributions

- Data Science Game 2017 Qualification phase: Music recommendation

- CTR prediction task from EDA



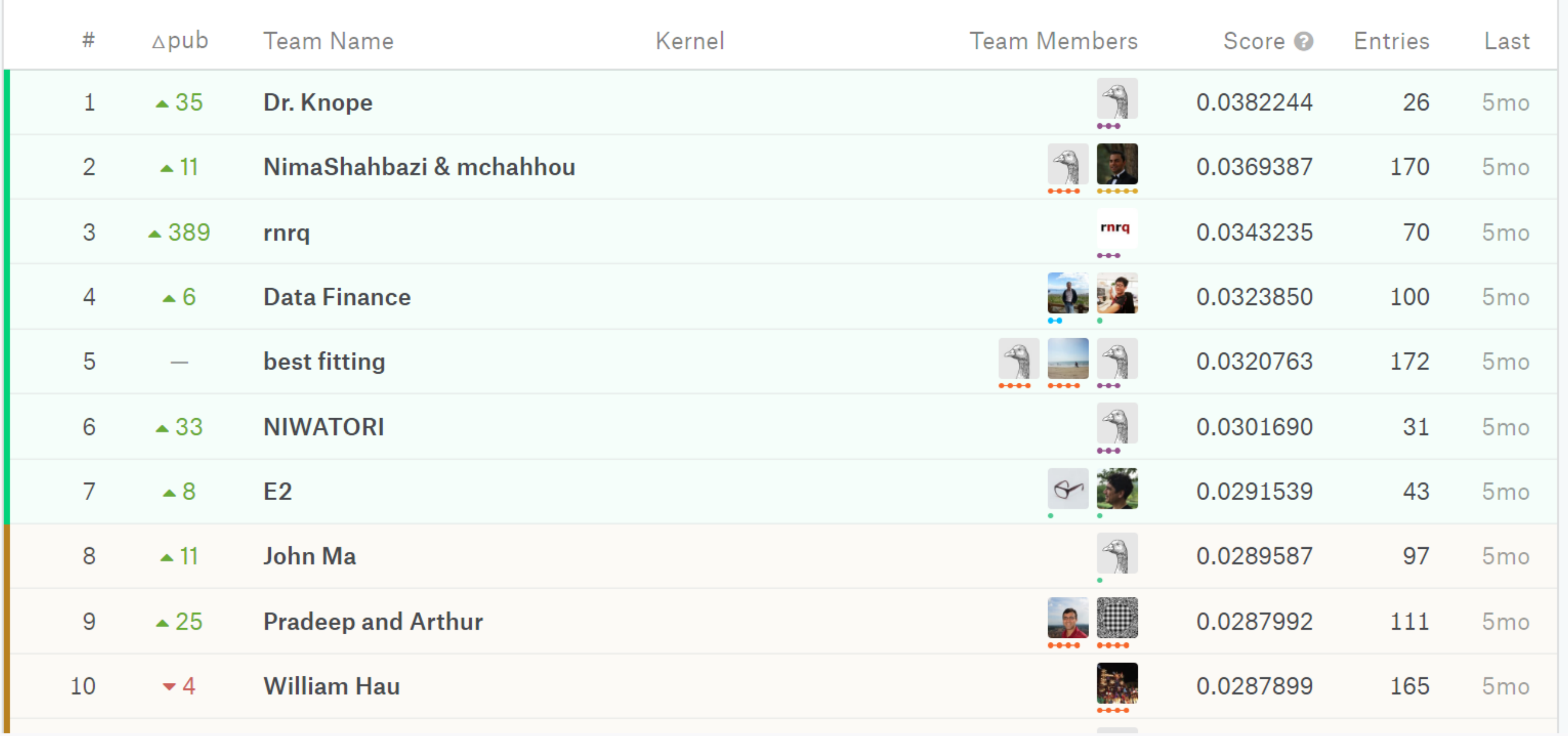
CTR. So, the train data, which basically was the history of displayed ads obviously didn't contain ads which were not shown. On the contrary, the test data consisted of every possible ad. Notice this is the exact case of different distributions in train and test.

Submission stage

Causes of validation problems:

- too little data in public leaderboard
- incorrect train/test split
- different distributions in train and test

LB shuffle



#	∆pub	Team Name	Kernel	Team Members	Score ?	Entries	Last
1	▲35	Dr. Knope			0.0382244	26	5mo
2	▲11	NimaShahbazi & mchahhou			0.0369387	170	5mo
3	▲389	rnrq		rnrq	0.0343235	70	5mo
4	▲6	Data Finance			0.0323850	100	5mo
5	—	best fitting			0.0320763	172	5mo
6	▲33	NIWATORI			0.0301690	31	5mo
7	▲8	E2			0.0291539	43	5mo
8	▲11	John Ma			0.0289587	97	5mo
9	▲25	Pradeep and Arthur			0.0287992	111	5mo
10	▼4	William Hau			0.0287899	165	5mo

Expect LB shuffle because of

- Randomness

- Little amount of data

- Different public/private distributions

Expect LB shuffle because of

- Randomness

2σ TWO SIGMA

Liberty Mutual®
INSURANCE

- Little amount of data

TFI
TAB
FOOD
INVESTMENTS

- Different public/private distributions

ROSSMANN

Conclusion

- If we have big dispersion of scores on validation stage, we should do extensive validation
  - Average scores from different KFold splits
  - Tune model on one split, evaluate score on the other
- If submission's score do not match local validation score, we should
  - Check if we have too little data in public LB
  - Check if we overfitted
  - Check if we chose correct splitting strategy
  - Check if train/test have different distibutions
- Expect LB shuffle because of
  - Randomness
  - Little amount of data
  - Different public/private distributions

Summary of Validation topic

1. Defined validation and its connection to overfitting
2. Described common validation strategies
3. Demonstrated major data splitting strategies
4. Analysed and learn how to tackle main validation problems