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

Validation stage

Causes of different scores and optimal parameters

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

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

- 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

Submission stage

0. We may already have quite different scores in Kfold

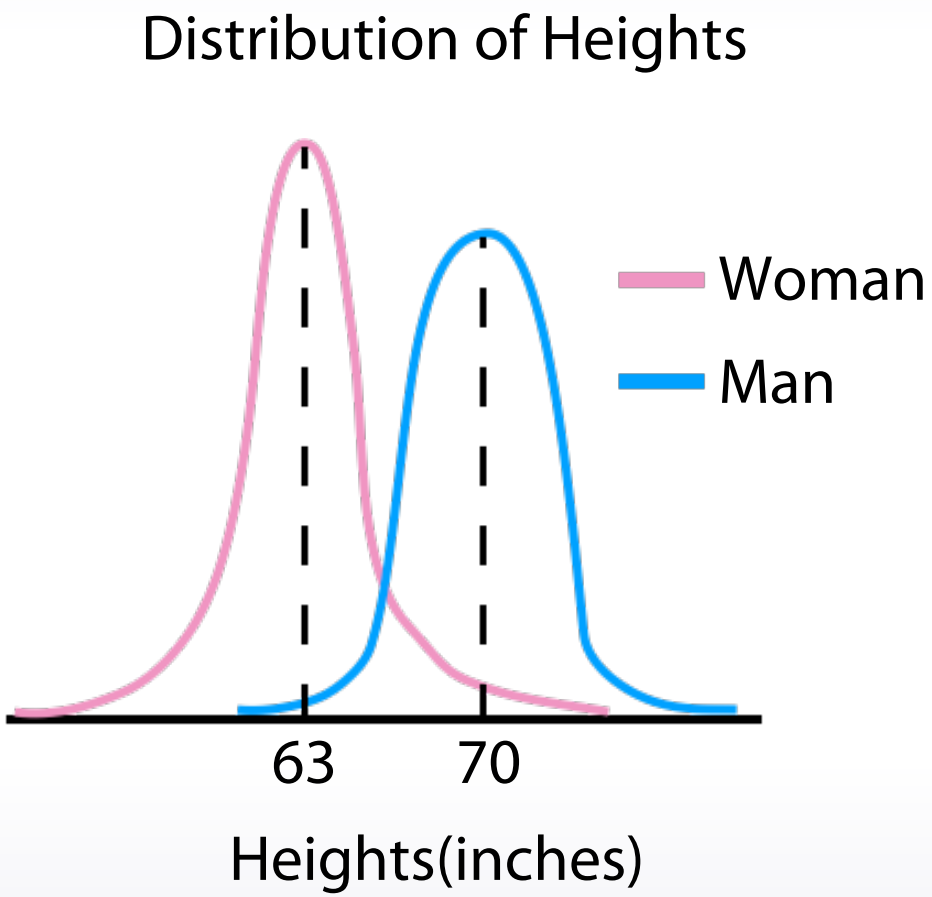
Submission stage

0. We may already have quite different scores in Kfold

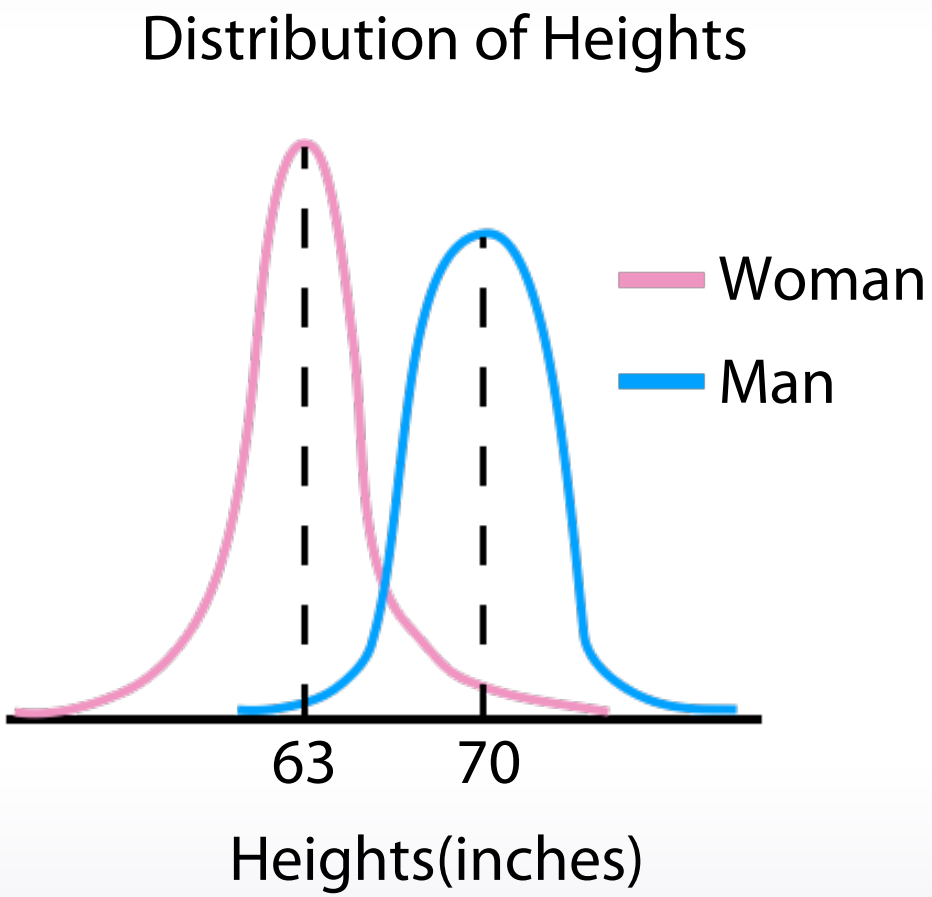
Other reasons:

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

Submission stage: different distributions

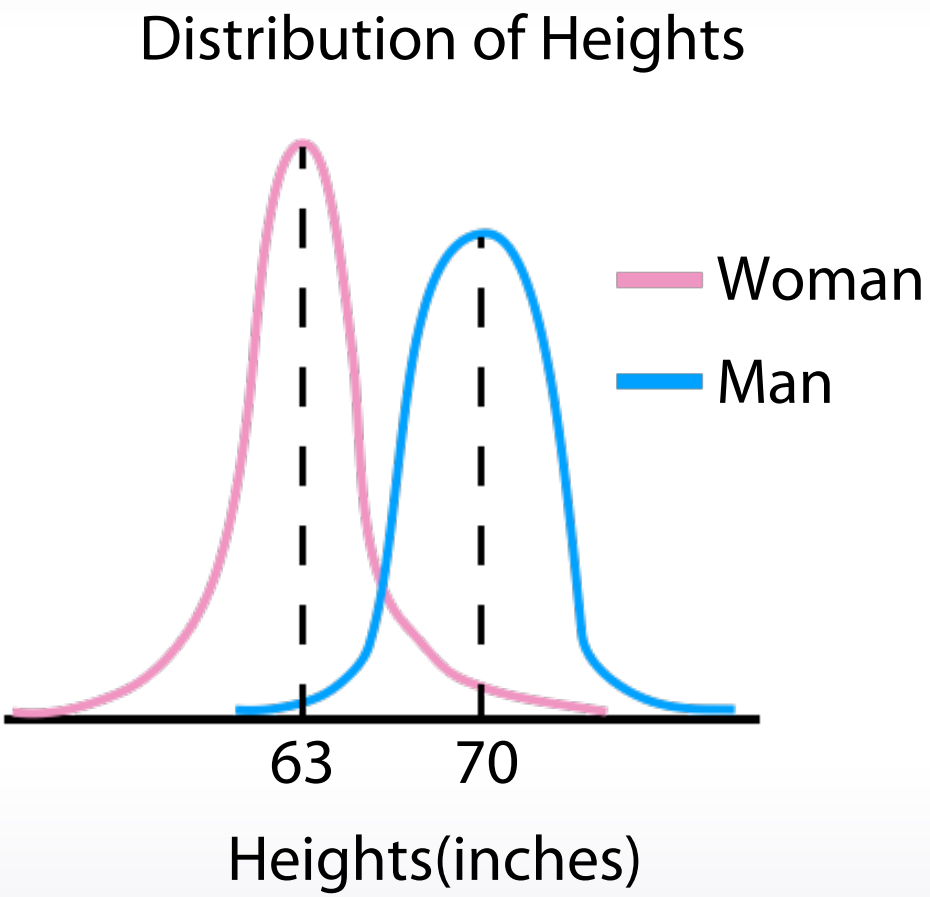


Submission stage: different distributions



- Mean for train: Calculate from the train data
- Mean for test: Leaderboard probing

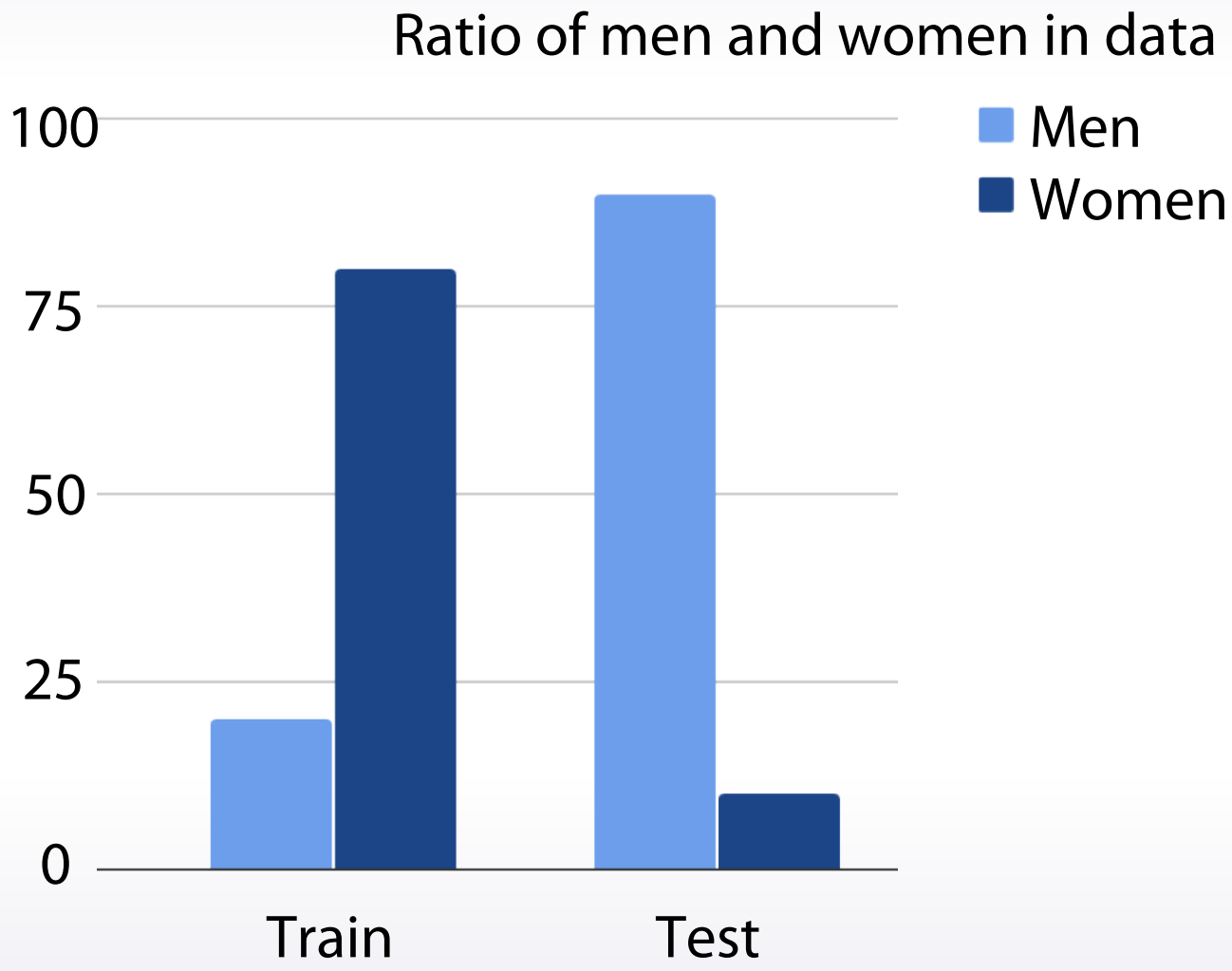
Submission stage: different distributions



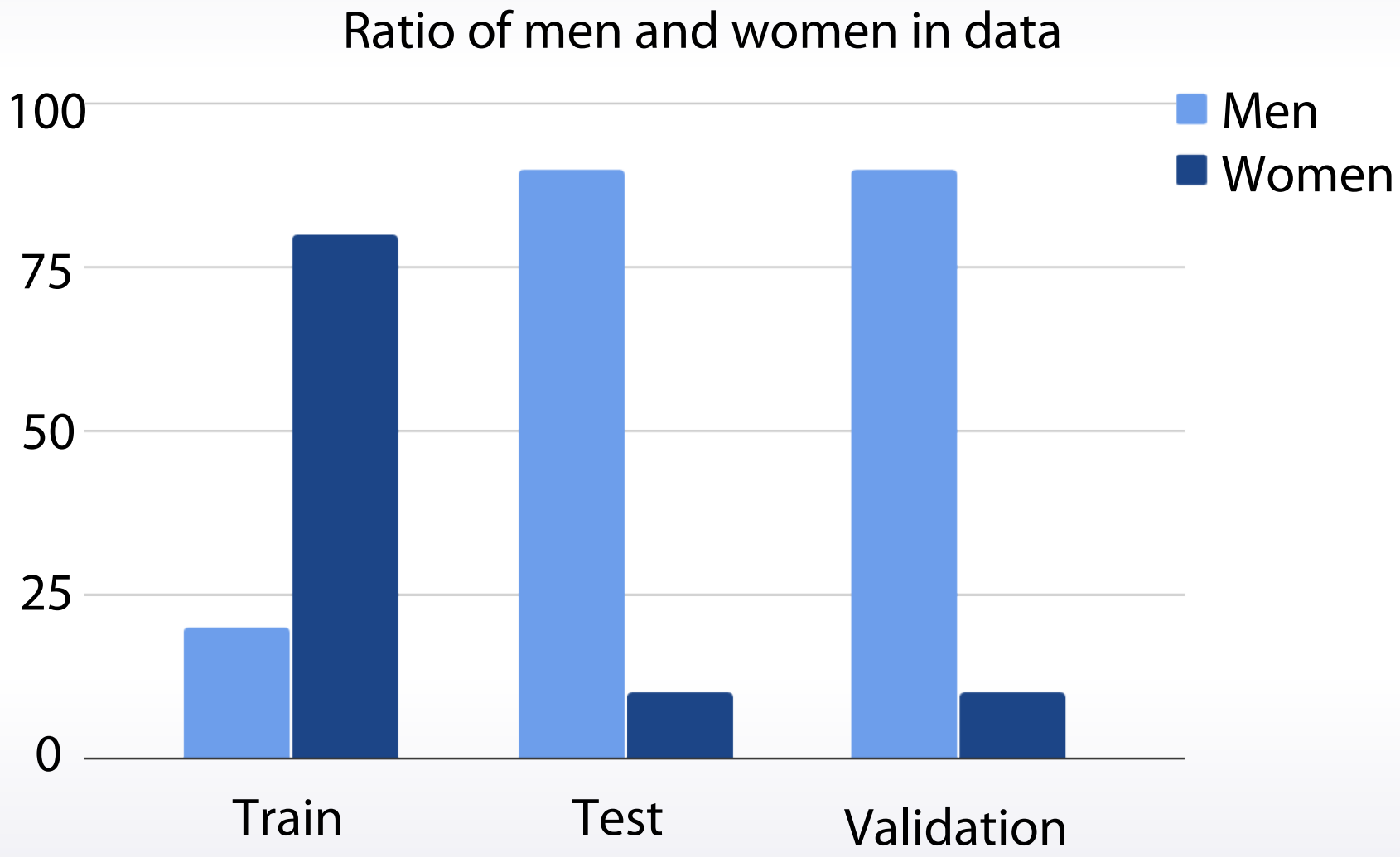
Quora Question Pairs

Quora

Submission stage: different distributions



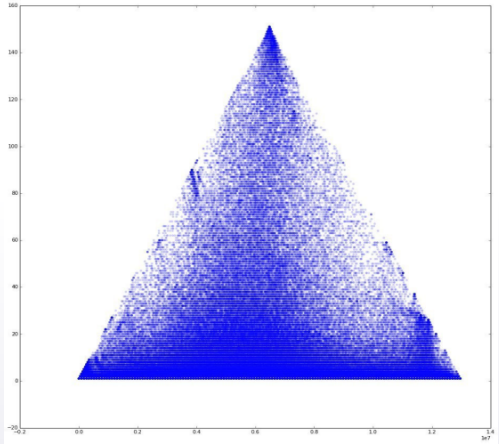
Submission stage: different distributions



Submission stage: different distributions

- Data Science Game 2017 Qualification phase: Music recommendation

- CTR prediction task from EDA



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			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

- Little amount of data

- Different public/private distributions

Expect LB shuffle because of

- Randomness

- Little amount of data

- Different public/private distributions

Expect LB shuffle because of

- Randomness

- Little amount of data

- Different public/private distributions

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

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

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