# Validation and overfitting



#### **Restaurant Revenue Prediction**

Predict annual restaurant sales based on objective measurements

\$30,000 · 2,257 teams · 2 years ago

**Public Leaderboard** 

Private Leaderboard

This leaderboard is calculated with approximately 30% of the test data.

The final results will be based on the other 70%, so the final standings may be different.



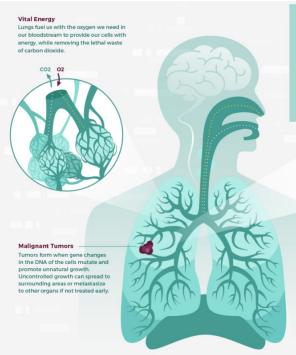
# △priv	Team Name	Kernel	Team Members	Score 2	Entries	Last
1 ▼19	BAYZ, M.D.		9 😂 9	0.00000	115	2у
2 ▼16	Will lam		-	710063.76	116	2у
3 ▼10	Scott Lowe		<b>10</b>	1462479.4	106	2у
4 ▼ 935	AMAR_PREM_AnandAkela_Teja			1464692.1	97	2у
5 ▼ 683	Analytic Bastard		9 9	1492787.0	115	2у

### **Next videos**

- 1. We will understand the concept of validation and overfitting
- 2. We will identify the number of splits that should be done to establish stable validation
- 3. We will break down most frequent ways to repeat train test split
- 4. We will discuss most often validation problems

## Validation: example





Lung Cancer is the most common type of cancer with...

225,000

\$12 billion
were accounted for in healthcare
costs in the U.S. every year<sup>2</sup>



Low-Dose CT scans help assess if a person is at risk of lung cancer or other pulmonary disease. Scientific research reports...

#### 20%

of lung cancer deaths can be reduced with early detection<sup>3</sup>

However, the image assessments in use today are identifying lung lesions as potentially cancerous that later turn out to not be cancer.

#### High false positive rates

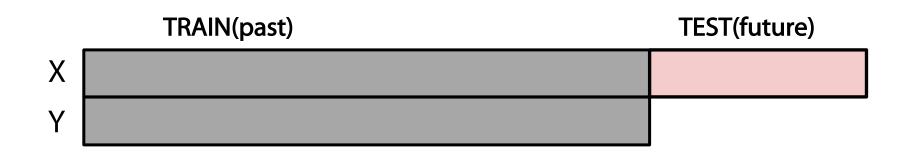
lead to unnecessary patient anxiety, additional follow-up imaging and interventional treatments 3.4

#### SOURCE

<sup>1</sup>Siegel RL, Miller KD, Jemai A. "Cancer Statistics," 2016. CA: A Cancer Journal for Clinicians. 2016; 66:7-30.

\*National Institutes of Health, "Cancer costs projected to reach at least \$150 billion in 2020, "https://www.nit.gov/news-events/news-releases/cancer-costs-projected-reach-least-\$50 billion-2020, (January 12, 2011).
\*Aborte DR, Adama AH, Berg CD, et al. "Reduced lang-cancer mortality with low dose computed tomographic screening." New Feet J Med. 2011;36:539-409.
\*\*Leve Owser CT has historially resulted in high face positive sets of oranged 25% facilities 4.0, New England J Med. 2011;36:539-409.

# Validation: example



# Validation: example

	TRAIN(past)	VALIDATION (past)	TEST(future)
X			
Υ			

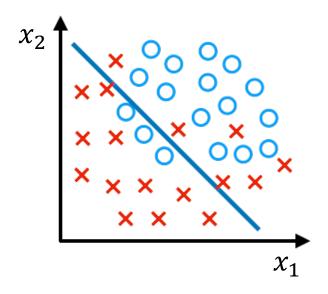
# Validation: competitions

	TRAIN	VALIDATION	TEST
X			
Υ			

# Validation: competitions

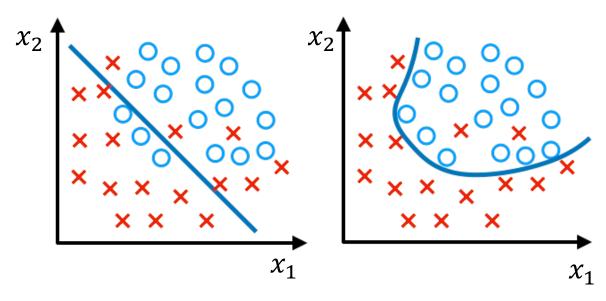
			TEST	
			<b>LEADERBOARD</b>	
	TRAIN	VALIDATION	<b>PUBLIC</b>	PRIVATE
X				
Υ				

#### **UNDERFITTING**



$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$
  
(g = sigmoid function)

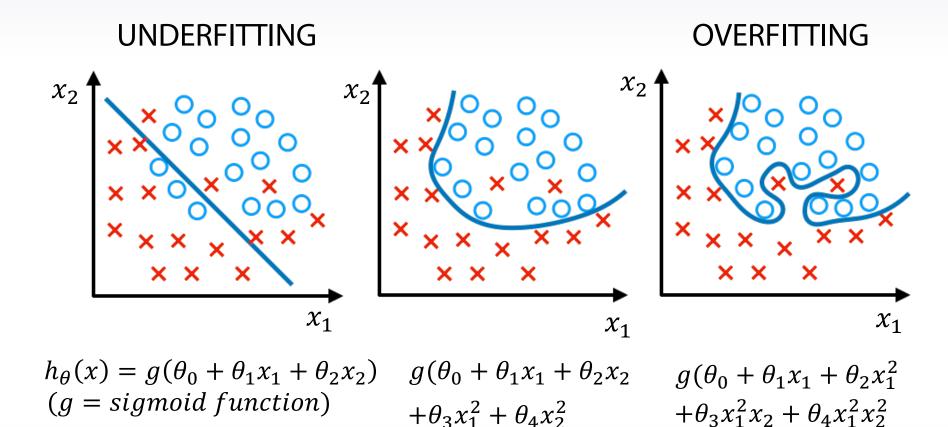
#### UNDERFITTING



$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2) \qquad g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

$$(g = sigmoid function) \qquad +\theta_3 x_1^2 + \theta_4 x_2^2$$

$$+\theta_5 x_1 x_2)$$

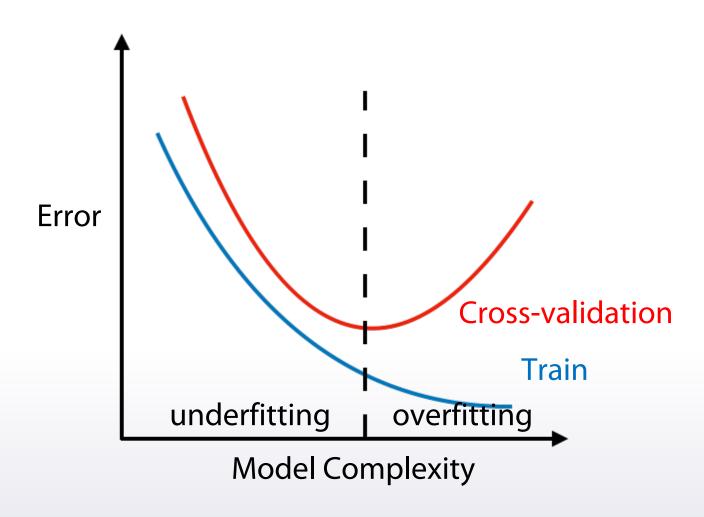


 $+\theta_5 x_1 x_2$ 

 $+\theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2$ 

Overfitting in general != overfitting in competitions

Overfitting in general != overfitting in competitions



### Conclusion

- 1. Validation helps us evaluate a quality of the model
- 2. Validation helps us select the model which will perform best on the unseen data
- 3. Underfitting refers to not capturing enough patterns in the data
- 4. Generally, overfitting refers to
  - a. capturing noize
  - b. capturing patterns which do not generalize to test data
- 5. In competitions, overfitting refers to
  - a. low model's quality on test data, which was unexpected due to validation scores