

Validation and overfitting



TAB
FOOD
INVESTMENTS

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.

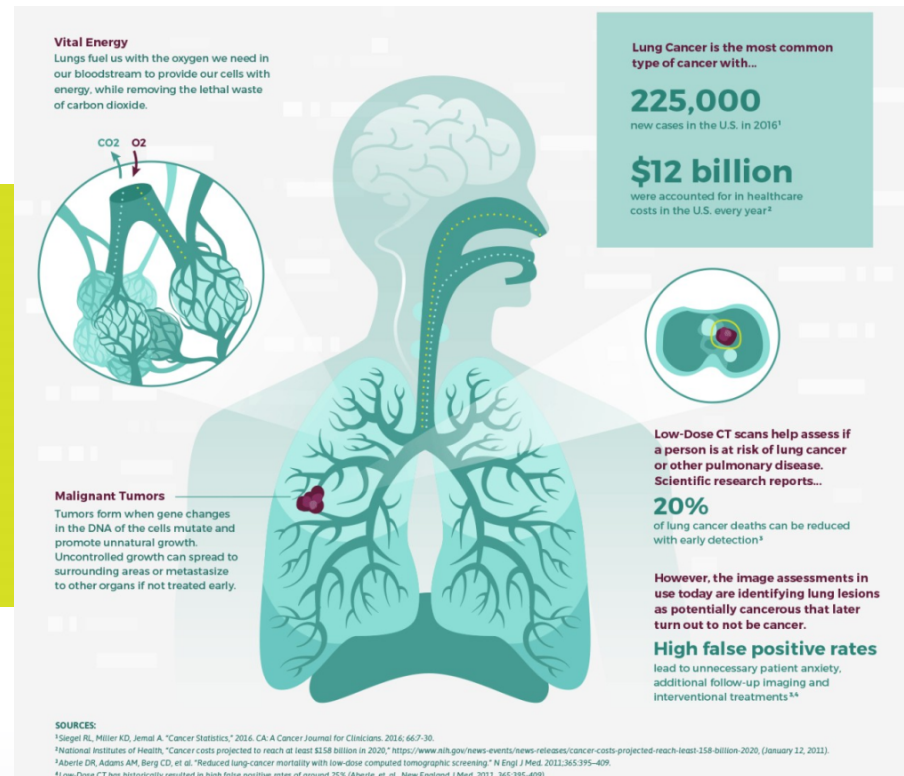
[Raw Data](#) [Refresh](#)

| # | Δpriv | Team Name | Kernel | Team Members | Score ? | Entries | Last |
|---|---------|---------------------------|--------|---|--------------|---------|------|
| 1 | ▼ 19... | BAYZ, M.D. | |    | 0.00000 | 115 | 2y |
| 2 | ▼ 16... | Will lam | |  | 710063.76... | 116 | 2y |
| 3 | ▼ 10... | Scott Lowe | |  | 1462479.4... | 106 | 2y |
| 4 | ▼ 935 | AMAR_PREM_AnandAkela_Teja | |     | 1464692.1... | 97 | 2y |
| 5 | ▼ 683 | Analytic Bastard | |    | 1492787.0... | 115 | 2y |

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

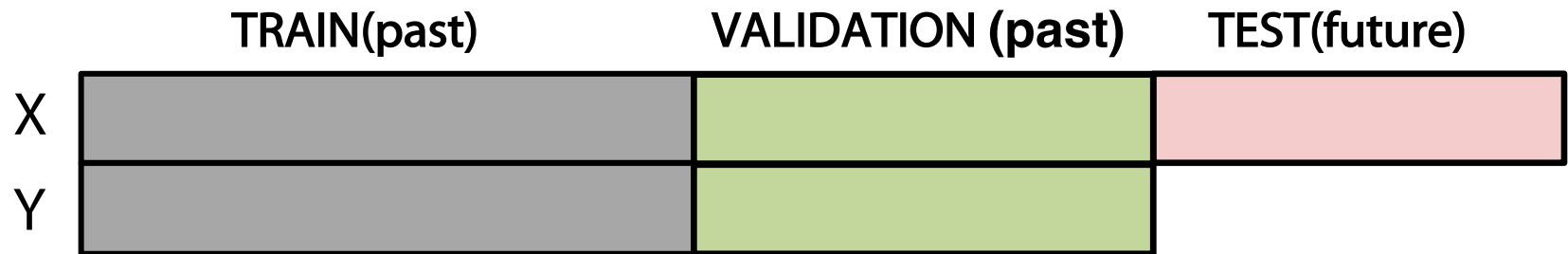
Validation: example



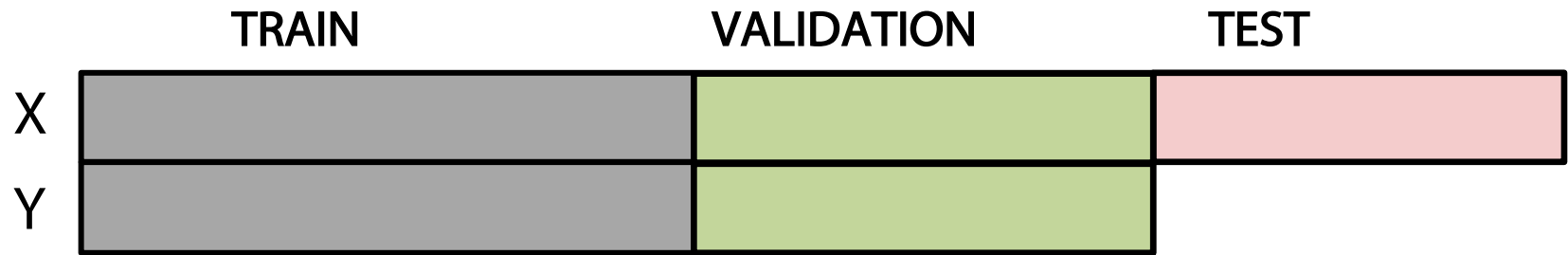
Validation: example



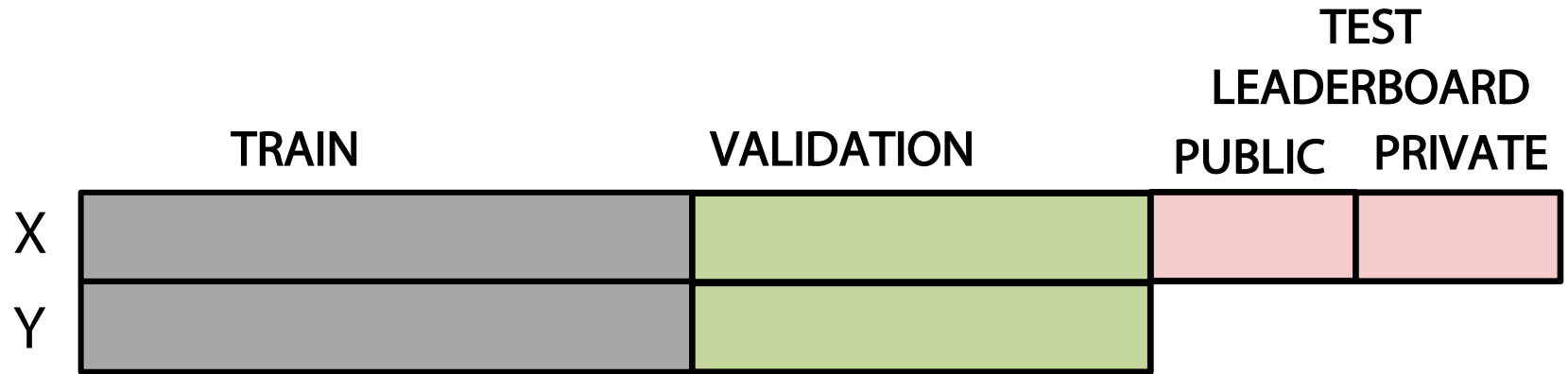
Validation: example



Validation: competitions

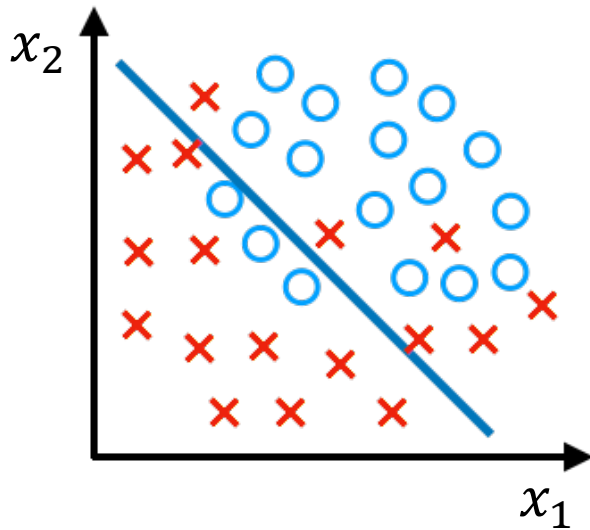


Validation: competitions



Validation: underfitting and overfitting

UNDERFITTING

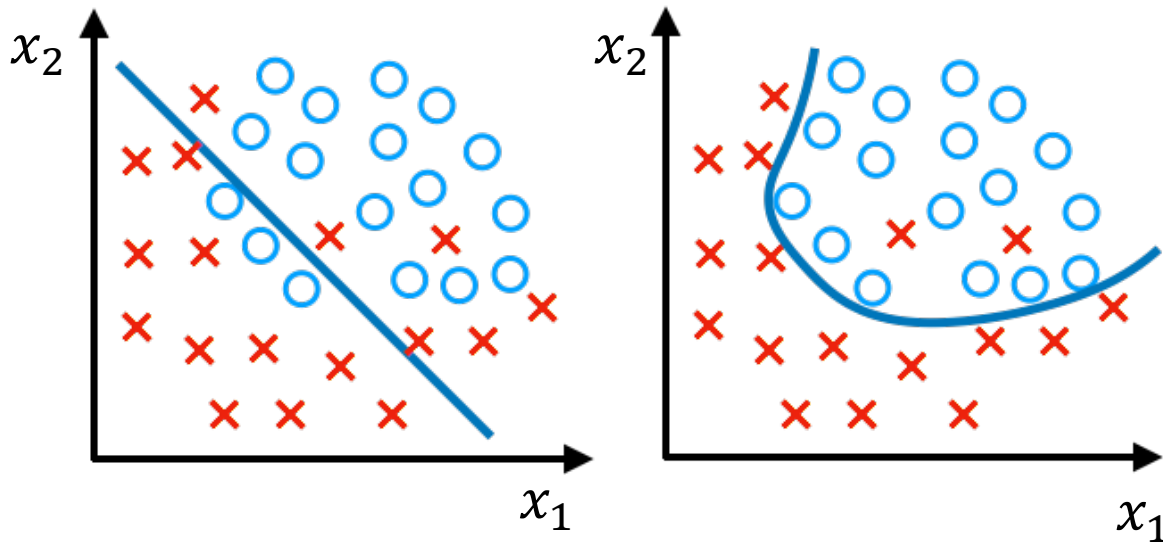


$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

(g = sigmoid function)

Validation: underfitting and overfitting

UNDERFITTING



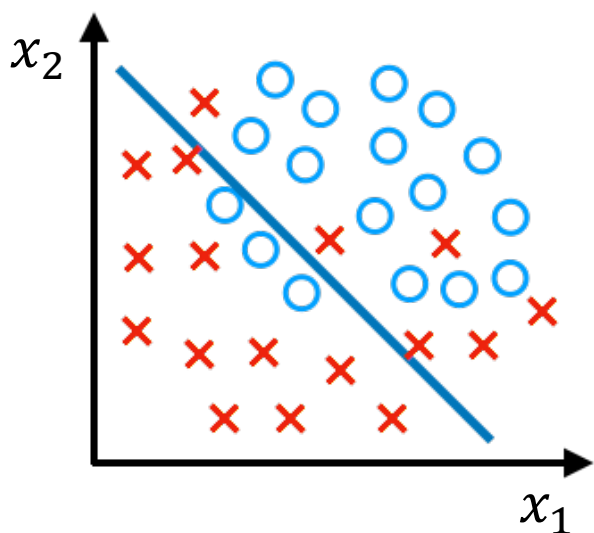
$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

($g = \text{sigmoid function}$)

$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$

Validation: underfitting and overfitting

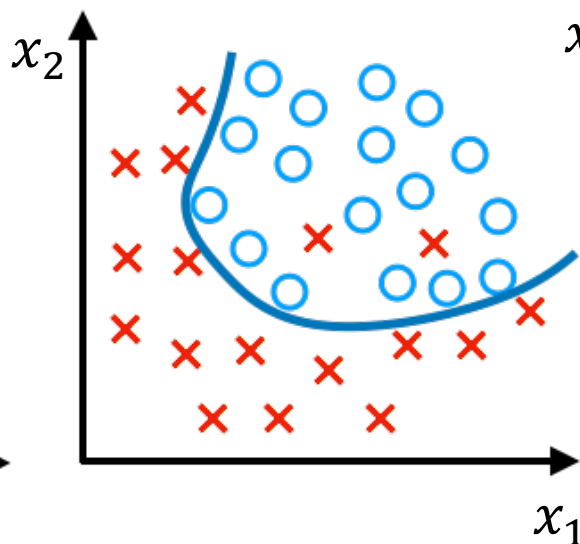
UNDERFITTING



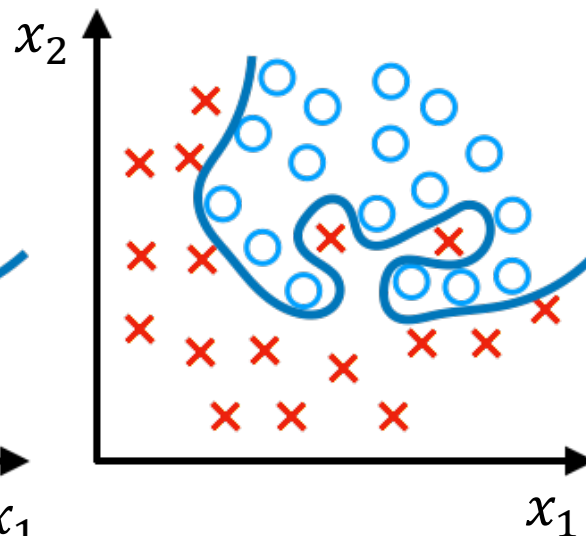
$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

($g = \text{sigmoid function}$)

OVERFITTING



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$



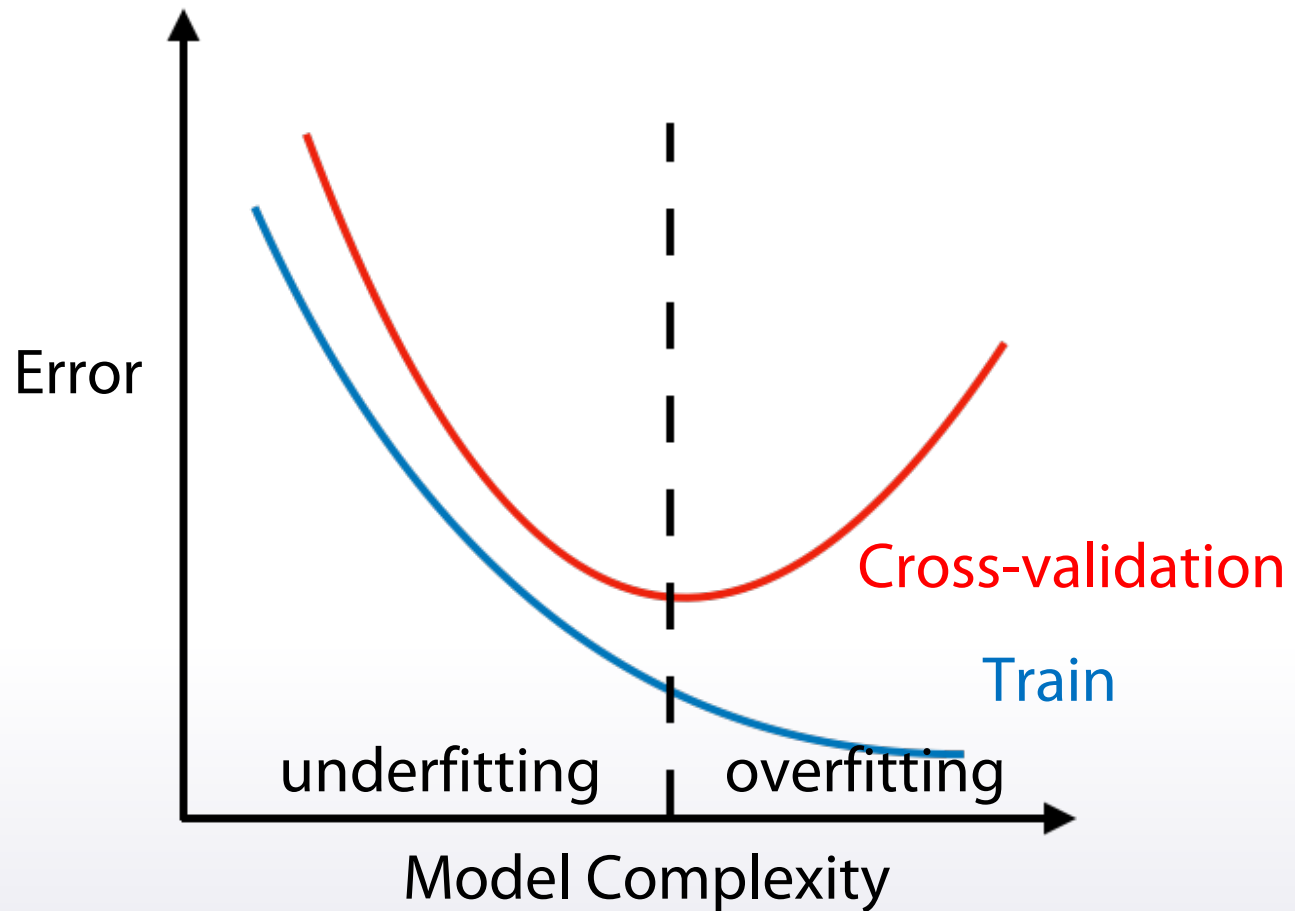
$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2)$$

Validation: underfitting and overfitting

Overfitting in general \neq overfitting in competitions

Validation: underfitting and overfitting

Overfitting in general \neq 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 noise
 - 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