Quantitative Data Analysis

SOCIOL 212B
Winter 2025

Lecture 4. Data-Driven Selection of an Estimator

Learning goals for today

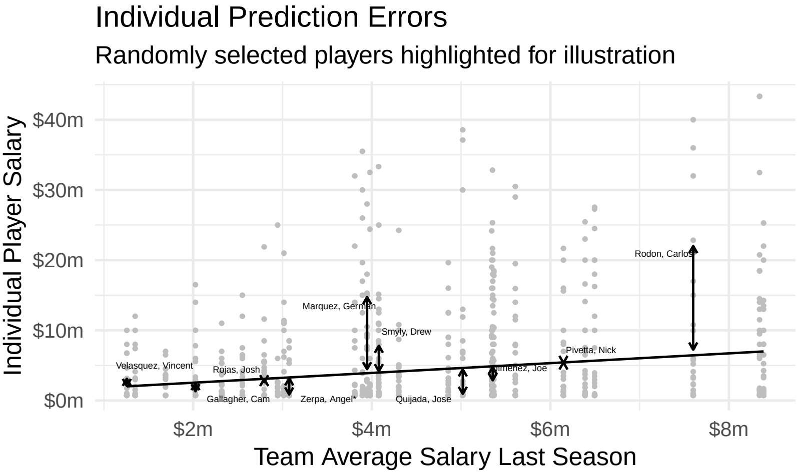
By the end of class, you will be able to

- Use individual prediction error as a metric to choose an algorithm that makes good group-level estimates
- Understand why we predict out of sample
- Carry out a sample split
- Explain the idea of cross-validation

On a different topic, we will close with a writing task.

Model player salaries this year as a linear function of team average salaries last year

$$\text{E(Salary} \mid \text{Team)} = \alpha + \beta\text{(Team Average Salary Last Season)}$$


Individual Prediction Errors
Randomly selected players highlighted for illustration
Individual Player Salary
$40m
$30m
$20m
$10m
$0m
Rodon, Carlos
Marquez, German
Smyly, Drew
Pivetta, Nick
Velasquez, Vincent
Rojas, Josh
imenez, Joe
Gallagher, Cam
Zerpa, Angel*
Quijada, Jose
$2m
$4m
$6m
$8m
Team Average Salary Last Season

$$R^2 = 1 - \frac{\overbrace{\mathsf{V}\left(Y - \hat{Y}\right)}^{\text{Variance of Errors}}}{\underbrace{\mathsf{V}(Y)}_{\text{Variance of Y}}}$$

Limiting cases:

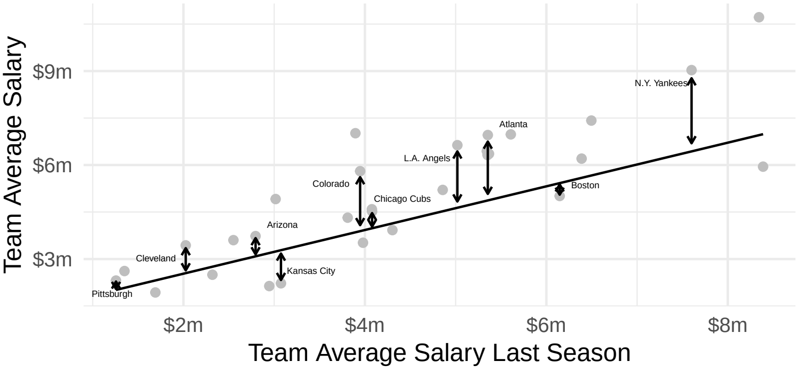
- 1 = perfect prediction
- 0 = predicted same value for all cases

This example: $R^2 = 0.058$

What if the goal is to estimate for subgroups instead of to predict for individuals?

Group−Level Estimation Errors

Dots are true team mean salaries, excluding players from the learning sample in which the line was estimated



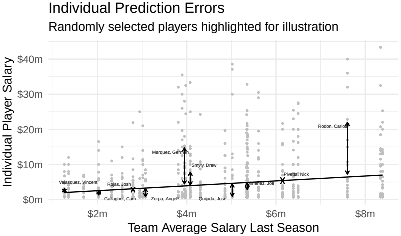
$$R^2_{\text{Group}} = 1 - \frac{\text{V}\left(\hat{Y}_{\text{Team}} - \bar{Y}_{\text{Team}}\right)}{\text{V}(\bar{Y}_{\text{Team}})}$$

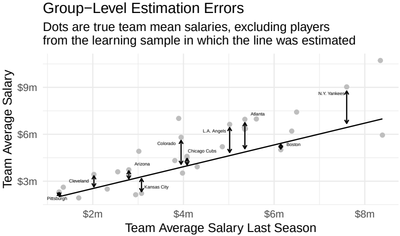
Variance of team-level prediction errors
over variance of team-level means[1]

In this example: $R^2_{\text{Group}} = 0.672$

[1] R^2_{Group} is not widely used, but I think it is useful here.

Individual prediction and subgroup estimation: Two very different goals?





$R^2 = 0.058$

$R^2_{\text{Group}} = 0.672$

Expected squared prediction error

Individual-level errors

$$\underbrace{\text{ESPE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Prediction Error}}} = \text{E}\left[\left(Y - \hat{f}(\vec{X})\right)^2\right]$$

In words:

The squared prediction error that occurs on average for a unit sampled at random from the population

Expected squared estimation error

Errors for estimating subgroup means

$$\underbrace{\text{ESEE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Estimation Error}}} = \text{E}\left[\left(\text{E}(Y \mid \vec{X}) - \hat{f}(\vec{X})\right)^2\right]$$

In words:

The squared error when estimating conditional means given $\vec{X}$, averaged over the distribution of the predictors $\vec{X}$

Within-group variance

Spread of individual outcomes within groups

$$\mathrm{E}\left[\mathrm{V}\left(Y \mid \vec{X})\right)\right]$$

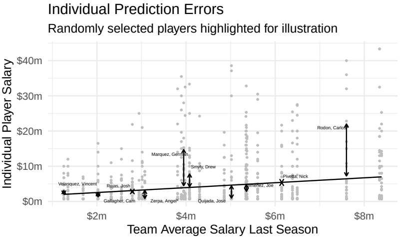
In words:
Within each subgroup defined by $\vec{X}$, outcomes vary.
This is the average value of the within-group variance,
with the average taken over $\vec{X}$

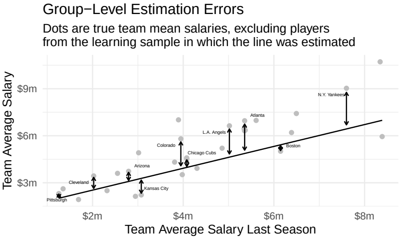
Decomposition of individual prediction error

$$\underbrace{\text{ESPE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Prediction Error}}} = \underbrace{\text{ESEE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Estimation Error}}} + \underbrace{\text{E}\left[\text{V}(Y \mid \vec{X})\right]}_{\substack{\text{Expected Within-}\\ \text{Group Variance}}}$$

Decomposition of individual prediction error

$$\underbrace{\text{ESPE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Prediction Error}}} = \underbrace{\text{ESEE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Estimation Error}}} + \underbrace{\text{E}\left[\text{V}(Y \mid \vec{X})\right]}_{\substack{\text{Expected Within-}\\ \text{Group Variance}}}$$





Decomposition of individual prediction error

$$\underbrace{\text{ESPE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Prediction Error}}} = \underbrace{\text{ESEE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Estimation Error}}} + \underbrace{\text{E}\left[\text{V}(Y \mid \vec{X})\right]}_{\substack{\text{Expected Within-}\\ \text{Group Variance}}}$$

Suppose $\hat{f}_1$ and $\hat{f}_2$ are prediction functions

Decomposition of individual prediction error

$$\underbrace{\text{ESPE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Prediction Error}}} = \underbrace{\text{ESEE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Estimation Error}}} + \underbrace{\text{E}\left[\text{V}(Y \mid \vec{X})\right]}_{\substack{\text{Expected Within-}\\ \text{Group Variance}}}$$

Suppose $\hat{f}_1$ and $\hat{f}_2$ are prediction functions
Suppose $\text{ESPE}(\hat{f}_1) < \text{ESPE}(\hat{f}_2)$

Decomposition of individual prediction error

$$\underbrace{\text{ESPE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Prediction Error}}} = \underbrace{\text{ESEE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Estimation Error}}} + \underbrace{\text{E}\left[\text{V}(Y \mid \vec{X})\right]}_{\substack{\text{Expected Within-}\\ \text{Group Variance}}}$$

Suppose $\hat{f}_1$ and $\hat{f}_2$ are prediction functions
Suppose $\text{ESPE}(\hat{f}_1) < \text{ESPE}(\hat{f}_2)$
Then

$$\text{ESEE}(\hat{f}_1) \quad ? \quad \text{ESEE}(\hat{f}_2)$$

Decomposition of individual prediction error

$$\underbrace{\text{ESPE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Prediction Error}}} = \underbrace{\text{ESEE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Estimation Error}}} + \underbrace{\text{E}\left[\text{V}(Y \mid \vec{X})\right]}_{\substack{\text{Expected Within-}\\ \text{Group Variance}}}$$

Suppose $\hat{f}_1$ and $\hat{f}_2$ are prediction functions
Suppose $\text{ESPE}(\hat{f}_1) < \text{ESPE}(\hat{f}_2)$
Then

$$\text{ESEE}(\hat{f}_1) < \text{ESEE}(\hat{f}_2)$$

Decomposition of individual prediction error

$$\underbrace{\text{ESPE}(\hat{f})}_{\substack{\text{Expected Squared} \\ \text{Prediction Error}}} = \underbrace{\text{ESEE}(\hat{f})}_{\substack{\text{Expected Squared} \\ \text{Estimation Error}}} + \underbrace{\text{E}\left[\text{V}(Y \mid \vec{X})\right]}_{\substack{\text{Expected Within-} \\ \text{Group Variance}}}$$

Suppose $\hat{f}_1$ and $\hat{f}_2$ are prediction functions
Suppose $\text{ESPE}(\hat{f}_1) < \text{ESPE}(\hat{f}_2)$
Then

$$\text{ESEE}(\hat{f}_1) \quad < \quad \text{ESEE}(\hat{f}_2)$$

Strategy:

Decomposition of individual prediction error

$$\underbrace{\text{ESPE}(\hat{f})}_{\substack{\text{Expected Squared}\\\text{Prediction Error}}} = \underbrace{\text{ESEE}(\hat{f})}_{\substack{\text{Expected Squared}\\\text{Estimation Error}}} + \underbrace{\text{E}\left[\text{V}(Y \mid \vec{X})\right]}_{\substack{\text{Expected Within-}\\\text{Group Variance}}}$$

Suppose $\hat{f}_1$ and $\hat{f}_2$ are prediction functions
Suppose $\text{ESPE}(\hat{f}_1) < \text{ESPE}(\hat{f}_2)$
Then

$$\text{ESEE}(\hat{f}_1) < \text{ESEE}(\hat{f}_2)$$

Strategy:

- With many candidate algorithms $\hat{f}_1, \hat{f}_2, \ldots$

Decomposition of individual prediction error

$$\underbrace{\text{ESPE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Prediction Error}}} = \underbrace{\text{ESEE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Estimation Error}}} + \underbrace{\text{E}\left[\text{V}(Y \mid \vec{X})\right]}_{\substack{\text{Expected Within-}\\ \text{Group Variance}}}$$

Suppose $\hat{f}_1$ and $\hat{f}_2$ are prediction functions
Suppose $\text{ESPE}(\hat{f}_1) < \text{ESPE}(\hat{f}_2)$
Then

$$\text{ESEE}(\hat{f}_1) \quad < \quad \text{ESEE}(\hat{f}_2)$$

Strategy:

- With many candidate algorithms $\hat{f}_1, \hat{f}_2, \ldots$
- choose the one that minimizes ESPE (individual prediction)

Decomposition of individual prediction error

$$\underbrace{\text{ESPE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Prediction Error}}} = \underbrace{\text{ESEE}(\hat{f})}_{\substack{\text{Expected Squared}\\ \text{Estimation Error}}} + \underbrace{\text{E}\left[\text{V}(Y \mid \vec{X})\right]}_{\substack{\text{Expected Within-}\\ \text{Group Variance}}}$$

Suppose $\hat{f}_1$ and $\hat{f}_2$ are prediction functions
Suppose $\text{ESPE}(\hat{f}_1) < \text{ESPE}(\hat{f}_2)$
Then

$$\text{ESEE}(\hat{f}_1) \quad < \quad \text{ESEE}(\hat{f}_2)$$

Strategy:

- With many candidate algorithms $\hat{f}_1, \hat{f}_2, \ldots$
- choose the one that minimizes ESPE (individual prediction)
- It will also minimize ESEE (group estimation)

Prediction and Estimation

Why Out-of-Sample Prediction

Sample Splitting

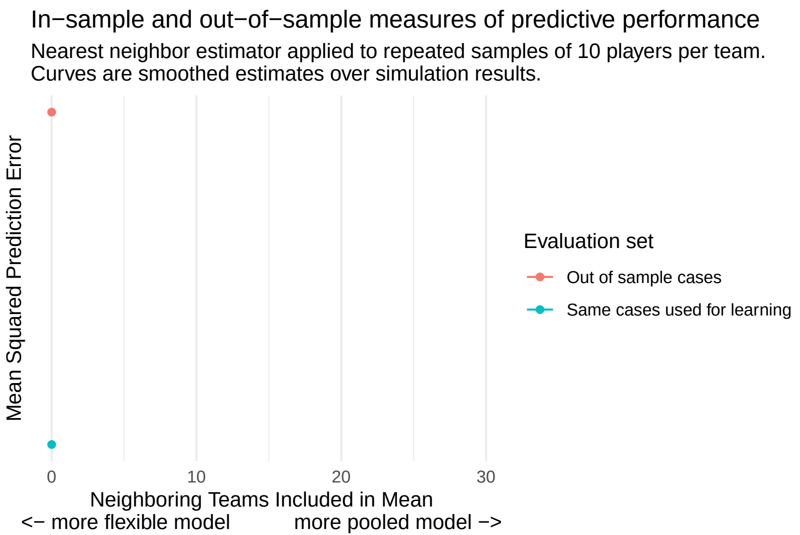
Cross Validation

Writing Task

k-nearest neighbors

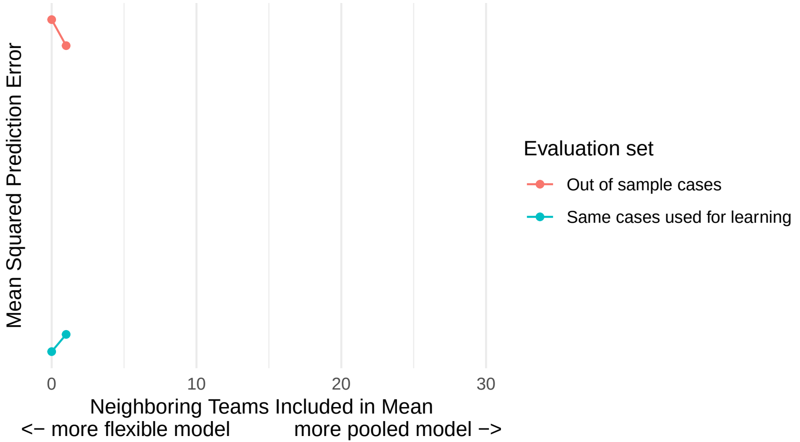
10 sampled players per team

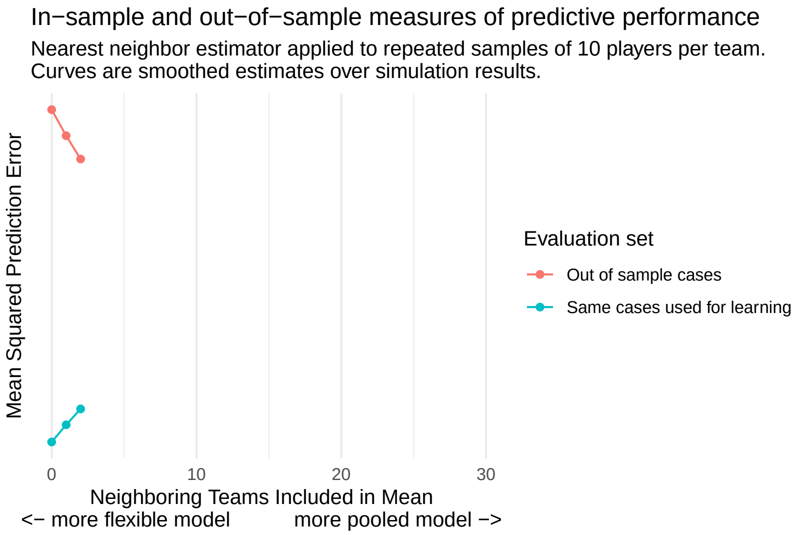
- Dodger sample mean might be noisy
- Could pool with similar teams defined by past mean salary
  - Dodgers: 8.39m
  - 1st-nearest neighbor. NY Mets: 8.34m
  - 2nd-nearest neighbor. NY Yankees: 7.60m
  - 3rd-nearest neighbor. Philadelphia: 6.50m
- How does performance change with the number of neighbors included?
  - measured by mean squared prediction error


In−sample and out−of−sample measures of predictive performance
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.
Mean Squared Prediction Error
0
10
20
30
Neighboring Teams Included in Mean
<− more flexible model
more pooled model −>
Evaluation set
Out of sample cases
Same cases used for learning

In−sample and out−of−sample measures of predictive performance

Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.

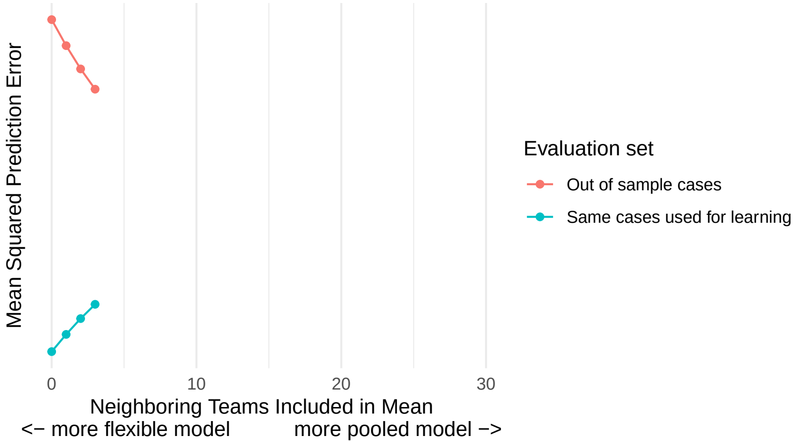




In−sample and out−of−sample measures of predictive performance
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.
Mean Squared Prediction Error
0
10
20
30
Neighboring Teams Included in Mean
<− more flexible model
more pooled model −>
Evaluation set
Out of sample cases
Same cases used for learning

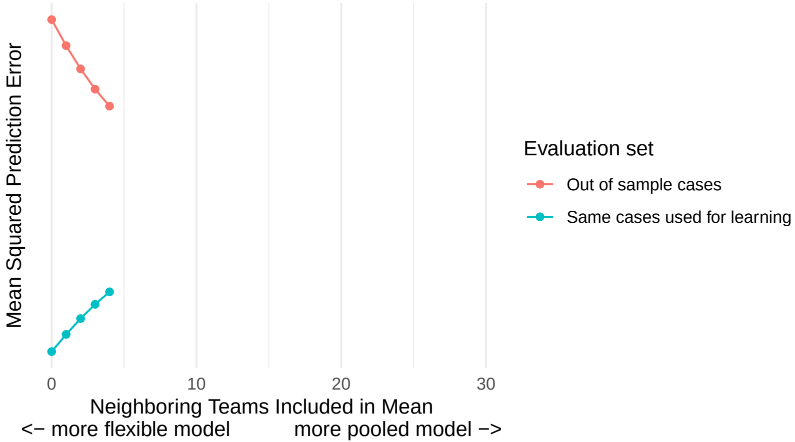
In−sample and out−of−sample measures of predictive performance

Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.



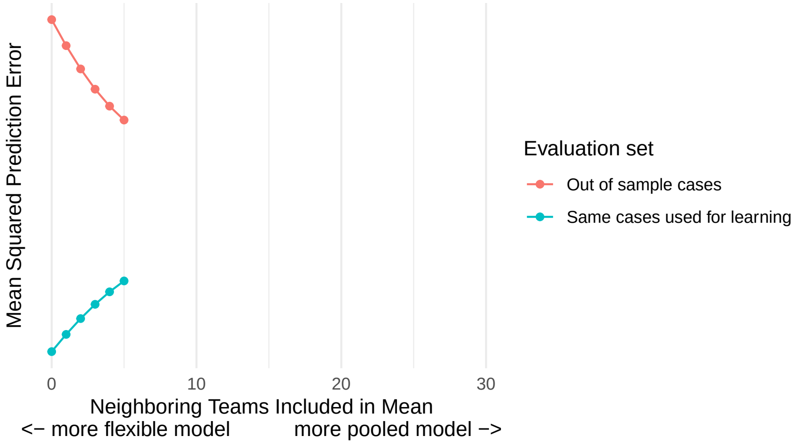
In−sample and out−of−sample measures of predictive performance

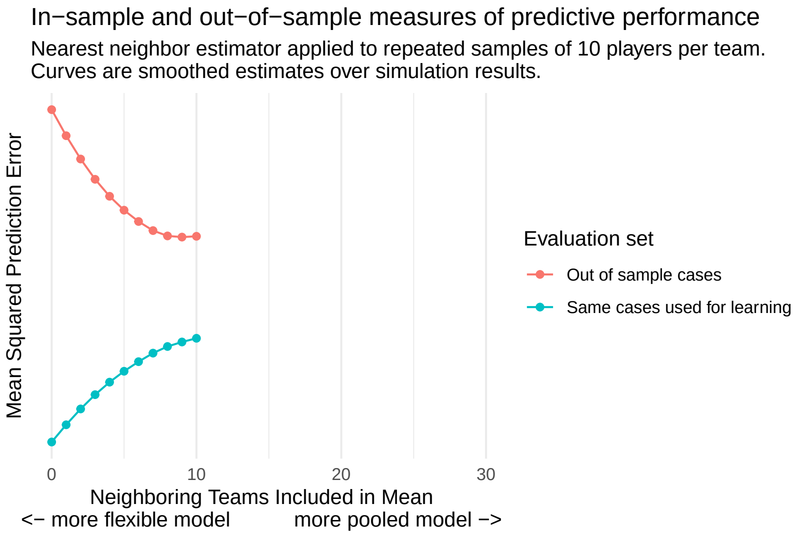
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.

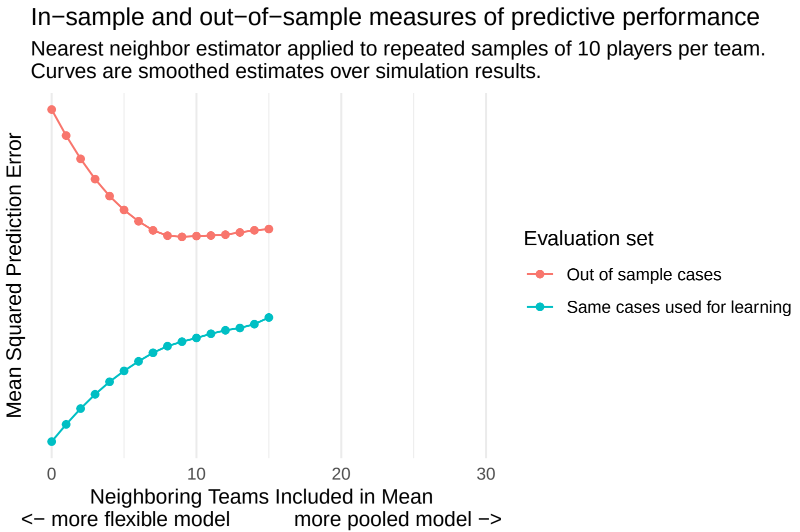


In−sample and out−of−sample measures of predictive performance

Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.

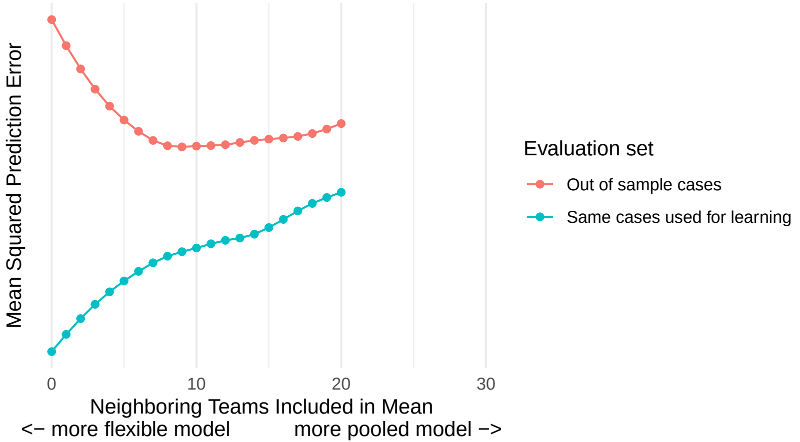



In−sample and out−of−sample measures of predictive performance
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.
Mean Squared Prediction Error
0
10
20
30
Neighboring Teams Included in Mean
<− more flexible model
more pooled model −>
Evaluation set
Out of sample cases
Same cases used for learning


In−sample and out−of−sample measures of predictive performance
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.
Mean Squared Prediction Error
0
10
20
30
Neighboring Teams Included in Mean
<− more flexible model
more pooled model −>
Evaluation set
Out of sample cases
Same cases used for learning

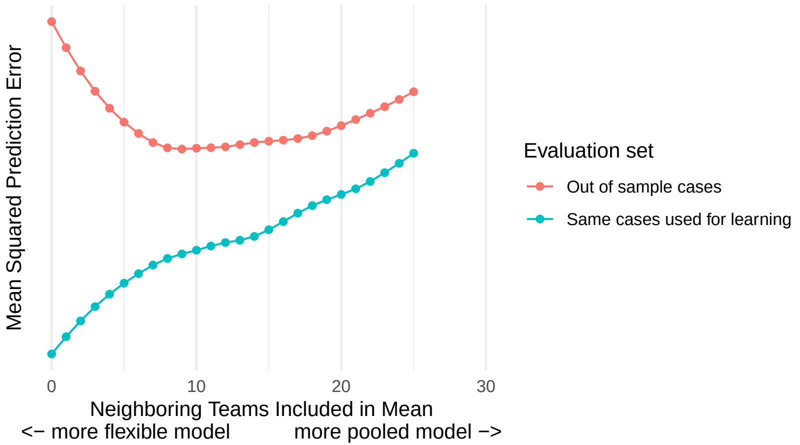
In−sample and out−of−sample measures of predictive performance

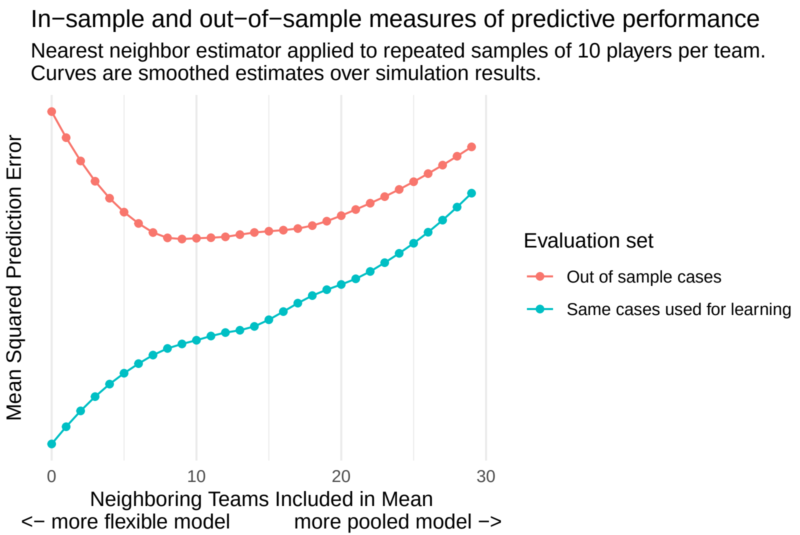
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.



In−sample and out−of−sample measures of predictive performance

Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.




In−sample and out−of−sample measures of predictive performance
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.
Mean Squared Prediction Error
0
10
20
30
Neighboring Teams Included in Mean
<− more flexible model
more pooled model −>
Evaluation set
Out of sample cases
Same cases used for learning

Prediction and Estimation

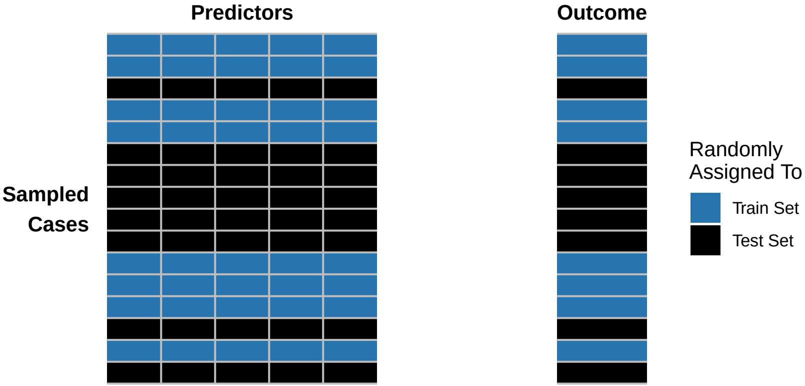
Why Out-of-Sample Prediction

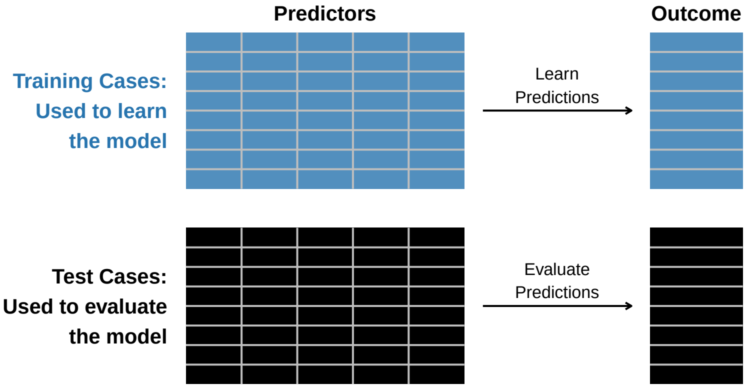
Sample Splitting

Cross Validation

Writing Task

You have one sample.
How do you estimate out-of-sample performance?


Predictors
Outcome
Sampled
Cases
Randomly
Assigned To
Train Set
Test Set


Predictors
Outcome
Training Cases:
Used to learn
the model
Learn
Predictions
Test Cases:
Used to evaluate
the model
Evaluate
Predictions

Exercise: Conduct a sample split in code

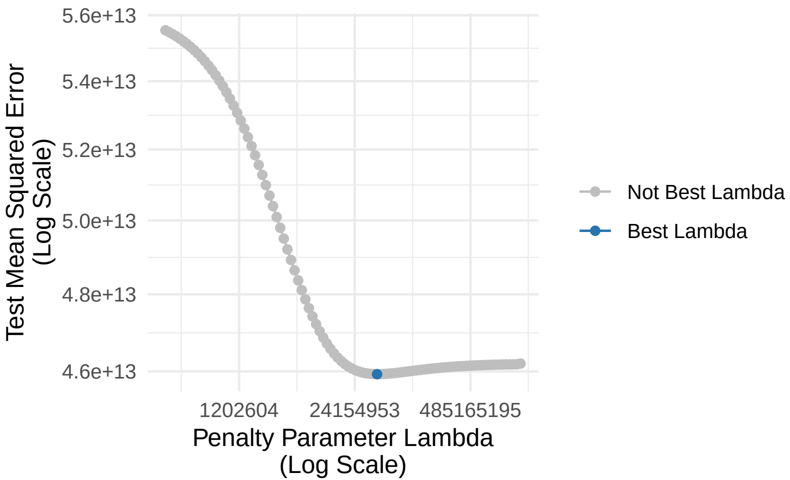
1. Sample 10 players per team
2. Take a 50-50 sample split stratified by team
3. Fit a linear regression in the train set
4. Predict in the test set
5. Report mean squared error

Sample splitting for parameter tuning: Ridge regression

$$\text{E}(\text{Player Salary} \mid \text{Team} = t) = \alpha + \beta_t$$

$$\left\{\hat{\alpha}, \hat{\vec{\beta}}\right\} = \underset{\tilde{\alpha}, \tilde{\vec{\beta}}}{\text{arg min}} \underbrace{\sum_{i=1}^{n} \left(Y_i - \left(\tilde{\alpha} + \tilde{\beta}_{t(i)}\right)\right)}_{\text{Prediction Error}} + \lambda \underbrace{\sum_{t} \tilde{\beta}_t^2}_{\text{Penalty}}$$

How do we choose λ?


5.6e+13
5.4e+13
5.2e+13
5.0e+13
4.8e+13
4.6e+13
Test Mean Squared Error
(Log Scale)
1202604
24154953
485165195
Penalty Parameter Lambda
(Log Scale)
Not Best Lambda
Best Lambda

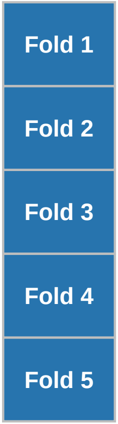
Cross Validation

A train test split loses lots of data to testing.

Is there a way to bring it back?

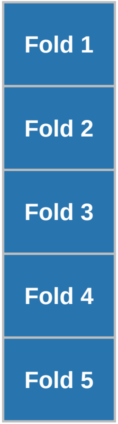
Cross Validation

Randomize
to 5 folds



Cross Validation

Randomize to 5 folds

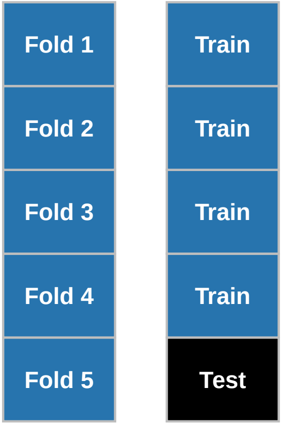


Iteratively use each as the test set

Cross Validation

Randomize to 5 folds

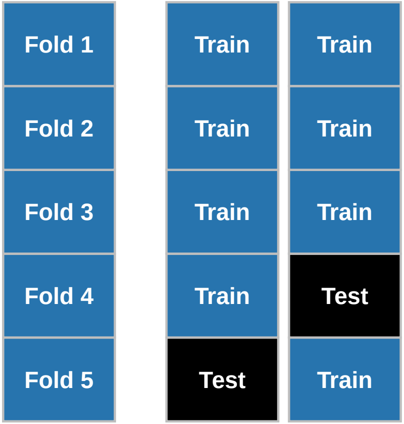
Iteratively use each as the test set



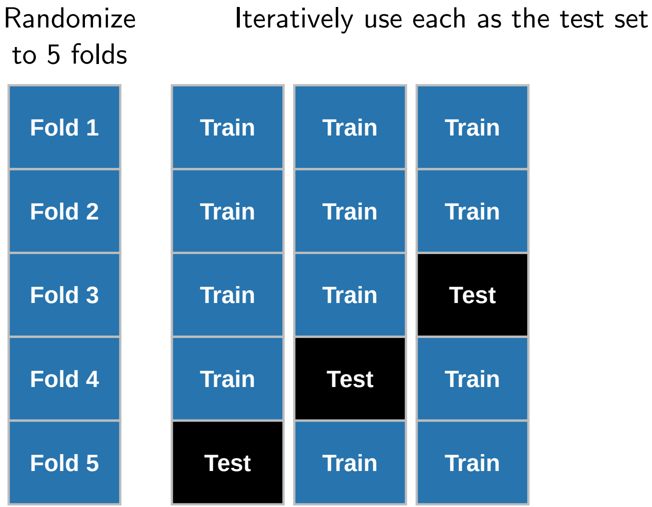
Cross Validation

Randomize
to 5 folds

Iteratively use each as the test set



Cross Validation



Cross Validation

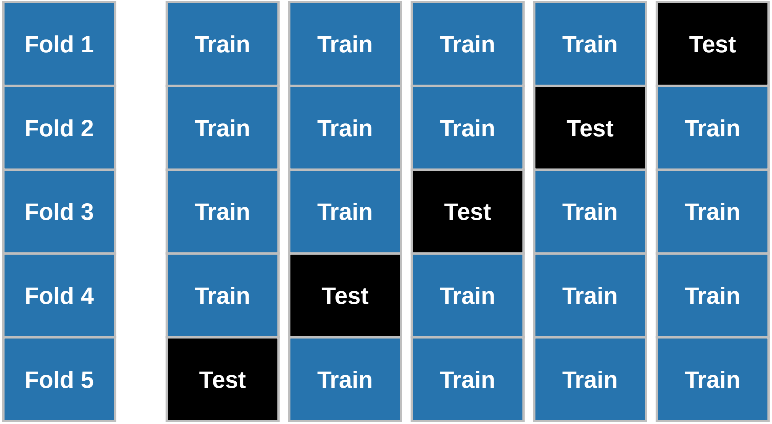
Randomize
to 5 folds

Iteratively use each as the test set

Cross Validation

Randomize to 5 folds

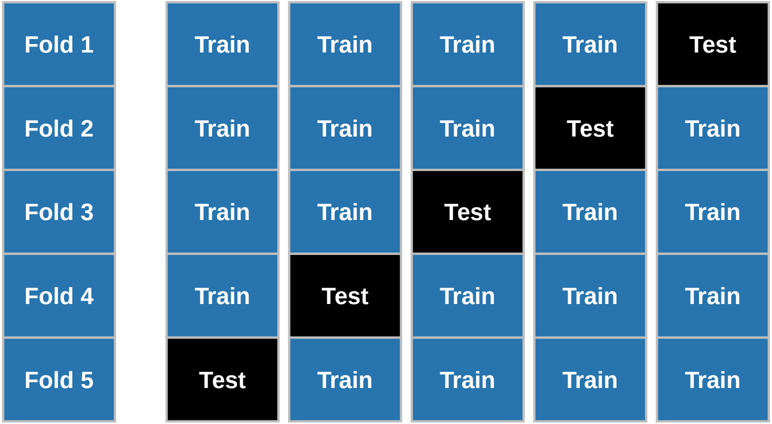
Iteratively use each as the test set



Cross Validation

Randomize to 5 folds

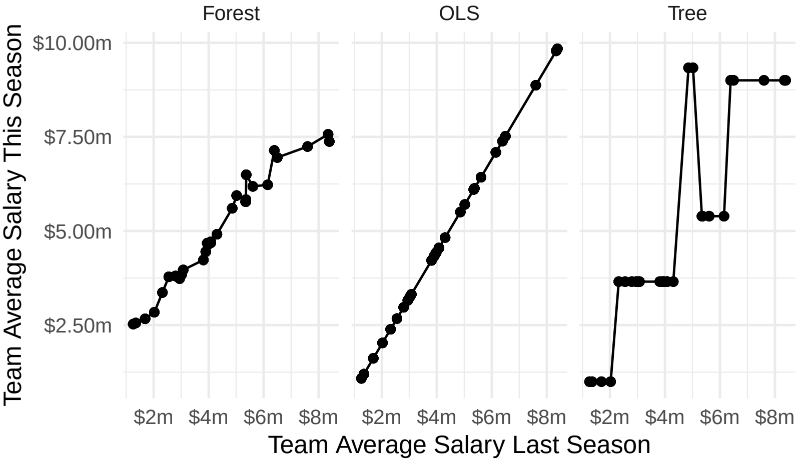
Iteratively use each as the test set

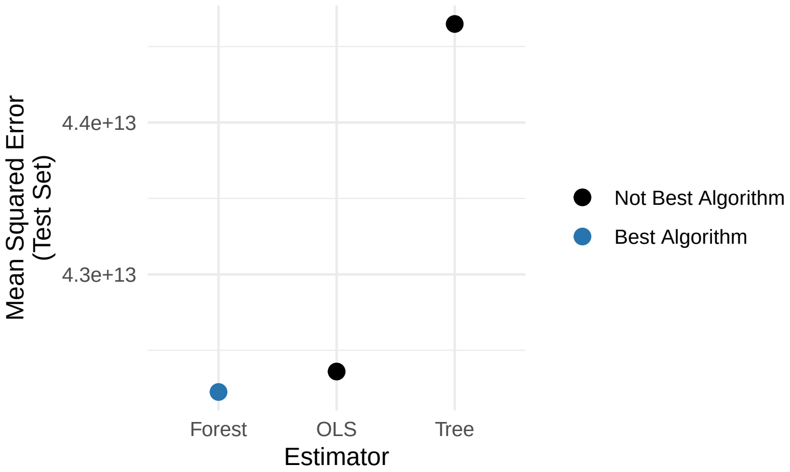


Average prediction error over folds

Out-of-sample predictive performance is not just for tuning parameters.

It can help you choose your algorithm.


Forest
OLS
Tree
Team Average Salary This Season
$10.00m
$7.50m
$5.00m
$2.50m
$2m
$4m
$6m
$8m
Team Average Salary Last Season


Mean Squared Error
(Test Set)
4.4e+13
4.3e+13
Forest
OLS
Tree
Estimator
Not Best Algorithm
Best Algorithm

Prediction and Estimation

Why Out-of-Sample Prediction

Sample Splitting

Cross Validation

Writing Task

Writing Task: A possible abstract

Imagine that your results came out really amazing.
Write the abstract of your paper with these imaginary results.

- Minimize jargon. Write for a New York Times reader.
- Emphasize big claims, not buried in statistics

Goal is to ask a high-impact question.
If possible, also write the abstract if results were opposite.

If your abstract is not compelling,
you might consider finding a new research question.

Learning goals for today

By the end of class, you will be able to

- Use individual prediction error as a metric to choose an algorithm that makes good group-level estimates
- Understand why we predict out of sample
- Carry out a sample split
- Explain the idea of cross-validation

On a different topic, we will close with a writing task.