Social Data Science

Soc 114
Winter 2026

Data-Driven Estimator Selection

Learning goals

- k-nearest-neighbors estimator
- bias-variance tradeoff
- sample splitting
- cross validation

A running example

- Sample 10 players from each MLB team
- Estimate sample average salary on each team
- Produces data where
  - Unit of analysis i is a team
  - Outcome y_i is average salary
  - Predictor x_i is prior year average salary

Goal: Predict mean salary of all Dodgers (sampled and unsampled)

Task

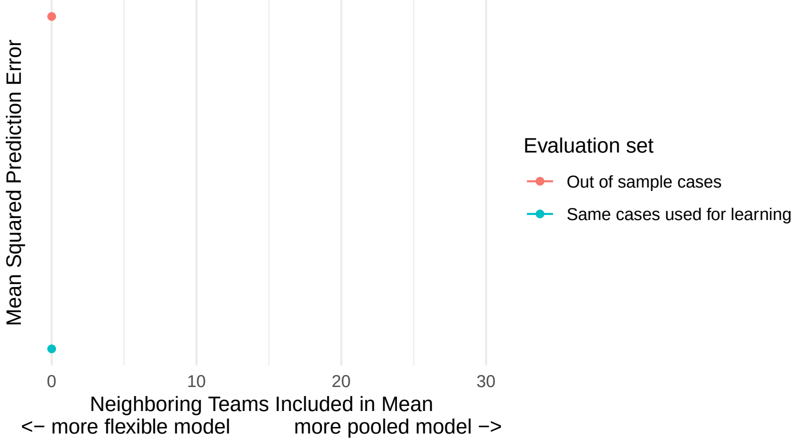
Estimator: k-nearest neighbors

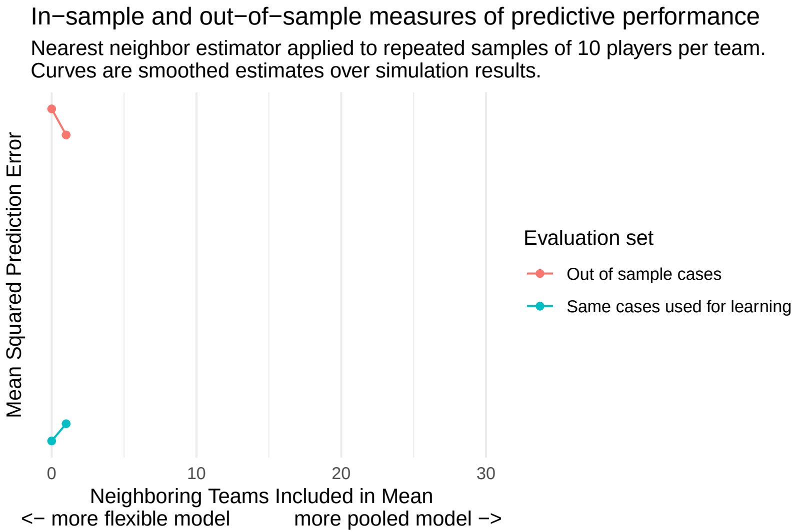
10 sampled players per team

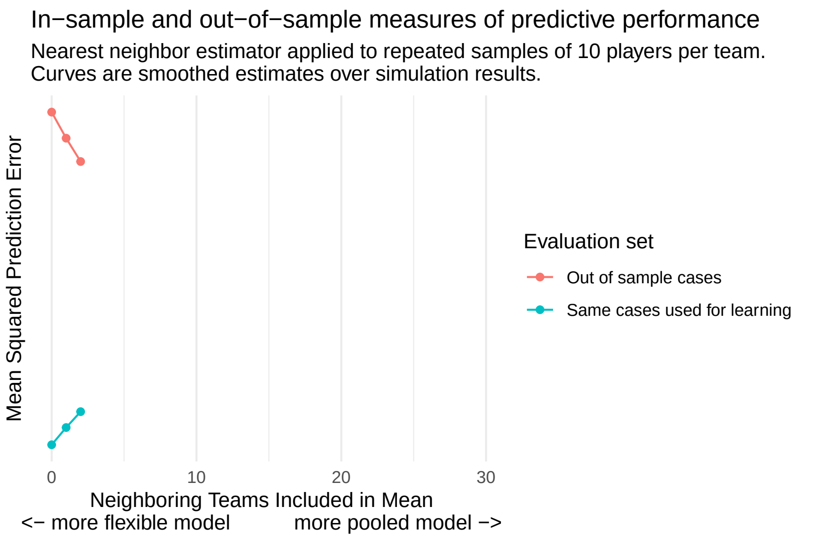
- Dodger sample mean might be noisy
- Could pool with similar teams defined by past mean salary
  - Dodgers: 8.39m
  - 1st-nearest neighbor. NY Mets: 8.34m
  - 2nd-nearest neighbor. NY Yankees: 7.60m
  - 3rd-nearest neighbor. Philadelphia: 6.50m
- How does performance change with the number of neighbors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

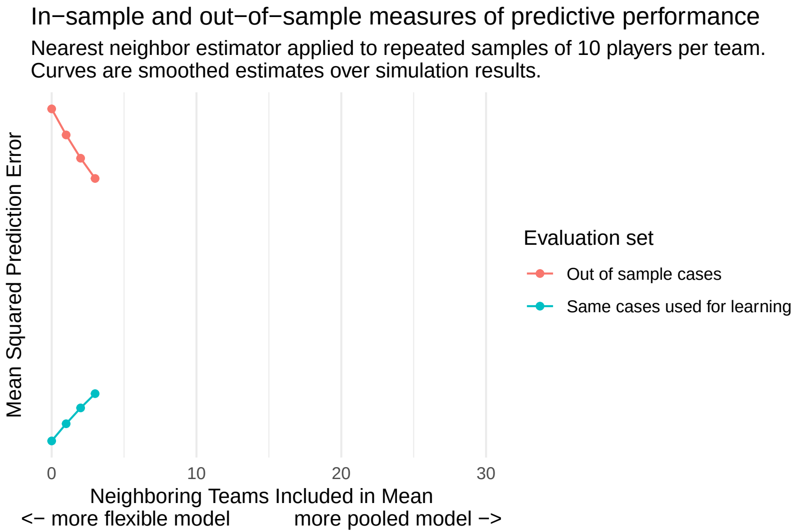
In−sample and out−of−sample measures of predictive performance

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

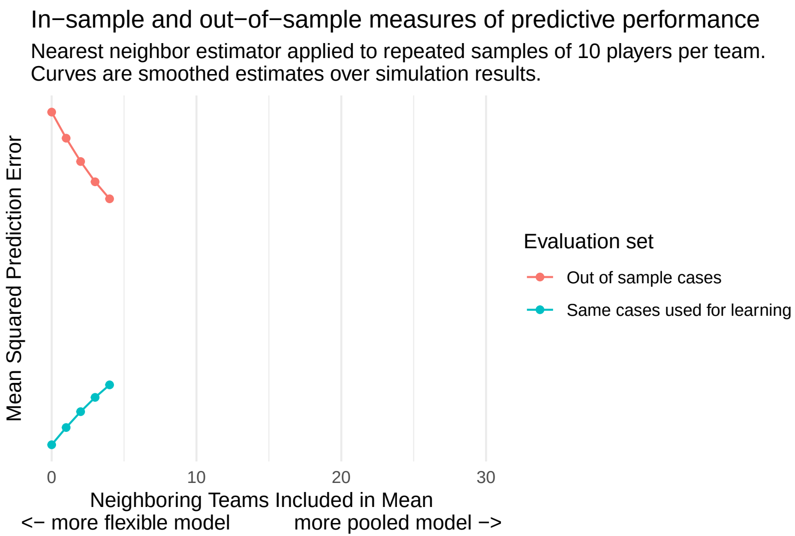


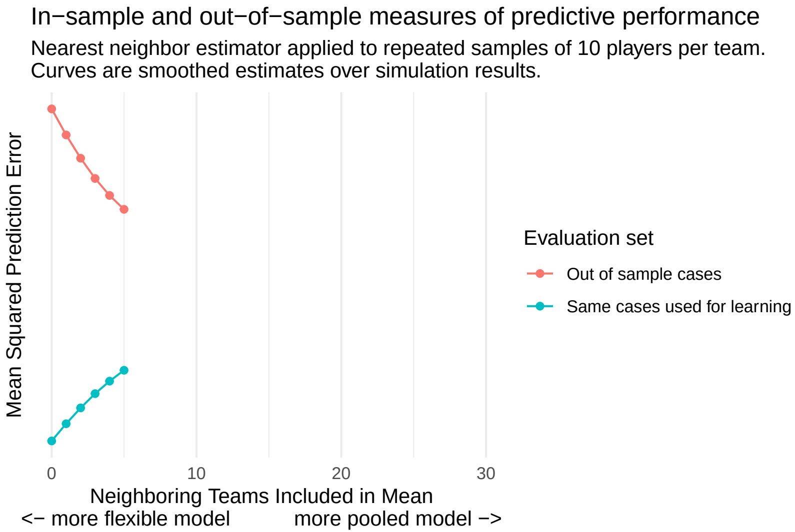

In−sample and out−of−sample measures of predictive performance
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.
Mean Squared Prediction Error
0
10
20
30
Neighboring Teams Included in Mean
<− more flexible model
more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

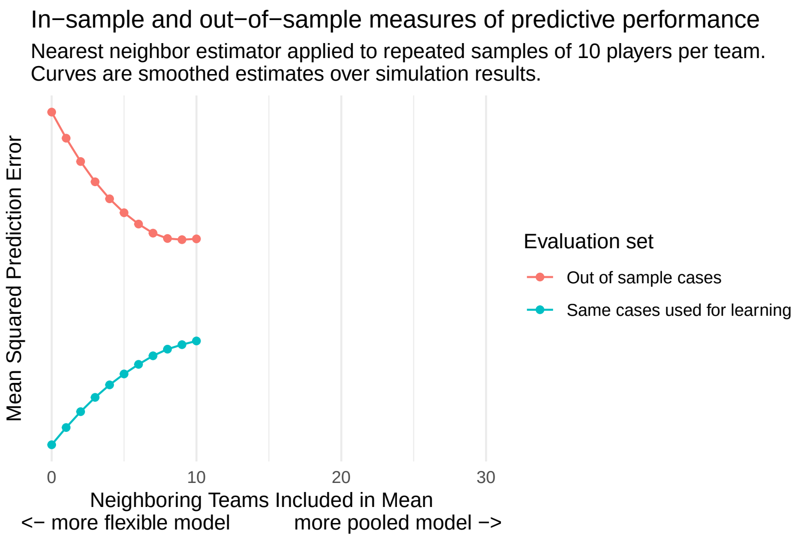

In−sample and out−of−sample measures of predictive performance
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.
Mean Squared Prediction Error
0
10
20
30
Neighboring Teams Included in Mean
<− more flexible model
more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

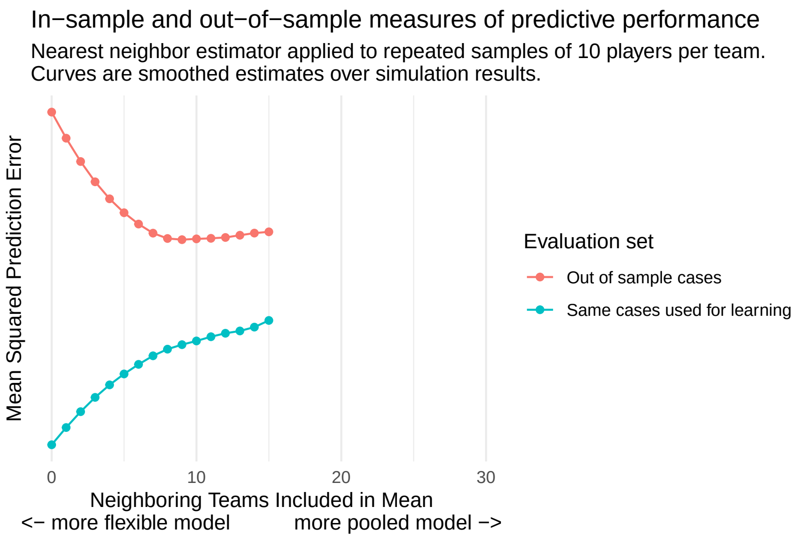


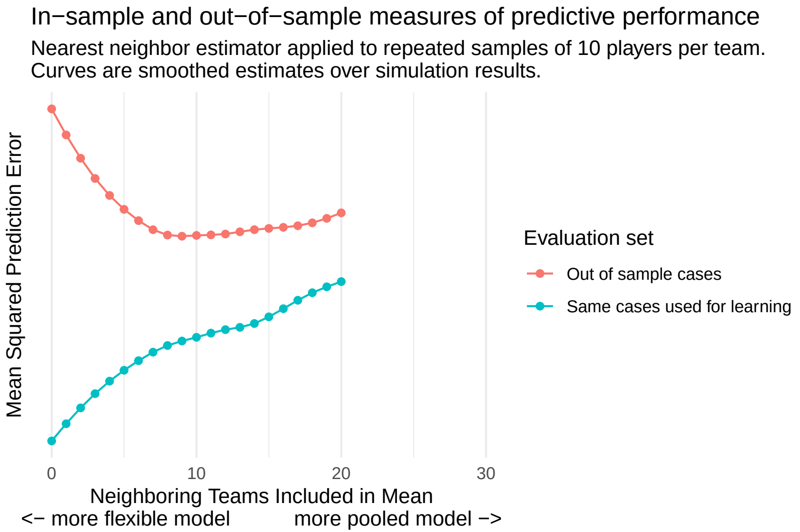
In−sample and out−of−sample measures of predictive performance
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.
Mean Squared Prediction Error
0
10
20
30
Neighboring Teams Included in Mean
<− more flexible model
more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

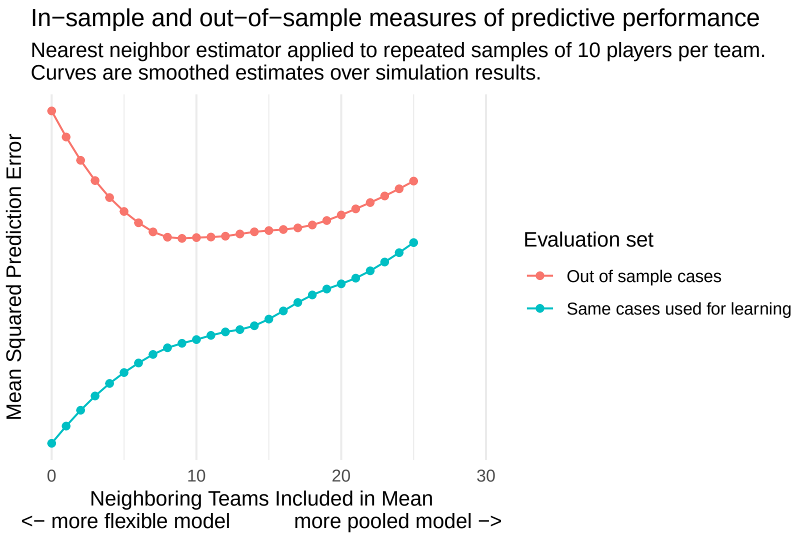

In−sample and out−of−sample measures of predictive performance
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.
Mean Squared Prediction Error
0
10
20
30
Neighboring Teams Included in Mean
<− more flexible model
more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

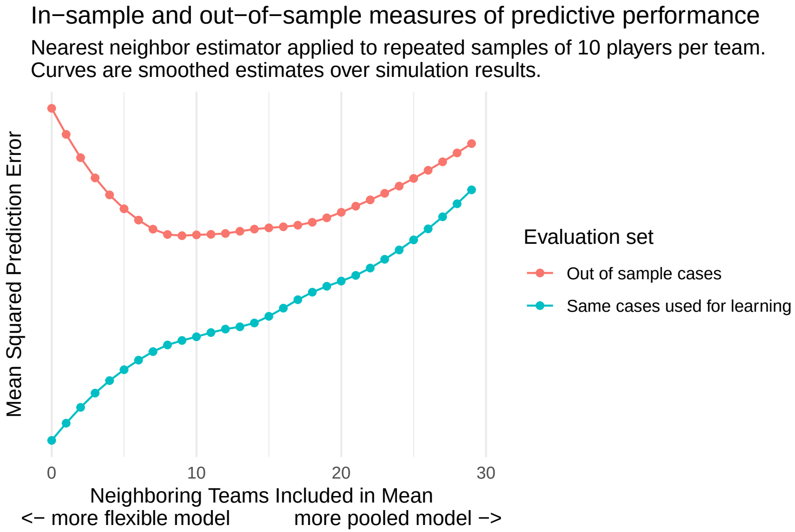

In−sample and out−of−sample measures of predictive performance
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.
Mean Squared Prediction Error
0
10
20
30
Neighboring Teams Included in Mean
<− more flexible model
more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

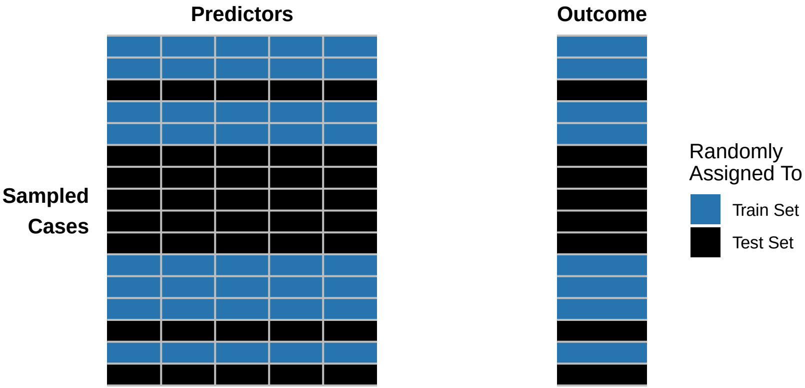

In−sample and out−of−sample measures of predictive performance
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.
Mean Squared Prediction Error
0
10
20
30
Neighboring Teams Included in Mean
<− more flexible model
more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

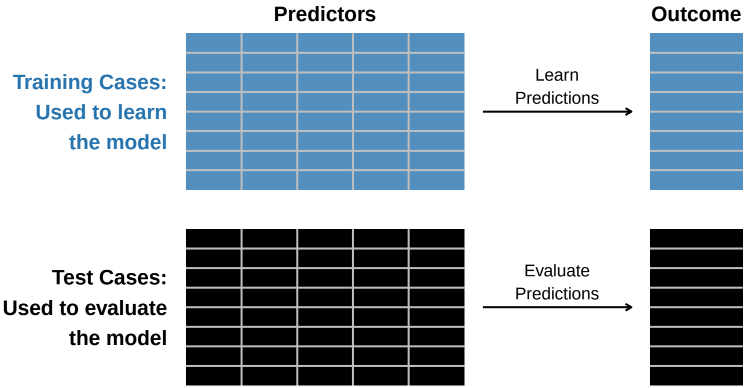


In−sample and out−of−sample measures of predictive performance
Nearest neighbor estimator applied to repeated samples of 10 players per team.
Curves are smoothed estimates over simulation results.
Mean Squared Prediction Error
0
10
20
30
Neighboring Teams Included in Mean
<− more flexible model
more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

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

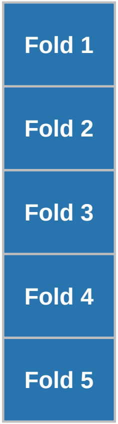
Cross Validation

A train test split loses lots of data to testing.

Is there a way to bring it back?

Cross Validation

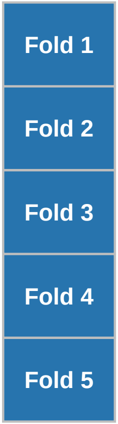
Randomize
to 5 folds



Cross Validation

Randomize
to 5 folds

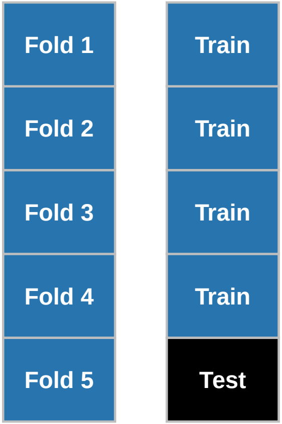
Iteratively use each as the test set



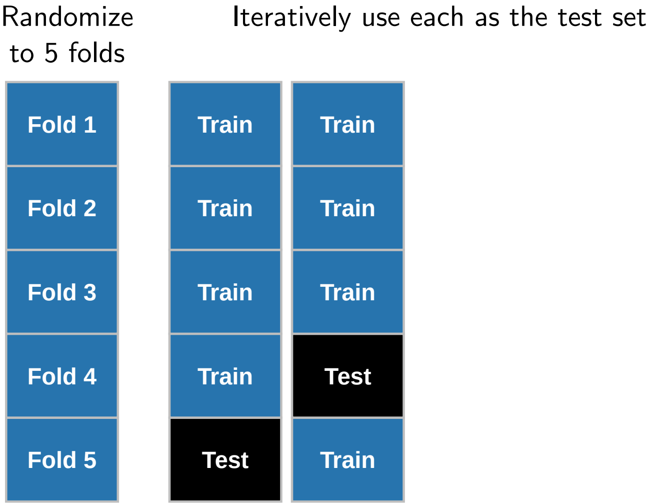
Cross Validation

Randomize to 5 folds

Iteratively use each as the test set



Cross Validation



Cross Validation

Randomize to 5 folds

Iteratively use each as the test set

Cross Validation

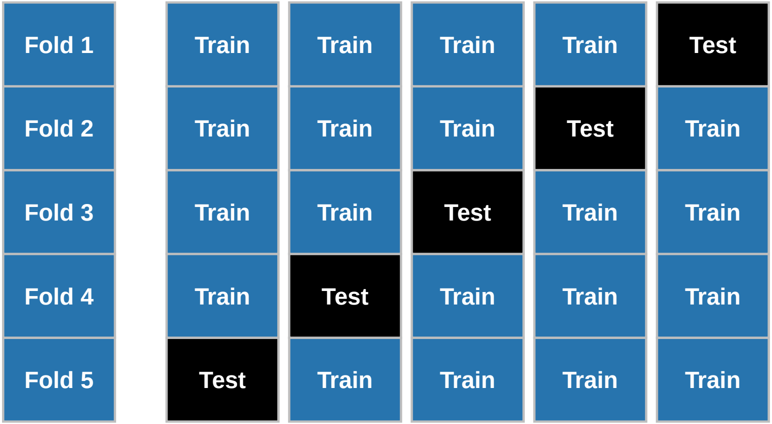
Randomize to 5 folds

Iteratively use each as the test set

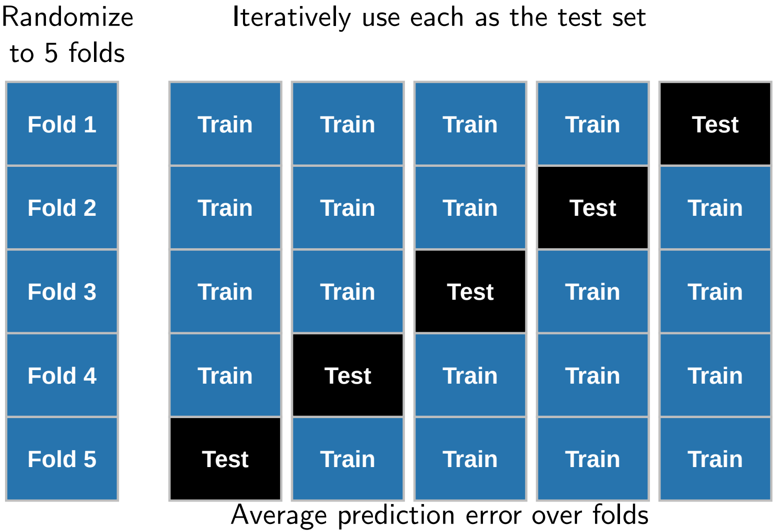
Cross Validation

Randomize to 5 folds

Iteratively use each as the test set

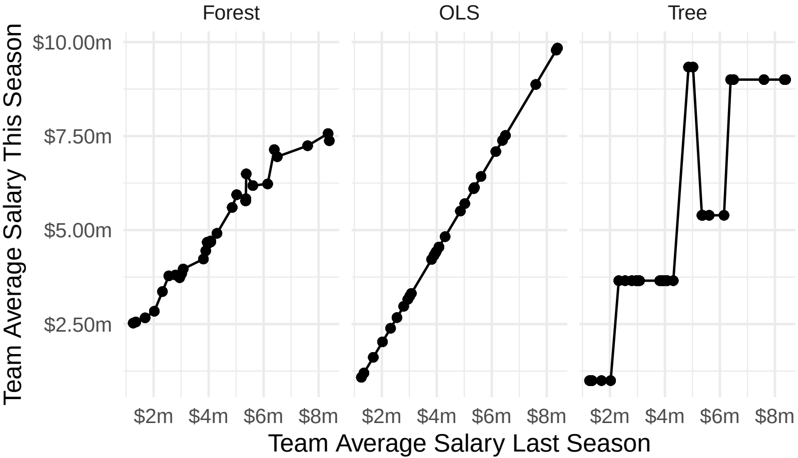


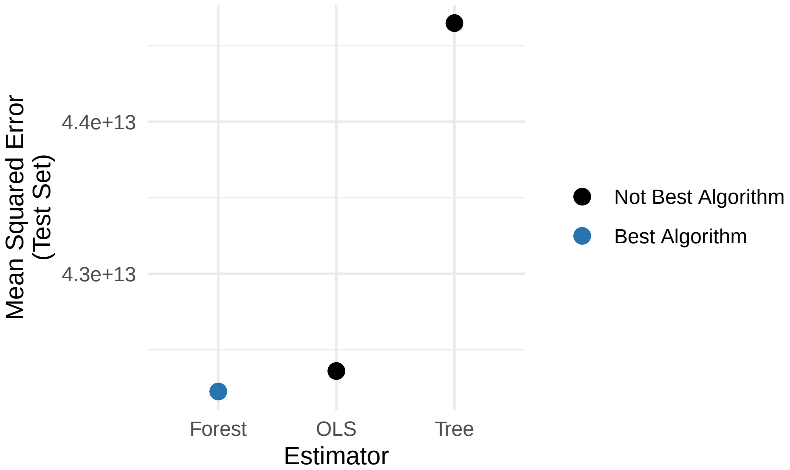
Cross Validation



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

Learning goals for today

By the end of class, you will be able to

- understand sample splitting: a common data science procedure for choosing among many candidate estimators