



Please Sit With Your Group

L-5

Model Evaluation

MAIB7002
Fundamentals of AI Models and Technologies

Peer Feedback

Work with your assigned partner group below

Group	Partner Group
A	X
B	N*
C	Alpha
D	S*
E	H
F	G
I	Delta
J	O

Group	Partner Group
K	Eta
L	P
M	R
Q	V
T	W
U	Epsilon
Y	Gamma
Z	Zeta

* Pairings assigned to resolve conflicting L4 reports

Group B: Please check in with the TA

Peer Feedback

5 min · Reconnect and Recap

Explain your current project and progress since L4

10 min · Exchange Feedback

Five minutes per group: discuss strengths, concerns and improvements

10 min · Plan the Evidence

Choose one change, a demo case, a baseline comparison and a measure of success

Ask your partners to challenge the plan

5 min · Refine and Submit

Record your feedback, planned change and evidence needed

Optional Coding Support Sessions

If you struggle with coding, attend a tutorial

Course code walkthroughs and Q&A on existing material

12:30–14:00, Block B, Cyberport 4

Date	Venue
30 September Completed ✓	CP-EFG, 2/F
7 October	CP4-J (Classroom J), 1/F, updated venue
14 October	CP-ABC, 3/F

Updated Assignment Dates

A1 Due Today: 2 October

A2 Released Tomorrow: 3 October

A2 covers L1–L5 and is due on 13 October

Predicting Hotel Booking Outcomes

One Booking

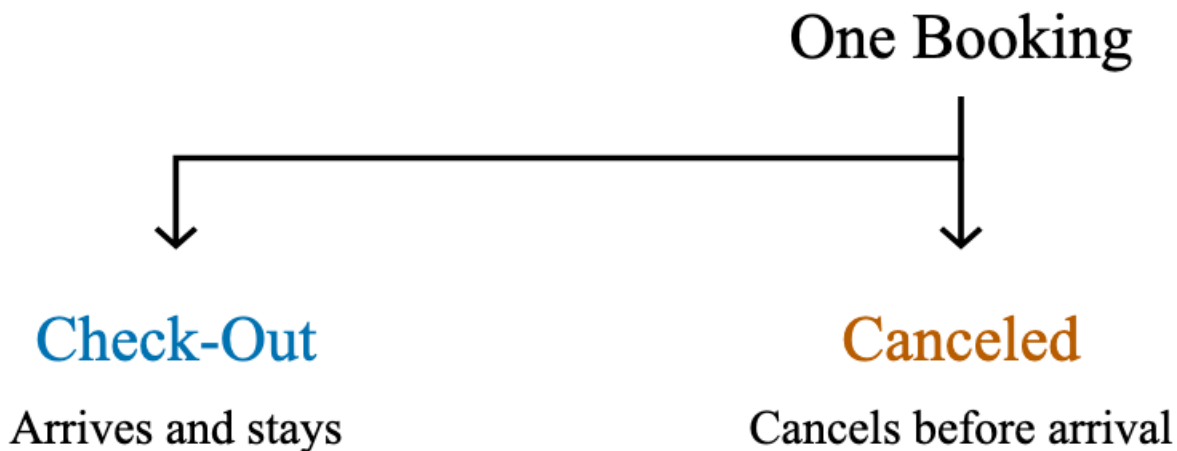
Real hotel data: <https://doi.org/10.1016/j.dib.2018.11.126>

Predicting Hotel Booking Outcomes



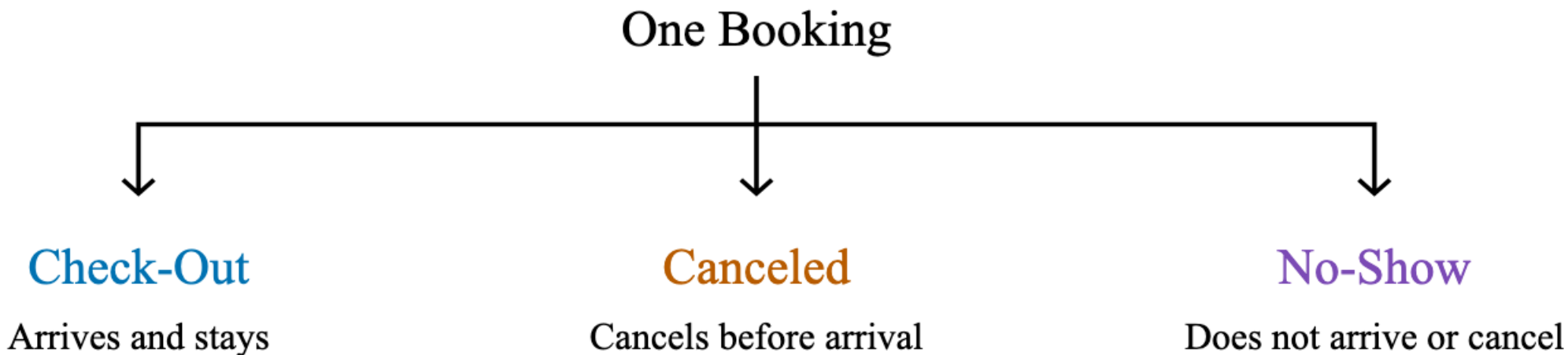
Real hotel data: <https://doi.org/10.1016/j.dib.2018.11.126>

Predicting Hotel Booking Outcomes



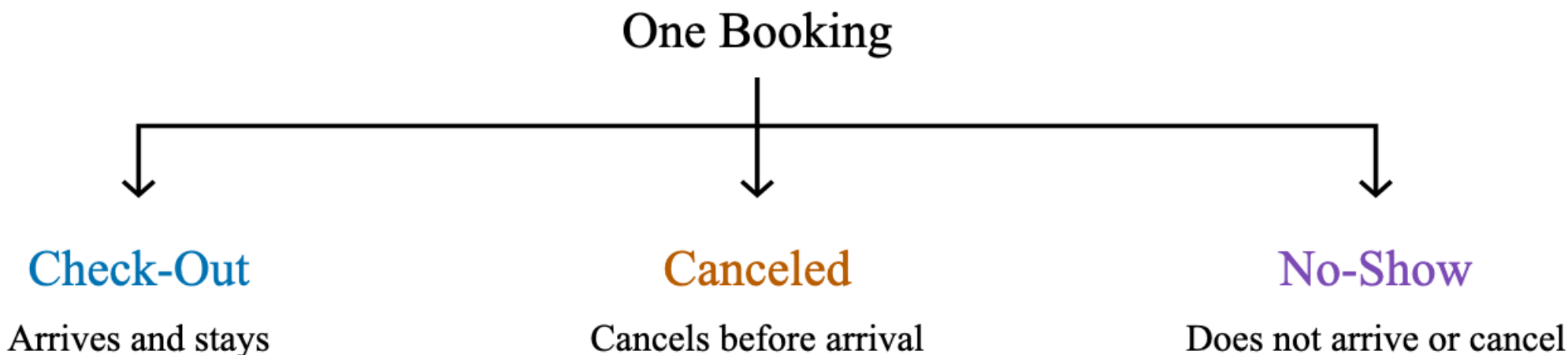
Real hotel data: <https://doi.org/10.1016/j.dib.2018.11.126>

Predicting Hotel Booking Outcomes



Real hotel data: <https://doi.org/10.1016/j.dib.2018.11.126>

Predicting Hotel Booking Outcomes



How can we fill more rooms without turning guests away?

Real hotel data: <https://doi.org/10.1016/j.dib.2018.11.126>

Inside the Hotel Dataset

119,390 rows (bookings) · 32 columns (variables)

Inputs: Features **X**

Days Ahead	Nights	Adults	Market Segment	Room Type	
17	3	2	Direct	F	
85	3	2	Online TA	A	
2	3	2	Online TA	G	
0	3	1	Corporate	A	
3	3	3	Online TA	A	
69	2	2	Offline TA/TO	A	

Inside the Hotel Dataset

119,390 rows (bookings) · 32 columns (variables)

Inputs: Features X					Outcome: Label y
Days Ahead	Nights	Adults	Market Segment	Room Type	reservation_status
17	3	2	Direct	F	Check-Out
85	3	2	Online TA	A	Canceled
2	3	2	Online TA	G	No-Show
0	3	1	Corporate	A	Check-Out
3	3	3	Online TA	A	Check-Out
69	2	2	Offline TA/TO	A	Check-Out

Explore the Hotel Bookings

Hotel	Days Ahead	Market Segment	Outcome	Nights	Adults	
All Hotels ▾	All ▾	All Segments ▾	All Outcomes ▾	All ▾	All ▾	Reset
Columns	Booking ▾	Previous	Next	119,390 matching bookings		
Days Ahead	Nights	Adults	Room	Segment	Outcome	
342	0	2	C	Direct	Check-Out	<div><div></div></div>
737	0	2	C	Direct	Check-Out	<div><div></div></div>
7	1	1	A	Direct	Check-Out	<div><div></div></div>
13	1	1	A	Corporate	Check-Out	<div><div></div></div>
14	2	2	A	Online TA	Check-Out	<div><div></div></div>
14	2	2	A	Online TA	Check-Out	<div><div></div></div>
0	2	2	C	Direct	Check-Out	<div><div></div></div>

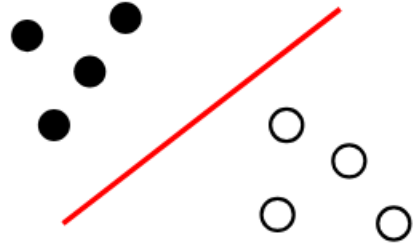
Four Ways to Classify a Booking

The same booking features **X**

Four Ways to Classify a Booking

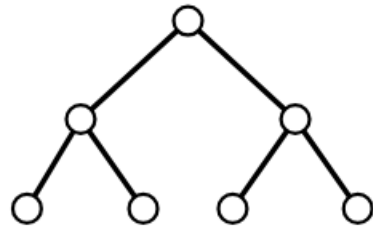
The same booking features $\mathbf{X}$

Perceptron



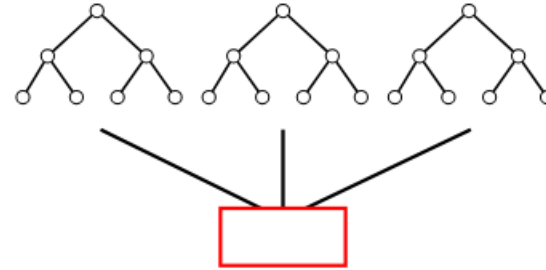
A linear boundary

Decision Tree



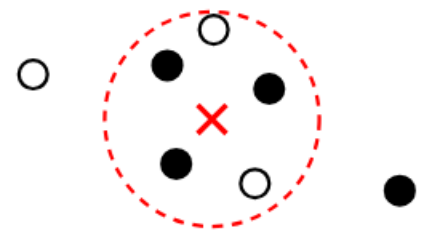
A sequence of questions

Random Forest



Combine many trees

kNN

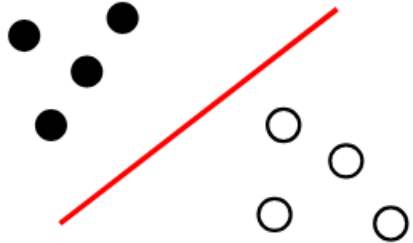


Look at nearby examples

Four Ways to Classify a Booking

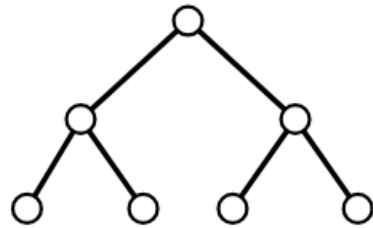
The same booking features $\mathbf{X}$

Perceptron



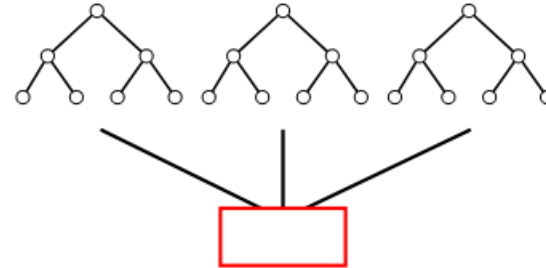
A linear boundary

Decision Tree



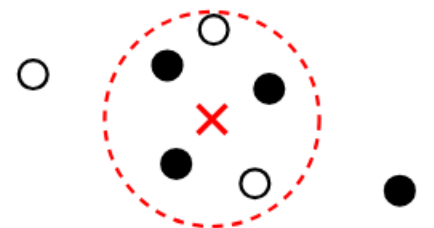
A sequence of questions

Random Forest



Combine many trees

kNN



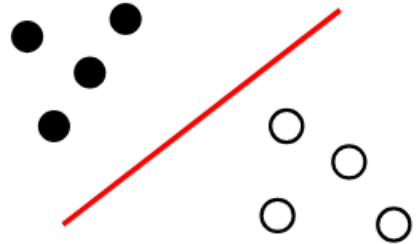
Look at nearby examples

Predict: Check-Out · Canceled · No-Show

Four Ways to Classify a Booking

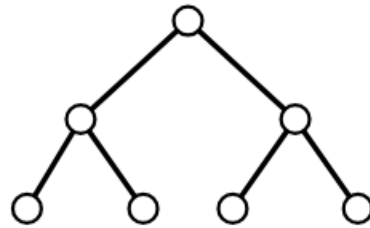
The same booking features $\mathbf{X}$

Perceptron



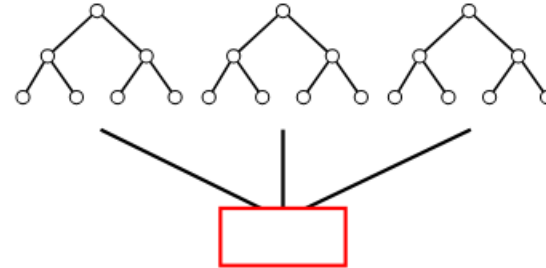
A linear boundary

Decision Tree



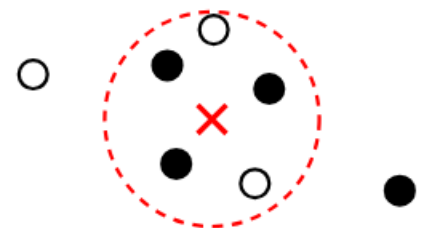
A sequence of questions

Random Forest



Combine many trees

kNN



Look at nearby examples

Predict: Check-Out · Canceled · No-Show

How can we compare predictions from different models?

How to Compare?

Confusion Matrix

Start with a naïve baseline: Always predict “Check-Out”

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out			
	Canceled			
	No-Show			

How to Compare?

Confusion Matrix

Start with a naïve baseline: Always predict “Check-Out”

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	75,166	0	0
	Canceled			
	No-Show			

How to Compare?

Confusion Matrix

Start with a naïve baseline: Always predict “Check-Out”

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	75,166	0	0
	Canceled	43,017	0	0
	No-Show			

How to Compare?

Confusion Matrix

Start with a naïve baseline: Always predict “Check-Out”

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	75,166	0	0
	Canceled	43,017	0	0
	No-Show	1,207	0	0

How to Compare?

Confusion Matrix

Start with a naïve baseline: Always predict “Check-Out”

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	75,166	0	0
	Canceled	43,017	0	0
	No-Show	1,207	0	0

Acc: 63%

Accuracy: correct predictions / all predictions

$$75,166 / 119,390 = 63\%$$

How to Compare?

Confusion Matrix

Start with a naïve baseline: Always predict “Check-Out”

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	75,166	0	0
	Canceled	43,017	0	0
	No-Show	1,207	0	0
		Acc: 63%		Err: 37%

Error Rate: incorrect predictions / all predictions

$$44,224 / 119,390 = 37\%$$

How to Compare?

Confusion Matrix

Start with a naïve baseline: Always predict “Check-Out”

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	75,166	0	0
	Canceled	43,017	0	0
	No-Show	1,207	0	0

Acc: 63% Err: 37%

Do All Errors Cost the Same?

Illustrative predictions for 100 bookings

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	55	7	1
	Canceled	8	26	2
	No-Show	1	0	0

Do All Errors Cost the Same?

Illustrative predictions for 100 bookings

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	55	7	1
	Canceled	8	26	2
	No-Show	1	0	0
		Acc: 81%		Err: 19%

Do All Errors Cost the Same?

Illustrative predictions for 100 bookings

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	55	7	1
	Canceled	8	26	2
	No-Show	1	0	0
		Acc: 81%		Err: 19%

1 actual No-Show predicted as Check-Out
A room held for the guest may remain empty

Do All Errors Cost the Same?

Illustrative predictions for 100 bookings

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	55	7	1
	Canceled	8	26	2
	No-Show	1	0	0
		Acc: 81%		Err: 19%

7 actual Check-Outs predicted as Canceled
If rooms are resold, arriving guests may have no room

Do All Errors Cost the Same?

Illustrative predictions for 100 bookings

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	55	7	1
	Canceled	8	26	2
	No-Show	1	0	0
		Acc: 81%		Err: 19%

Accuracy counts both as errors
Should the hotel treat them equally?

Do All Errors Cost the Same?

Illustrative predictions for 100 bookings

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	55	7	1
	Canceled	8	26	2
	No-Show	1	0	0
		Acc: 81%		Err: 19%

Select a source, then click a destination to move one booking

Reset Example

Quiz: Reading a Confusion Matrix

Illustrative predictions for 60 emails

<https://edstem.org/au/courses/41222/lessons/120632/slides/831167>

		Predicted		
		Important	Routine	Spam
Actual	Important	24	3	1
	Routine	2	16	2
	Spam	1	3	8

Accuracy

Correct predictions $\div$ all predictions

Error Rate

Incorrect predictions $\div$ all predictions

Precision for a class

Correct predictions for that class
 $\div$ its predicted total (**column**)

Recall for a class

Correct predictions for that class
 $\div$ its actual total (**row**)

Correct predictions lie on the diagonal

Generalisation

Decision Tree · maib7002.dirk.hk/L5.zip

Demo: hotel_training_fit.ipynb

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	71,597	3,540	29
	Canceled	2,743	40,268	6
	No-Show	343	60	804

Training Accuracy

94%

Err: 6%

112,669 correct / 119,390 bookings

Measured on the records used for training

Generalisation

Decision Tree · maib7002.dirk.hk/L5.zip

Demo: hotel_training_fit.ipynb

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	71,597	3,540	29
	Canceled	2,743	40,268	6
	No-Show	343	60	804

Training Accuracy

94%

Err: 6%

112,669 correct / 119,390 bookings

Measured on the records used for training

Do we expect 94% accuracy on new bookings?

Generalisation

Decision Tree · maib7002.dirk.hk/L5.zip

Demo: hotel_training_fit.ipynb

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	71,597	3,540	29
	Canceled	2,743	40,268	6
	No-Show	343	60	804

Training Accuracy

94%

Err: 6%

112,669 correct / 119,390 bookings

Measured on the records used for training

119,390 bookings used to fit · 0 unseen bookings tested

We predicted the same bookings used to train the model

Generalisation

Decision Tree · maib7002.dirk.hk/L5.zip

Demo: hotel_training_fit.ipynb

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	71,597	3,540	29
	Canceled	2,743	40,268	6
	No-Show	343	60	804

Training Accuracy

94%

Err: 6%

112,669 correct / 119,390 bookings

Measured on the records used for training

Training performance is measured · Generalisation is unknown

Generalisation means predicting unseen examples well

What should we set aside before training?

Holdout Evaluation

119,390 bookings · Stratified random split

Training · 80%
95,512 bookings

Holdout · 20%
23,878 bookings

Holdout Evaluation

119,390 bookings · Stratified random split

Training · 80%
95,512 bookings

Holdout · 20%
23,878 bookings

1 · Learn

Training bookings
Inputs + known outcomes



Fitted Tree

Holdout Evaluation

119,390 bookings · Stratified random split

Training · 80%
95,512 bookings

Holdout · 20%
23,878 bookings

1 · Learn

Training bookings
Inputs + known outcomes



Fitted Tree

2 · Predict

Holdout bookings
Inputs only



Same Tree

Illustrative predictions

Booking	Prediction	Actual Outcome
A	Check-Out	?
B	Canceled	?

Holdout Evaluation

119,390 bookings · Stratified random split

Training · 80%
95,512 bookings

Holdout · 20%
23,878 bookings

1 · Learn

Training bookings
Inputs + known outcomes



Fitted Tree

2 · Predict

Holdout bookings
Inputs only



Same Tree

Illustrative predictions

Booking	Prediction	Actual Outcome	3 · Compare
A	Check-Out	Check-Out	✓
B	Canceled	Check-Out	✗

Can the model predict outcomes it has not seen?

Does Training Performance Transfer?

Will accuracy be as high on the held-out bookings?

Demo: `hotel_holdout.ipynb`

Does Training Performance Transfer?

Will accuracy be as high on the held-out bookings?

Demo: hotel_holdout.ipynb

Training · 80% · 95,512 bookings

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	57,455	2,657	21
	Canceled	1,946	32,464	3
	No-Show	266	43	657

Acc: 95% Err: 5%

Does Training Performance Transfer?

Will accuracy be as high on the held-out bookings?

Demo: hotel_holdout.ipynb

Training · 80% · 95,512 bookings

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	57,455	2,657	21
	Canceled	1,946	32,464	3
	No-Show	266	43	657
		Acc: 95% Err: 5%		

Holdout · 20% · 23,878 bookings

		Predicted		
		Check-Out	Canceled	No-Show
Actual	Check-Out	12,152	2,745	136
	Canceled	2,540	6,030	34
	No-Show	147	53	41
		Acc: 76% Err: 24%		

Fresh fit on 80% · Same tree settings · A 19 percentage-point gap

Training, Validation and Test

Which tree depth should we choose?

119,390 bookings · Stratified random split

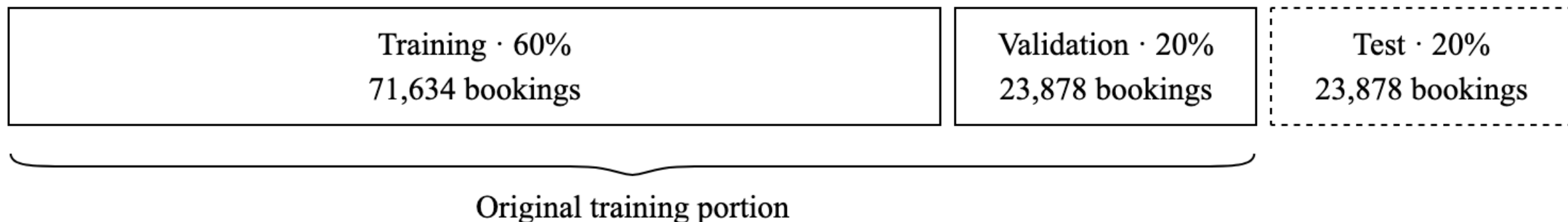
Training · 80%
95,512 bookings

Holdout · 20%
23,878 bookings

Training, Validation and Test

Which tree depth should we choose?

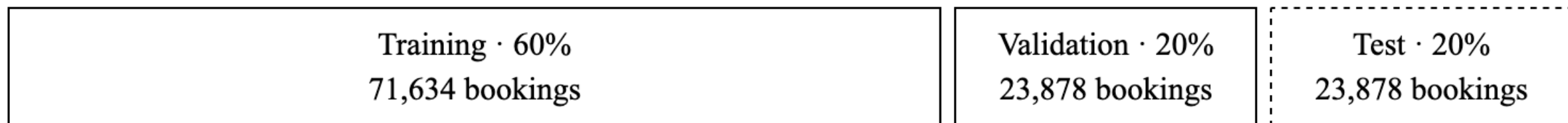
119,390 bookings · Stratified random split



Training, Validation and Test

Which tree depth should we choose?

119,390 bookings · Stratified random split



Original training portion

Fit

Try different tree depths

Choose

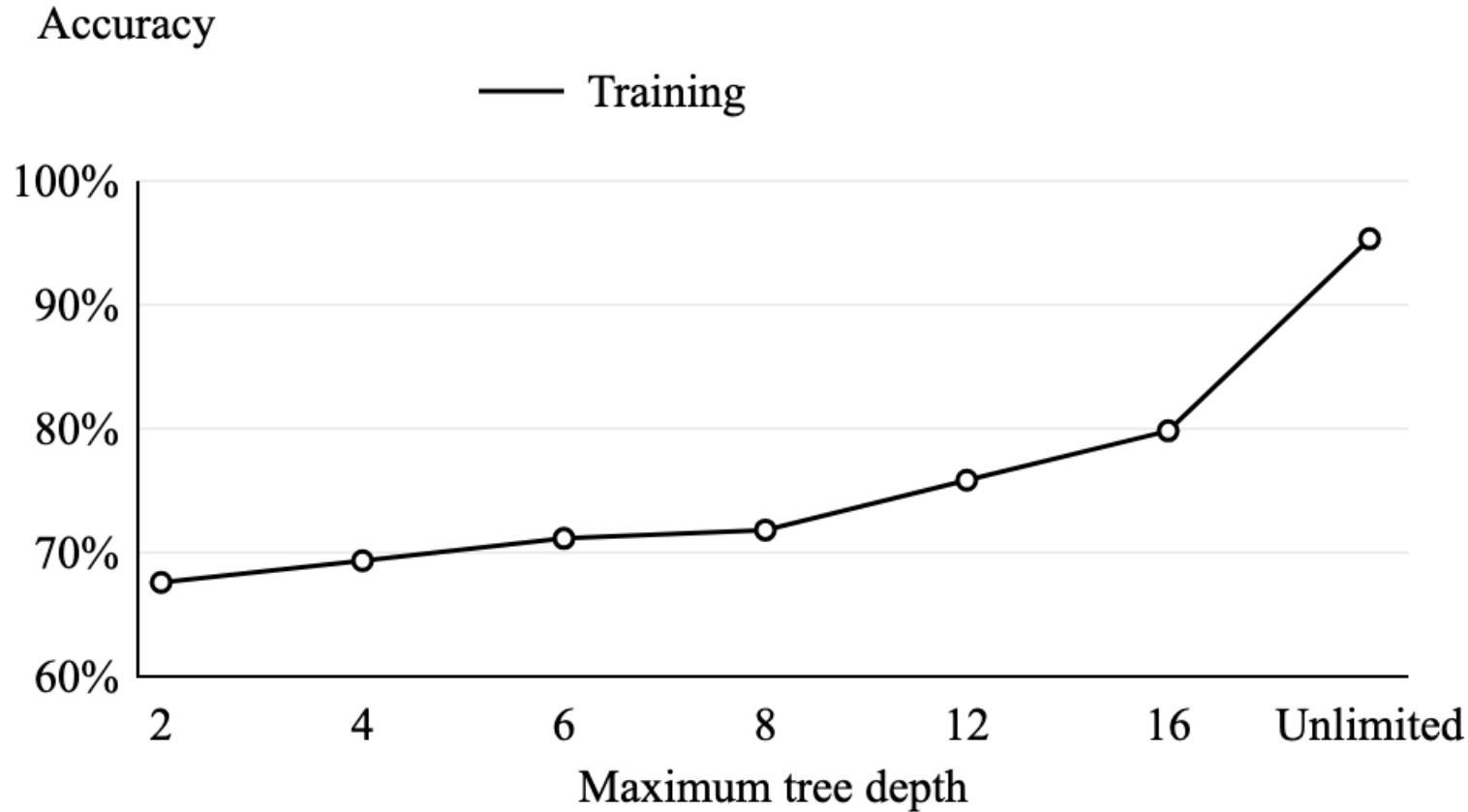
Select a depth

Evaluate

Assess the
chosen model

Validation guides the choice · Test assesses the chosen model

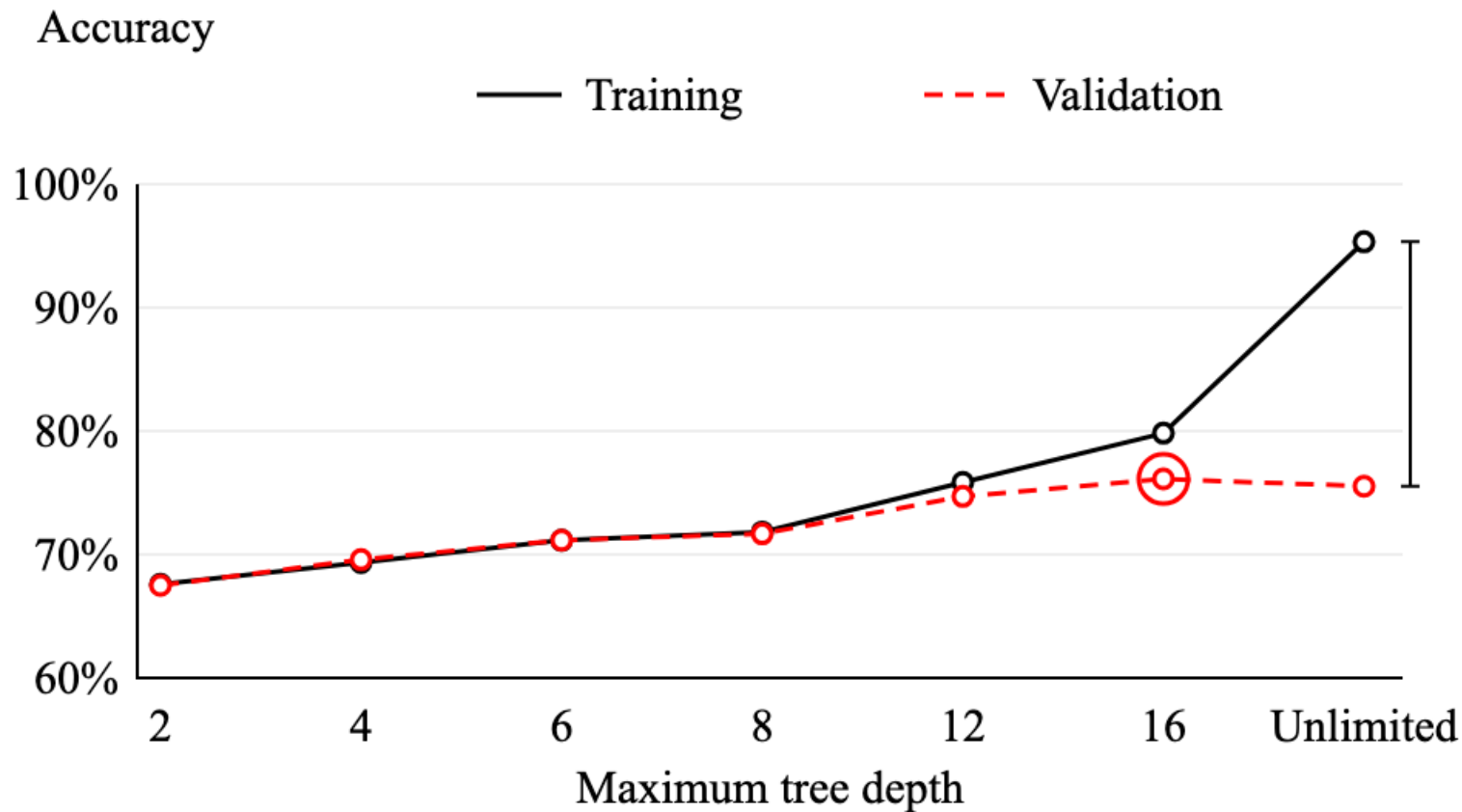
Choosing Tree Depth



Does better training accuracy mean a better model?

Demo: `hotel_validation.ipynb`

Choosing Tree Depth



Unlimited depth

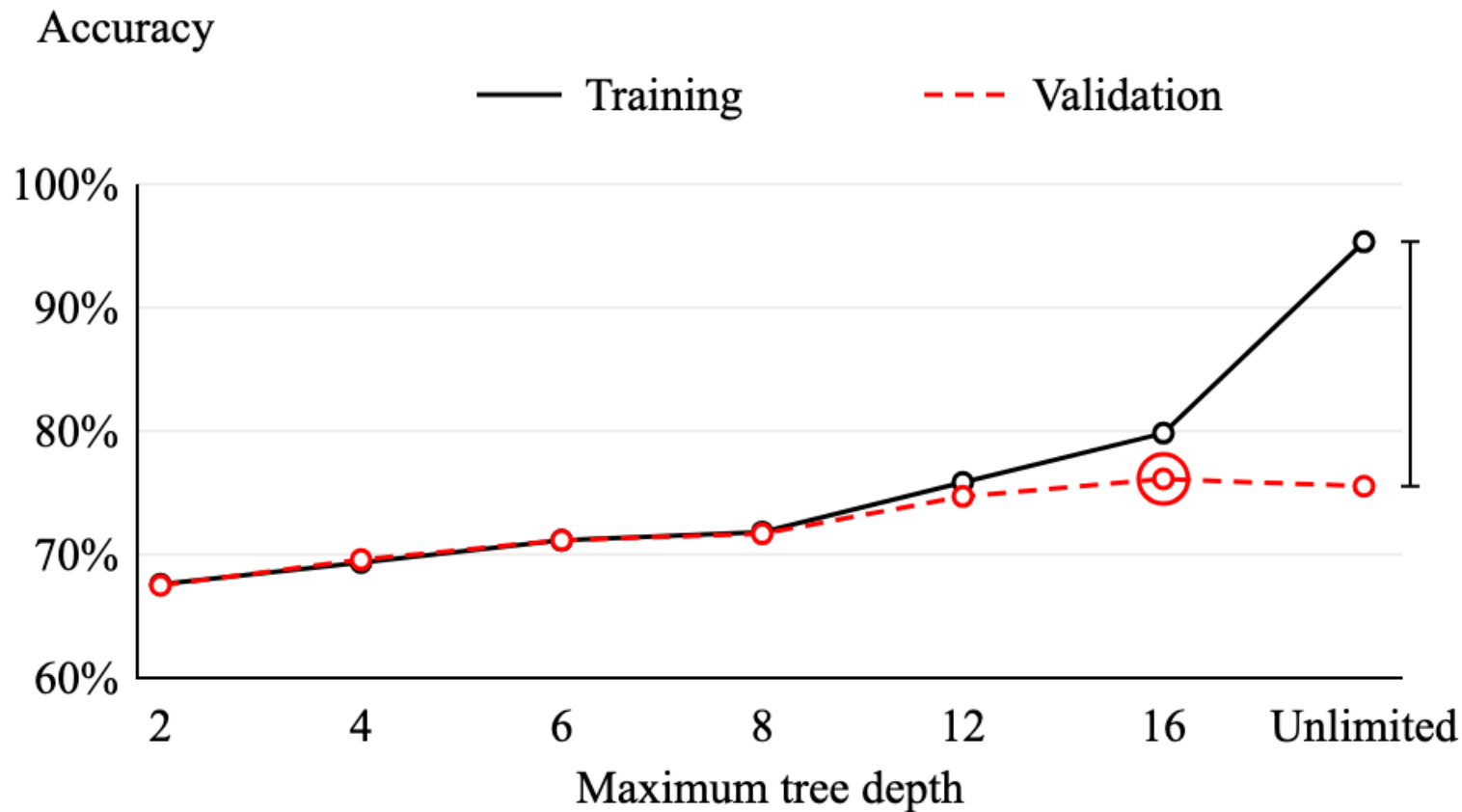
95% training

76% validation

Depth 16 gets **141 more** validation bookings correct than unlimited depth

Demo: `hotel_validation.ipynb`

Choosing Tree Depth



Choose depth 16

↓ Refit on training
+ validation

Test accuracy: **76%**

Choose depth using validation performance

Demo: `hotel_validation.ipynb`

Your Turn: Choose a Depth

Activity: `hotel_depth_activity.ipynb`

Run Step 1, then go to Step 2: Fill Four Blanks

1. In Step 2, replace the four _____ placeholders
2. Add **10** to the depth list; rerun the comparison and chart
3. Choose a depth using the validation results
4. Explain your choice and submit on Ed

Submit on Ed: depth-10 accuracy, chosen depth with its correct validation count,
and why training accuracy alone is insufficient

Hyperparameter Search

You chose a depth limit before fitting each tree

Chosen Before Fitting

Maximum depth
A hyperparameter

Learned During Fitting

Questions and thresholds
Fitted parameters

Hyperparameter Search

You chose a depth limit before fitting each tree

Chosen Before Fitting

Maximum depth
A hyperparameter

Learned During Fitting

Questions and thresholds
Fitted parameters

Candidate depth limits

2 4 6 8 10 12 16 Unlimited

Hyperparameter Search

You chose a depth limit before fitting each tree

Chosen Before Fitting

Maximum depth
A hyperparameter

Learned During Fitting

Questions and thresholds
Fitted parameters

Candidate depth limits

2

4

6

8

10

12

16

Unlimited

Fit on training → Compare on validation → Choose

The comparison loop performed hyperparameter search

Automatic Tree Search

GridSearchCV compares eight depth limits

2 4 6 8 10 12 16 Unlimited

Demo: `random_forest_search.ipynb`

Automatic Tree Search

GridSearchCV compares eight depth limits

Fit each tree on the same 71,634 training bookings

2



4



6



8



10



12



16



Unlimited

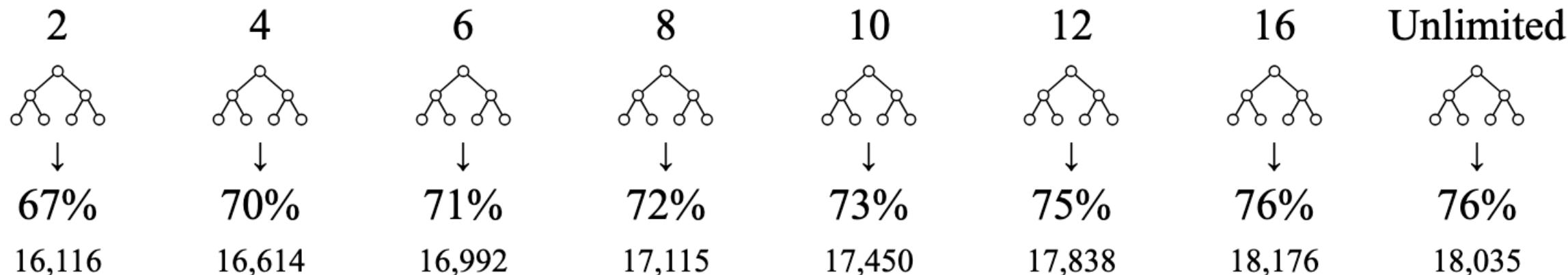


Demo: `random_forest_search.ipynb`

Automatic Tree Search

GridSearchCV compares eight depth limits

Fit each tree on the same 71,634 training bookings



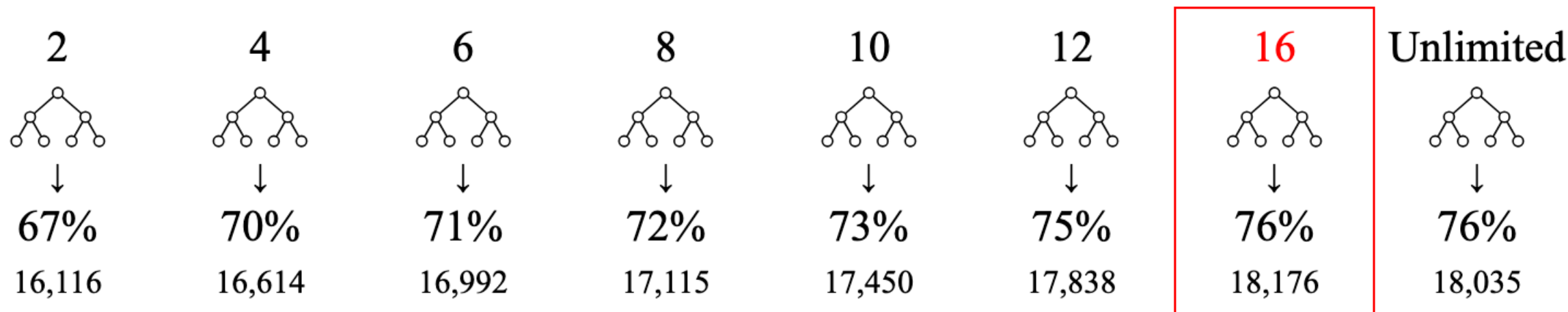
Validation accuracy and correct predictions out of 23,878 bookings

Demo: `random_forest_search.ipynb`

Automatic Tree Search

GridSearchCV compares eight depth limits

Fit each tree on the same 71,634 training bookings



Validation accuracy and correct predictions out of 23,878 bookings

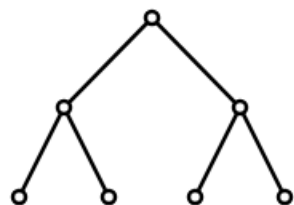
Select depth 16 · Best among these candidates

Demo: random_forest_search.ipynb

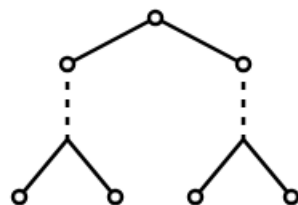
Random Forest Settings

Maximum Depth

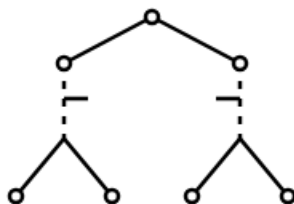
`max_depth`



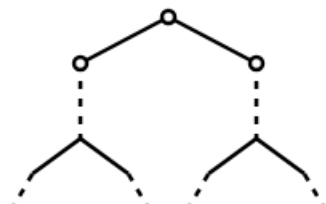
2



8



16



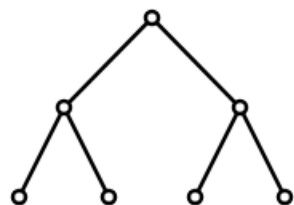
Unlimited

Schematic drawings · Labels specify the depth limits and tree counts

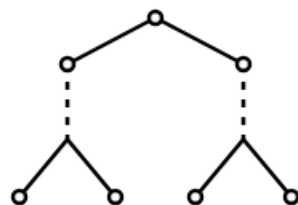
Random Forest Settings

Maximum Depth

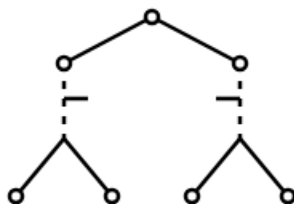
`max_depth`



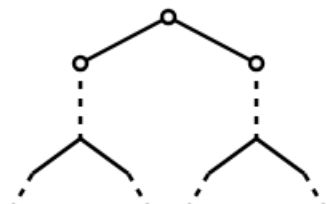
2



8



16



Unlimited

Number of Trees

`n_estimators`



1 Tree



10 Trees



100 Trees



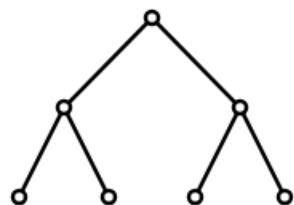
300 Trees

Schematic drawings · Labels specify the depth limits and tree counts

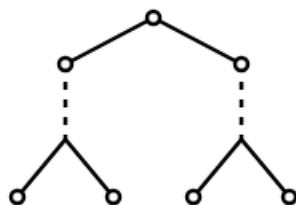
Random Forest Settings

Maximum Depth

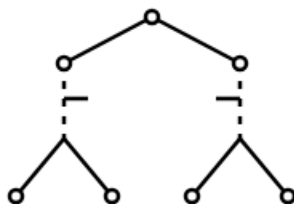
`max_depth`



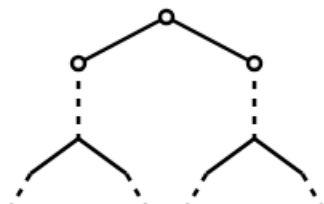
2



8



16



Unlimited

Number of Trees

`n_estimators`



1 Tree



10 Trees



100 Trees

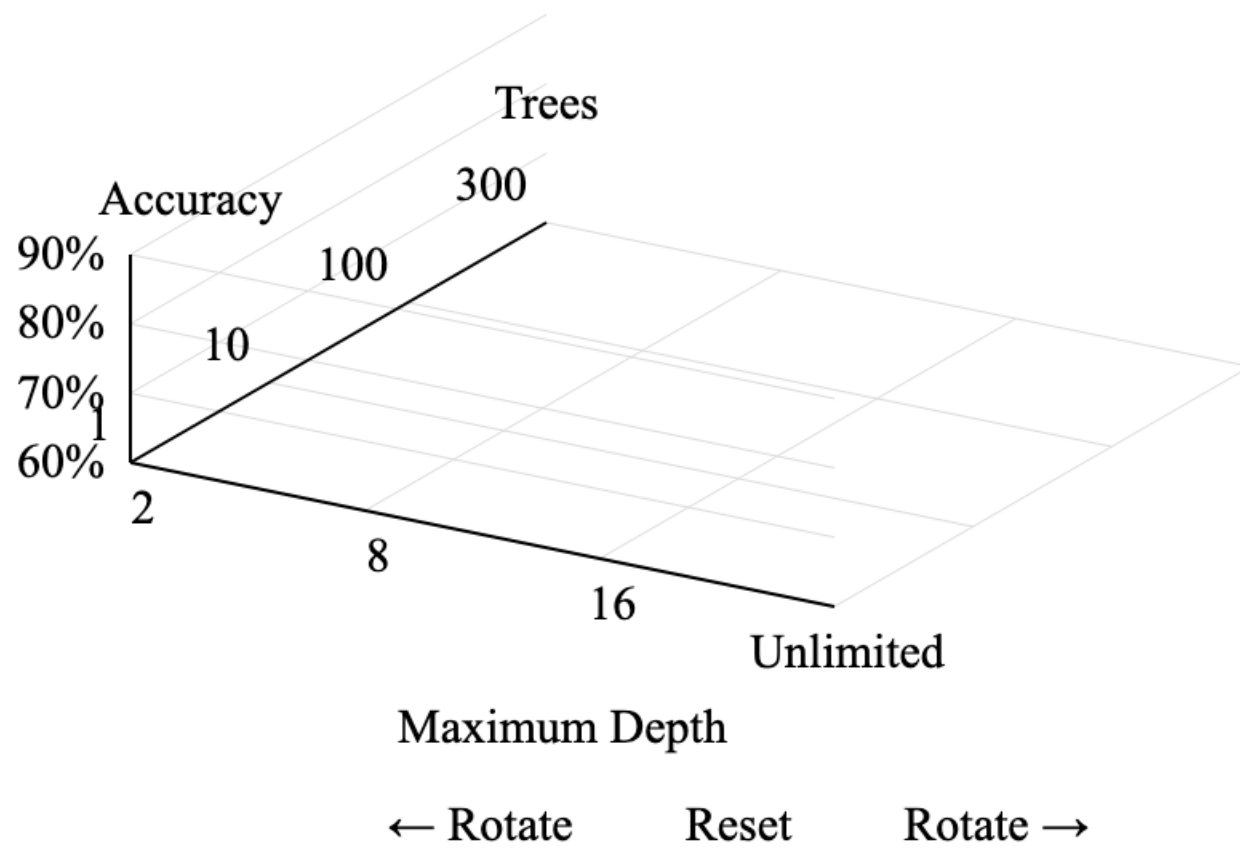


300 Trees

Schematic drawings · Labels specify the depth limits and tree counts

4 Depth Choices × 4 Tree Counts = **16 Combinations**

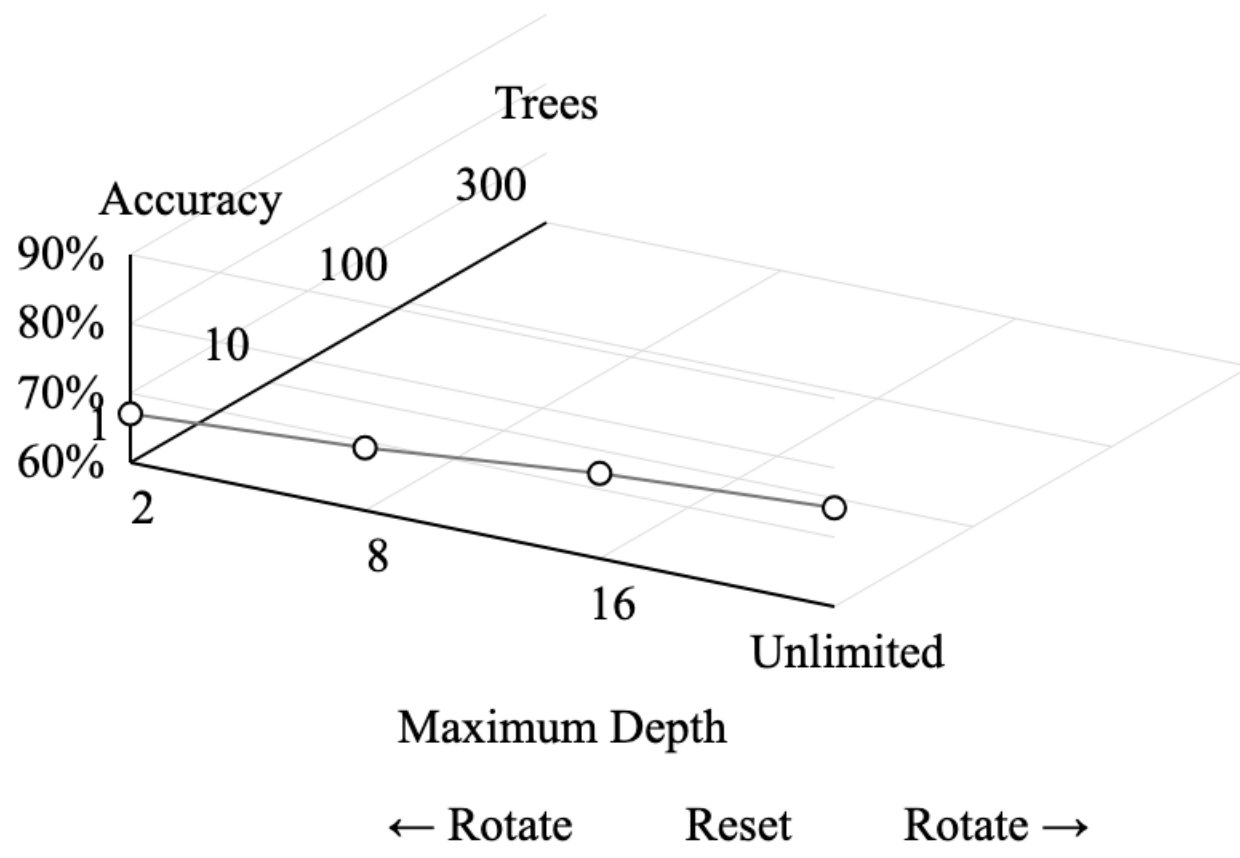
Random Forest Search



Which settings would you choose?

Tested choices are equally spaced

Random Forest Search

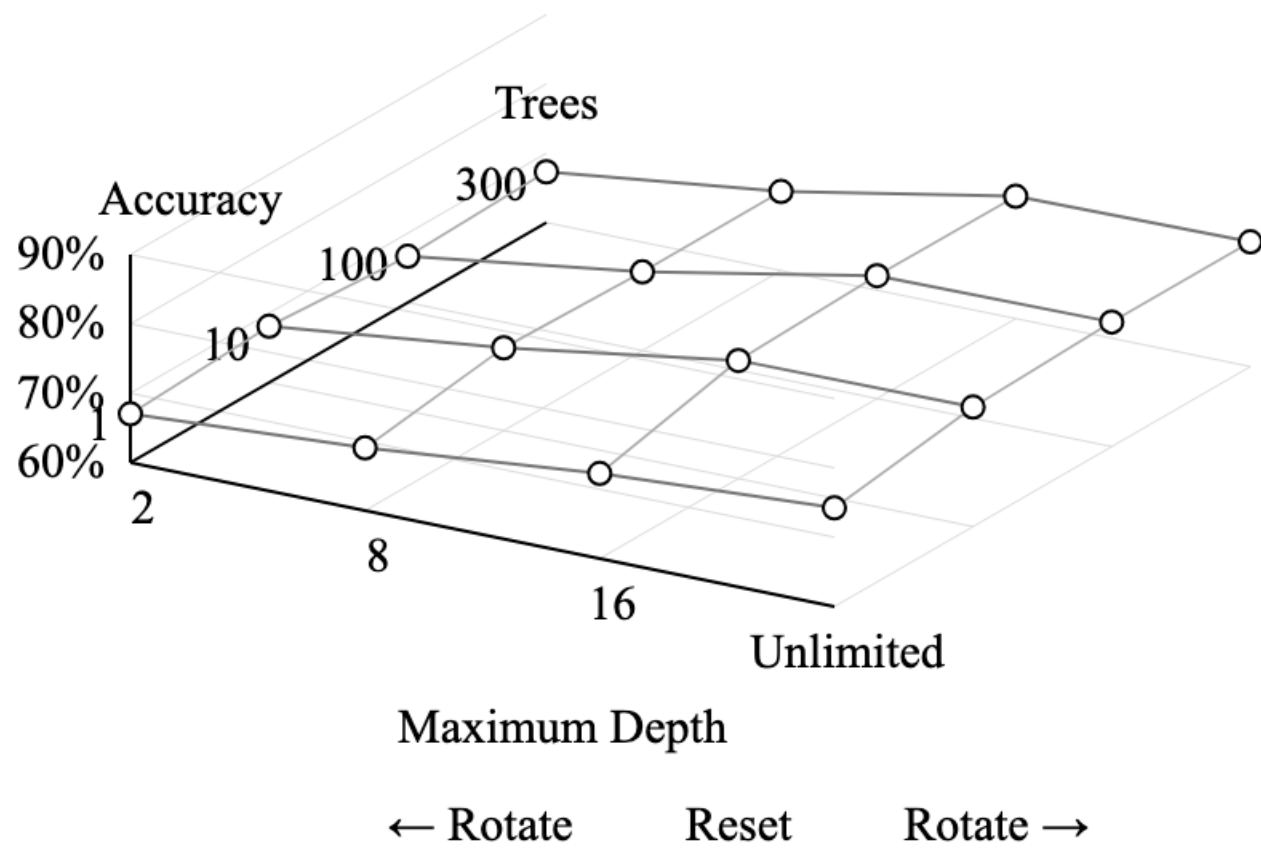


Tested choices are equally spaced

Which settings would you choose?

Select a point to inspect

Random Forest Search

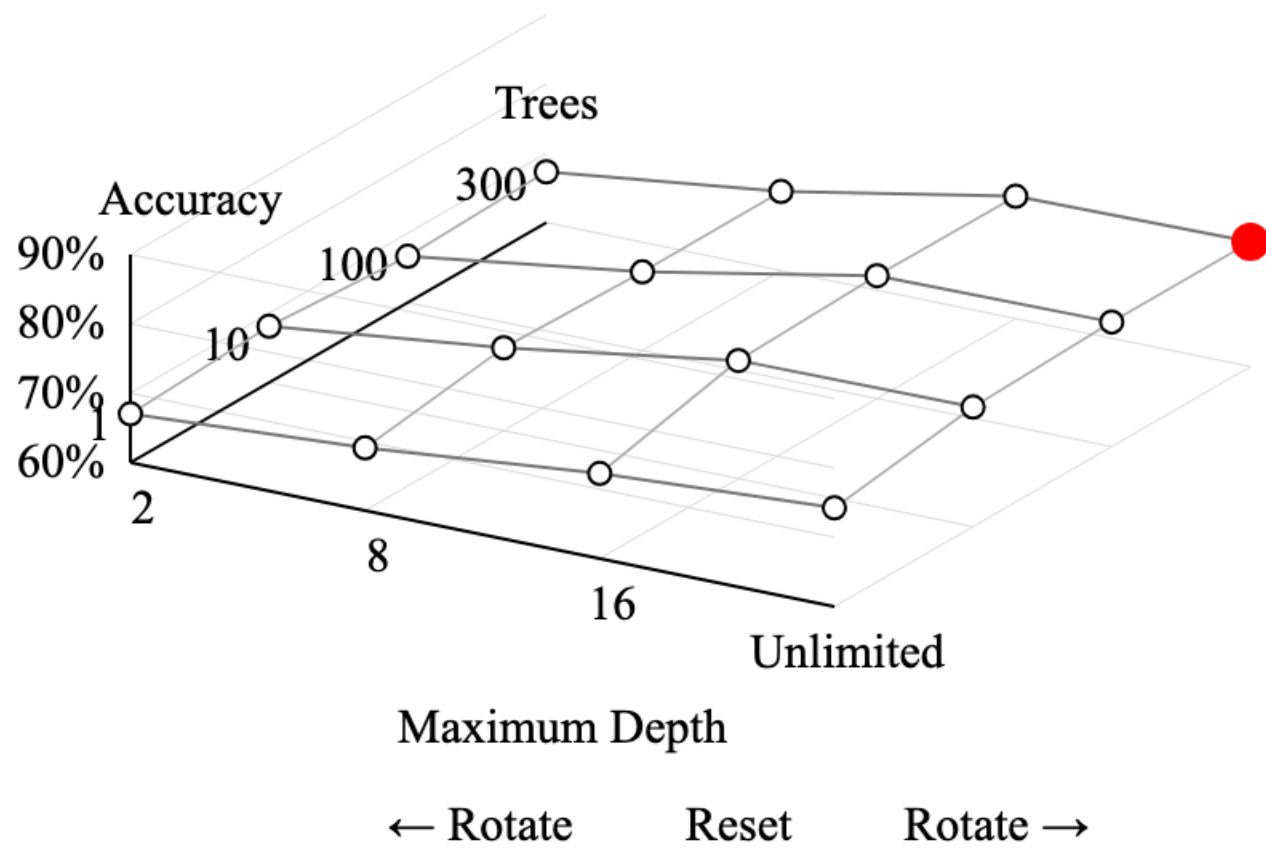


Tested choices are equally spaced

Which settings would you choose?

Select a point to inspect

Random Forest Search



Tested choices are equally spaced

Validation Accuracy

78%

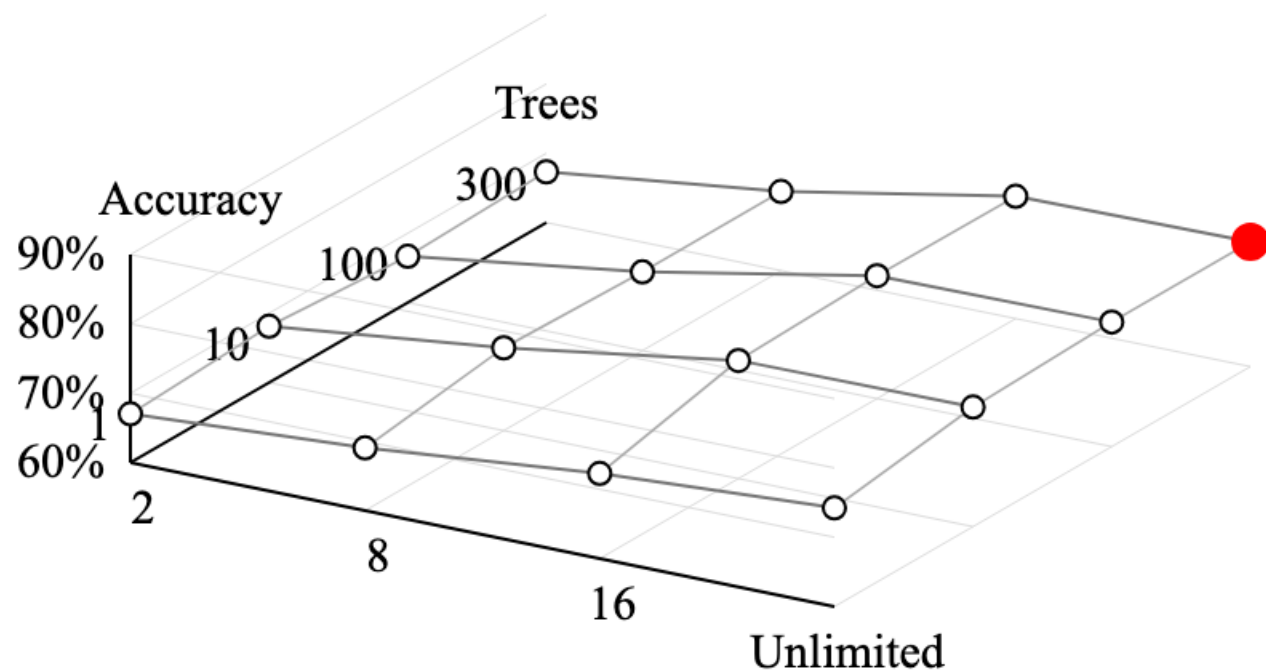
300 Trees

Unlimited Depth

18,621 / 23,878 correct

Select a point to inspect

Random Forest Search



Validation Accuracy

78%

300 Trees

Unlimited Depth

18,621 / 23,878 correct

Select a point to inspect

← Rotate

Reset

Rotate →

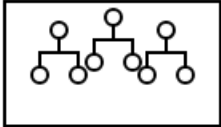
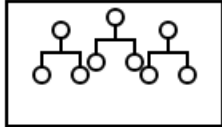
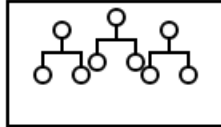
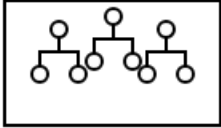
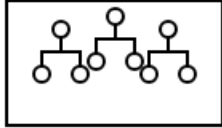
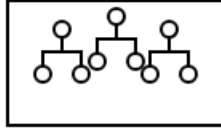
Tested choices are equally spaced

Demo: random_forest_search.ipynb

A Larger Search

Setting	Choices
Tree Count	100, 300
Maximum Depth	16, 32, Unlimited

Refine the Earlier Grid

	Maximum Depth		
	16	32	Unlimited
100 Trees			
300 Trees			

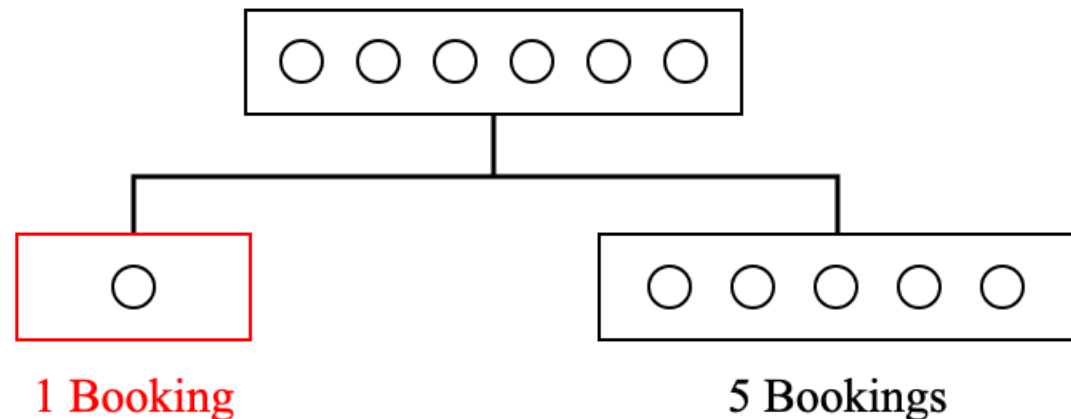
Keep larger forests; refine the depth choices

$$2 \times 3 = 6 \text{ Combinations}$$

A Larger Search

Setting	Choices
Tree Count	100, 300
Maximum Depth	16, 32, Unlimited
Minimum Leaf Size	1, 2, 5

Example: Minimum Leaf Size = 2



Reject This Split

Every leaf needs at least 2 bookings

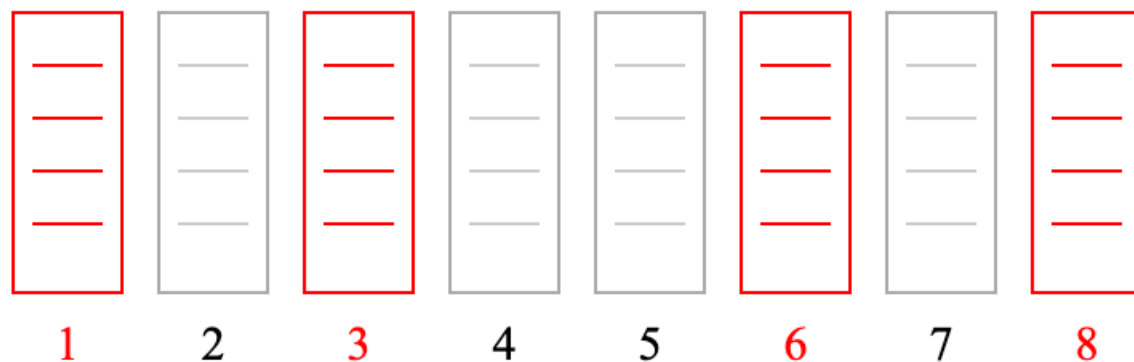
$$2 \times 3 \times 3 = 18 \text{ Combinations}$$

A Larger Search

Setting	Choices
Tree Count	100, 300
Maximum Depth	16, 32, Unlimited
Minimum Leaf Size	1, 2, 5
Features per Split	Square Root, Half, All

Example: Half the Features

Consider 4 of 8 encoded features at this split



A new subset is drawn at each split

$$2 \times 3 \times 3 \times 3 = 54 \text{ Combinations}$$

A Larger Search

Setting	Choices
Tree Count	100, 300
Maximum Depth	16, 32, Unlimited
Minimum Leaf Size	1, 2, 5
Features per Split	Square Root, Half, All
Bootstrap Draws	70%, 100%

Example: 10 Training Bookings

Training



100%: 10 Draws



70%: 7 Draws

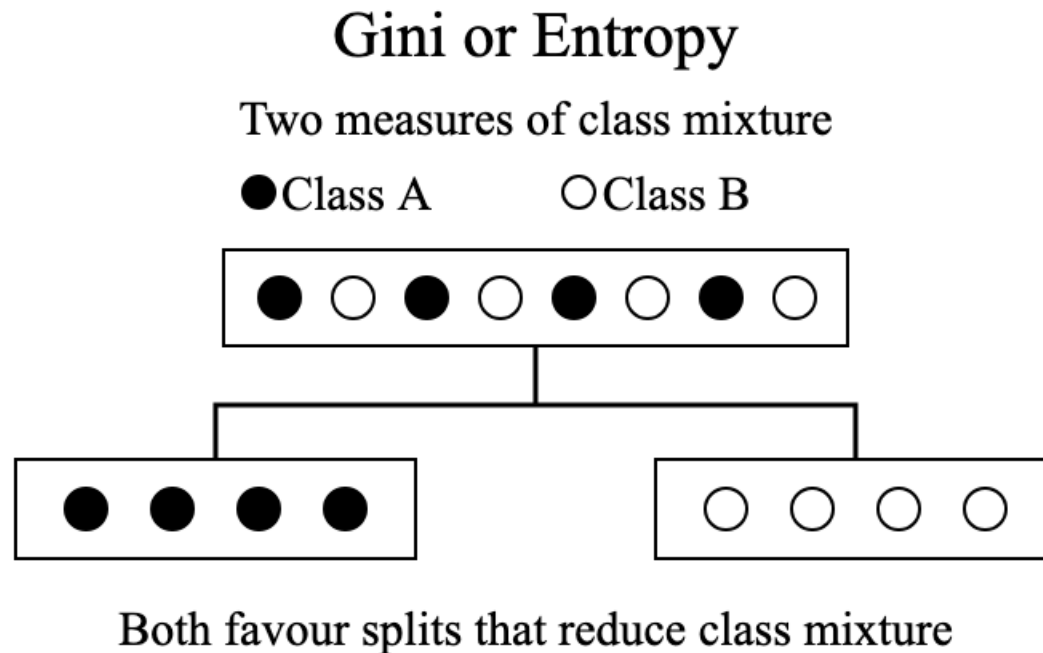


Draw with replacement: repeats and omissions are possible

$$2 \times 3 \times 3 \times 3 \times 2 = 108 \text{ Combinations}$$

A Larger Search

Setting	Choices
Tree Count	100, 300
Maximum Depth	16, 32, Unlimited
Minimum Leaf Size	1, 2, 5
Features per Split	Square Root, Half, All
Bootstrap Draws	70%, 100%
Split Rule	Gini, Entropy



$$2 \times 3 \times 3 \times 3 \times 2 \times 2 = 216 \text{ Combinations}$$

A Larger Search

Setting	Choices
Tree Count	100, 300
Maximum Depth	16, 32, Unlimited
Minimum Leaf Size	1, 2, 5
Features per Split	Square Root, Half, All
Bootstrap Draws	70%, 100%
Split Rule	Gini, Entropy

Compare Every Combination

Fit on Training



Validation Accuracy



Select Best Settings

Same validation set for every candidate

$$2 \times 3 \times 3 \times 3 \times 2 \times 2 = 216 \text{ Combinations}$$

More parallel fits need more memory

Larger Search Results

Validation accuracy · Same 23,878 bookings

Earlier Search · 16 Candidates



18,621 / 23,878 correct

0% 20% 40% 60% 80% 100%

Larger Search Results

Validation accuracy · Same 23,878 bookings

Earlier Search · 16 Candidates



18,621 / 23,878 correct

Larger Search · 216 Candidates



19,049 / 23,878 correct

0% 20% 40% 60% 80% 100%

Larger Search Results

Validation accuracy · Same 23,878 bookings

Earlier Search · 16 Candidates



18,621 / 23,878 correct

Larger Search · 216 Candidates



19,049 / 23,878 correct

0% 20% 40% 60% 80% 100%

Selected Settings

Trees	300
Depth	32
Minimum Leaf	2
Features	Half
Sample Size	100%
Split Rule	Entropy

428 additional correct predictions

Larger Search Results

Validation accuracy · Same 23,878 bookings

Earlier Search · 16 Candidates



18,621 / 23,878 correct

Larger Search · 216 Candidates



19,049 / 23,878 correct

0% 20% 40% 60% 80% 100%

Selected Settings

Trees	300
Depth	32
Minimum Leaf	2
Features	Half
Sample Size	100%
Split Rule	Entropy

428 additional correct predictions

Best among the tested settings · Test data excluded from this search

Demo: random_forest_tuning.ipynb

Accuracy and Errors

Earlier Forest

78%

Accuracy

Tuned Forest

80%

Accuracy

		Predicted			Earlier Forest 18,621 correct predictions 23,878 validation bookings
		Stay	Cancel	No-Show	
Actual	Stay	12,833	2,145	55	
	Cancel	2,844	5,739	20	
	No-Show	162	31	49	

Select a forest, then explore rows, columns or cells

Accuracy and Errors

Earlier Forest

78%

Accuracy

Tuned Forest

80%

Accuracy

		Predicted			
		Stay	Cancel	No-Show	
Actual	Stay	13,673	1,348	12	Tuned Forest 19,049 correct predictions 23,878 validation bookings
	Cancel	3,248	5,351	4	
	No-Show	184	33	25	

Select a forest, then explore rows, columns or cells

Accuracy and Errors

Earlier Forest

78%

Accuracy

Tuned Forest

80%

Accuracy

		Predicted			Actual No-Show
		Stay	Cancel	No-Show	
Actual	Stay	13,673	1,348	12	25 / 242 correctly found 217 missed
	Cancel	3,248	5,351	4	
	No-Show	184	33	25	

Select a forest, then explore rows, columns or cells

Accuracy and Errors

Earlier Forest

78%

Accuracy

Tuned Forest

80%

Accuracy

		Predicted			Predicted No-Show
		Stay	Cancel	No-Show	
Actual	Stay	13,673	1,348	12	25 / 41 flags correct 16 false alarms
	Cancel	3,248	5,351	4	
	No-Show	184	33	25	

Select a forest, then explore rows, columns or cells

Accuracy and Errors

Earlier Forest

78%

Accuracy

Tuned Forest

80%

Accuracy

		Predicted			Predicted No-Show
		Stay	Cancel	No-Show	
Actual	Stay	13,673	1,348	12	25 / 41 flags correct 16 false alarms
	Cancel	3,248	5,351	4	
	No-Show	184	33	25	

Which errors matter for the hotel's proposed action?

Beyond Accuracy

Did we find the
actual no-shows?

Were our
no-show flags correct?

Beyond Accuracy

Did we find the
actual no-shows?

Were our
no-show flags correct?

Recall

Beyond Accuracy

Did we find the
actual no-shows?

Recall

Were our
no-show flags correct?

Precision

Beyond Accuracy

Did we find the
actual no-shows?

Recall

Were our
no-show flags correct?

Precision

How can we summarise both?

A high score should require both to be strong

Recall, Precision and F1

217

Missed No-Shows

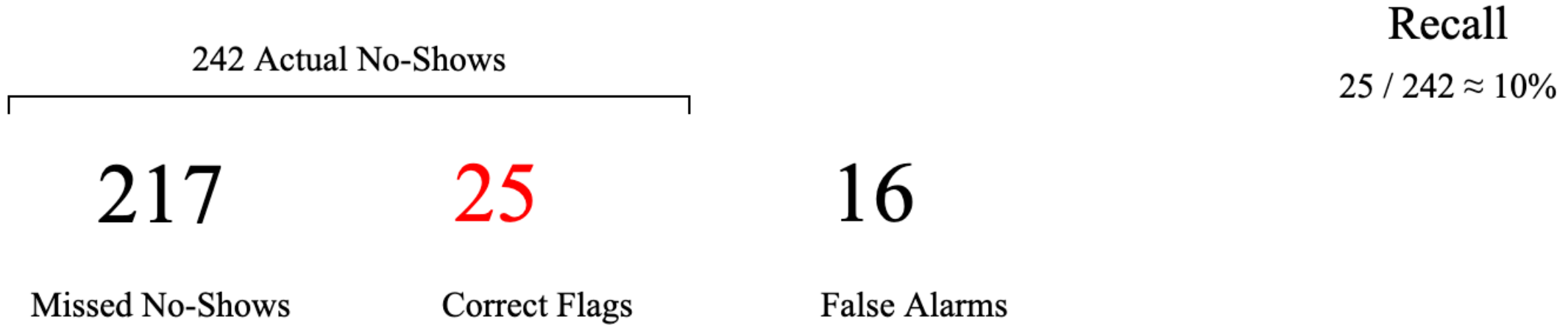
25

Correct Flags

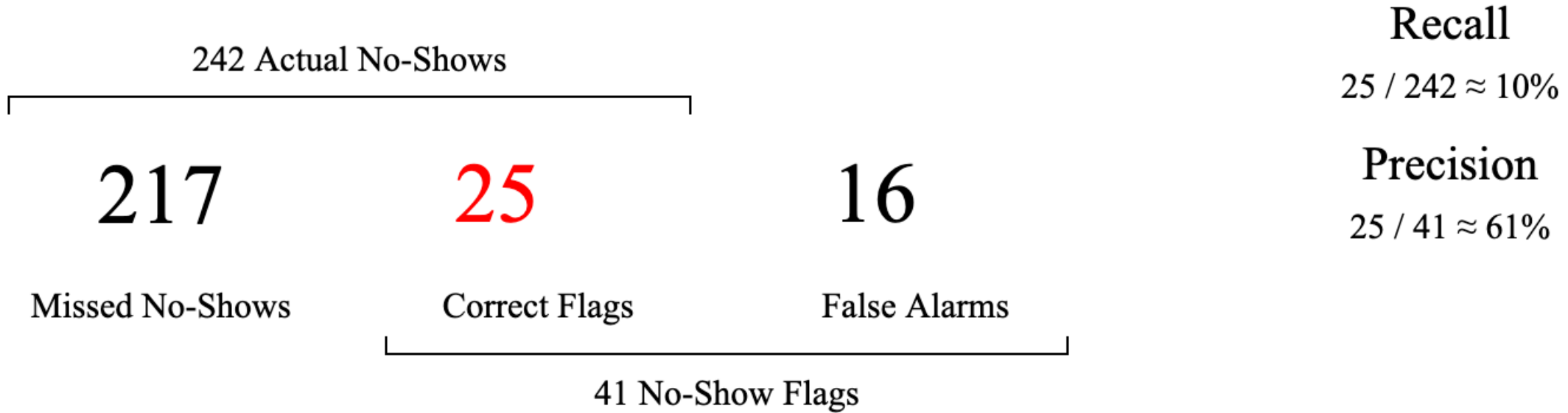
16

False Alarms

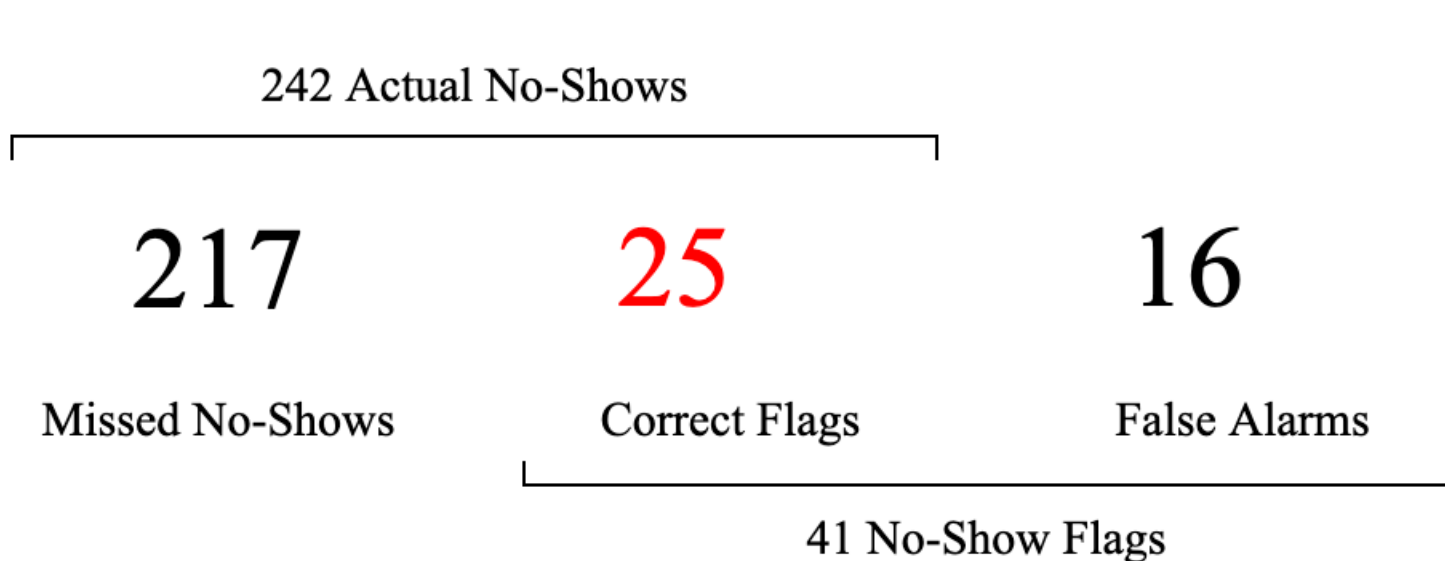
Recall, Precision and F1



Recall, Precision and F1



Recall, Precision and F1



Recall
 $25 / 242 \approx 10\%$

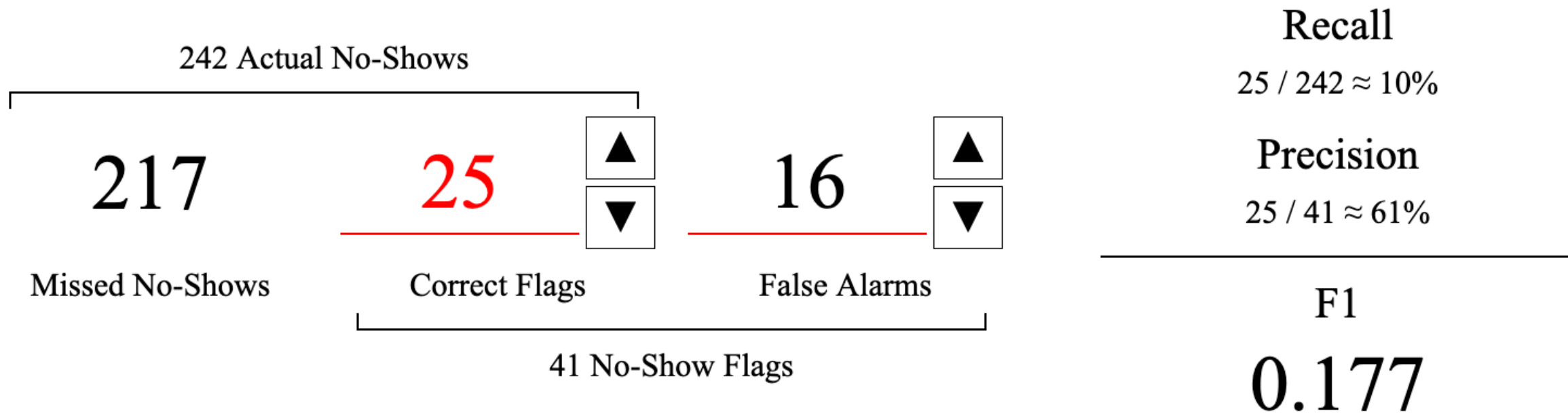
Precision
 $25 / 41 \approx 61\%$

F1
0.177

F1 Uses the Harmonic Mean

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Recall, Precision and F1



F1 Uses the Harmonic Mean

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Edit found cases or false alarms

Hotel Result

Flag Everyone

Perfect Predictions

Macro-F1

One F1 score per class, with equal weight

$$\begin{array}{ccccc} \text{Stay} & & \text{Cancel} & & \text{No-Show} \\ & + & & + & \\ 0.851 & & 0.698 & & 0.177 \end{array}$$

Macro-F1

One F1 score per class, with equal weight

$$\begin{array}{ccccc} \text{Stay} & & \text{Cancel} & & \text{No-Show} \\ & + & & + & \\ 0.851 & & 0.698 & & 0.177 \end{array}$$

$$\frac{0.851 + 0.698 + 0.177}{3} \approx 0.575$$

Macro-F1

One F1 score per class, with equal weight

$$\begin{array}{ccc} \text{Stay} & & \text{Cancel} & & \text{No-Show} \\ & + & & + & \\ 0.851 & & 0.698 & & 0.177 \end{array}$$

$$\frac{0.851 + 0.698 + 0.177}{3} \approx 0.575$$

The rare class matters as much as each common class

Changing the Metric

Rescore the same 216 candidates

Selected By	Accuracy	F1 by Outcome			Macro-F1
		Stay	Cancel	No-Show	
Accuracy	80%	0.851	0.698	0.177	0.575

Changing the Metric

Rescore the same 216 candidates

Selected By	Accuracy	F1 by Outcome			Macro-F1
		Stay	Cancel	No-Show	
Accuracy	80%	0.851	0.698	0.177	0.575
Macro-F1	79%	0.842	0.701	0.302	0.615

Changing the Metric

Rescore the same 216 candidates

Selected By	Accuracy	F1 by Outcome			Macro-F1
		Stay	Cancel	No-Show	
Accuracy	80%	0.851	0.698	0.177	0.575
Macro-F1	79%	0.842	0.701	0.302	0.615

No-shows found: 25 → 50 out of 242

Changing the Metric

Rescore the same 216 candidates

Selected By	Accuracy	F1 by Outcome			Macro-F1
		Stay	Cancel	No-Show	
Accuracy	80%	0.851	0.698	0.177	0.575
Macro-F1	79%	0.842	0.701	0.302	0.615

No-shows found: 25 → 50 out of 242

Demo: random_forest_f1.ipynb

Validation and Model Choice

Our search compared 216 settings

Would different bookings change which settings we choose?

Training Bookings
Fit Each Candidate

Validation Bookings
Score and Choose

The winner may partly reflect quirks of this validation group

Compare candidates across several validation groups before choosing



Training



Validation



Test Set: Final Assessment

Cross-Validation

100 trees, unlimited depth, minimum leaf size 1

Same settings, a fresh forest each time

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Macro-F1
Fit 1	<div>Validate</div>	<div></div>	<div></div>	<div></div>	<div></div>	0.607

Training Validation Test Set: Final Assessment

Cross-Validation

100 trees, unlimited depth, minimum leaf size 1

Same settings, a fresh forest each time

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Macro-F1
Fit 1	<div>Validate</div>	<div></div>	<div></div>	<div></div>	<div></div>	0.607
Fit 2	<div></div>	<div>Validate</div>	<div></div>	<div></div>	<div></div>	0.576

Training Validation Test Set: Final Assessment

Cross-Validation

100 trees, unlimited depth, minimum leaf size 1

Same settings, a fresh forest each time

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Macro-F1
Fit 1	Validate					0.607
Fit 2		Validate				0.576
Fit 3			Validate			0.596



Training



Validation



Test Set: Final Assessment

Cross-Validation

100 trees, unlimited depth, minimum leaf size 1

Same settings, a fresh forest each time

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Macro-F1
Fit 1	Validate					0.607
Fit 2		Validate				0.576
Fit 3			Validate			0.596
Fit 4				Validate		0.611



Training



Validation



Test Set: Final Assessment

Cross-Validation

100 trees, unlimited depth, minimum leaf size 1

Same settings, a fresh forest each time

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Macro-F1
Fit 1	Validate					0.607
Fit 2		Validate				0.576
Fit 3			Validate			0.596
Fit 4				Validate		0.611
Fit 5					Validate	0.596

Which single score should represent these settings?



Training



Validation



Test Set: Final Assessment

Cross-Validation

100 trees, unlimited depth, minimum leaf size 1

Same settings, a fresh forest each time

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Macro-F1
Fit 1	Validate					0.607
Fit 2		Validate				0.576
Fit 3			Validate			0.596
Fit 4				Validate		0.611
Fit 5					Validate	0.596

Mean Macro-F1: 0.597



Training



Validation



Test Set: Final Assessment

Cross-Validation

100 trees, unlimited depth, minimum leaf size 1

Same settings, a fresh forest each time

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Macro-F1
Fit 1	Validate					0.607
Fit 2		Validate				0.576
Fit 3			Validate			0.596
Fit 4				Validate		0.611
Fit 5					Validate	0.596

Mean Macro-F1: 0.597

Repeat for each candidate on the same folds, then compare means



Training



Validation



Test Set: Final Assessment

Search Across Folds

GridSearchCV: 4 settings $\times$ 5 folds = 20 fits

Mean Validation Macro-F1

Depth 16, Leaf 1



0.545

Search Across Folds

GridSearchCV: 4 settings $\times$ 5 folds = 20 fits

Mean Validation Macro-F1



Search Across Folds

GridSearchCV: 4 settings $\times$ 5 folds = 20 fits

Mean Validation Macro-F1



Demo: random_forest_f1.ipynb

Information Leakage

What is available when the hotel makes the prediction?

At Booking Time

Requested stay
Number of guests

Information Leakage

What is available when the hotel makes the prediction?

At Booking Time

Requested stay
Number of guests

After the Outcome

Final status
Cancellation outcome

Information Leakage

What is available when the hotel makes the prediction?

At Booking Time

Requested stay
Number of guests

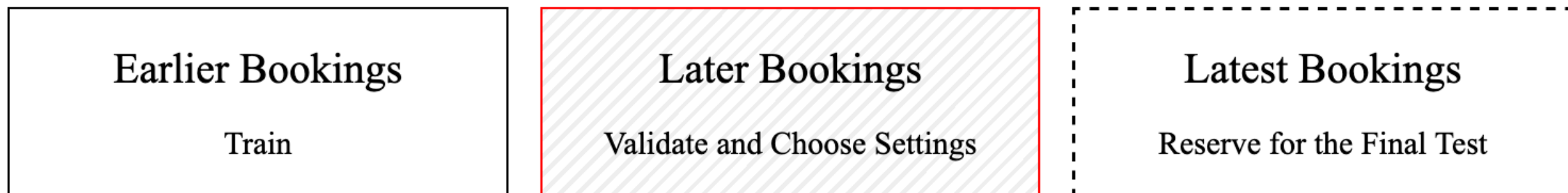
After the Outcome

Final status
Cancellation outcome

Using later information gives away the answer

Choose, Refit and Test

For future bookings, evaluate in time order



Earlier

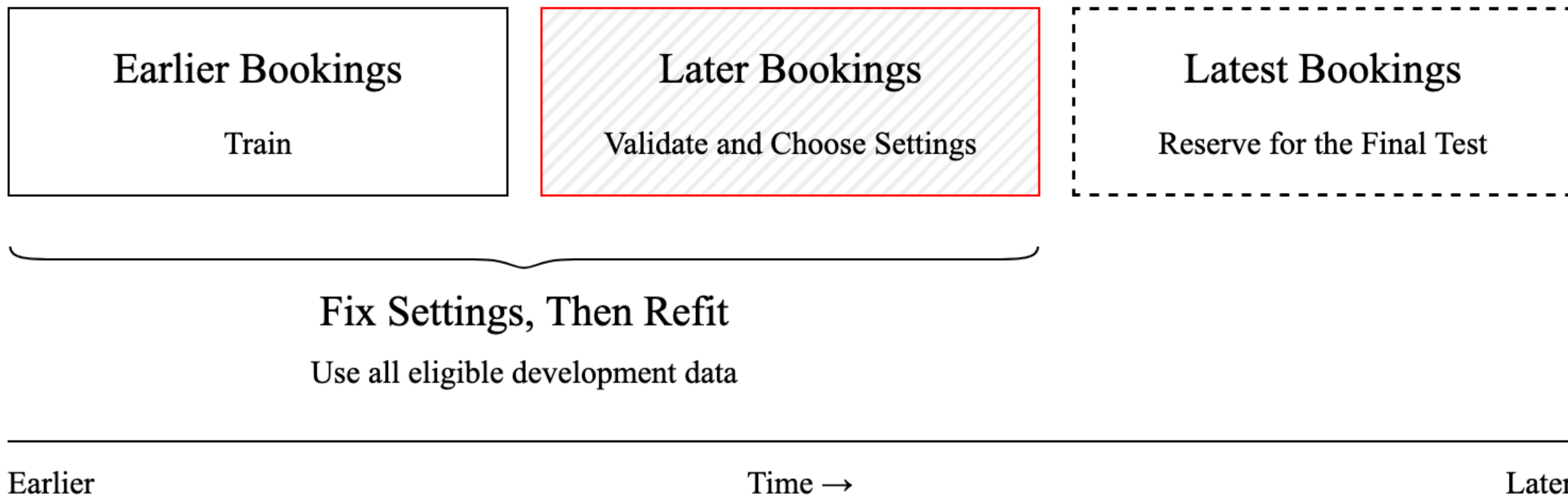
Time →

Later

Train only on outcomes already known at the cutoff

Choose, Refit and Test

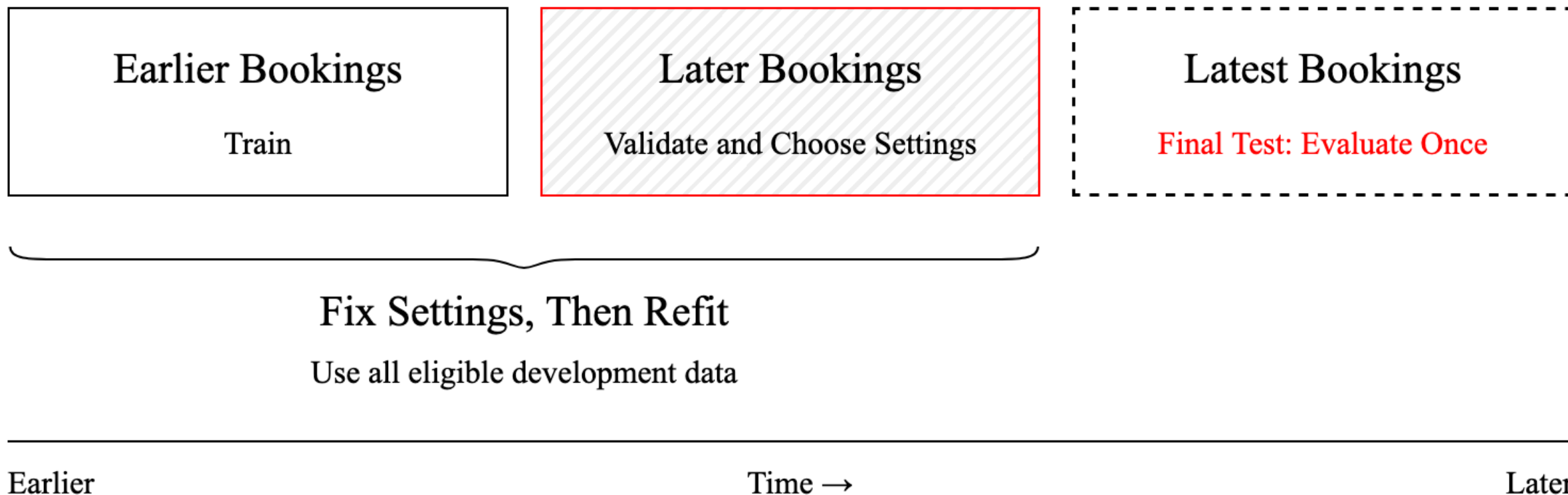
For future bookings, evaluate in time order



Train only on outcomes already known at the cutoff

Choose, Refit and Test

For future bookings, evaluate in time order



Train only on outcomes already known at the cutoff

Exit Ticket

Your Project Evaluation Plan

1. What project output or feature will you evaluate, and what baseline will you compare it against?
2. Which metric or scoring rubric will you use, and why does it fit your project's goal?
3. Which data or test cases will you keep separate during development for the final evaluation?
4. What could make your evaluation misleading, and how will you prevent or check for it?

Complete With Your Project Group Before You Leave

Submit One Response per Group on Ed · Include Your Group ID