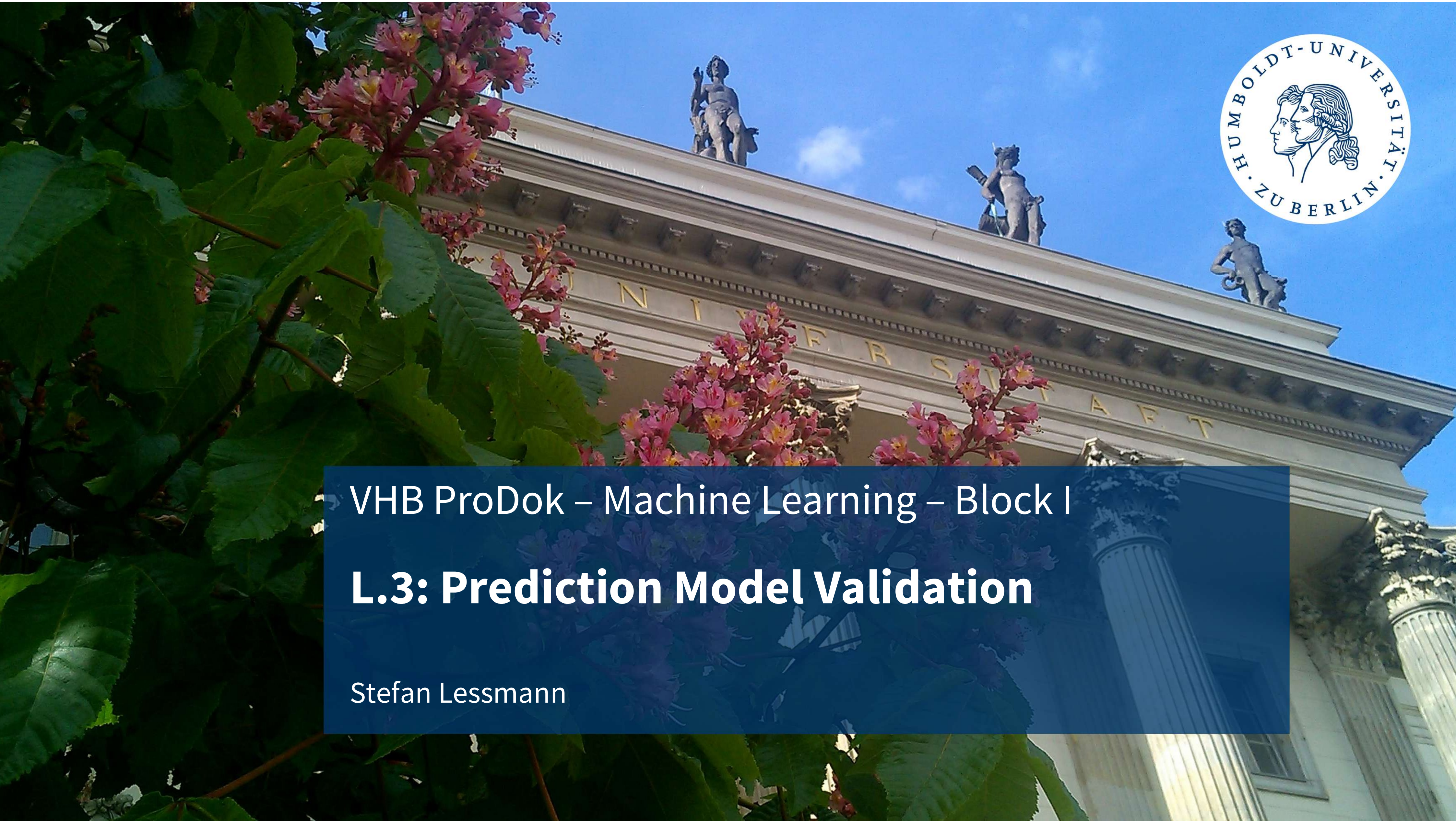


VHB ProDok – Machine Learning – Block I

L.3: Prediction Model Validation

Stefan Lessmann

Agenda

Introduction
Motivation and dimensions of model performance

Performance Indicators for Regression and Classification
Error measures, confusion matrix, and ROC analysis

Simulation of prediction model application
Resubstitution estimate, split sampling, and cross-validation

Summary

Appendix (self-study)
Model selection process and strategies

Introduction

Motivation and dimensions of model performance

Dimensions of Model Performance

Many factors determine the *value* of a machine learning model

Accuracy

How well does the model predict? For example, is it able to distinguish good and bad risks with high accuracy?

Scalability

How much time is needed to build and to apply the model? Does it scale to large data sets?

Robustness/Safety

Can the model cope with noise, missing values, or multicollinearity? How does it react to unusual input data (e.g., outliers)?

Comprehensibility

Can we understand the model? Is it clear how it transforms attribute values into predictions of the response variable?

Justifiability

Is the use of attributes within the model in line with business rules/ understanding?

Compliance/Fairness

Do the model & modeling process comply with relevant regulation? Do model predictions suggest a disparate treatment of social groups?

Assessing Predictive Accuracy – Intuition and Ingredients
Comparing model-based forecasts to actual outcomes

- **The more forecasts agree with true values of the target better the model**
- **Question 1: How to measure agreement between forecasts & actuals?**
  - Say we know that the actual price of a stock is \$125
  - Say a model predicted the price to be \$98. How good or bad is that forecast?
- **Question 2: How to gather the data for this comparison**
  - The point of developing a predictive model is to forecast future target values
  - We don't know actual target values a priori
  - So how to assess a model's predictive accuracy before using it in real life?
- **Two core ingredients of forecast accuracy evaluation**
  - Measures for predictive performance
  - Practices to organize the available data (see later)

Y	$\hat{Y}$
$\dots$	$\dots$
…	…
…	…
…	…
…	…
…	…
…	…
…	…
…	…

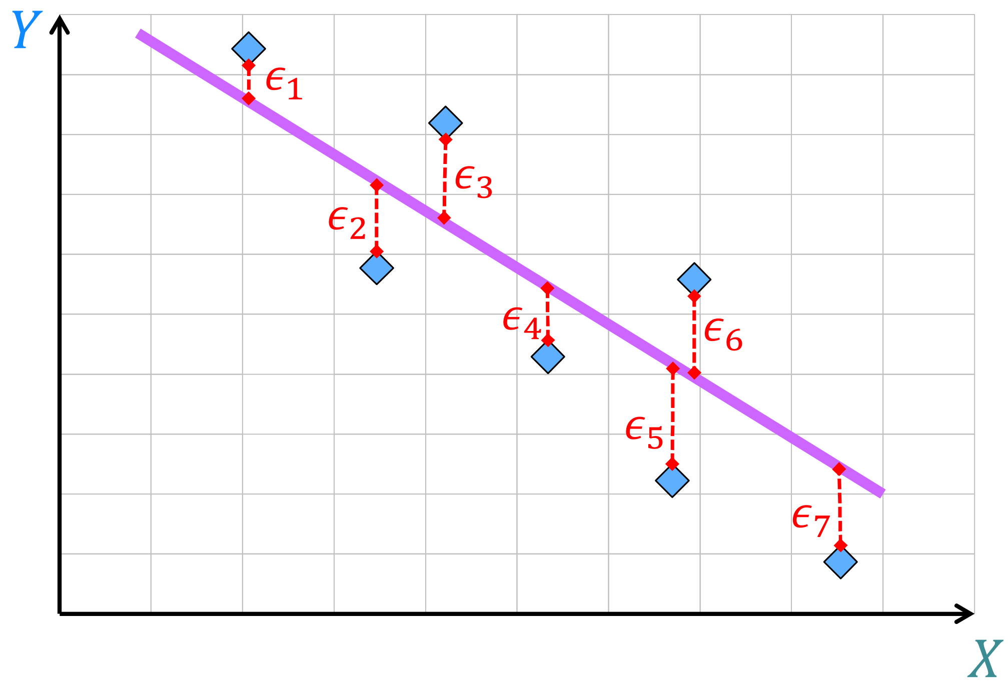
Performance Indicators for Regression and Classification

Error measures, confusion matrix, and ROC analysis

Measuring Forecast Accuracy in Regression

Compare model-based forecasts to true realizations of the target variable

- **Model residuals capture the difference between a true outcome and a forecast**
- **Error measures aggregate residuals into an overall measure of forecast error**
- **Forecast error and accuracy are just two sides of one coin**



$$\begin{bmatrix} \epsilon_1 = y_1 - \hat{y}_1 \\ \epsilon_2 = y_2 - \hat{y}_2 \\ \epsilon_3 = y_3 - \hat{y}_3 \\ \epsilon_4 = y_4 - \hat{y}_4 \\ \epsilon_5 = y_5 - \hat{y}_5 \\ \epsilon_6 = y_6 - \hat{y}_6 \\ \epsilon_7 = y_7 - \hat{y}_7 \end{bmatrix}$$

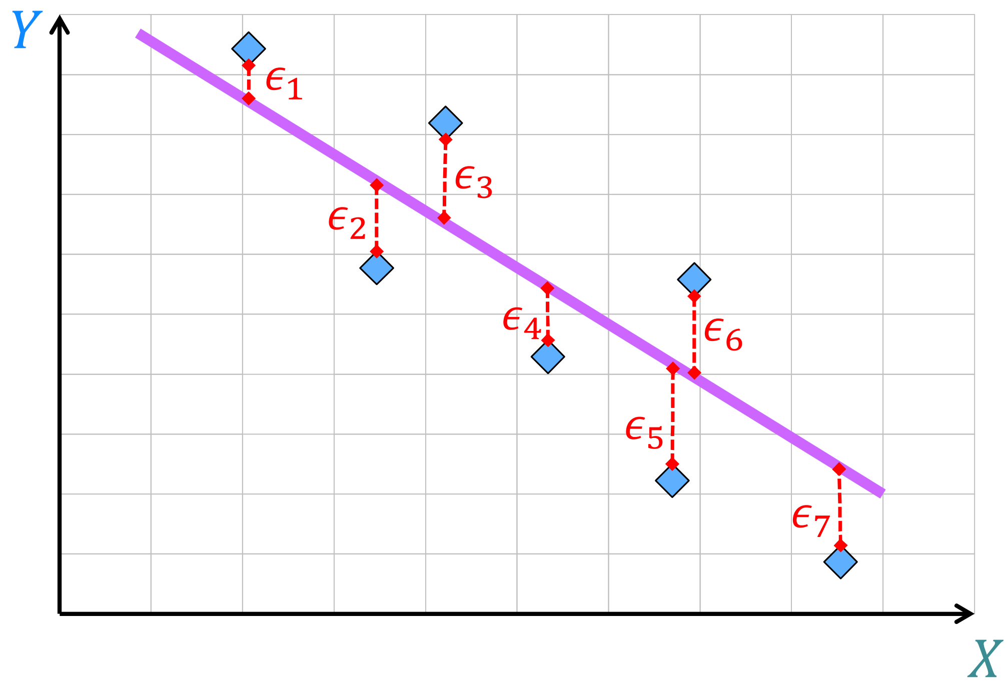
$$TE = \sum_{i=1}^{n=7} \epsilon_i = \sum_{i=1}^{n=7} (y_i - \hat{y}_i)$$

Total error (TE)

- Positive and negative residuals even out
  - Can be used as a measure of model bias (see later)
  - Less useful for error/accuracy measures
- Magnitude depends on the number of data points

Common Error Measures for Regression
Squared error measures

- **Measures of squared errors emphasizes large residuals**
- **RMSE is of the same scale as the target → easy to interpret**
  - For example, target is measured in USD
  - MSE is measured in USD^2 whereas RMSE is measures in USD



Squared error (SE)

$$SE = \sum_{i=1}^{n=7} \epsilon_i^2 = \sum_{i=1}^{n=7} (y_i - \hat{y}_i)^2$$

Mean squared-error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

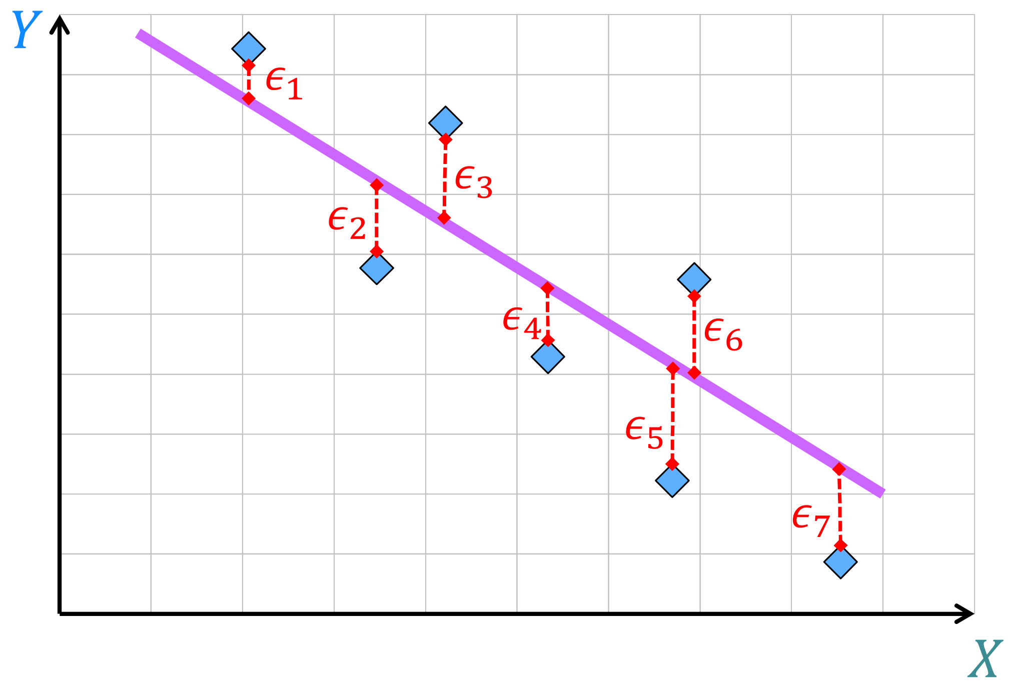
Root-mean squared-error (RMSE)

$$RMSE = \sqrt{MSE}$$

Common Error Measures for Regression

Absolute error measures

- **Measures of absolute errors are perhaps easiest to understand**
- **Mathematically, they are less convenient to work with**
  - No easy derivative c.f. squared error
  - Matters if we use a measure for both, model training and model evaluation



Absolute error (AE)

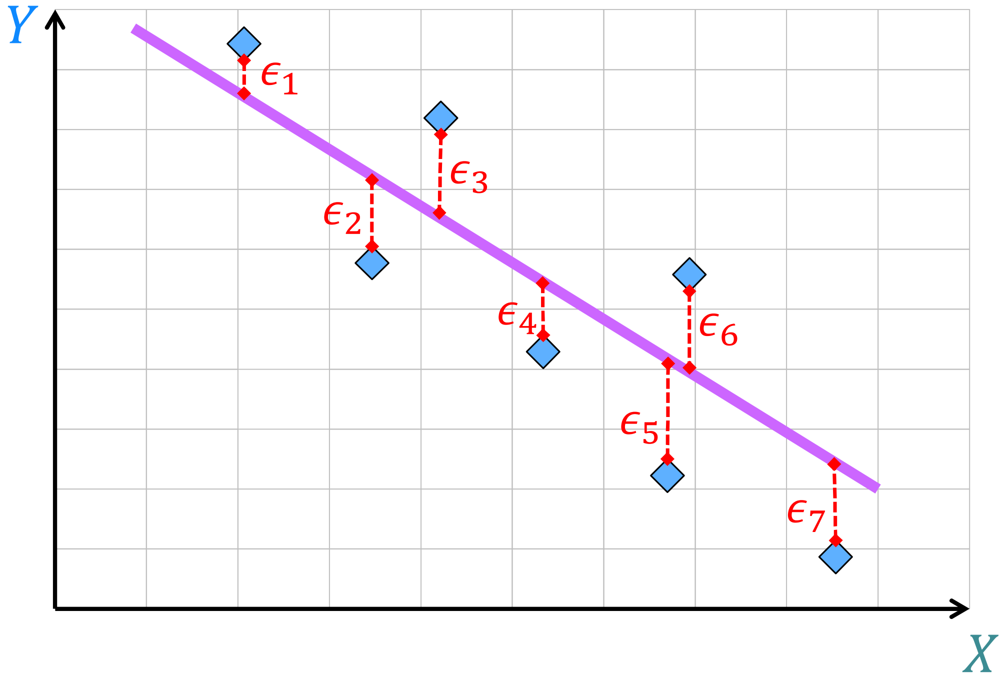
$$AE = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Mean absolute error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Common Error Measures for Regression
Percentage error measures

- **Consider ratio of the error to actual value**
- **Support comparing models for different outcomes**
  - Stock price forecasting model with actual prices in USD
  - Sales forecasting model with outcome in units sold
  - But be careful with comparisons of different models



Mean percentage error

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{y_i}$$

Mean absolute percentage error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Symmetric MAPE

$$sMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

Classification Models Predict Discrete Targets

Sufficient to focus on binary settings → Relevant and Representative

- **Common examples from the decision support literature**
  - Predict if loan applicants would repay → Approve if probability of repayment is high enough
  - Predict if customer is at risk of churning → proactively contact high-risk customers to prevent churn
  - Predict if an insurance claim is fraudulent → route suspicious claims to fraud analyst; pay the others
  - Predict if a machine is about to break → proactive maintenance will reduce cost and downtimes
- **Binary classification (i.e., two outcomes) settings are omnipresent**
  - Any medical test involves binary classification → positive or negative test outcome
  - Many decision problems translate to yes/no questions (i.e., act or do not act)
  - Any problem with multiple outcomes can be broken down into a chain of binary classifications
- **Binary classification models predict outcomes / estimate their probability**

Common Performance Indicators for Binary Classification

Confusion matrix summarizes test outcomes

		Actual Class	
		Positive ($Y = 1$)	**Negative** ($Y = 0$)
Predicted Class	**Positive** ($\hat{Y} = 1$)	True Positive (TP)	False Positive (FP)
	Negative ($\hat{Y} = 0$)	False Negative (FN)	True Negative (TN)

- **Classification accuracy** $\dfrac{TP + TN}{TP + TN + FP + FN}$
- **Classification error** $\dfrac{FP + FN}{TP + TN + FP + FN}$
- **F-Score (balanced)** $2 \times \dfrac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \dfrac{2TP}{2TP + FP + FN}$
- **Specificity** $\dfrac{TN}{TN + FP}$
- **Sensitivity / Recall** $\dfrac{TP}{TP + FN}$
- **Precision** $\dfrac{TP}{TP + FP}$

Common Performance Indicators for Classification

Confusion matrix is a function of the *classification cut-off*

i	Y	$\hat{p}(Y = 1 \mid \boldsymbol{X})$
1	1	0.9
2	1	0.7
3	1	0.6
4	0	0.6
5	0	0.2

$\tau = 0.5$

	Positive ($Y = 1$)	**Negative** ($Y = 0$)
Positive $(\hat{Y} = 1)$	3	1
Negative $(\hat{Y} = 0)$	0	1

To obtain a **discrete class prediction**, compare $\hat{p}(Y = 1 \mid \boldsymbol{X})$ to **cut-off** τ:

predict $\hat{Y} = 1$ if $\hat{\boldsymbol{p}}(Y = 1 \mid \boldsymbol{X}) > \tau$,

and $\hat{Y} = 0$ otherwise.

Common Performance Indicators for Classification

Receiver Operating Characteristic (ROC) Curve

- **Generalization of the confusion matrix**
  - One confusion matrix corresponds to one cut-off
  - ROC curve depicts classifier performance across **all cut-offs**
- **Two-dimensional graph of sensitivity (TP rate) vs. 1-specificity (FP rate)**
  - Passes through the points (0,0) where all cases are classified as Negative
  - And the point (1,1) where all cases are classified as Positive
  - Guessing classes at random produces a straight line through (0,0) and (1,1)
    - Naïve benchmark
    - Every classifier's ROC curve should be above the diagonal
  - Optimal point (0,1), classifier makes no errors
  - The more the ROC curve approaches the optimal point, the better the classifier

Construction of the ROC Curve

Visualization of classifier performance across all cut-offs

i	Y	$\hat{p}(Y = 1\mid\boldsymbol{X})$
1	1	0.9
2	1	0.7
3	1	0.6
4	0	0.6
5	0	0.2

Compare $\hat{p}(Y = 1|\boldsymbol{X})$ to **cut-off** τ:
predict $\hat{Y} = 1$ if $\hat{p}(Y = 1|\boldsymbol{X}) > \tau$,
and $\hat{Y} = 0$ otherwise.

$\xrightarrow{\tau = 0.9}$

	Positive ($Y = 1$)	**Negative** ($Y = 0$)
Positive ($\hat{Y} = 1$)	0	0
Negative ($\hat{Y} = 0$)	3	2

$\xrightarrow{\tau = 0.7}$

	Positive ($Y = 1$)	**Negative** ($Y = 0$)
Positive ($\hat{Y} = 1$)	1	0
Negative ($\hat{Y} = 0$)	2	2

$\xrightarrow{\tau = 0.6}$

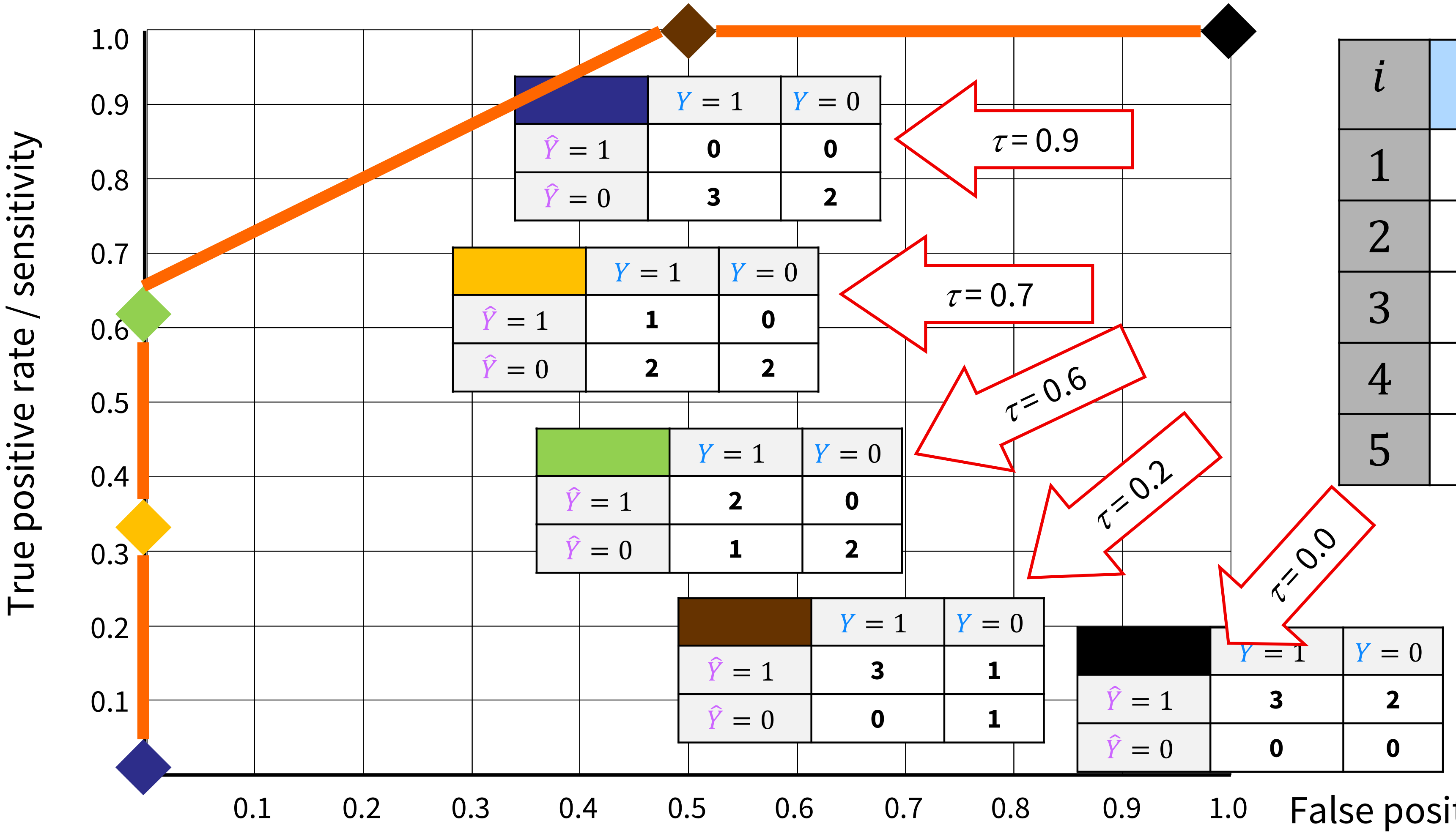
	Positive ($Y = 1$)	**Negative** ($Y = 0$)
Positive ($\hat{Y} = 1$)	2	0
Negative ($\hat{Y} = 0$)	1	2

$\xrightarrow{\tau = 0.2}$ …

$\xrightarrow{\tau = 0.0}$ …

Construction of the ROC Curve

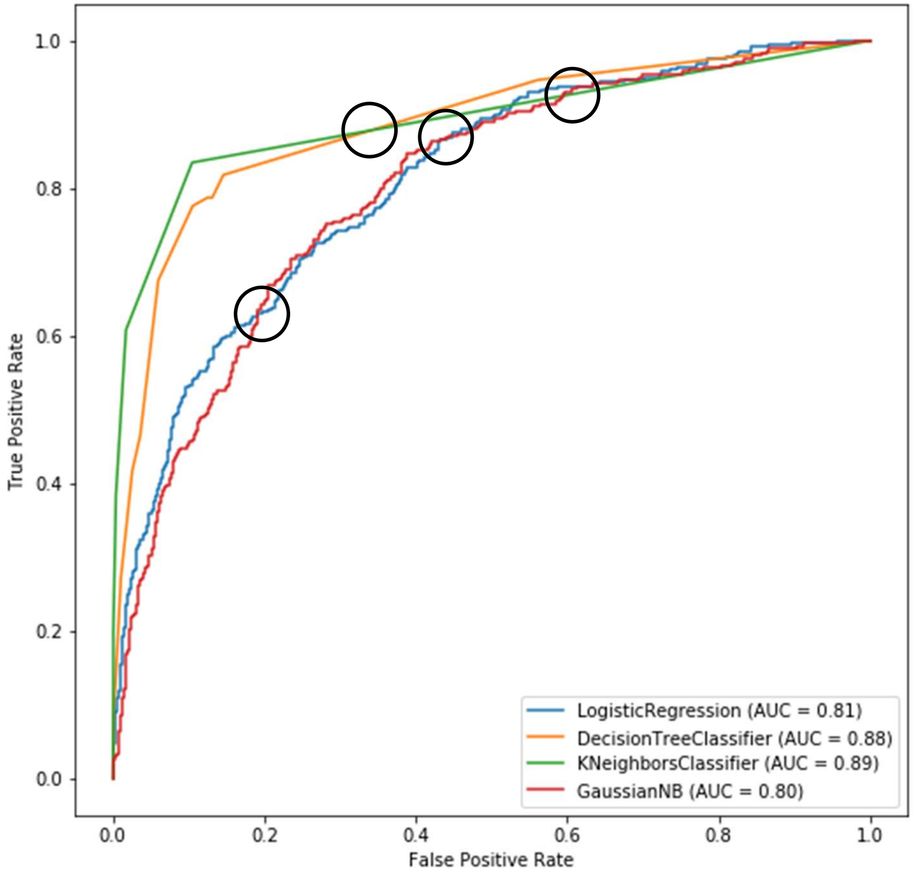
Visualization of classifier performance across all cut-offs



i	Y	$\hat{p}(Y = 1 \mid \boldsymbol{X})$
1	1	0.9
2	1	0.7
3	1	0.6
4	0	0.6
5	0	0.2

The Area Under the ROC Curve
Summarizes the ROC curve in a single number

- **Useful to compare intersecting ROC curves**
- **The higher the better**
  - ☐ Classifier is on average closer to the optimum
  - ☐ Good classifier: AUC well above 0.5
- **Equivalent to Wilcoxon or Mann-Whitney or U- statistic**
  - ☐ The AUC estimates the probability that a randomly chosen positive instance is correctly ranked higher than a randomly chosen negative (Hanley and McNeil, 1982)
  - ☐ Assesses classifier's ability to discriminate between positives and negatives?
  - ☐ AUC is a **ranking indicator**
  - ☐ Ranking based on classifier's **score distribution**
- **See Fawcett (2006) for a good introduction**



Further Indicators of Predictive Accuracy
A vast set of other generic and application-specific measures exist

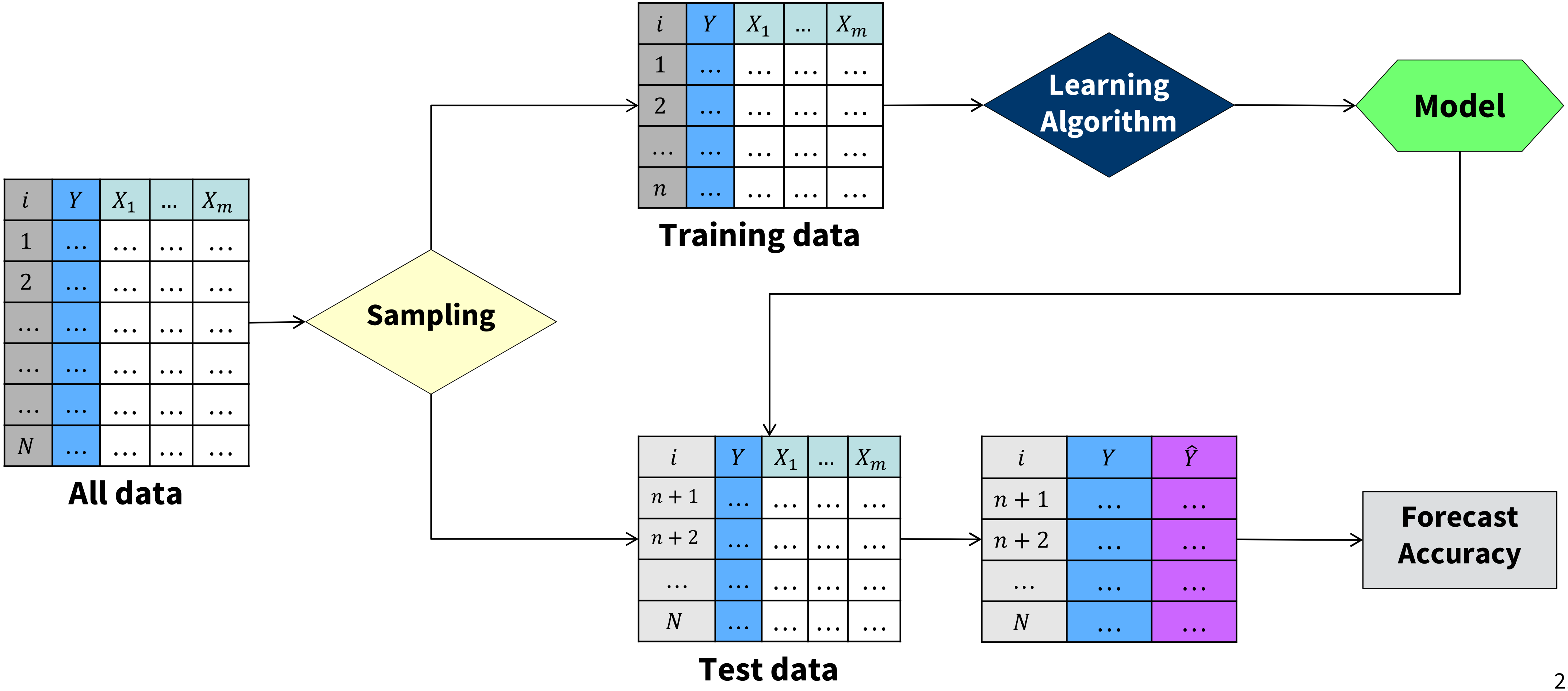
- **Predictive accuracy of classification models**
  - Precision & recall, precision-recall curve, area under the PR-curve (e.g., Saito & Rehmsmeier 2015)
  - Brier score, log-loss, cross-entropy (see, e.g., neural network part)
  - H-measure (Hand & Anagnostopoulos 2013, 2014; Hand 2009)
  - Cost- and Brier curves (Hernández-Orallo et al. 2011, Drummond & Holte 2006)
- **Predictive accuracy of regression models**
  - Theil's U, MSE decomposition, skill scores (e.g., Nikolopoulos et al. 2007, Wheatcroft 2019)
  - (Asymmetric) error costs (e.g., Dress et al. 2018)
- **Examples of application specific measures**
  - Lift-/Gain analysis, uplift-/qini curves (e.g., Surry & Radcliffe 2011, Devriendt et al. 2021)
  - Expected maximum profit criterion for churn/credit scoring (Verbraken et. al. 2012, 2014)

Simulation of prediction model application

Resubstitution estimate, split sampling, and cross-validation

Question 2 (see above): How to Know the True Target Values?

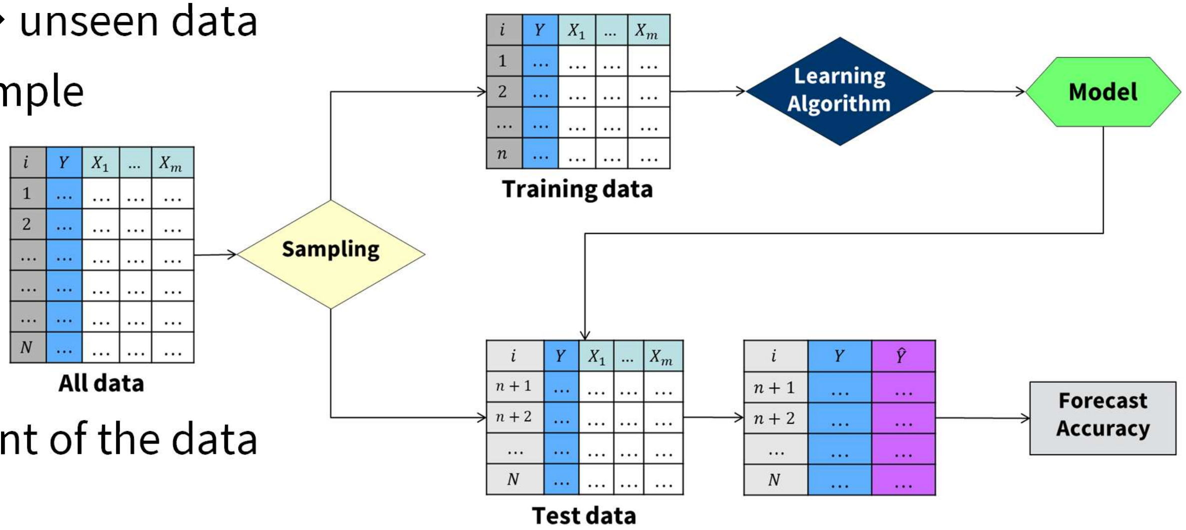
Holdout evaluation reserves some of the historical data for model testing



Holdout Evaluation Under the Microscope

Splitting data in only two samples can be inefficient and unstable

- **Splitting data into train & test sets simulates real-world application of model**
  - Test data not used during model training → unseen data
  - But training & test data stem from same sample
  - Assumes a stable environment (i.e., data generation process)
  - Ideally use out-of-time validation



- **Data splitting is wasteful**
  - Train / test set often comprise 70 / 30 percent of the data
  - Much data lost for training and testing
- **High variance / risk of drawing a 'lucky' test sample**
- **Alternatives aim to increase robustness & efficiency of performance estimate**
  - Repeat the random sampling of the data into train and test set
  - Cross-validation (see next), jackknifing, bootstrapping, …

K-Fold Cross Validation (CV)

Repeat model training & hold-out evaluation K times on different subsets

- **Say we have a data set with 10 observations and set K=5**
- **We split the data into K=5 partitions of equal size (i.e., two observations)**

i	Product	List price [$]	Age [month]	Industry	…	Resale price [$]	
1	Dell XPS 15’	2,500	36	Mining	…	347	Fold 1
2	Dell XPS 15’	2,500	24	Health	…	416	
3	Dell XPS 17’	3,000	36	Manufacturing	…	538	Fold 2
4	HP Envy 17’	1,300	24	Office	…	121	
5	HP EliteBook 850	1,900	36	Manufacturing	…	172	Fold 3
6	Lenovo Yoga 11’	799	12	Office	…	88	
7	Lenovo Yoga 13’	1,100	12	Office	…	266	Fold 4
8	Dell Inspiron 15'	1,499	12	Manufacturing	…	189	
9	HP Envy 15’	2,300	24	Health	…	235	Fold 5
10	MacBook	2,750	12	Office	…	1,125	

K-Fold Cross Validation (CV)

Repeat model training & hold-out evaluation K times on different subsets

- **Say we have a data set with 10 observations and set K=5**
- **We split the data into K=5 partitions of equal size (i.e., two observations)**
- **We use one partition for hold-out validation of a model, which we train on the union of the other partitions**

i	Product	List price [$]	Age [month]	Industry	…	Resale price [$]
1	Dell XPS 15’	2,500	36	Mining	…	347
2	Dell XPS 15’	2,500	24	Health	…	416
3	Dell XPS 17’	3,000	36	Manufacturing	…	538
4	HP Envy 17’	1,300	24	Office	…	121
5	HP EliteBook 850	1,900	36	Manufacturing	…	172
6	Lenovo Yoga 11’	799	12	Office	…	88
7	Lenovo Yoga 13’	1,100	12	Office	…	266
8	Dell Inspiron 15'	1,499	12	Manufacturing	…	189
9	HP Envy 15’	2,300	24	Health	…	235
10	MacBook	2,750	12	Office	…	1,125

Iteration 1

Training data

Validation data

K-Fold Cross Validation (CV)

Repeat model training & hold-out evaluation K times on different subsets

- **Say we have a data set with 10 observations and set K=5**
- **We split the data into K=5 partitions of equal size (i.e., two observations)**
- **We use one partition for hold-out validation of a model, which we train on the union of the other partitions**
- **We repeat this K times each time using a different partition for hold-out validation**

i	Product	List price [$]	Age [month]	Industry	…	Resale price [$]
1	Dell XPS 15’	2,500	36	Mining	…	347
2	Dell XPS 15’	2,500	24	Health	…	416
3	Dell XPS 17’	3,000	36	Manufacturing	…	538
4	HP Envy 17’	1,300	24	Office	…	121
5	HP EliteBook 850	1,900	36	Manufacturing	…	172
6	Lenovo Yoga 11’	799	12	Office	…	88
7	Lenovo Yoga 13’	1,100	12	Office	…	266
8	Dell Inspiron 15'	1,499	12	Manufacturing	…	189
9	HP Envy 15’	2,300	24	Health	…	235
10	MacBook	2,750	12	Office	…	1,125

Iteration 2

Training data

Validation data

K-Fold Cross Validation (CV)

Repeat model training & hold-out evaluation K times on different subsets

- **Say we have a data set with 10 observations and set K=5**
- **We split the data into K=5 partitions of equal size (i.e., two observations)**
- **We use one partition for hold-out validation of a model, which we train on the union of the other partitions**
- **We repeat this K times each time using a different partition for hold-out validation**

i	Product	List price [$]	Age [month]	Industry	…	Resale price [$]
1	Dell XPS 15’	2,500	36	Mining	…	347
2	Dell XPS 15’	2,500	24	Health	…	416
3	Dell XPS 17’	3,000	36	Manufacturing	…	538
4	HP Envy 17’	1,300	24	Office	…	121
5	HP EliteBook 850	1,900	36	Manufacturing	…	172
6	Lenovo Yoga 11’	799	12	Office	…	88
7	Lenovo Yoga 13’	1,100	12	Office	…	266
8	Dell Inspiron 15'	1,499	12	Manufacturing	…	189
9	HP Envy 15’	2,300	24	Health	…	235
10	MacBook	2,750	12	Office	…	1,125

Iteration 3

Training data

Validation data

K-Fold Cross Validation (CV)

Repeat model training & hold-out evaluation K times on different subsets

- **Say we have a data set with 10 observations and set K=5**
- **We split the data into K=5 partitions of equal size (i.e., two observations)**
- **We use one partition for hold-out validation of a model, which we train on the union of the other partitions**
- **We repeat this K times each time using a different partition for hold-out validation**

i	Product	List price [$]	Age [month]	Industry	…	Resale price [$]
1	Dell XPS 15’	2,500	36	Mining	…	347
2	Dell XPS 15’	2,500	24	Health	…	416
3	Dell XPS 17’	3,000	36	Manufacturing	…	538
4	HP Envy 17’	1,300	24	Office	…	121
5	HP EliteBook 850	1,900	36	Manufacturing	…	172
6	Lenovo Yoga 11’	799	12	Office	…	88
7	Lenovo Yoga 13’	1,100	12	Office	…	266
8	Dell Inspiron 15'	1,499	12	Manufacturing	…	189
9	HP Envy 15’	2,300	24	Health	…	235
10	MacBook	2,750	12	Office	…	1,125

Iteration 4

Training data

Validation data

K-Fold Cross Validation (CV)

Repeat model training & hold-out evaluation K times on different subsets

- **Say we have a data set with 10 observations and set K=5**
- **We split the data into K=5 partitions of equal size (i.e., two observations)**
- **We use one partition for hold-out validation of a model, which we train on the union of the other partitions**
- **We repeat this K times each time using a different partition for hold-out validation**

i	Product	List price [$]	Age [month]	Industry	…	Resale price [$]
1	Dell XPS 15’	2,500	36	Mining	…	347
2	Dell XPS 15’	2,500	24	Health	…	416
3	Dell XPS 17’	3,000	36	Manufacturing	…	538
4	HP Envy 17’	1,300	24	Office	…	121
5	HP EliteBook 850	1,900	36	Manufacturing	…	172
6	Lenovo Yoga 11’	799	12	Office	…	88
7	Lenovo Yoga 13’	1,100	12	Office	…	266
8	Dell Inspiron 15'	1,499	12	Manufacturing	…	189
9	HP Envy 15’	2,300	24	Health	…	235
10	MacBook	2,750	12	Office	…	1,125

Iteration 5

Training data

Validation data

K-Fold Cross Validation (CV)

Each (sub-)model gives forecasts for the corresponding validation fold

i	Product	List price [$]	Age [month]	Industry	…	Resale price [$]
1	Dell XPS 15'	2,500	36	Mining	…	347
2	Dell XPS 15'	2,500	24	Health	…	416
3	Dell XPS 17'	3,000	36	Manufacturing	…	538
4	HP Envy 17'	1,300	24	Office	…	121
5	HP EliteBook 850	1,900	36	Manufacturing	…	172
6	Lenovo Yoga 11'	799	12	Office	…	88
7	Lenovo Yoga 13'	1,100	12	Office	…	266
8	Dell Inspiron 15'	1,499	12	Manufacturing	…	189
9	HP Envy 15'	2,300	24	Health	…	235
10	MacBook	2,750	12	Office	…	1,125

i	Product	List price [$]	Age [month]	Industry	…	Resale price [$]
1	Dell XPS 15'	2,500	36	Mining	…	347
2	Dell XPS 15'	2,500	24	Health	…	416
3	Dell XPS 17'	3,000	36	Manufacturing	…	538
4	HP Envy 17'	1,300	24	Office	…	121
5	HP EliteBook 850	1,900	36	Manufacturing	…	172
6	Lenovo Yoga 11'	799	12	Office	…	88
7	Lenovo Yoga 13'	1,100	12	Office	…	266
8	Dell Inspiron 15'	1,499	12	Manufacturing	…	189
9	HP Envy 15'	2,300	24	Health	…	235
10	MacBook	2,750	12	Office	…	1,125

i	Product	List price [$]	Age [month]	Industry	…	Resale price [$]
1	Dell XPS 15'	2,500	36	Mining	…	347
2	Dell XPS 15'	2,500	24	Health	…	416
3	Dell XPS 17'	3,000	36	Manufacturing	…	538
4	HP Envy 17'	1,300	24	Office	…	121
5	HP EliteBook 850	1,900	36	Manufacturing	…	172
6	Lenovo Yoga 11'	799	12	Office	…	88
7	Lenovo Yoga 13'	1,100	12	Office	…	266
8	Dell Inspiron 15'	1,499	12	Manufacturing	…	189
9	HP Envy 15'	2,300	24	Health	…	235
10	MacBook	2,750	12	Office	…	1,125

i	Product	List price [$]	Age [month]	Industry	…	Resale price [$]
1	Dell XPS 15'	2,500	36	Mining	…	347
2	Dell XPS 15'	2,500	24	Health	…	416
3	Dell XPS 17'	3,000	36	Manufacturing	…	538
4	HP Envy 17'	1,300	24	Office	…	121
5	HP EliteBook 850	1,900	36	Manufacturing	…	172
6	Lenovo Yoga 11'	799	12	Office	…	88
7	Lenovo Yoga 13'	1,100	12	Office	…	266
8	Dell Inspiron 15'	1,499	12	Manufacturing	…	189
9	HP Envy 15'	2,300	24	Health	…	235
10	MacBook	2,750	12	Office	…	1,125

i	Product	List price [$]	Age [month]	Industry	…	Resale price [$]
1	Dell XPS 15'	2,500	36	Mining	…	347
2	Dell XPS 15'	2,500	24	Health	…	416
3	Dell XPS 17'	3,000	36	Manufacturing	…	538
4	HP Envy 17'	1,300	24	Office	…	121
5	HP EliteBook 850	1,900	36	Manufacturing	…	172
6	Lenovo Yoga 11'	799	12	Office	…	88
7	Lenovo Yoga 13'	1,100	12	Office	…	266
8	Dell Inspiron 15'	1,499	12	Manufacturing	…	189
9	HP Envy 15'	2,300	24	Health	…	235
10	MacBook	2,750	12	Office	…	1,125

Model 1 **Model 2** **Model 3** **Model 4** **Model 5**

i	Resale price [$]	Forecast
1	347	325
2	416	398

i	Resale price [$]	Forecast
3	538	612
4	121	101

i	Resale price [$]	Forecast
5	172	214
6	88	59

i	Resale price [$]	Forecast
7	266	307
8	189	182

i	Resale price [$]	Forecast
9	235	231
10	1,125	875

K-Fold Cross Validation (CV)

Stacking the validation fold gives out-of-sample predictions for the entire data

i	Resale price [$]	Forecast
1	347	325
2	416	398

i	Resale price [$]	Forecast
3	538	612
4	121	101

i	Resale price [$]	Forecast
5	172	214
6	88	59

i	Resale price [$]	Forecast
7	266	307
8	189	182

i	Resale price [$]	Forecast
9	235	231
10	1,125	875

i	Resale price [$]	Forecast
1	347	325
2	416	398
3	538	612
4	121	101
5	172	214
6	88	59
7	266	307
8	189	182
9	235	231
10	1,125	875

Thanks to cross-validation, we obtain hold-out forecasts for the entire data set. We can assess our model based on these hold-out forecast using any forecast accuracy indicator.
Unlike the basic hold-out method, no data is lost for either training **or** validation. Instead, each observations contributes information to both steps, training **and** validation.

The disadvantage or ‘cost’ of cross-validation is that we have to train K models. Training an advanced model on a large data set can consume a significant amount of time and computer resources. However, whenever this is feasible, cross-validation will give a more robust estimate of forecast accuracy and model performance.

Which Model to Use After K-Fold Cross Validation (CV)?

CV is not about training but finding a best modeling option. Use best option to build final model.

- **Performance assessment aims at answering modeling questions**
  - Which learning algorithm to use? Which features to use? How to set hyperparameters? …
  - Assessment practices like cross-validation answer offer empirical answers to such questions
- **Train a final model with *best* spefication afterwards using all available data**

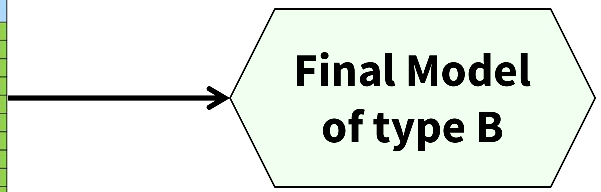
Modeling option A

MSE = 0.12 (0.04)

Modeling option B

MSE = 0.09 (0.025) Lower MSE→better

i	Product	List price [$]	Age [month]	Industry	…	Resale price [$]
1	Dell XPS 15'	2,500	36	Mining	…	347
2	Dell XPS 15'	2,500	24	Health	…	416
3	Dell XPS 17'	3,000	36	Manufacturing	…	538
4	HP Envy 17'	1,300	24	Office	…	121
5	HP EliteBook 850	1,900	36	Manufacturing	…	172
6	Lenovo Yoga 11'	799	12	Office	…	88
7	Lenovo Yoga 13'	1,100	12	Office	…	266
8	Dell Inspiron 15'	1,499	12	Manufacturing	…	189
9	HP Envy 15'	2,300	24	Health	…	235
10	MacBook	2,750	12	Office	…	1,125

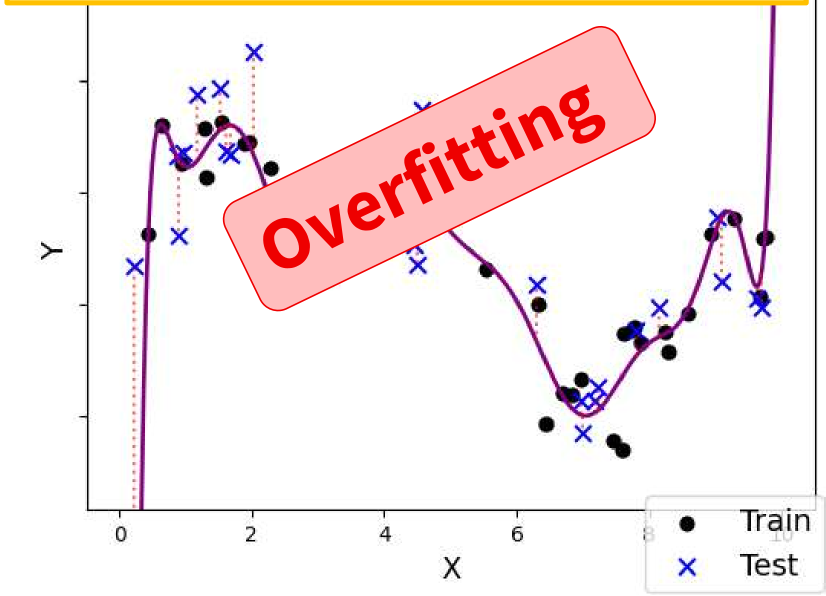


Holdout Evaluation & Overfitting

Detect overfitting by comparing training to test set performance

- **Overfitting implies high accuracy on training and low accuracy on test data**
- **Holdout validation facilitate detecting overfitting**
  - Also measure accuracy on train set and compare
  - Applies to any holdout strategy (sample splitting, CV, etc.)
- **Very different from preventing overfitting or correcting overfit models**
  - Change learning algorithm or its configuration
  - Regularization, early-stopping, ensembling, …
  - We will learn about these approaches later
- **Can we detect overfitting reliably?**
  - Maybe… answer depends on several factors
  - Random partitioning is not ideal

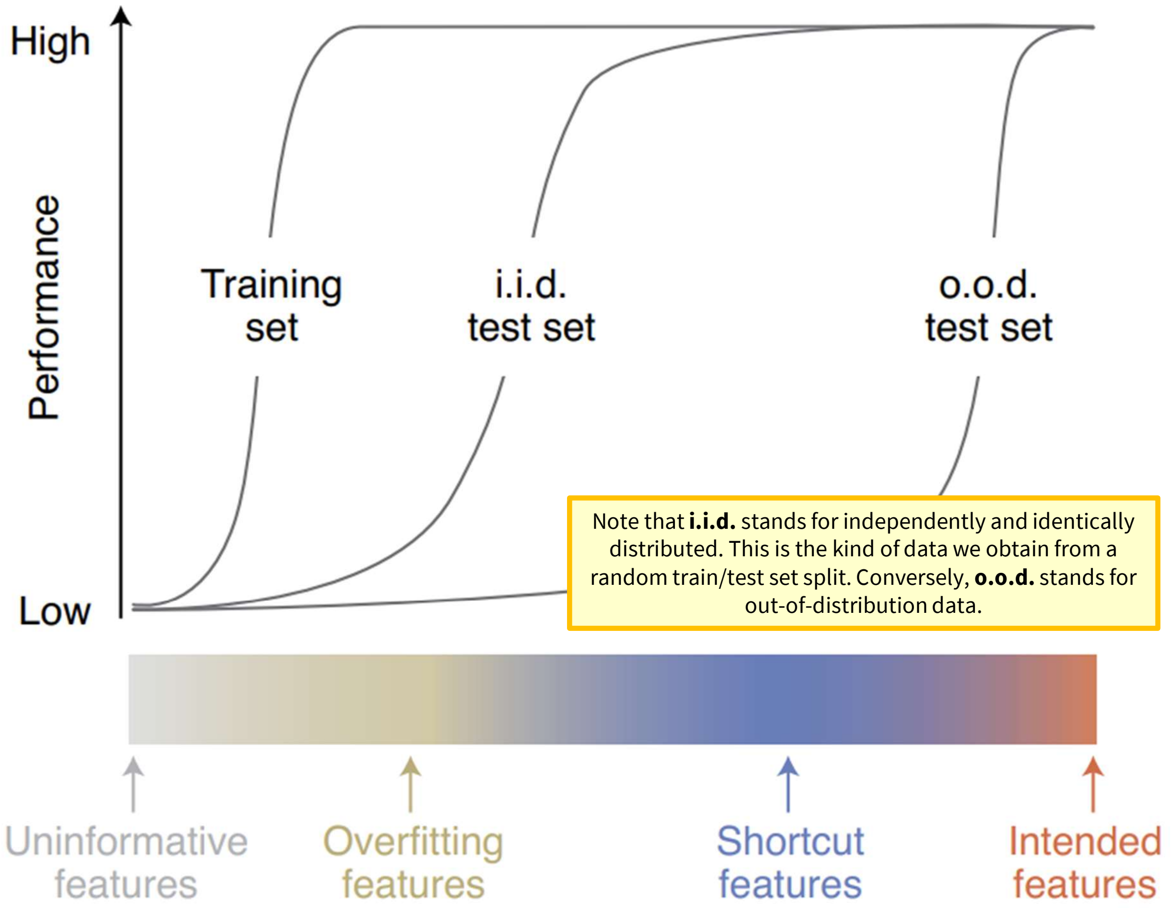
An overfitting model shows high (low) accuracy on the training (test).



Out-of-Sample, Out-of-Time, Out-of-Domain

Proving generalization from a finite, historical sample is hard

- **Dynamic environments**
  - Data generation processes evolve
  - $P_t(Y \mid X) \neq P_{t+\Delta}(Y \mid X)$
    - Covariate shift alters $P(X)$
    - Label shift alters $P(Y)$
    - Conditional drift alters $P(Y \mid X)$
  - Prediction model becomes misspecified
  - Need out-of-time evaluation and mode monitoring
- **Out-of-domain is more difficult**
- **Interesting read: Geirhos, et al. (2020)**



Note that **i.i.d.** stands for independently and identically distributed. This is the kind of data we obtain from a random train/test set split. Conversely, **o.o.d.** stands for out-of-distribution data.

Geirhos, et al. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence, 2(11), 665-673.* https://doi.org/10.1038/s42256-020-00257-z

Illustration of Data Partitioning Options

Use case: Forecast companies' sales revenue

TIME	COMPANY	…	SALES REVENUES
Jan 2023	Alphabet	…	…
Jan 2023	Amazon	…	…
Jan 2023	Apple	…	…
Jan 2023	…	…	…
Jan 2023	Unilever	…	…
Jan 2023	Walmart	…	…
Jan 2023	Xerox	…	…
Feb 2023	Alphabet	…	…
Feb 2023	Amazon	…	…
Feb 2023	Apple	…	…
Feb 2023	…	…	…
Feb 2023	Unilever	…	…
Feb 2023	Walmart	…	…
Feb 2023	Xerox	…	…
…	…	…	…

Training data

Test data

Illustration of Data Partitioning Options

Use case: Forecast companies' sales revenue

- **Random sampling**

TIME	COMPANY	…	SALES REVENUES
Jan 2023	Alphabet	…	…
Jan 2023	Amazon	…	…
Jan 2023	Apple	…	…
Jan 2023	…	…	…
Jan 2023	Unilever	…	…
Jan 2023	Walmart	…	…
Jan 2023	Xerox	…	…
Feb 2023	Alphabet	…	…
Feb 2023	Amazon	…	…
Feb 2023	Apple	…	…
Feb 2023	…	…	…
Feb 2023	Unilever	…	…
Feb 2023	Walmart	…	…
Feb 2023	Xerox	…	…
…	…	…	…

Training data **Test data**

Illustration of Data Partitioning Options

Use case: Forecast companies' sales revenue

- **Random sampling**
- **Temporal sampling**

TIME	COMPANY	…	SALES REVENUES
Jan 2023	Alphabet	…	…
Jan 2023	Amazon	…	…
Jan 2023	Apple	…	…
Jan 2023	…	…	…
Jan 2023	Unilever	…	…
Jan 2023	Walmart	…	…
Jan 2023	Xerox	…	…
Feb 2023	Alphabet	…	…
Feb 2023	Amazon	…	…
Feb 2023	Apple	…	…
Feb 2023	…	…	…
Feb 2023	Unilever	…	…
Feb 2023	Walmart	…	…
Feb 2023	Xerox	…	…
…	…	…	…

Training data **Test data**

Illustration of Data Partitioning Options

Use case: Forecast companies' sales revenue

- **Random sampling**
- **Temporal sampling**
- **Entity-based sampling**

TIME	COMPANY	…	SALES REVENUES
Jan 2023	Alphabet	…	…
Jan 2023	Amazon	…	…
Jan 2023	Apple	…	…
Jan 2023	…	…	…
Jan 2023	Unilever	…	…
Jan 2023	Walmart	…	…
Jan 2023	Xerox	…	…
Feb 2023	Alphabet	…	…
Feb 2023	Amazon	…	…
Feb 2023	Apple	…	…
Feb 2023	…	…	…
Feb 2023	Unilever	…	…
Feb 2023	Walmart	…	…
Feb 2023	Xerox	…	…
…	…	…	…

Training data **Test data**

Illustration of Data Partitioning Options

Use case: Forecast companies' sales revenue

- **Random sampling**
- **Temporal sampling**
- **Entity-based sampling**
- **Temporal & entity-based sampling**

TIME	COMPANY	…	SALES REVENUES
Jan 2023	Alphabet	…	…
Jan 2023	Amazon	…	…
Jan 2023	Apple	…	…
Jan 2023	…	…	…
Jan 2023	Unilever	…	…
Jan 2023	Walmart	…	…
Jan 2023	Xerox	…	…
Feb 2023	Alphabet	…	…
Feb 2023	Amazon	…	…
Feb 2023	Apple	…	…
Feb 2023	…	…	…
Feb 2023	Unilever	…	…
Feb 2023	Walmart	…	…
Feb 2023	Xerox	…	…
…	…	…	…

Training data **Test data**

Closing Remarks

Evaluating generalization ability is hard and there is no silver bullet

- **Many issues can jeopardize holdout evaluation**
  - Changes in the data generation process (drifts, structural breaks, etc.)
  - Changes in business strategy
- **Panel setting is useful to illustrate real-life challenges**
  - Cross-sectional tabular data used in many textbooks and courses (same here ☺)
  - Most real-life data has temporal structure
- **Crucial add-ons to evaluation**
  - Continuous model monitoring
  - Human judgement and oversight

TIME	COMPANY	…	SALES REVENUES
Jan 2023	Alphabet	…	…
Jan 2023	Amazon	…	…
Jan 2023	Apple	…	…
Jan 2023	…	…	…
Jan 2023	Unilever	…	…
Jan 2023	Walmart	…	…
Jan 2023	Xerox	…	…
Feb 2023	Alphabet	…	…
Feb 2023	Amazon	…	…
Feb 2023	Apple	…	…
Feb 2023	…	…	…
Feb 2023	Unilever	…	…
Feb 2023	Walmart	…	…
Feb 2023	Xerox	…	…
…	…	…	…

Summary

Summary

Learning goals

- Experimental designs to assess predictive models
- Accuracy indicators for regression & classification

Findings

- Model performance has facets beyond accuracy
- Confusion matrix, classification cut-off, and ROC
- No in-sample evaluation. Hold-out data is crucial
- CV is more robust than split-sample
- Verifying generalization is challenging
  - Changes in DGP, business strategy, etc. cause data drift
  - Advanced sampling schemes can help, but no silver bullet

What next

- More sophisticated learning algorithms
- Ensembles, random forest, gradient boosting

A

Appendix

Materials are not discussed in course and are provided for self-study

Model Selection

Search strategies and process perspective

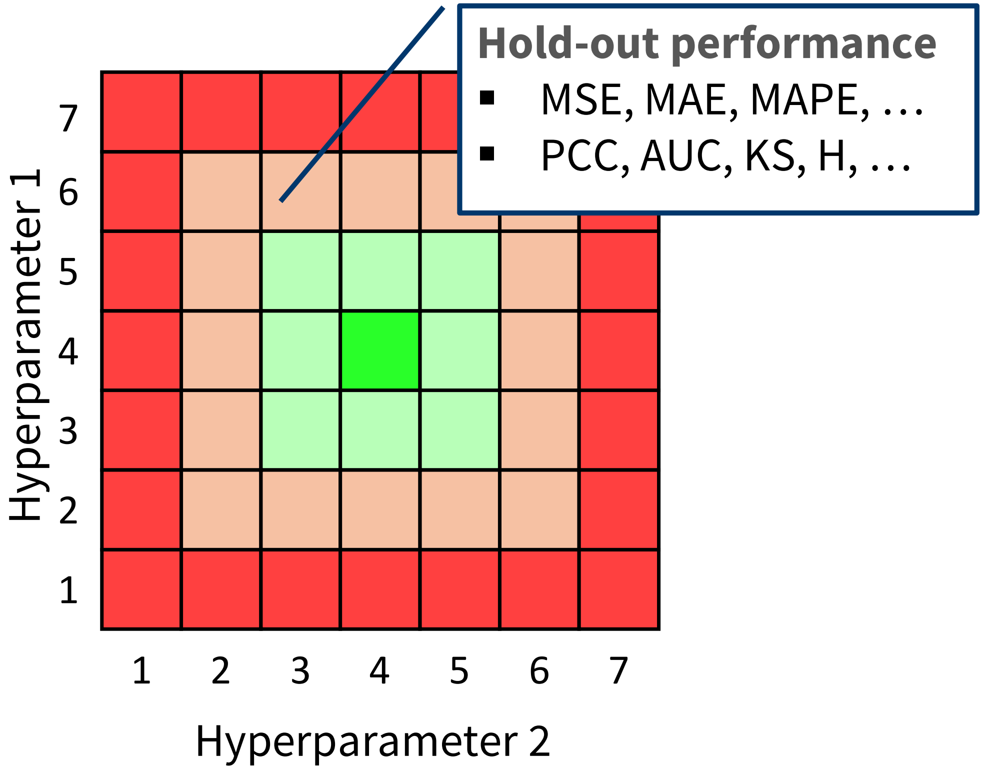
Model Selection
Tuning of algorithmic hyperparameters

- **Advanced classifiers offer hyperparameters (also called meta-parameters)**
  - Facilitate adapting the classifier to a given data set
  - Need to be set by the data scientist
- **Similar to feature selection (in regression modeling)**
  - Manually decide which features to use in a model
  - Try out candidate settings using heuristic search (forward/backward, stagewise regression)
- **How to take corresponding decisions?**
  - Default settings / rules of thumb (not a good idea!)
  - Experience (may work, may fail as well)
  - Empirically, in a model selection process (common practice)

Grid Search

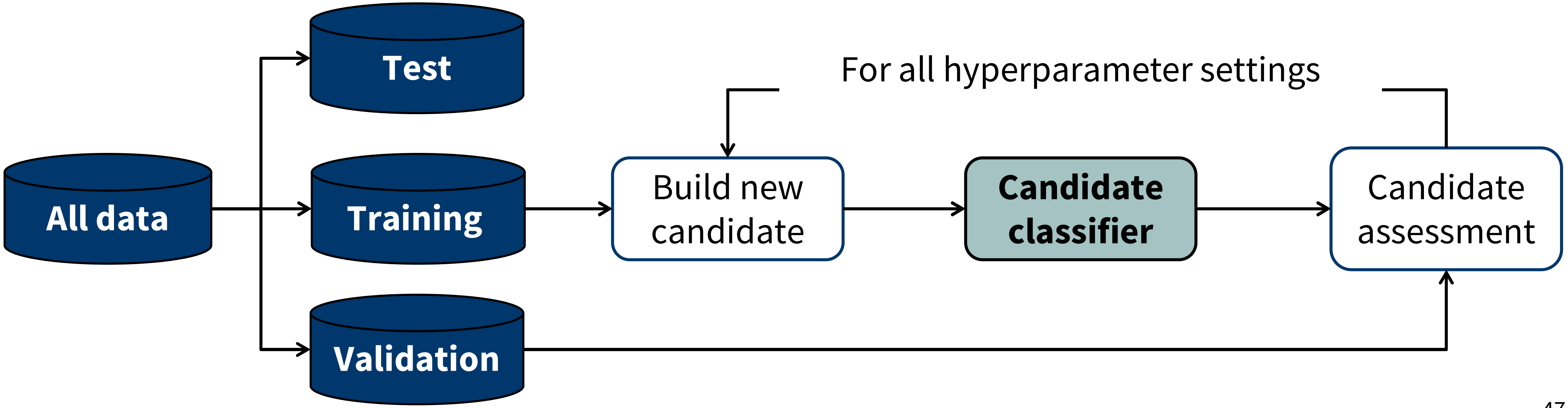
A versatile approach toward model selection

- **Fully enumerative search through all possible combinations of candidate hyperparameter settings**
- **Algorithm**
  - ☐ Define candidate range for each hyperparameter
  - ☐ Enumerate combinations of candidate values
  - ☐ Train model with given configuration
  - ☐ Assess model performance on hold-out data
  - ☐ Repeat with next configuration
- **Magnify grid resolution in promising regions of the search space**



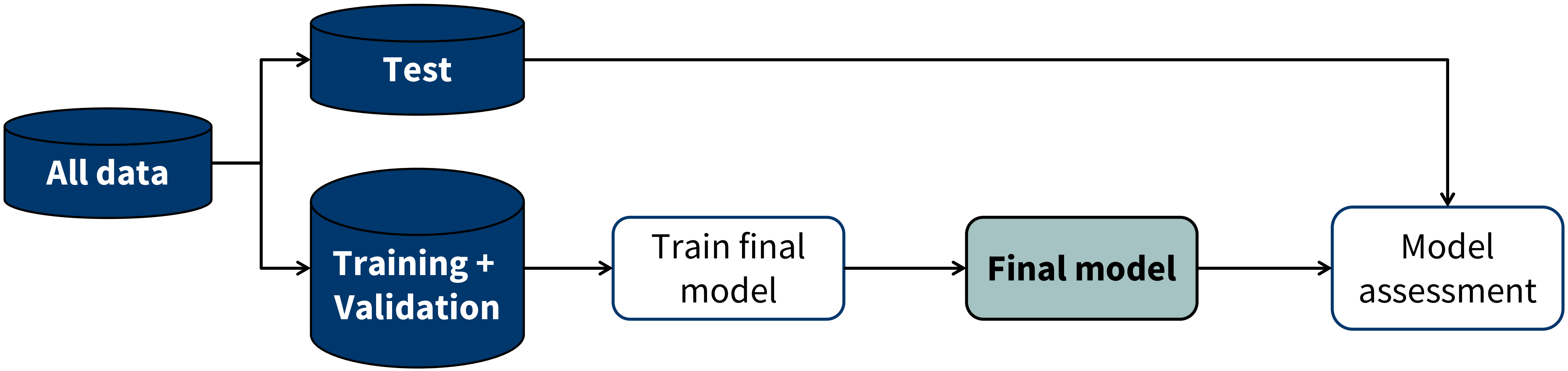
Model Selection Process

- **Additional modeling step to tune hyperparameters**
  - Rules of accuracy assessment apply to model selection
  - Need ‘fresh’ set of hold-out data to assess candidate models with different hyperparameters
- **Generalization of the split-sample approach**
- **Can also involve cross-validation**



Model Selection Process (cont.)

- **Identify best hyperparameter values**
- **Build final classifier with best hyperparameters**
  - No need for auxiliary validation data anymore
  - Can train on the union of training and validation sample



Model Selection Process (cont.)
A note on computational efficiency

- **Model selection is costly**
  - Iterative estimation of different candidate models
  - As many as candidate hyperparameter values in grid-search
  - Potentially more if using cross-validation
  - Careful exploration of parameter space computationally challenging
- **Practical recommendation**
  - Check whether you reduce the among of data during model selection
  - Does the best hyperparameters depend on the size of the training sample?
  - If not (aggressively) down-sample the training set, determine best hyperparameters, and build a model with best hyperparameters on the full training set can give a major speed-up
  - Can start from a learning curve analysis (Perlich et al., 2003) to determine how much down-sampling is possible

Literature

- Caruana, R., Munson, A., & Niculescu-Mizil, A. (2006). Getting the Most Out of Ensemble Selection. In *Proceedings of the 6th International Conference on Data Mining* (pp. 828-833). Hong Kong, China.
- Dress, K., Lessmann, S., & von Mettenheim, H.-J. (2018). Residual value forecasting using asymmetric cost functions. International Journal of Forecasting, 34(4), 551–565.
- Devriendt, F., Belle, J. V., Guns, T., & Verbeke, W. (2021). Learning to rank for uplift modeling. IEEE Transactions on Knowledge and Data Engineering, to appear.
- Drummond, C., & Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance. Machine Learning, 65(1), 95-130.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* 27(8), 861-874.
- Flach, P. A., Hernández-Orallo, J., & Ramirez, C. F. (2011). A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance. In L. Getoor & T. Scheffer (Eds.). Proc. of the 28th Intern. Conf. on Machine Learning, Omnipress: Madison, pp. 657-664.
- Hand, D. J., & Anagnostopoulos, C. (2014). A better Beta for the H measure of classification performance. Pattern Recognition Letters, 40(0), 41-46.
- Hand, D. J., & Anagnostopoulos, C. (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? Pattern Recognition Letters, 34(5), 492-495.
- Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning, 77(1), 103-123.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology,* 143, 29-36.
- Hernández-Orallo, J., Flach, P. A., & Ramirez, C. F. (2011). Brier Curves: A New Cost-Based Visualisation of Classifier Performance. In L. Getoor & T. Scheffer (Eds.). Proceedings of the 28th International Conference on Machine Learning (ICML'11), Omnipress: Madison, pp. 585-592.
- Nikolopoulos, K., Goodwin, P., Patelis, A., & Assimakopoulos, V. (2007). Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches. European Journal of Operational Research, 180(1), 354-368.
- Perlich, C., Provost, F., Simonoff, J. S., & Cohen, W. W. (2003). Tree induction vs. logistic regression: A learning-curve analysis. Journal of Machine Learning Research, 4(2), 211-255.
- Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS One, 10(3), e011843.
- Surry, P. D., & Radcliffe, N. J. (2011). Quality measures for uplift models. Stochastic Solutions Working Paper. [Retrieved from http://www.stochasticsolutions.com/kdd2011late.html]
- Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research, 238(2), 505-513.
- Verbraken, T., Verbeke, W., & Baesens, B. (2012). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. IEEE Transactions on Knowledge and Data Engineering, 25(5), 961-973.
- Wheatcroft, E. (2019). Interpreting the skill score form of forecast performance metrics. International Journal of Forecasting, 35(2), 573-579.

Thank you for your attention!

Stefan Lessmann

Chair of Information Systems
School of Business and Economics
Humboldt-University of Berlin, Germany

Tel. +49.30.2093.5742
Fax. +49.30.2093.5741

stefan.lessmann@hu-berlin.de
http://bit.ly/hu-wi

www.hu-berlin.de