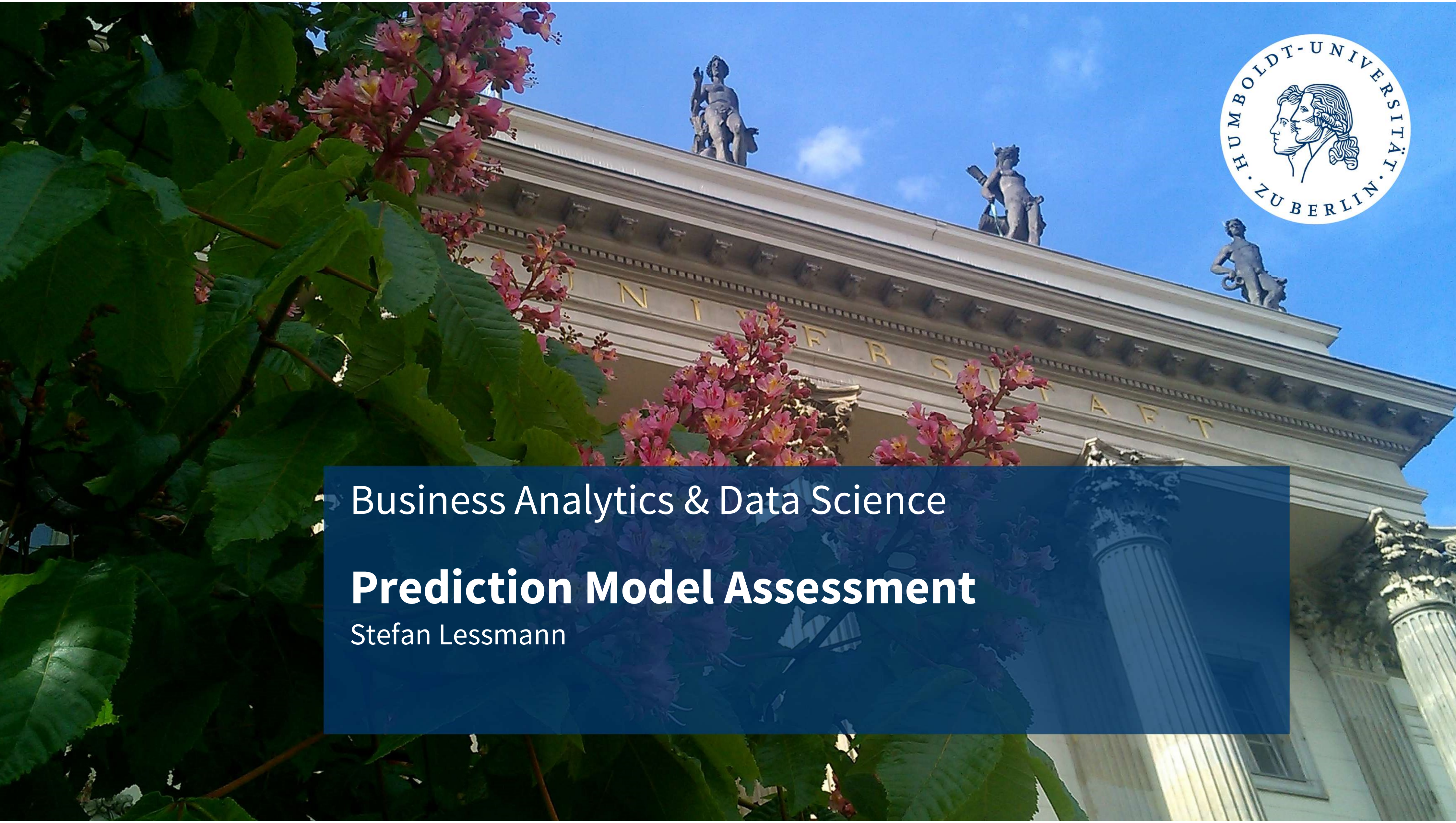


Business Analytics & Data Science

Prediction Model Assessment

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

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

Can the model cope with noise and missing values? How about irrelevant and correlated attributes?

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?

Calibration

For probability forecasts! Out of 100 events predicted to have 90% chance, about 90 should have occurred. True?

Assessing Forecast 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 resale price of a returned notebook is \$125
  - Say a model, based on notebook and client characteristics, forecasts the resale price to be \$98. How good or bad is that forecast?
- **Question 2: How to know the true values of the target variable?**
  - The point of developing a predictive model is to forecast future values of the target
  - We never know actual target values before we deploy the model
  - So how compare forecast and actual resale prices?
- **Two core ingredients of forecast accuracy evaluation**
  - Measures for predictive performance
  - Practices to organize the available data (see later)

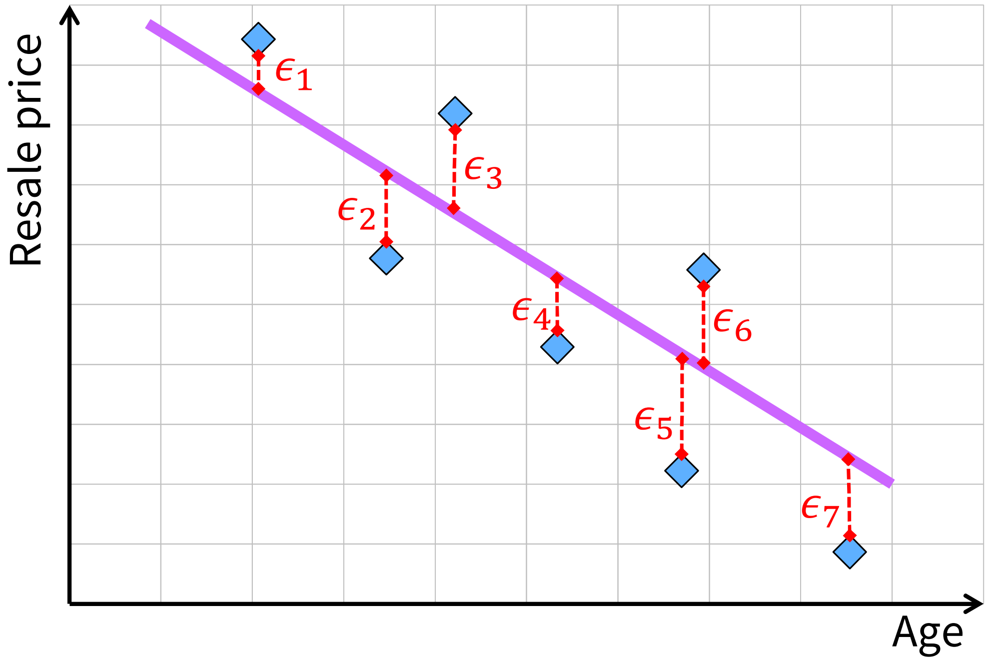
Y	$\hat{Y}$
$\ldots$	$\ldots$
...	...
...	...
...	...
...	...
...	...
...	...
...	...
...	...

Performance Indicators for Regression and Classification

Error measures, confusion matrix, and ROC analysis

Common Error Measures for Regression
Squared error measures

- **Measures of squared errors emphasizes large residuals**
- **Note that RMSE is of the same scale as the target variable**
  - For example, resale price 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)

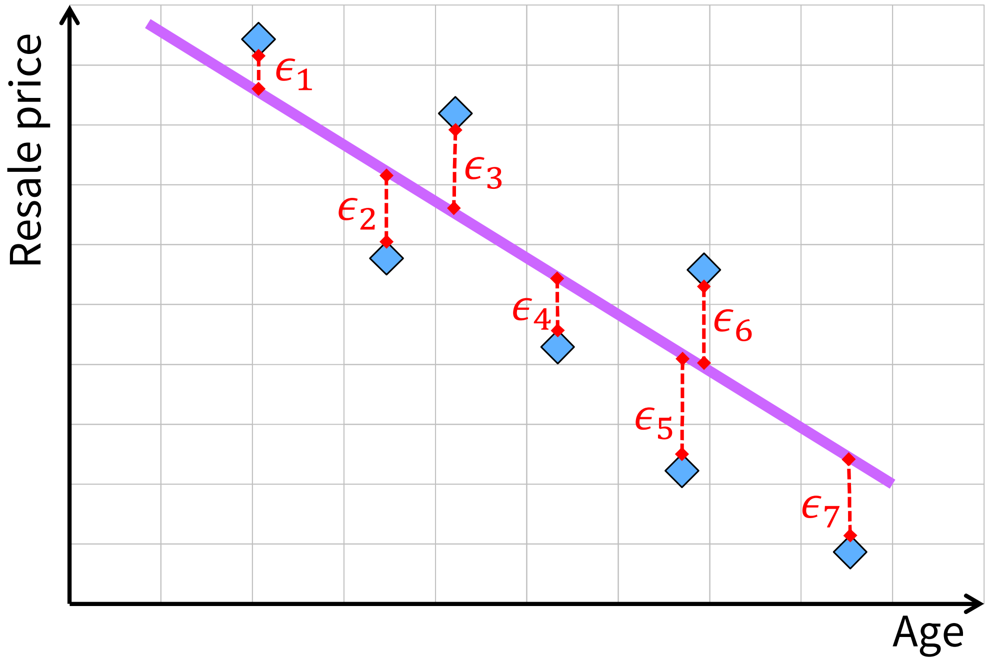
$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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)

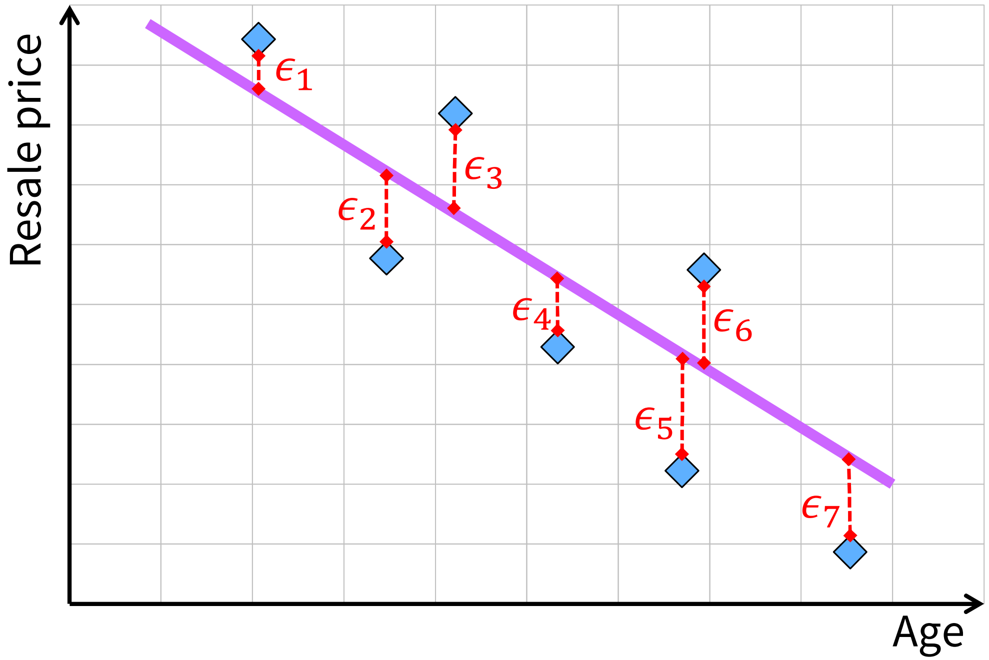
$$AE = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

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 ration of the error to actual value**
- **Supports comparing models for different outcomes**
  - Resale price forecasting model with actual prices in USD
  - Sales forecasting model with outcome in units sold
  - But always be careful with comparisons of different models



Mean percentage error

$$MPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{y_i}$$

Mean absolute percentage error

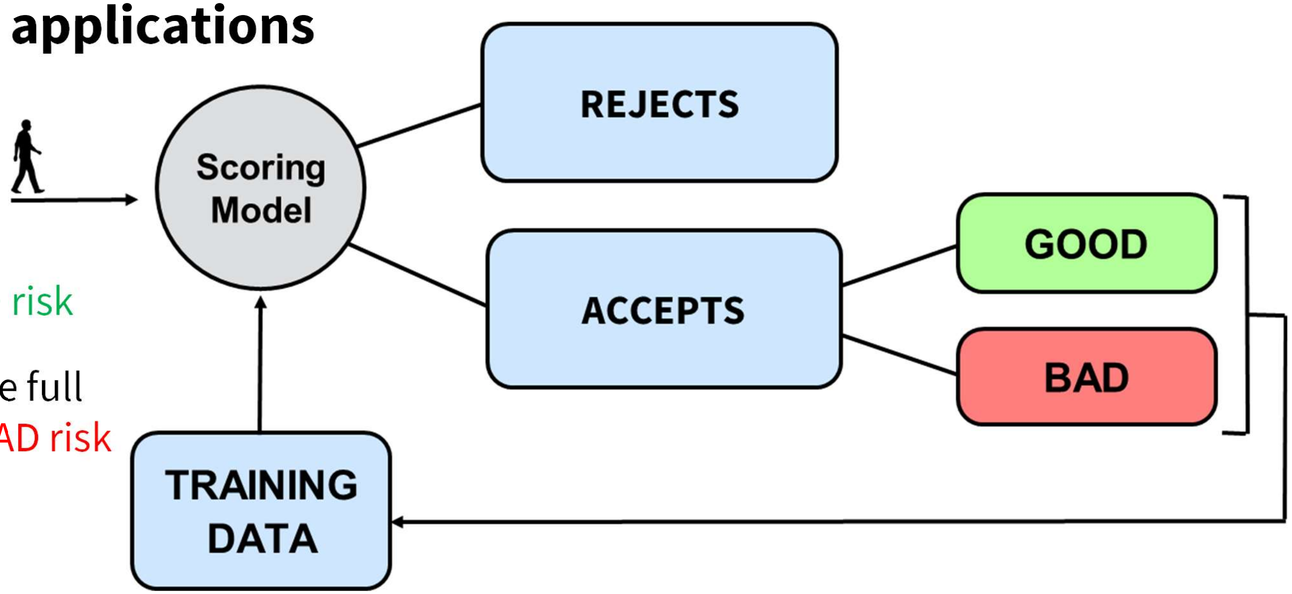
$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Symmetric MAPE

$$sMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}$$

Assessing Predictive Models for Binary Classification
Scorecard-Based Decision-Making in the Credit Industry

- **Scoring model decides incoming applications**
  - Data on borrower characteristics
  - Predict repayment behavior
    - All repaid according to contract → GOOD risk
    - Delayed payments, failure to pay back the full amount of the loan, other problems → BAD risk
  - Approve or reject applicant



- **Types of errors and their consequences**

The Confusion Matrix in a Credit Scoring Setting

Confusion matrix		**True repayment status**	
		GOOD payer	**BAD payer**
Predicted repayment status	**GOOD payer**	Accept a credit-worthy applicant.	Accidentally accept a bad payer.
	BAD payer	Erroneously reject an applicant who would repay.	Reject an applicant who would have defaulted.

The (General-Purpose) Confusion Matrix

- **Working with application-specific class labels can be cumbersome**
  - Strong commonalities across applications
  - We always have 2x2 possible cases
  - Two types of correct classification
  - Two types of errors
- **Abstraction**
  - Introduce standard names (i.e. labels)
  - Generic naming convention:
  - We call one class the positive class, and the other the negative class
  - It does not matter which class you label as „the positive“ class
  - However, many applications will know standards
    - For ex. medical diagnosis
    - Positive test outcome implies presence of a disease

		Actual Class	
		Positive	**Negative**
Predicted Class	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Standard Performance Indicators for Binary Classification

Confusion matrix for a binary classification problem

		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 / Percentage correctly classified** $\dfrac{TP + TN}{TP + TN + FP + FN}$
- **Classification error** $\dfrac{FP + FN}{TP + TN + FP + FN}$
- **Specificity** $\dfrac{TN}{TN + FP}$
- **Sensitivity / Recall** $\dfrac{TP}{TP + FN}$
- **Precision** $\dfrac{TP}{TP + FP}$

But How to Obtain a Confusion Matrix in the First Place?

The confusion matrix is a function of the classification cut-off

i	Y	$\hat{p}(Y=1\|X)$
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 $\hat{p}(Y=1|X)$ to **cut-off** τ:

$$\hat{Y} = \begin{cases} 1 & \hat{p}(Y=1|X) > \tau \\ 0 & \hat{p}(Y=1|X) \leq \tau \end{cases}$$

Common Performance Indicators for Classification
Receiver Operating Characteristic (ROC) Curve

- **Generalization of the confusion matrix**
  - One confusion matrix corresponds to **one cut-off**
  - The 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 Positive
  - And the point (1,1) where all cases are classified as Negative
  - 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	$\hat{p}(Y = 1 \mid X)$
1	1	0.9
2	1	0.7
3	1	0.6
4	0	0.6
5	0	0.2

Compare $\hat{p}(Y = 1|X)$ to **cut-off** τ:

$$\hat{Y} = \begin{cases} 1 & \hat{p}(Y = 1|X) > \tau \\ 0 & \hat{p}(Y = 1|X) \leq \tau \end{cases}$$

$\tau = 0.9$

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

$\tau = 0.7$

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

$\tau = 0.6$

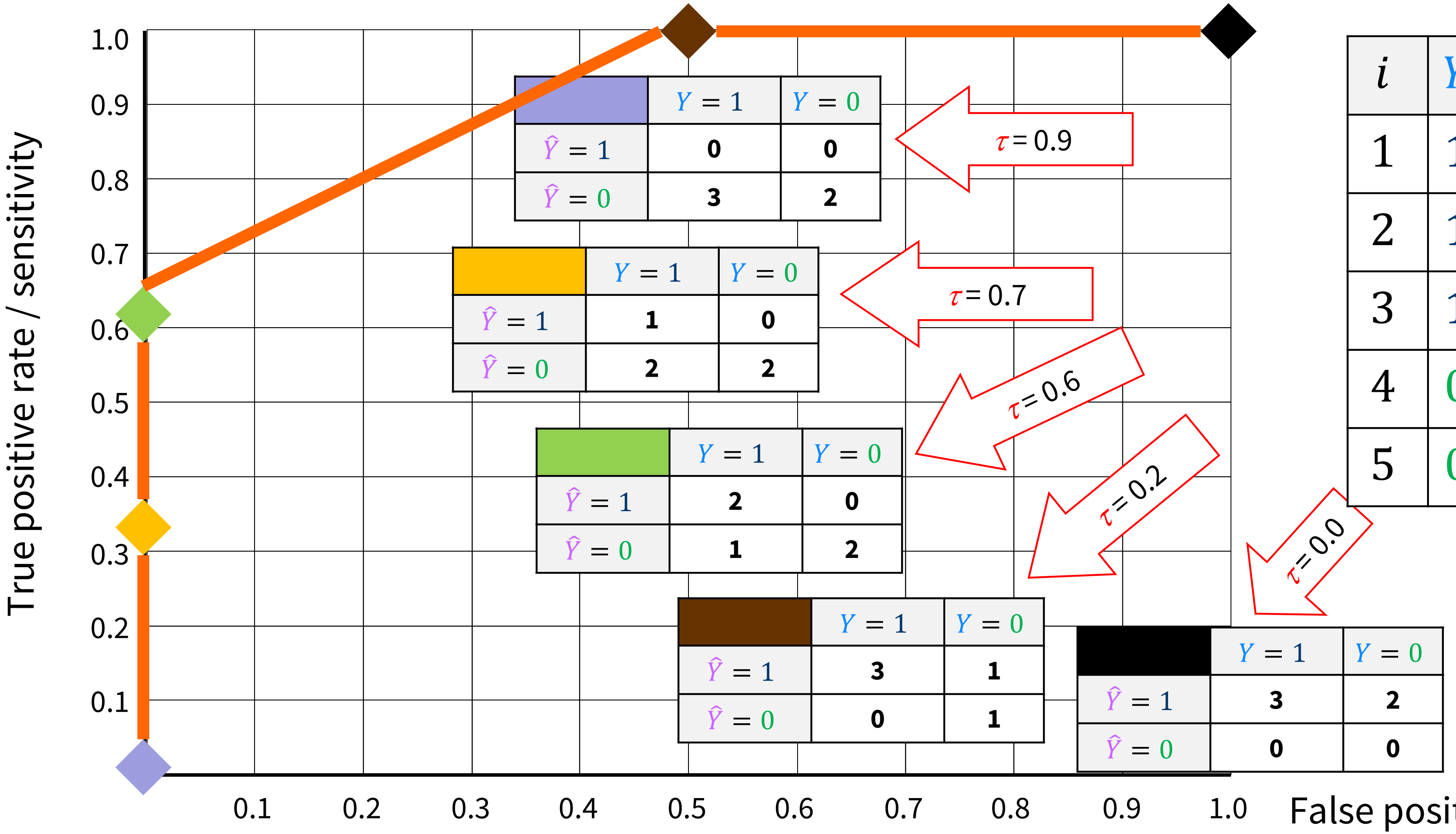
	Positive $(Y = 1)$	**Negative** $(Y = 0)$
Positive $(\hat{Y} = 1)$	2	0
Negative $(\hat{Y} = 0)$	1	2

$\tau = 0.2$ …

$\tau = 0.0$ …

Construction of the ROC Curve

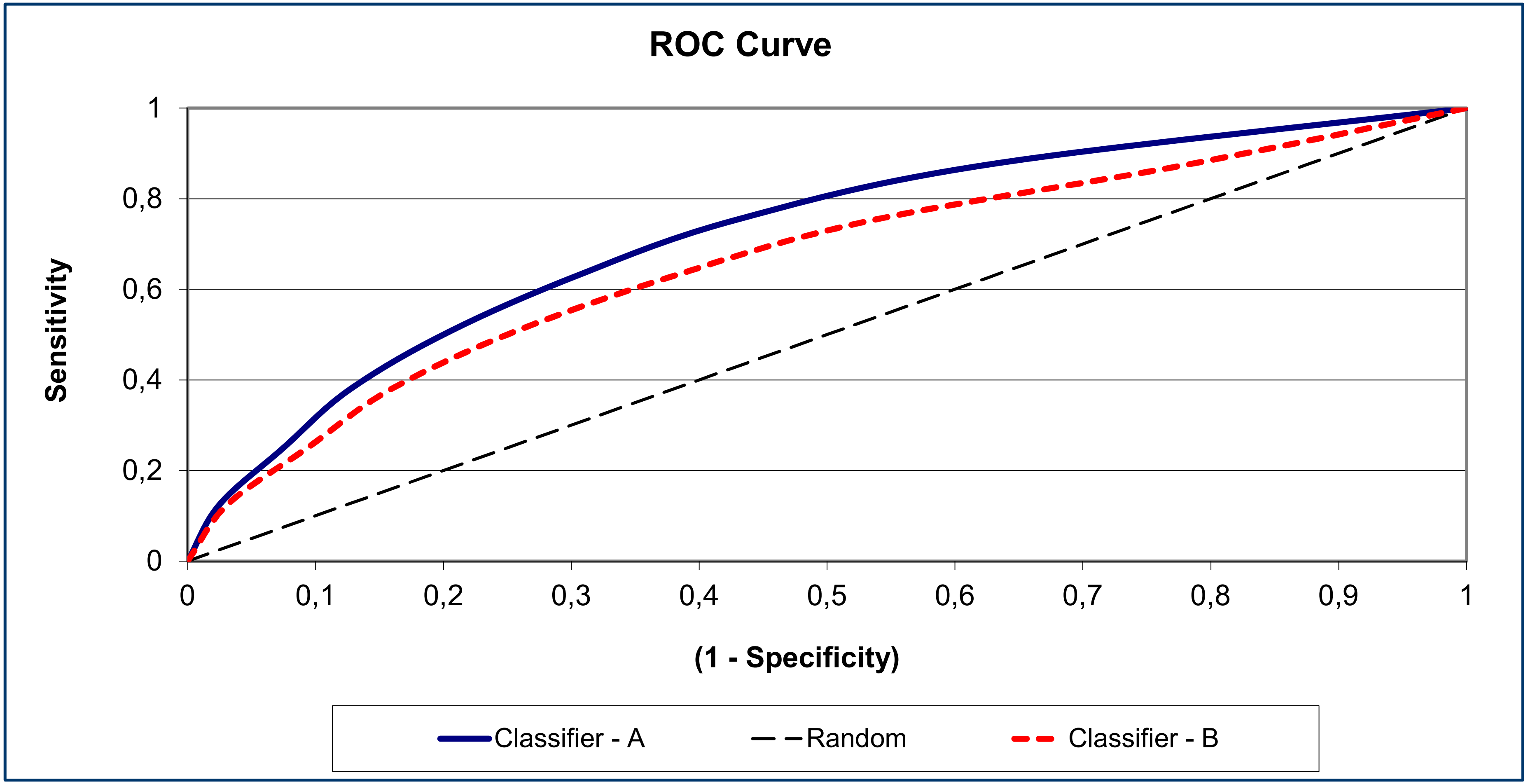
Visualization of classifier performance across all cut-offs



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

Construction of the ROC Curve

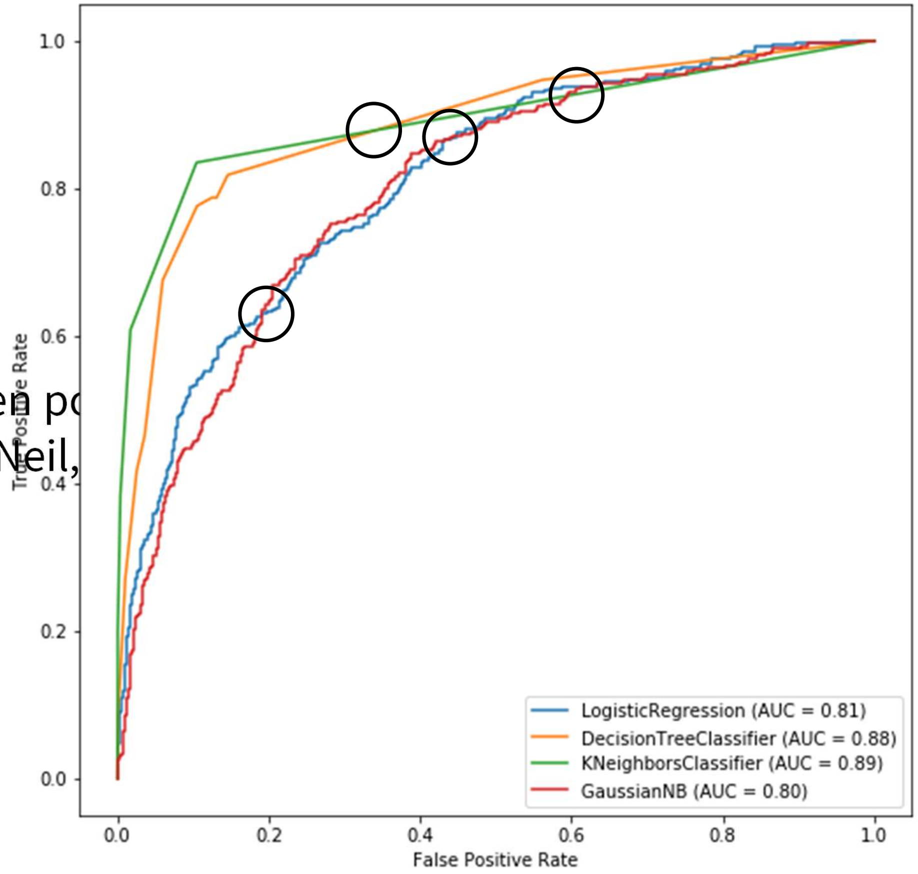
Comparing two classifiers (A and B) in ROC space



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 po
    higher than a randomly chosen negative (Hanley and McNeil,
  - 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
  - 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)

Data organization for supervised machine learning

Split sampling and cross-validation

Once upon a time…: Assessing Forecast Accuracy
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 resale price of a returned notebook is \$125
  - Say a model, based on notebook and client characteristics, forecasts the resale price to be \$98. How good or bad is that forecast?
- **Question 2: How to know the true values of the target variable?**
  - The point of developing a predictive model is to forecast future values of the target
  - We never know actual target values before we deploy the model
  - So how compare forecast and actual resale prices?
- **Two core ingredients of forecast accuracy evaluation**
  - Measures for predictive performance
  - Practices to organize the available data (up now)

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

Data Organization Intuition

Reserve some of the historical data for model testing

Learning Algorithm

Stage 1: Model Training

Data-driven development of a predictive model using labelled data $\mathcal{D} = \{Y_i, \boldsymbol{X}_i\}_{i=1}^{n}$

Historical data for training incl. Y

i	Y	X_1	X_2	…	X_m
1	…	…	…	…	…
2	…	…	…	…	…
…	…	…	…	…	…
n	…	…	…	…	…

Model

Stage 2: Model Testing

Apply trained model to the hold-out data to obtain prediction and compare to known actuals.

Historical data for testing incl. Y

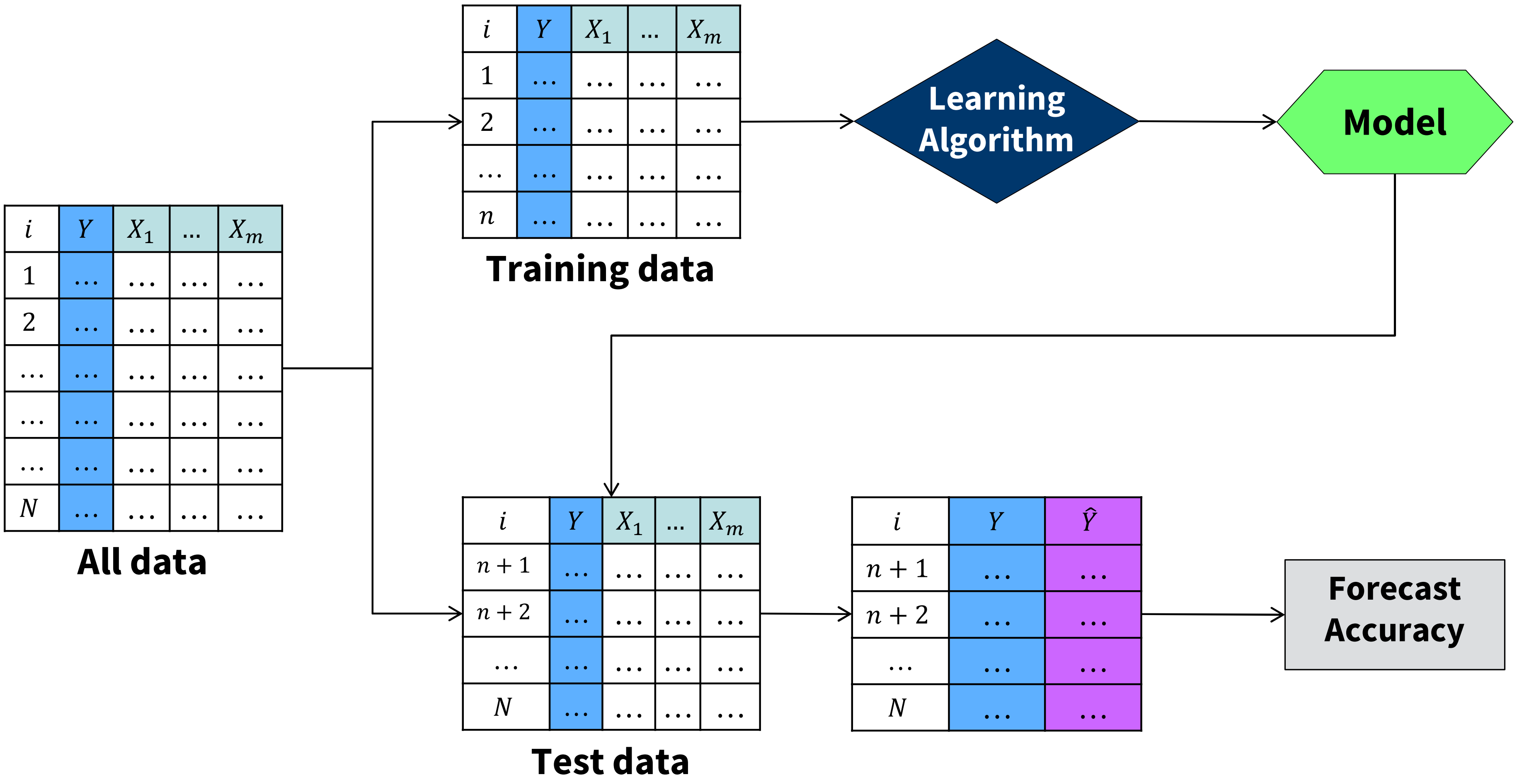
i	Y	X_1	…	X_m
$n+1$	…	…	…	…
$n+2$	…	…	…	…
…	…	…	…	…
N	…	…	…	…

Forecasts of Y

i	Y	$\hat{Y}$
$n+1$	…	…
$n+2$	…	…
…	…	…
N	…	…

Measuring Forecast Accuracy Needs 'Fresh' Data Not Used for Training

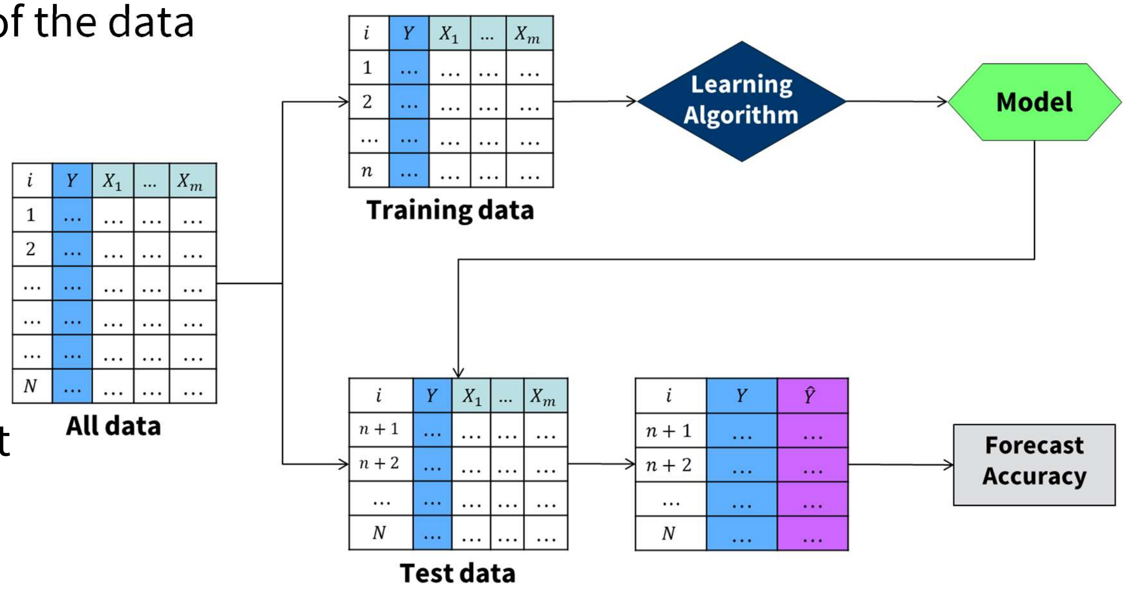
Hold-out method: split data in disjoint subsets for training & testing



Hold-Out Method Under the Microscope

Simple, easy-to-understand but somewhat inefficient approach

- **Data splitting is wasteful**
  - Train / test set often comprise 70 / 30 percent of the data
  - Much data lost for training; same for testing
  - Both training and test set should be large
- **'Lucky sample' problem**
  - We split the data randomly
  - It may be, we draw a test set in which performance seems fantastic, although it is not
  - (High variance)
- **Many alternatives exist**
  - Increase efficiency of data usage
  - Increase robustness of performance estimate



K-Fold Cross Validation

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	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

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 1

Training data

Validation data

K-Fold Cross Validation

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

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

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

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

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

Model 1

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

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 2

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

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 3

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

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 4

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

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 5

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

K-Fold Cross Validation

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.

Discussion

Do split-sampling/cross-validation simulate a real-world application of the prediction model?

Mind the Shortcut
Fallacies of the training/test set approach

- **Shortcut solutions in an ML context**
  - Model relies on simple characteristics of the data
  - Model does not learn the true essence of the relationship between features and the target
- **Problem with shortcut solutions**
  - Shortcut features facilitate accurate predictions for a specific data set
  - Splitting the data into training and test set does not help, as test set performance is still high
  - On novel data, however, the shortcut might no longer be accessible
  - This would break the model (i.e., poor generalization)

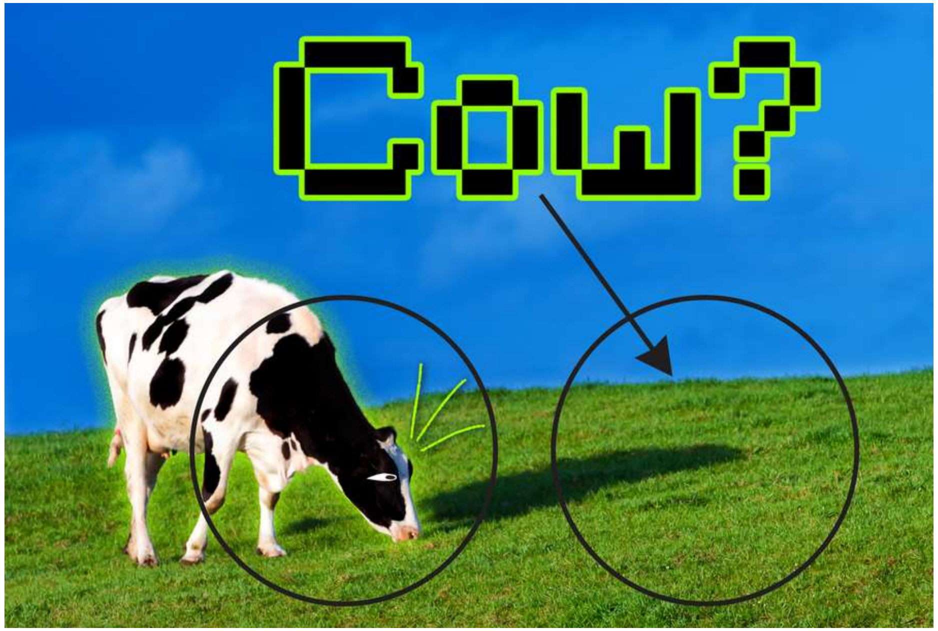


Image source:
MIT News https://news.mit.edu/2021/shortcut-artificial-intelligence-1102)

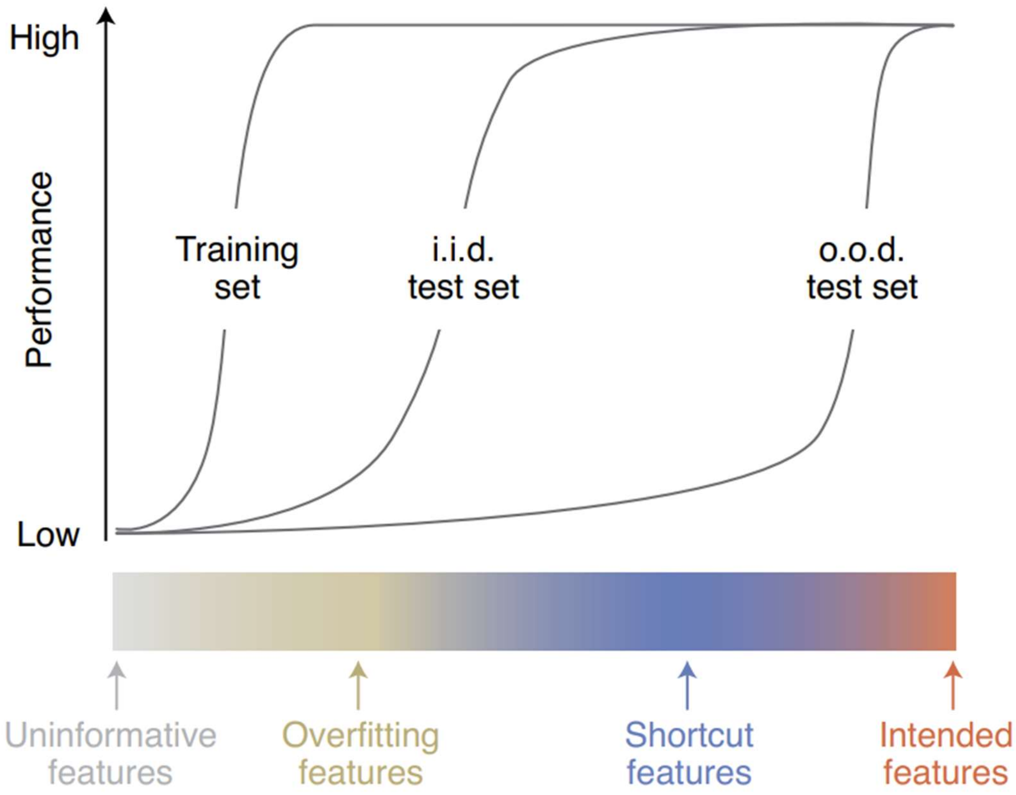
Geirhos, et al. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence, 2(11), 665-673. https://doi.org/10.1038/s42256-020-00257-z*

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.

Mind the Shortcut

Fallacies of the training/test set approach

- **Shortcut solutions in an ML context**
  - Model relies on simple characteristics of the data
  - Model does not learn the true essence of the relationship between features and the target
- **Problem with shortcut solutions**
  - Shortcut features facilitate accurate predictions for a specific data set
  - Splitting the data into training and test set does not help, as test set performance is still high
  - On novel data, however, the shortcut might no longer be accessible
  - This would break the model (i.e., poor generalization)



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

Summary

Summary

Learning goals

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

Findings

- Model performance has facets beyond accuracy
- Accuracy measures contrast actuals vs. forecasts
- Confusion matrix depends on classification cut-off
- ROC analysis generalizes the confusion matrix
- No in-sample evaluation! Hold-out data is crucial
- Pros and cons of cross-validation vs. split sample

What next

- Demo notebook on prediction model evaluation
- Some theory on supervised learning

Literature

- 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.
- 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

Appendix

Further dimensions of model performance

Dimensions of Model Performance

Many factors determine the value of an analytical 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

Can the model cope with noise and missing values? How about irrelevant and correlated attributes?

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?

Calibration

For probability forecasts! Out of 100 events predicted to have 90% chance, about 90 should have occurred. True?

Dimensions of Model Performance
Scalability

- **Consumption of time resources**
- **Time needed to build model (training time)**
  - Depends on number of cases and attributes
  - Run-time complexity
  - Importance depends on update frequency
- **Time needed to generate predictions**
  - Much less than training time
  - Critical in real-time settings (e.g., E-Commerce)
- **Both time factors differ substantially across algorithms**

- **Consumption of memory resources**
  - During model building
  - When storing final model
  - Big data prohibits keeping all training data in memory
- **Sensitivity with respect to hyperparameters**
  - Building one model is never enough
  - Some models need a lot more tuning than others
- **Parallelization important**
  - Model building
  - Model tuning

Dimensions of Model Performance
Robustness

- **Real-world data is noisy**
  - Missing values
  - Erroneous data entries
  - Wrong labels
  - Irrelevant / correlated attributes
- **Real-world phenomena change over time**
  - Concept drift
  - Model recalibration versus re-estimation
- **How to these factors affect the model?**
  - During model building
  - After model building

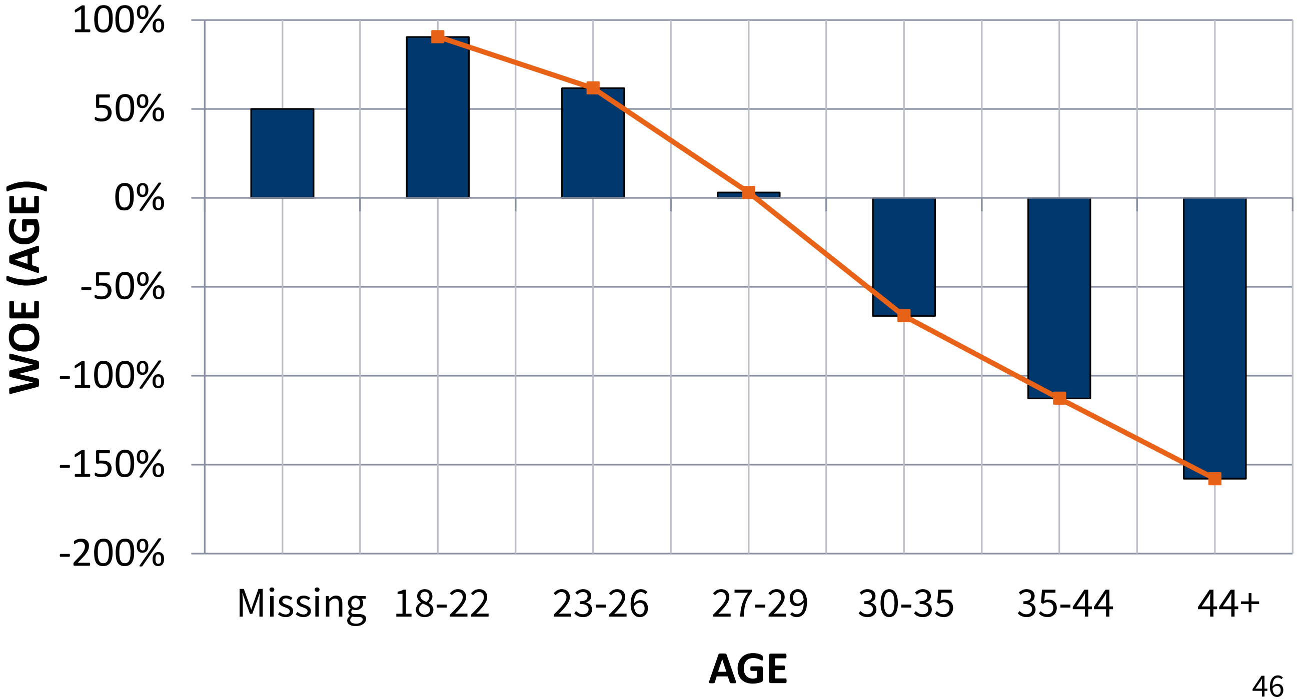
Dimensions of Model Performance
Comprehensibility: crucial and challenging to measure

- **Is it possible to understand how a model translates attribute values into prediction?**
  - Alternative terms: interpretability, transparency, white-box (vs. black-box) model
  - Becoming increasingly relevant with the raising popularity of machine learning
  - “Managers don’t trust black-box models”
- **New research fields on interpretable machine learning (see subsequent sessions)**
  - Global interpretability: equivalent to above point. How do covariates govern predictions
  - Local interpretability: how was the prediction of a specific observation determined by covariate values
- **Prediction versus insight and correlation versus causality**
  - Prediction: “Next month, we sell 100 laptops”
  - Insight: “Sales increase by 2% if we lower prices by €50”
  - Standard machine learning models are correlational

Dimensions of Model Performance

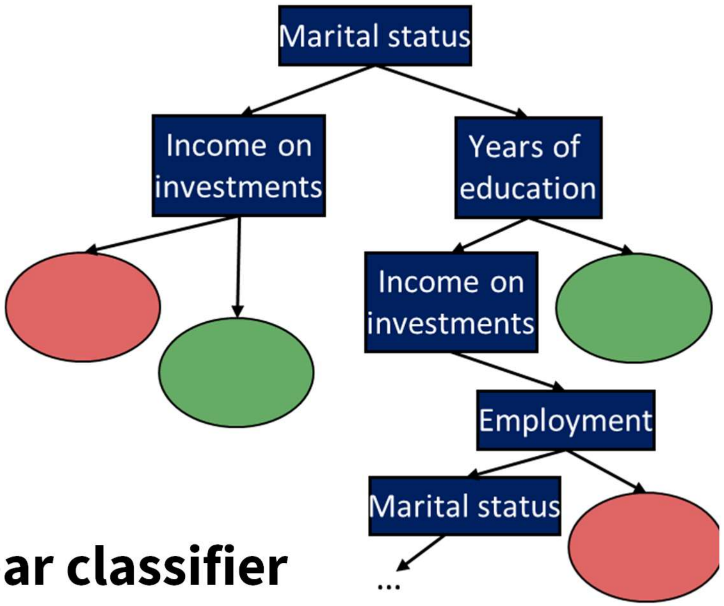
Justifiability: a key driver of model acceptance in industry

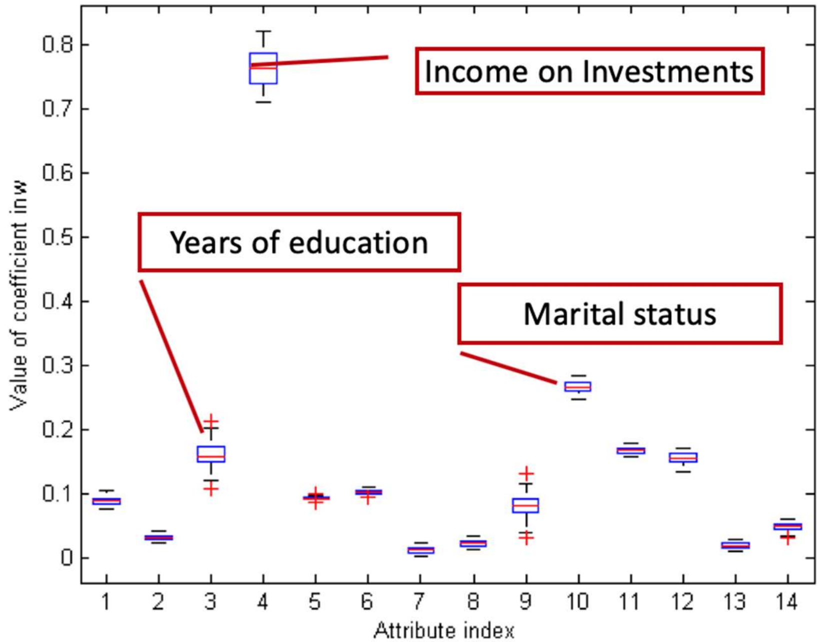
- **Does the way in which attribute values affect predictions agrees with prior beliefs or business rules?**
  - Exemplary business rules: sales decrease with price, long-term customers are more profitable than new customers, etc.
  - Requires interpretability
- **Credit risk example**
  - Business rule: credit risk decreases with age
  - Test: does WOE show this trend



Dimensions of Model Performance
Comprehensibility / Justifiability Example

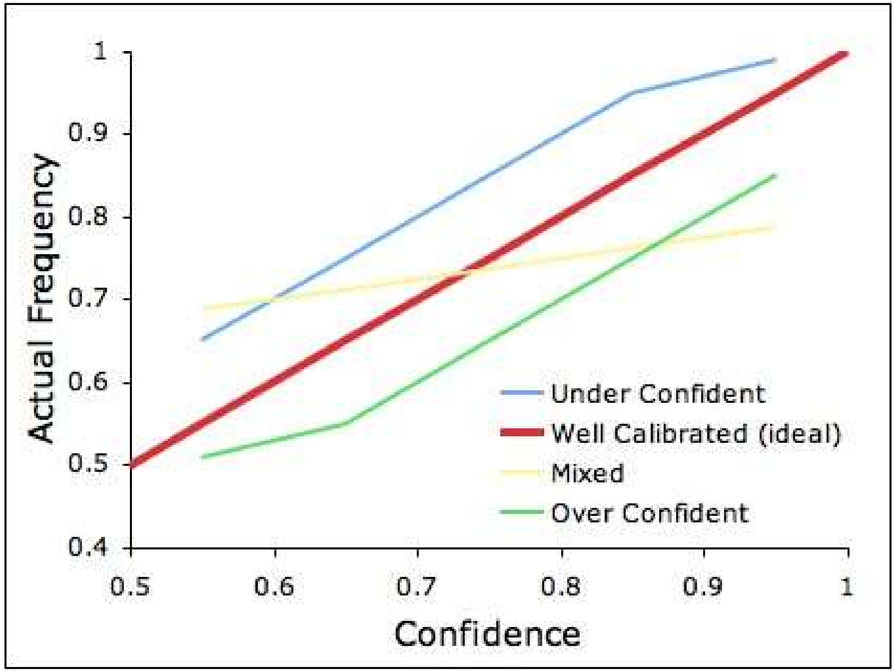
- **US Census data set from UCI library** (https://archive.ics.uci.edu/ml/datasets/Adult)
- **Classification task**
  - Is household income below or above $50,000 p.a.?
  - Fourteen attributes describing a household
    - Marital status
    - Working hours
    - Academic degree
    - Years of education
    - Country of origin
    - Income on investments
    - Employment
    - …
- **Result of tree and linear classifier**





Dimensions of Model Performance
Calibration

- **Feature of probabilistic predictions**
- **Credit Scoring Example**
  - Model makes risk forecasts for 100 credit applications
  - Forecasts are all the same and predict default of 90%
  - Then, we should eventually observe 90 actual defaults
- **For prediction models**
  - Calibration can be poor
  - Special treatment needed
  - See, e.g., Bequé et al. (2017)



[https://goodmorningeconomics.wordpress.com/2008/07/11/calibrated-probability-assessmentorg/]