EVALUATION OF PREDICTIVE ANALYTICS MODELS

Lecture

Topics:

- Evaluation of regression models
- Evaluation of classification models

Time: 2 hours

Co-funded by the European Union

TET – The Evolving Textbook
Project no: 2022-1-SI01-KA220-HED-000088975

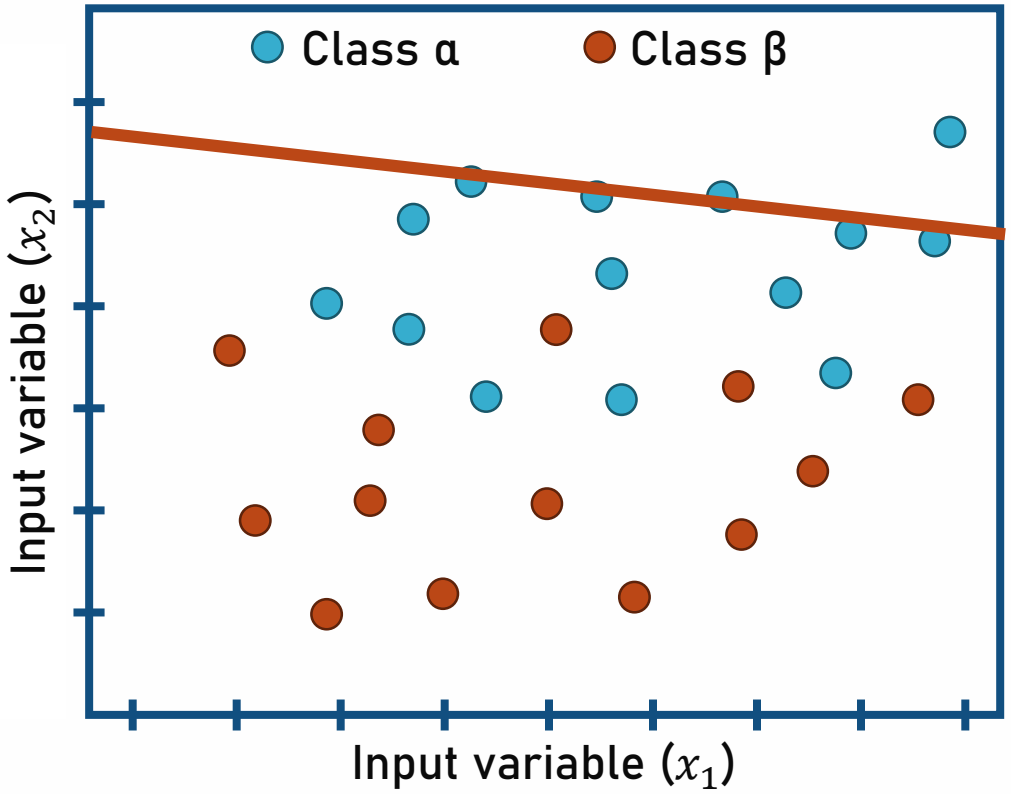
Łukasz Paśko, Rzeszów University of Technology

Underfitting

Comparison of regression and classification models

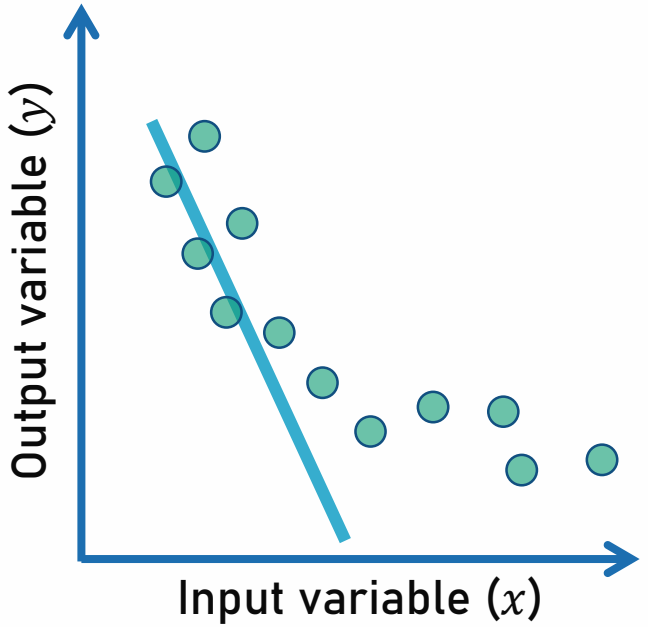
Classification

A line generated by a classification model that separates objects from two different classes



Regression

A line generated by a regression model, representing the relationship between x and y

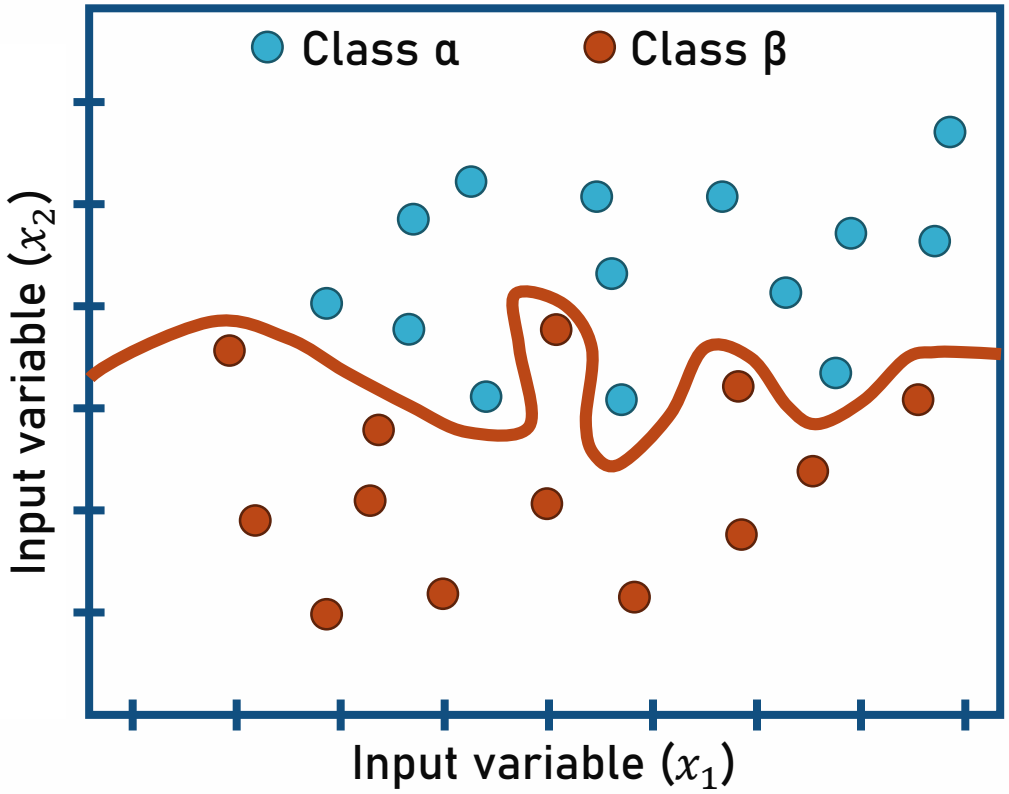


Overfitting

Comparison of regression and classification models

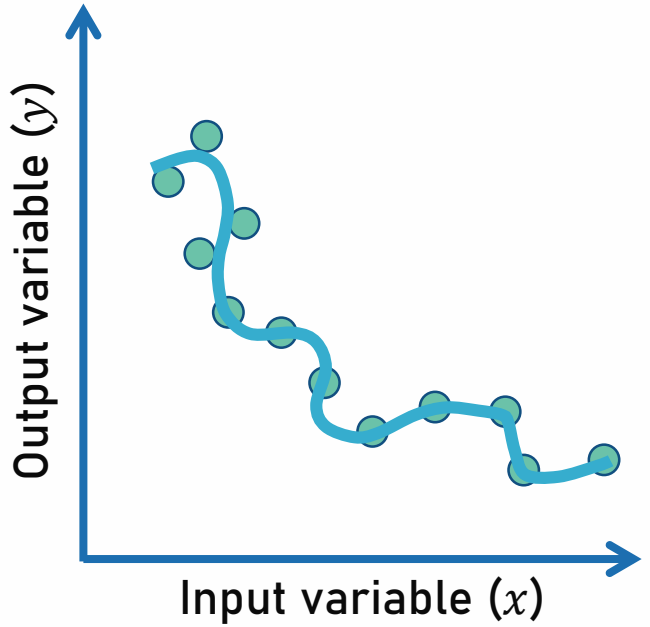
Classification

A line generated by a classification model that separates objects from two different classes



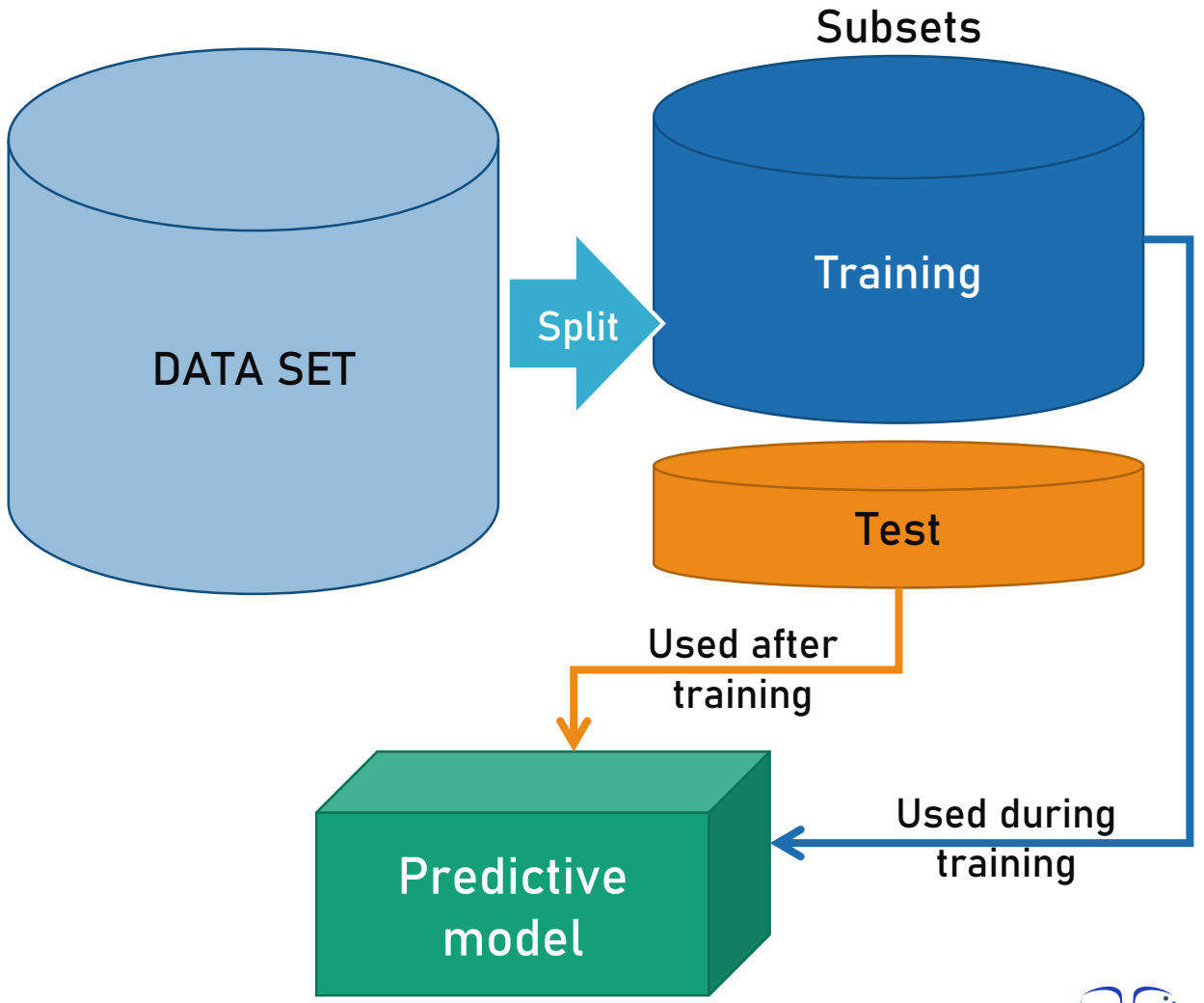
Regression

A line generated by a regression model, representing the relationship between x and y

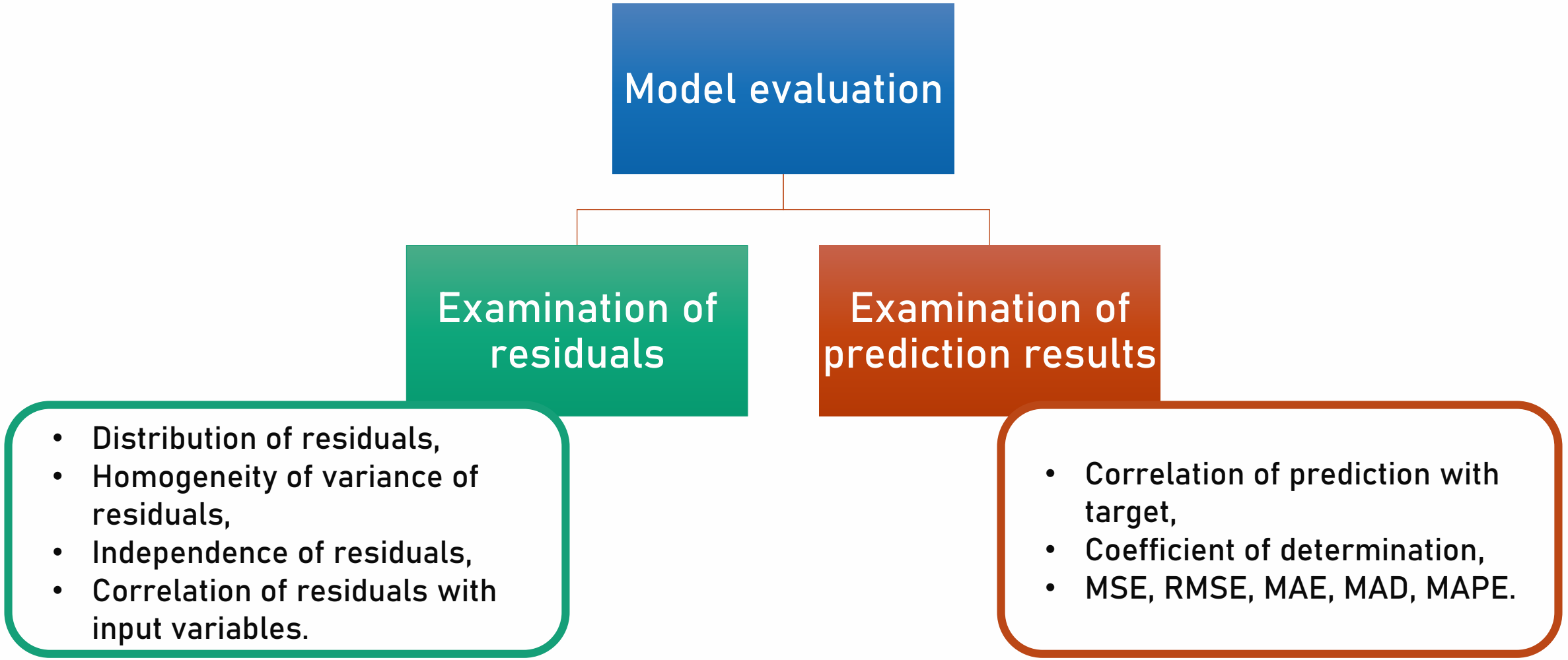


Test data set

- Both classification and regression models will be evaluated on a test subset extracted before training begins.
- Classification and regression quality measures will be calculated primarily on the test subset.
- For comparison, these measures can also be determined on the training subset – good prediction quality on the training set and poor on the test set indicates overfitting of the prediction model.

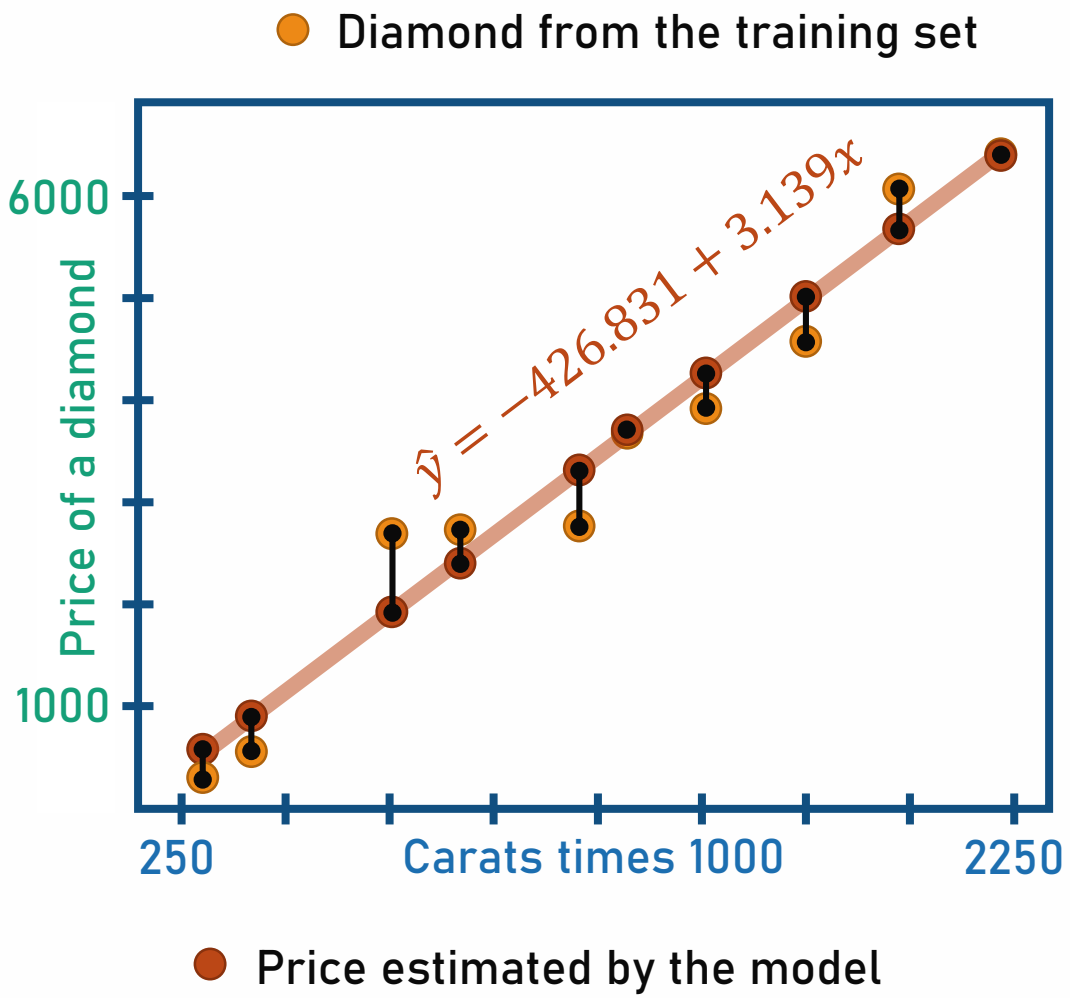


Evaluation of the regression model



Residuals

An example



Training set

x y

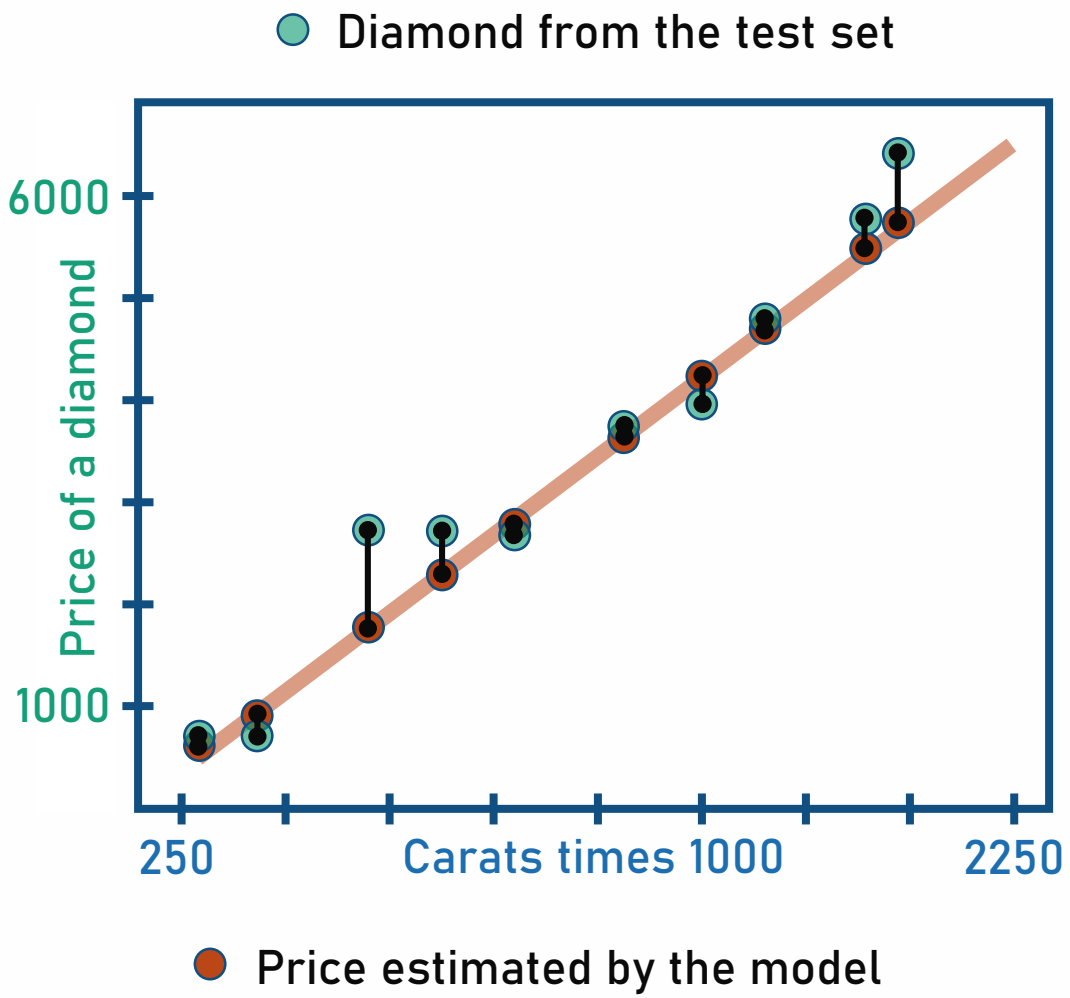
$\hat{y} = -426.831 + 3.139x$

Carats *1000	Price	Estimated price	Residuals $e = y - \hat{y}$
300	339	514.9	-175.9
410	561	860.2	-299.2
750	2760	1927.4	832.6
910	2763	2429.7	333.3
1200	2809	3340.0	-531.0
1310	3697	3685.3	11.7
1500	4022	4281.7	-259.7
1740	4677	5035.0	-358.0
1960	6147	5725.6	421.4
2210	6535	6510.4	24.6

Residuals calculated on the training set

Residuals

An example



Test set

x y

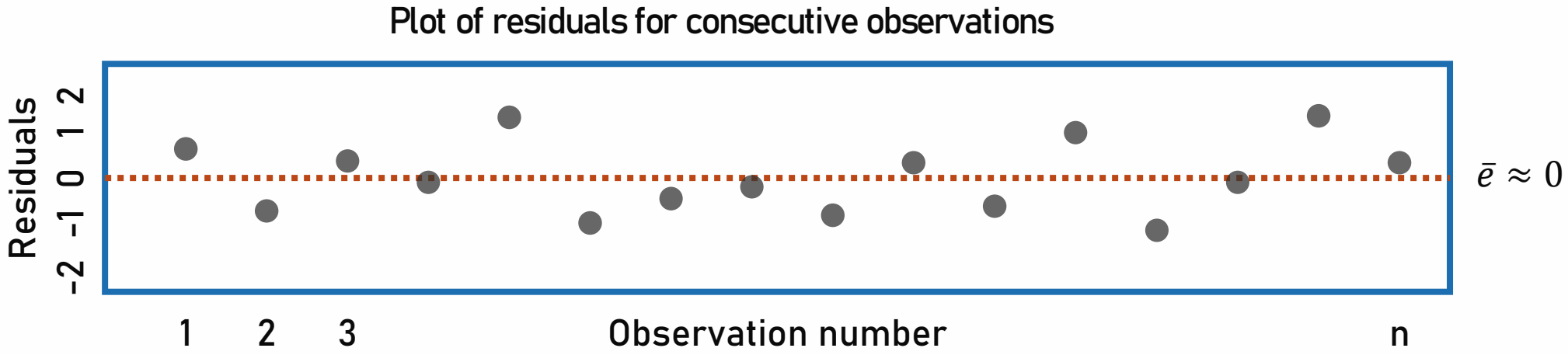
$$\hat{y} = -426.831 + 3.139x$$

Carats *1000	Price	Estimated price	Residuals $e = y - \hat{y}$
220	342	263.7	78.3
330	403	609.0	-206.0
710	2772	1801.9	970.1
810	2789	2115.8	673.2
1080	2869	2963.3	-94.3
1390	3914	3936.4	-22.4
1500	4022	4281.7	-259.7
1640	4849	4721.1	127.9
1850	5688	5380.3	307.7
1910	6632	5568.7	1063.3

Residuals calculated on the test set

Examination of residuals

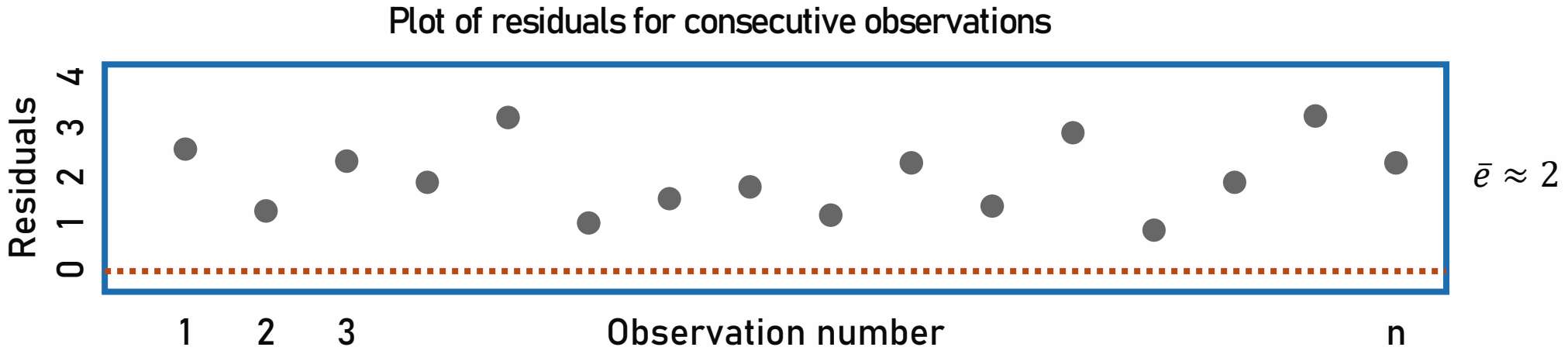
Arithmetic mean of residuals



- The arithmetic mean of the residuals should be close to zero (figure above).
- "Close to zero" depends on the value of the output variable. For example, if the output variable is expressed in millions, then the residuals expressed in tens or hundreds can be sufficiently close to zero.

Examination of residuals

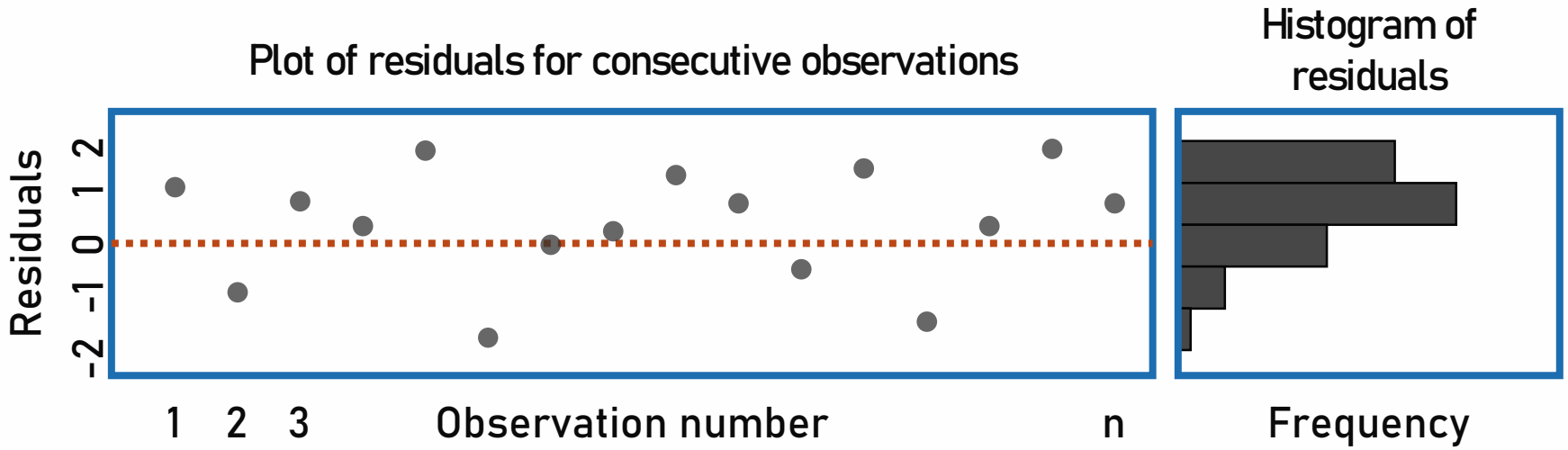
Arithmetic mean of residuals



- If the arithmetic mean of the residuals differs significantly from zero, it means that the prediction result is significantly underestimated or overestimated compared to the actual values of the output variable.
- The situation is particularly unfavorable when all the residuals are positive (figure above) or negative. This indicates an underestimation or overestimation of the prediction results.

Examination of residuals

Distribution of residuals

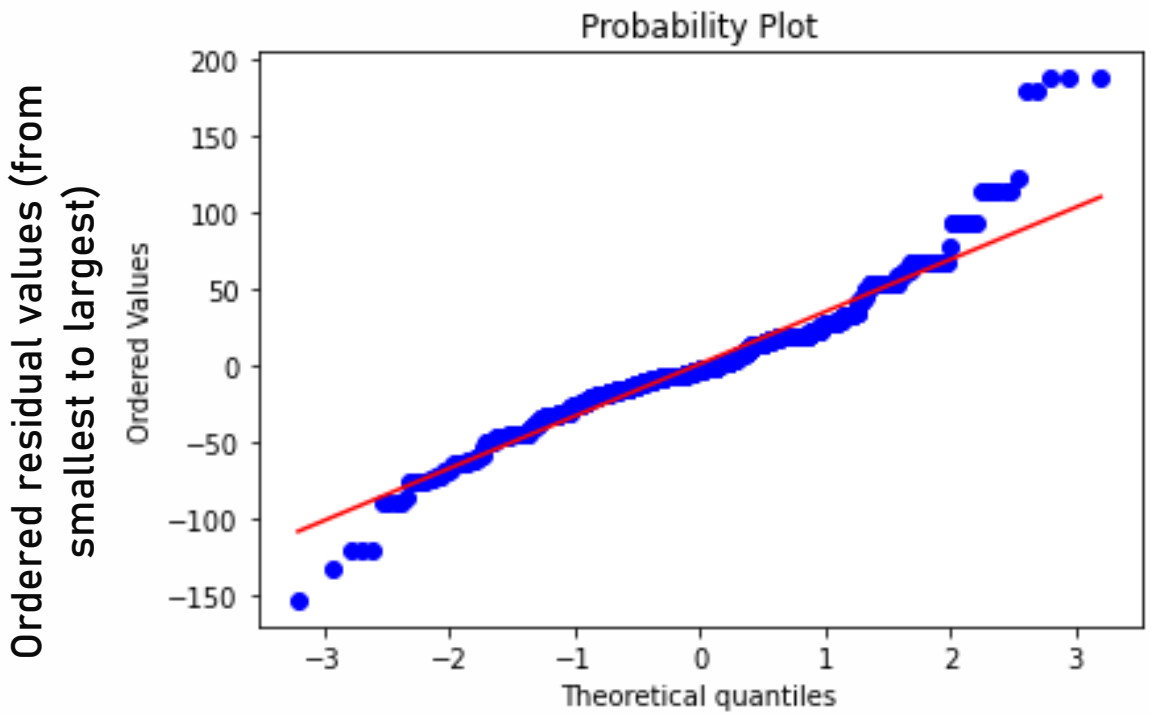


A non-normal distribution of residuals (figure above) may invalidate some of the technical requirements of regression.

Examination of residuals

Distribution of residuals

Quantile plot of residuals



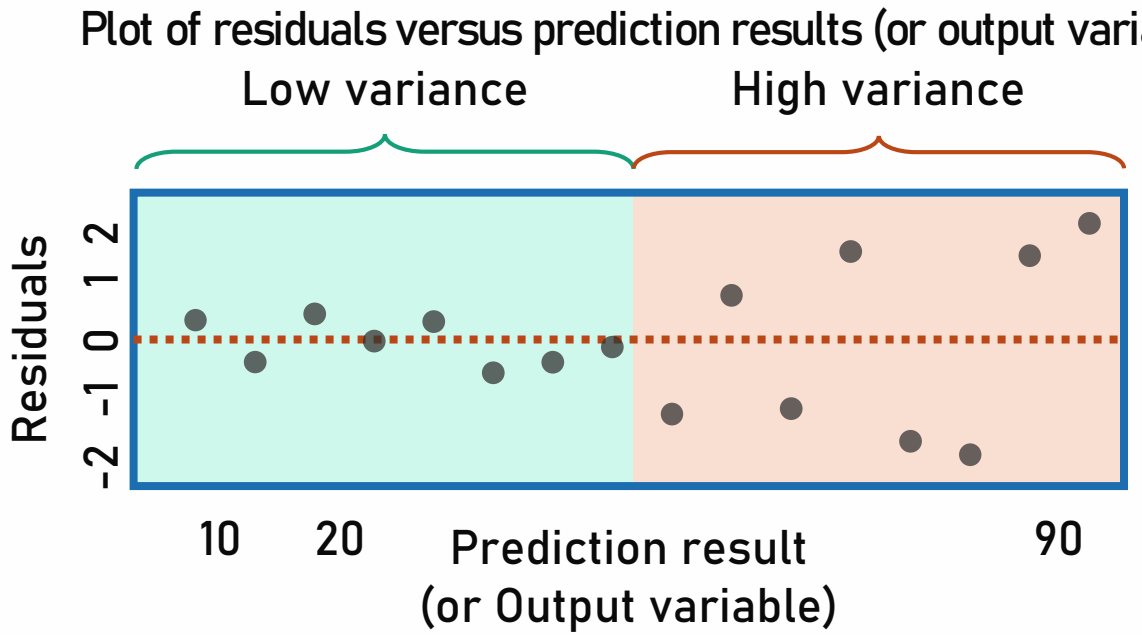
Theoretical quantiles for a normal distribution with mean 0 and standard deviation 1

Interpreting a quantile plot:

- When the residuals are normally distributed, the points should fall along the red line;
- When we see any pattern other than a straight line (e.g., an S shape or a shape resembling an exponential curve), this indicates that the distribution is not normal.

Examination of residuals

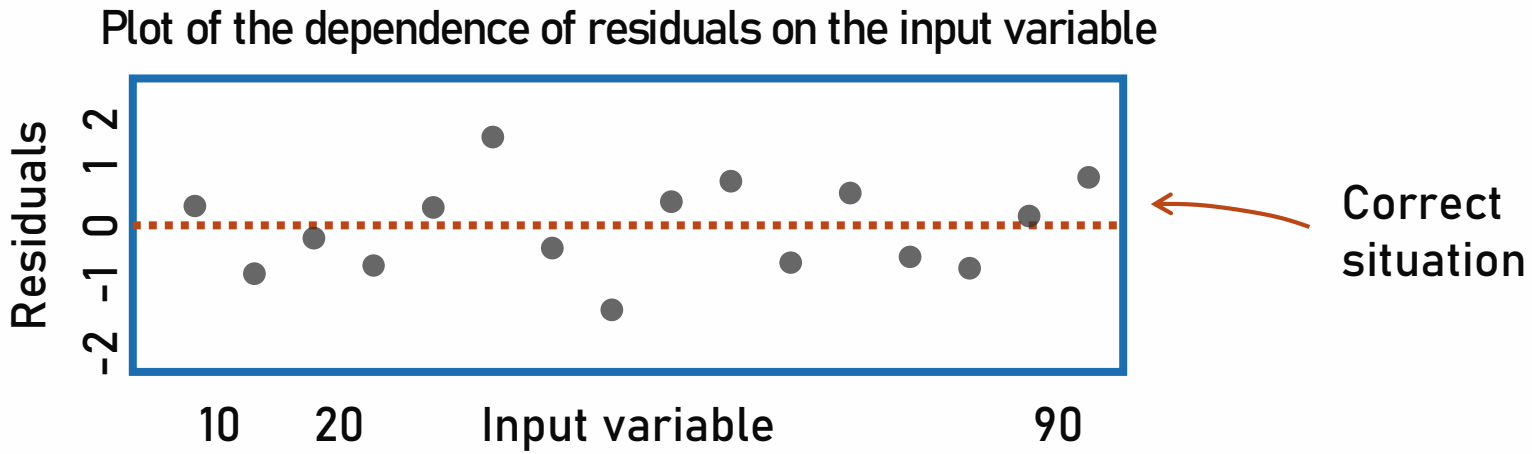
Homogeneity of variance



- The instability (variance heterogeneity) of residuals over the range of predicted values is a situation in which a larger variance of residuals is observed for certain ranges of results than in other ranges. This phenomenon is called heteroscedasticity of residuals.
- Similarly to the case of the lack of a normal distribution of residuals, heteroscedasticity of residuals is an obstacle to the validity of formal statistical assumptions (hypothesis tests) - heteroscedasticity increases the variance of the estimates of regression coefficients.

Examination of residuals

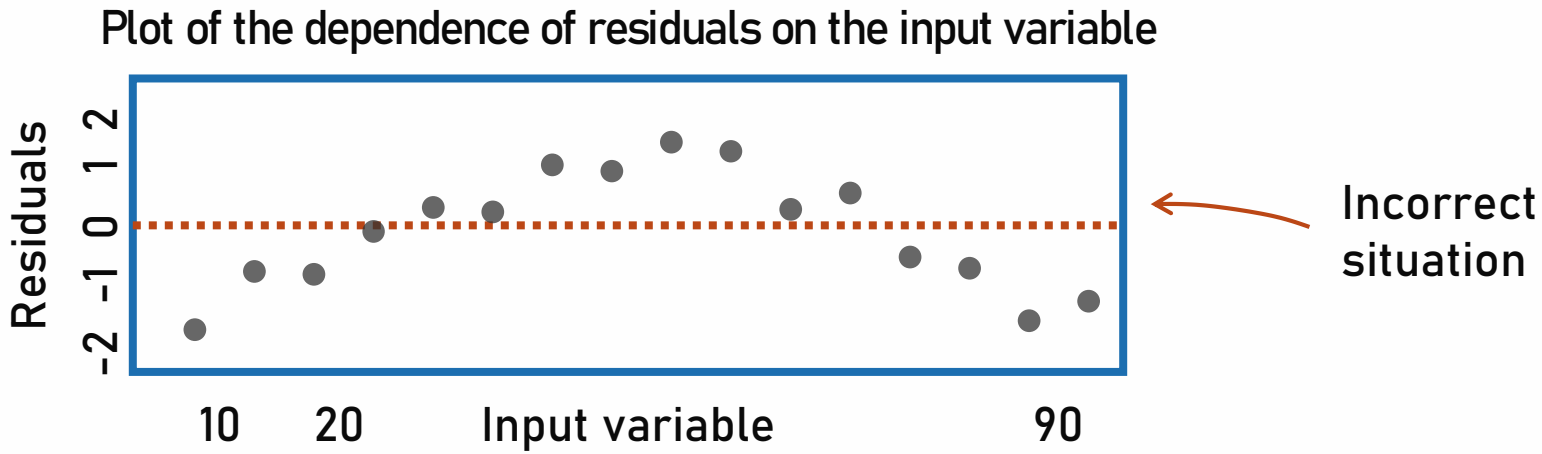
Correlation of residuals with input variables



- Residuals should not be correlated with the input variables.
- If the points on the residual plot are randomly distributed around zero, a linear regression model is appropriate for that input variable.

Examination of residuals

Correlation of residuals with input variables



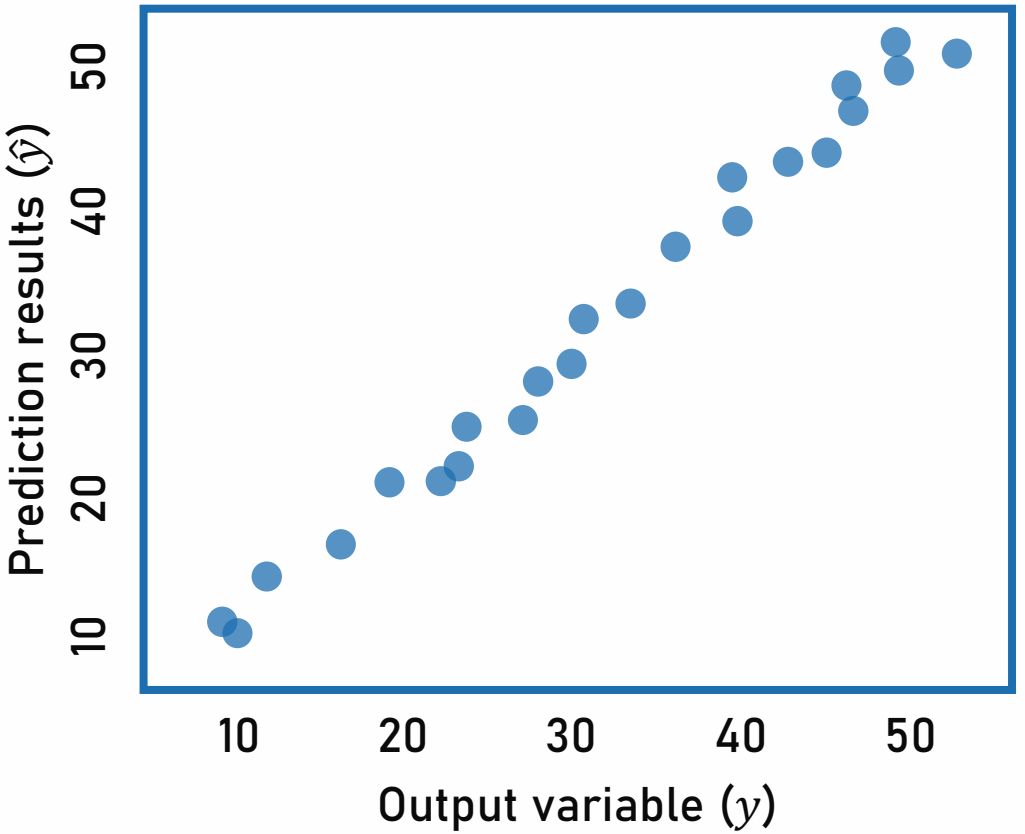
- When an input variable is correlated with the residuals, the coefficient on that variable in the regression model may be overestimated or underestimated, or even have the wrong sign.
- When the points on the residual plot follow a nonlinear pattern, then a nonlinear model may be more appropriate.

Examination of prediction results

Correlation of prediction results with target values

- Comparison of prediction results with target values on a scatterplot (target values are the values of the output variable);
- When the regression model can accurately predict the value of the output variable for each observation, then the points on the plot form a straight line;
- The plot can be made for the training and test sets, but checking the quality of the prediction on the test set is more important.

Plot of the dependence of prediction on the output variable



Examination of prediction results

Basic measures

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

Coefficient of determination

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n-1)}{n - m - 1}$$

Adjusted coefficient of determination

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

Mean Squared Error

$$RMSE = \sqrt{MSE}$$

Root Mean Squared Error

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

Mean Absolute Error

$$MAD = \text{mediana}(|y_i - \hat{y}_i|)$$

Median Absolute Deviation

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

Mean Absolute Percentage Error

n Number of cases

y_i The value of the output variable for the i-th case

$\hat{y}_i$ Prediction result for the i-th case

$\bar{y}$ Average value of the output variable

m Number of input variables

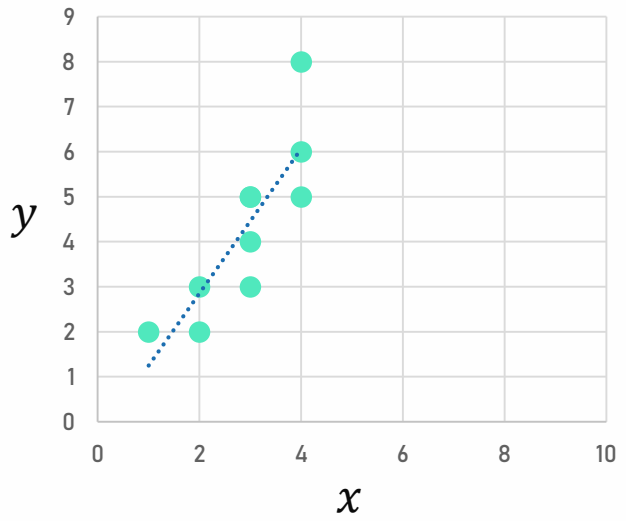
Examination of prediction results

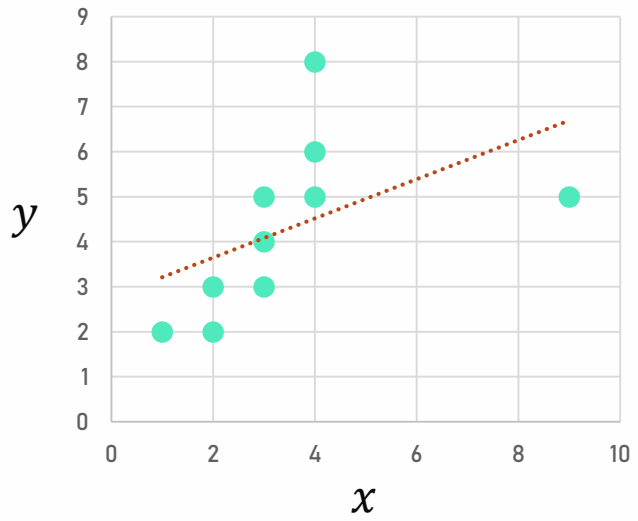
Comparison of measures on two data sets

x	y	$\hat{y}$
3	5	4.4605
4	6	6.0672
4	5	6.0672
2	3	2.8538
3	5	4.4605
1	2	1.2471
3	4	4.4605
2	2	2.8538
3	3	4.4605
4	8	6.0672

Two training data sets
differing by one observation

Line fitting the regression
model to the data





x	y	$\hat{y}$
3	5	4.0824
4	6	4.5177
4	5	4.5177
2	3	3.6471
9	5	6.6942
1	2	3.2118
3	4	4.0824
2	2	3.6471
3	3	4.0824
4	8	4.5177

Examination of prediction results

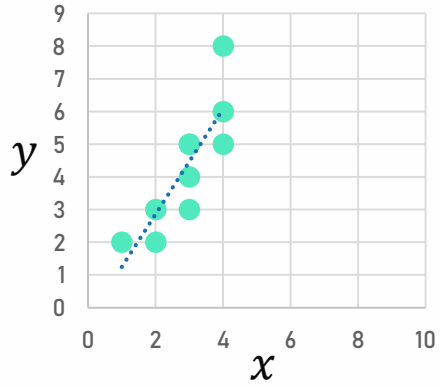
Coefficient of determination

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

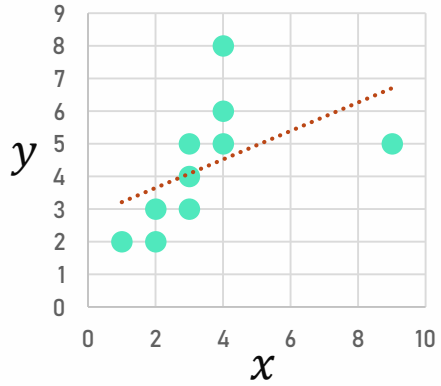
Coefficient of determination

- The higher the value,
  the better the model fit to the data.
- It tells us what percentage of the variability in the data is explained by the model (what part of the variability in the output variable is explained by the predictive model).
- It is usually assumed that a value of the coefficient of determination greater than 0.6 indicates a sufficient model fit, but this depends on the problem domain.

$R^2 = 0.72$



$R^2 = 0.25$



Examination of prediction results

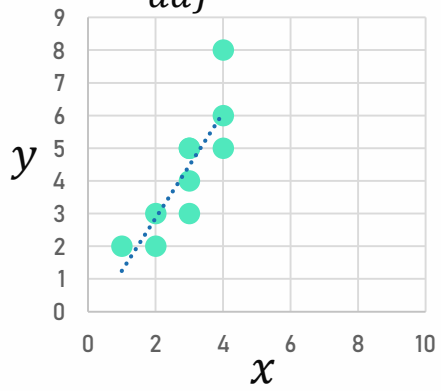
Adjusted coefficient of determination

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - m - 1}$$

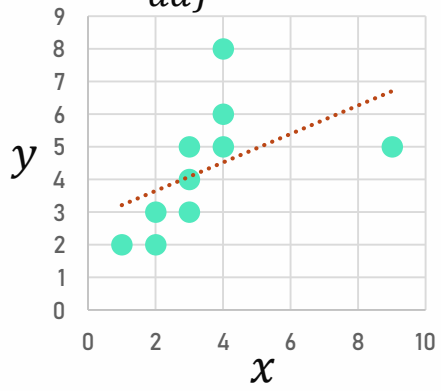
Adjusted coefficient of determination

- Value always less than or equal to R^2.
- The higher the value, the better the model fits the data.
- A modified version of R^2, taking into account the number of input variables in the model.
- Adjusted R^2 introduces a “penalty” for excess input variables (the penalty is to reduce the value of this measure by dividing by the number of variables m).

$R^2_{adj} = 0.68$



$R^2_{adj} = 0.16$



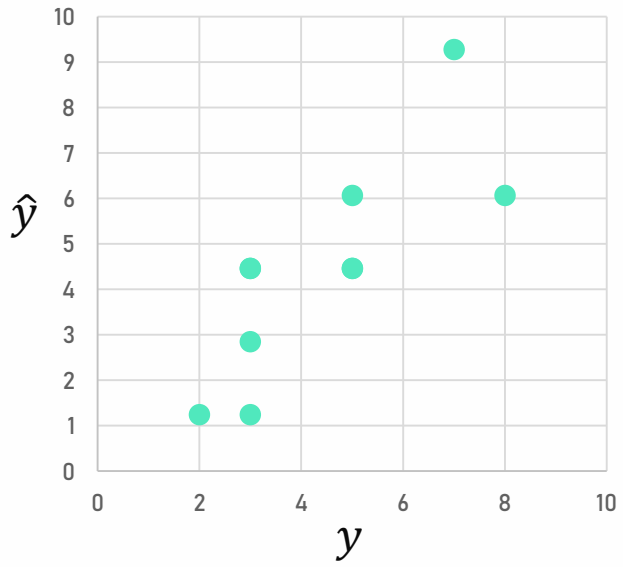
Examination of prediction results

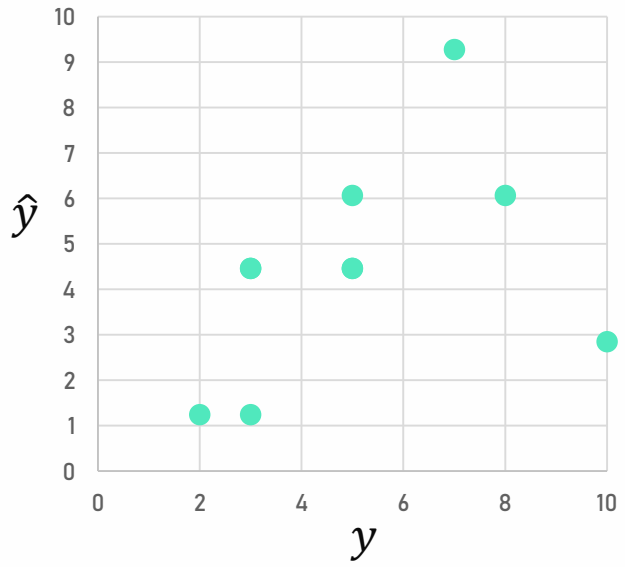
Comparison of measures on two data sets

x	y	$\hat{y}$
3	5	4.4605
4	5	6.0672
6	7	9.2806
2	3	2.8538
3	5	4.4605
1	3	1.2471
3	3	4.4605
4	8	6.0672
3	3	4.4605
1	2	1.2471

Two test data sets differing by one observation

Prediction results obtained with the same regression model





x	y	$\hat{y}$
3	5	4.4605
4	5	6.0672
6	7	9.2806
2	10	2.8538
3	5	4.4605
1	3	1.2471
3	3	4.4605
4	8	6.0672
3	3	4.4605
1	2	1.2471

Examination of prediction results

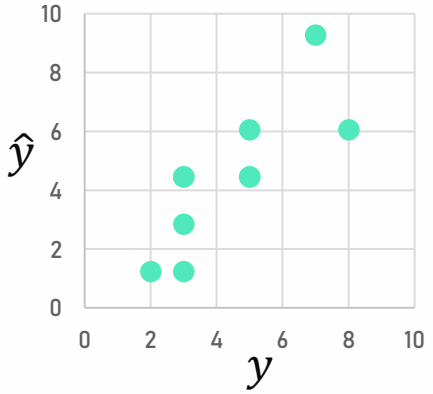
Mean Squared Error

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

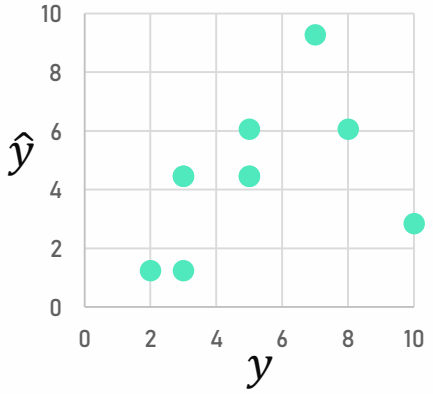
Mean Squared Error

- Average of squared errors of prediction results.
- The smaller the value, the better the model.
- Represents the mean squared difference between actual and predicted values.
- Measures the variance of residuals.
- Larger prediction errors are made by exponentiating residuals.

$MSE = 1.86$



$MSE = 6.96$



Examination of prediction results

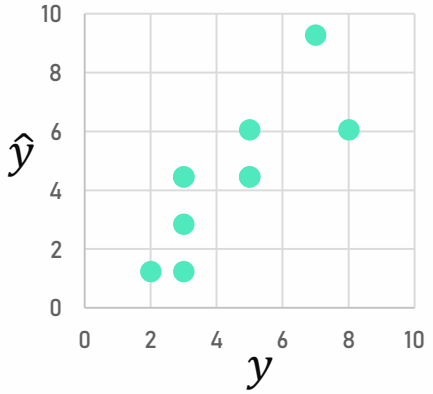
Root Mean Squared Error

$$RMSE = \sqrt{MSE}$$

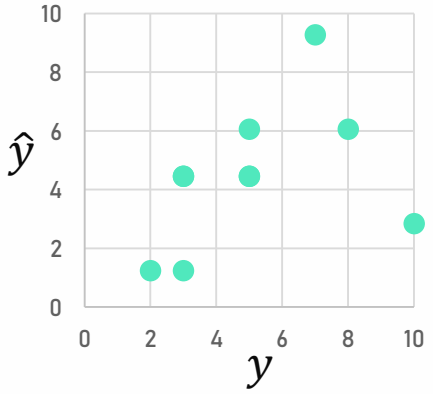
Root Mean Squared Error

- Root mean square error.
- The smaller the value, the better the model.
- The most commonly used metric for comparing regression models in data science.
- Expressed in the same unit as the dependent variable.
- Measures the standard deviation of the residuals.

$RMSE = 1.36$



$RMSE = 2.64$



Examination of prediction results

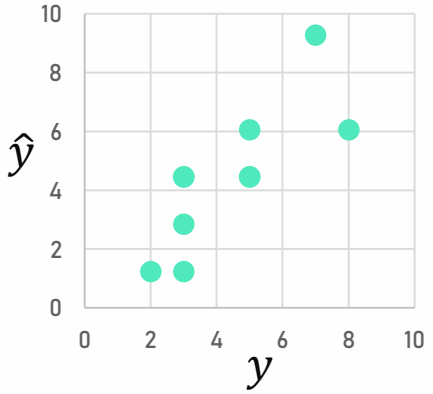
Mean Absolute Error

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

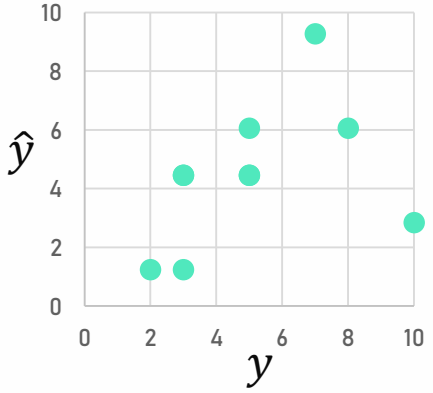
Mean Absolute Error

- Alsoknown as Mean Absolute Deviation.
- The smaller the value, the better the model.
- It represents the mean absolute difference between the actual and predicted values.
- It measures the average of the residuals in the data set.
- MAE is less “sensitive” to outliers compared to MSE, so if the goal is to train a model that focuses on reducing large outlier errors, the MSE measure should be used.

$MAE = 1.19$



$MAE = 1.89$



Examination of prediction results

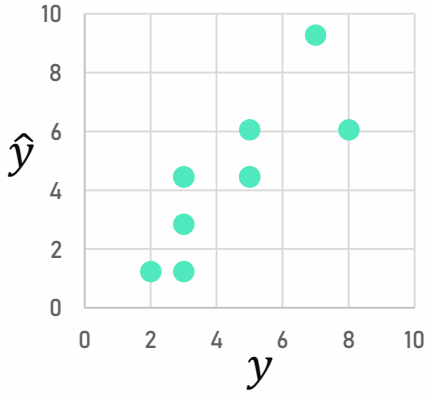
Median Absolute Deviation

$$MAD = \text{mediana}(|y_i - \hat{y}_i|)$$

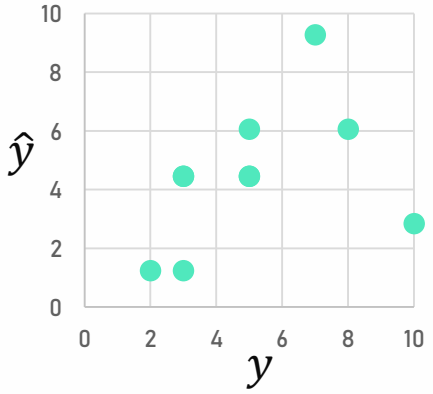
Median Absolute Deviation

- The smaller the value, the better the model.
- It measures the median of the residuals in the data set.
- Used when the test set may contain outliers that could distort the assessment of the quality of the regression model (MAD is less “sensitive” to outliers compared to MSE).

$MAD = 1.26$



$MAD = 1.46$



Examination of prediction results

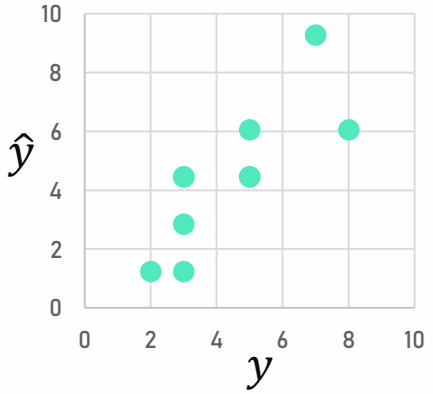
Mean Absolute Percentage Error

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

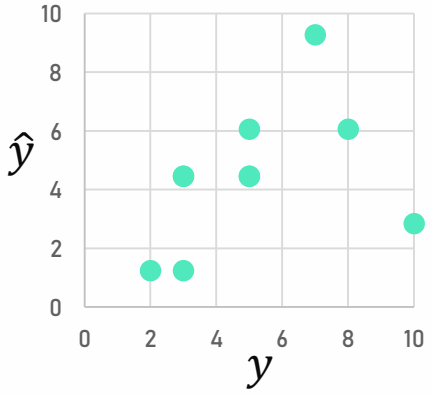
Mean Absolute Percentage Error

- Mean absolute error in percentage terms.
- The smaller the value, the better the model.
- It expresses the value of the model error in percentage terms, making it easy to interpret the size of the errors made by the regression model.
- Other names:
  - mean absolute percentage deviation (MAPD),
  - percentage absolute error (PAE).

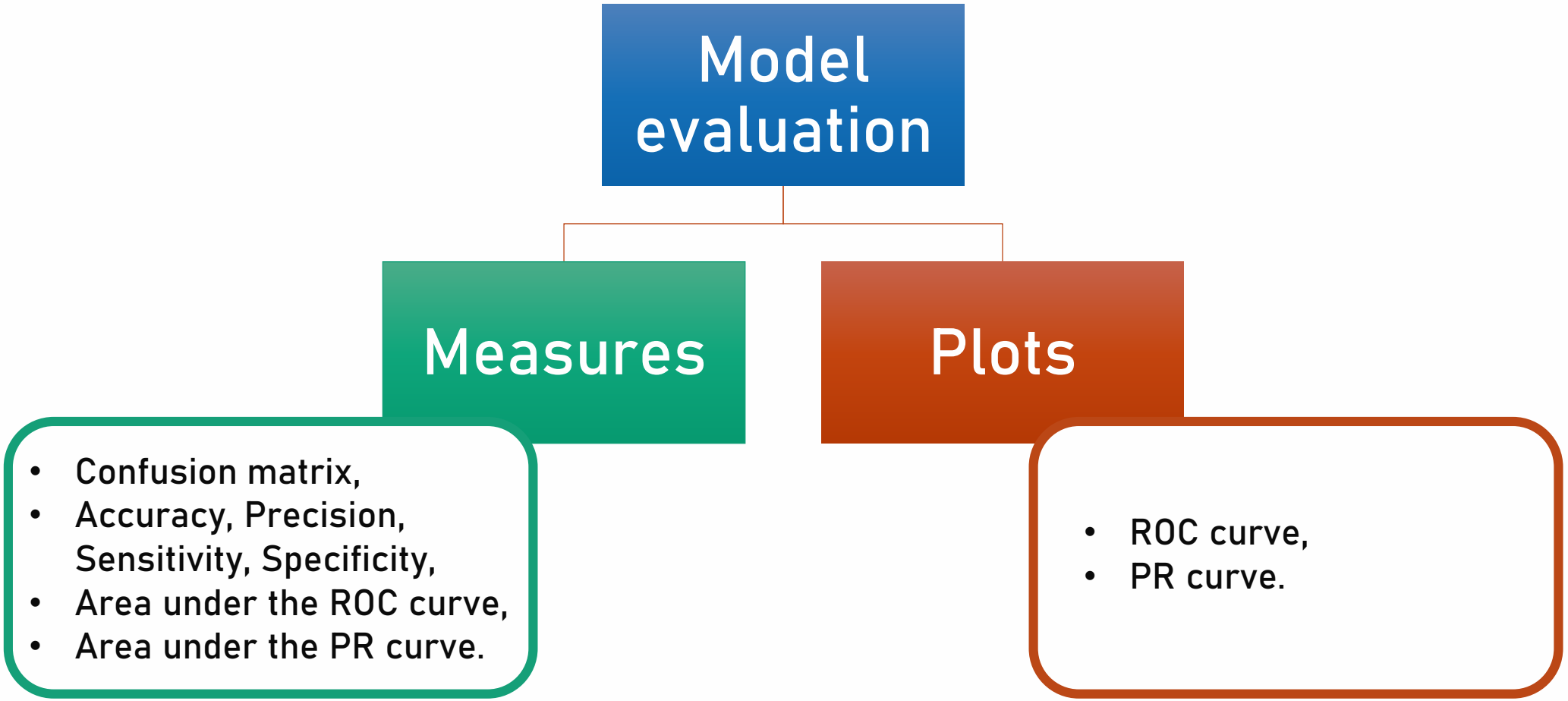
$MAPE = 29.80\%$



$MAPE = 36.46\%$



Evaluation of the classification model



Measures for evaluating the classification model

Confusion matrix

The confusion matrix (error matrix) is a matrix representation of classification results.

		Predicted classes	
		Positive (+)	Negative (-)
Actual classes	Positive (+)	True positive (TP)	False negative (FN)
	Negative (-)	False positive (FP)	True negative (TN)

+ Positive class – particularly interesting from the point of view of the prediction task and the research being conducted; denotes the occurrence of the phenomenon of interest to us.

– Negative class – lack of occurrence of the phenomenon of interest to us.

Measures for evaluating the classification model

Confusion matrix in multiclass problem

In the case of the k-class problem, the matrix has k columns and k rows.

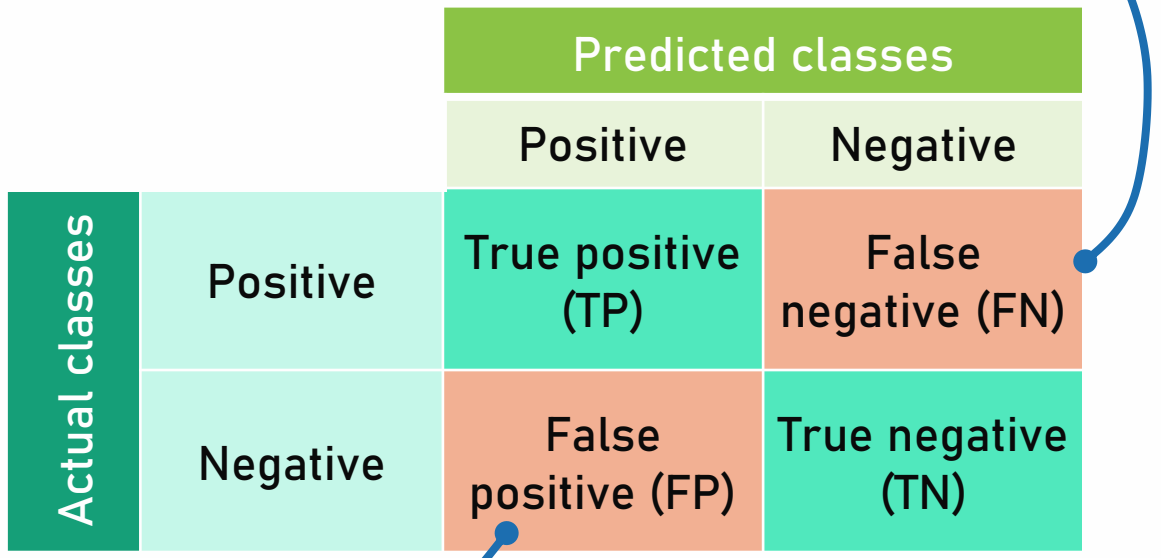
The main diagonal of the matrix (from the top left to the bottom right) shows the number of correct classifications. Off the main diagonal of the matrix are classification errors.

		Predicted classes		
		Class 1	Class 2	Class 3
Actual classes	Class 1			
	Class 2			
	Class 3			

Measures for evaluating the classification model

Types of errors

- The positive class is usually the class of interest, and the consequences of failing to recognize this class are potentially more damaging than failing to recognize the negative class.
- For this reason, type II errors are usually more costly than type I errors.

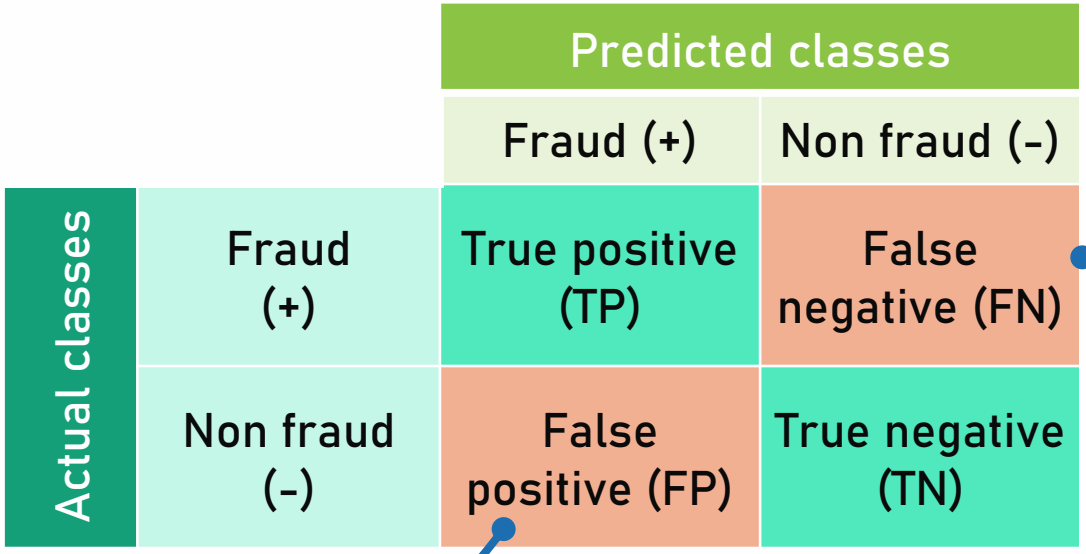


Measures for evaluating the classification model

Types of errors: an example

- In the problem of classifying insurance applications, the more interesting class is the "false application" (Fraud) than the "non-false application" (Non fraud).
- A fraudulent insurance application is considered non-false (**type II error**) - the insurance company has to pay compensation that is not due (often huge costs).
- A non-false insurance application is considered fraudulent (**type I error**) - the insurance company initiates an additional check of such an application, which generates costs, but these costs are lower than the cost of paying undue compensation.

Type II errors
(falsified insurance applications that the model found to be not fraudulent)



Type I errors
(non-false insurance applications that the model found to be fraudulent)

Measures for evaluating the classification model

Classifier quality indicators

		Predicted classes	
		(+)	(-)
Actual classes	(+)	TP	FN
	(-)	FP	TN

Acc	accuracy, correctness of fractions	$\frac{TP+TN}{TP+TN+FP+FN}$
Err	overall error rate	$\frac{FP+FN}{TP+TN+FP+FN}$
TPR	true positives rate, recall, sensitivity, hit rate	$\frac{TP}{TP+FN}$
TNR	true negatives rate, specificity	$\frac{TN}{TN+FP}$

PPV	positive predictive value, precision	$\frac{TP}{TP+FP}$
NPV	negative predictive value	$\frac{TN}{TN+FN}$
FPR	false positive rate, fall-out	$\frac{FP}{FP+TN} = 1 - TNR$
FNR	false negatives rate, miss rate	$\frac{FN}{TP+FN} = 1 - TPR$
F1	F1-score	$2\frac{PPV * TPR}{PPV + TPR}$

Measures for evaluating the classification model

Basic measures

		Predicted classes	
		(+)	(-)
Actual classes	(+)	TP	FN
	(-)	FP	TN

Accuracy
percentage of correct classifications

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity
percentage of positive cases correctly recognized by the model
(the "strength" of the model in recognizing positive cases)

$$TPR = \frac{TP}{TP + FN}$$

Precision
percentage of true positive cases out of all cases considered positive by the model

$$PPV = \frac{TP}{TP + FP}$$

Specificity
percentage of negative cases correctly recognized by the model
(the "strength" of the model in recognizing negative cases)

$$TNR = \frac{TN}{TN + FP}$$

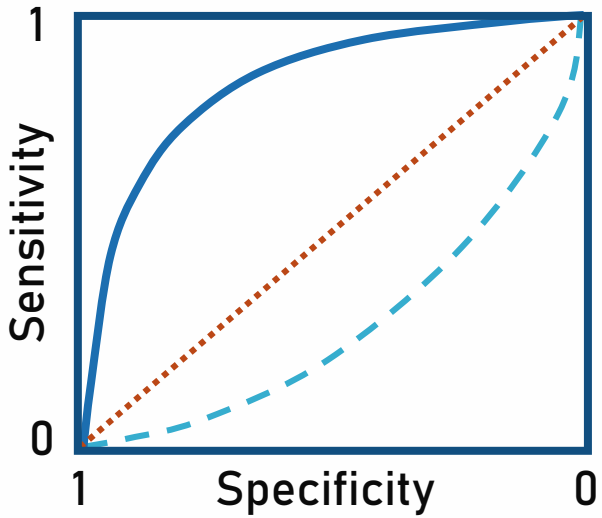
Plots for evaluating the classification model

ROC Curve (Receiver Operating Characteristics)

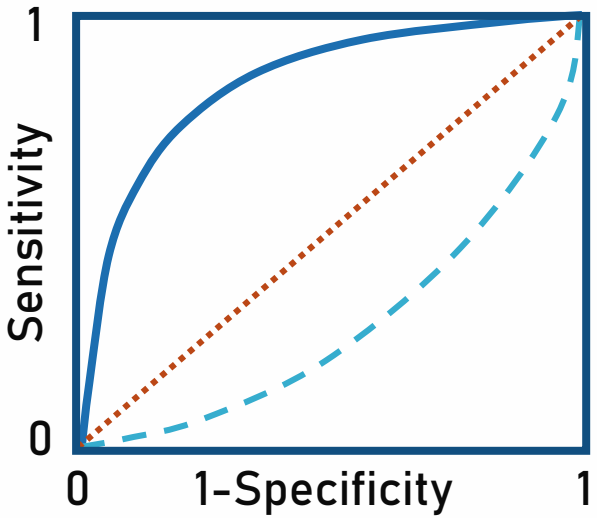
The ROC curve visualizes the trade-off between sensitivity and specificity when changing the classification threshold.

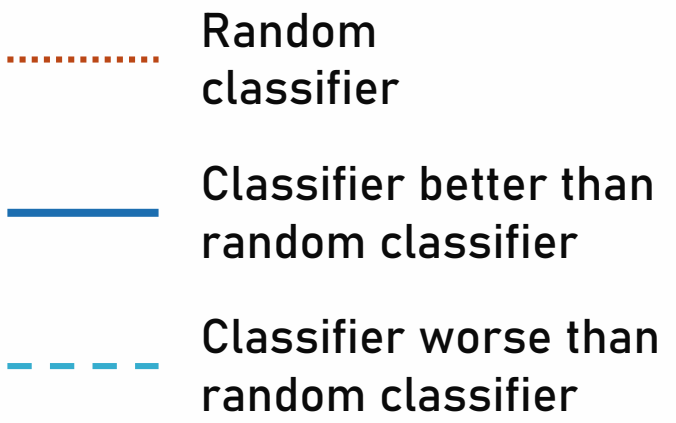
There are two versions of the ROC curve, differing only in the way the horizontal axis is drawn.

The first version of ROC chart



The second version of ROC chart

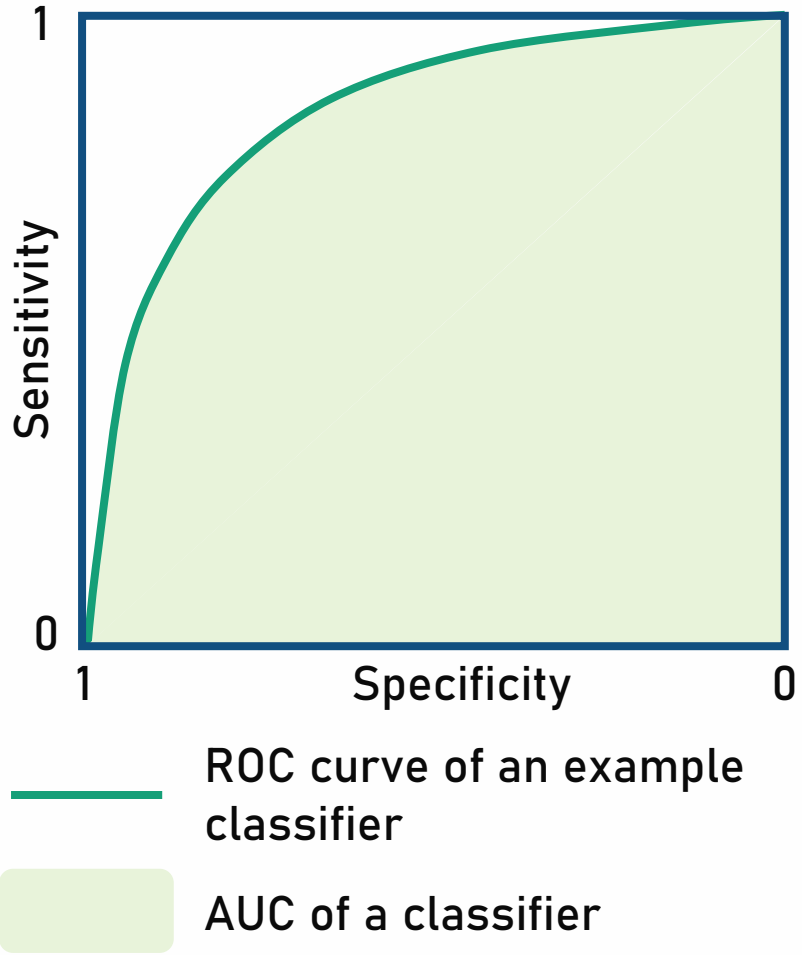




Plots for evaluating the classification model

Area Under Curve, AUC

- The area under the ROC curve is used as another indicator of classifier performance.
- The higher the AUC value, the better the classifier performance (ability to distinguish between positive and negative classes).
- The maximum AUC value is 1.



Plots for evaluating the classification model

Trade-off between sensitivity and specificity

- If we want the classifier to recognize more positive objects, we need to set the prediction threshold below 0.5, but this will result in an increase in the number of negative objects classified as positive. We will increase sensitivity (TPR), but decrease specificity (TNR).
- Moving the prediction threshold above 0.5 will have the opposite effect - we will increase specificity, but decrease sensitivity.

Threshold = 0.65

		Pred. (+)	Pred. (-)
Real	(+)	5	9
	(-)	0	13

TPR=0.36 TNR=1.00

Threshold = 0.5

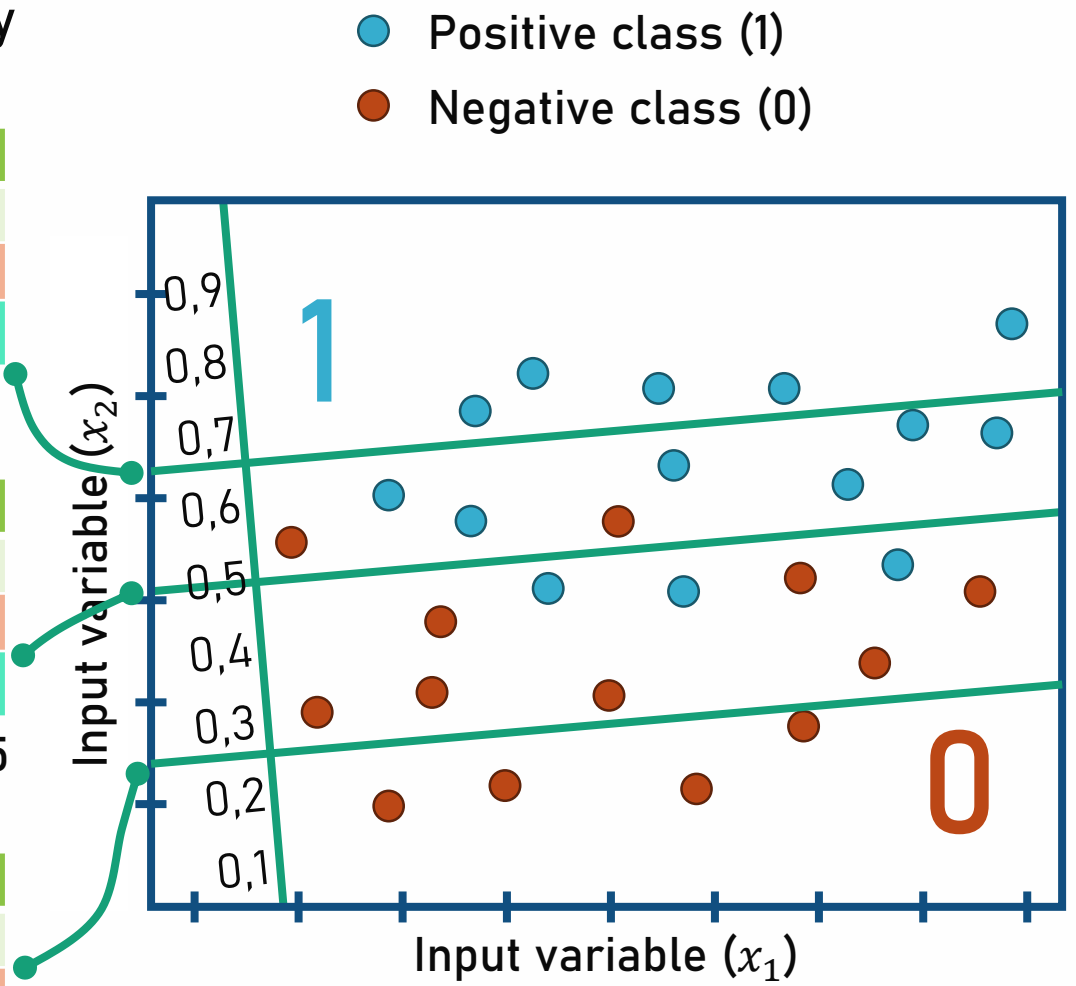
		Pred. (+)	Pred. (-)
Real	(+)	11	3
	(-)	2	11

TPR=0.79 TNR=0.85

Threshold = 0.25

		Pred. (+)	Pred. (-)
Real	(+)	14	0
	(-)	9	4

TPR=1.00 TNR=0.31

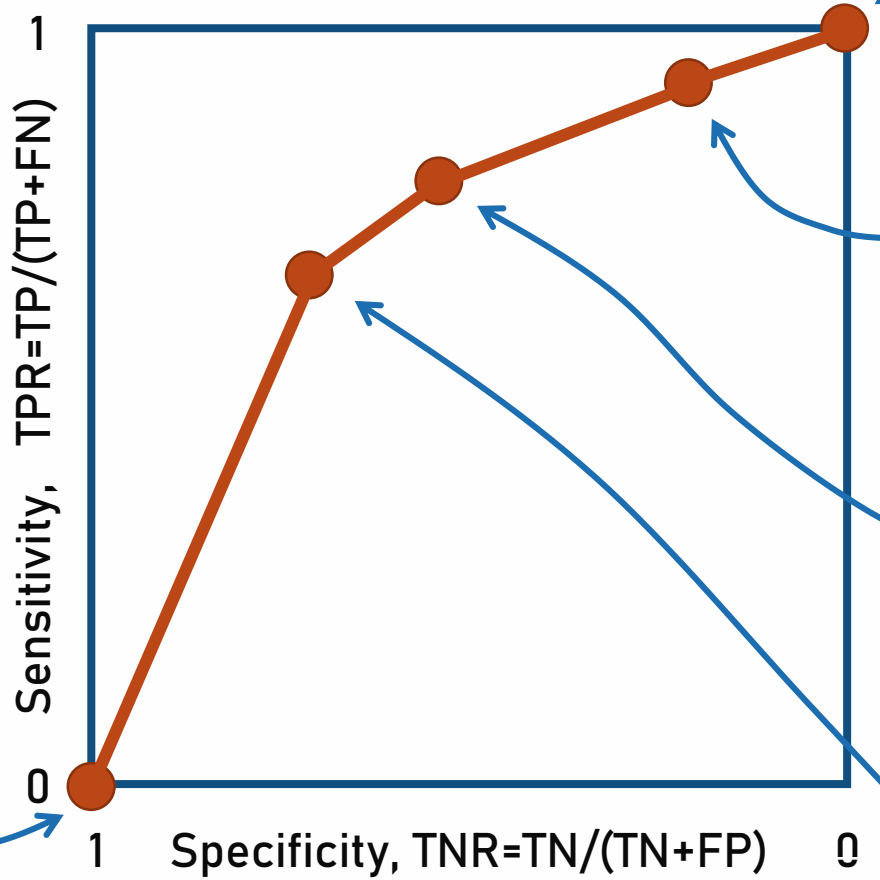


Plots for evaluating the classification model

ROC curve: an example

Analysis of the plot allows us to choose the appropriate classifier threshold.

When threshold=0.5, the classifier will correctly classify only 54% of cases from the negative class. If we want to correctly detect a larger number of positive cases, we need to reduce the threshold (but this will reduce the detection of negative cases).



Threshold = 0.00

		Pred. (+)	Pred. (-)
Real	(+)	30	0
	(-)	70	0

TPR=1.00
TNR=0.00

Threshold = 0.25

		Pred. (+)	Pred. (-)
Real	(+)	28	2
	(-)	55	15

TPR=0.93
TNR=0.21

Threshold = 0.50

		Pred. (+)	Pred. (-)
Real	(+)	24	6
	(-)	32	38

TPR=0.80
TNR=0.54

Threshold = 0.75

		Pred. (+)	Pred. (-)
Real	(+)	20	10
	(-)	20	50

TPR=0.67
TNR=0.71

Threshold = 1.00

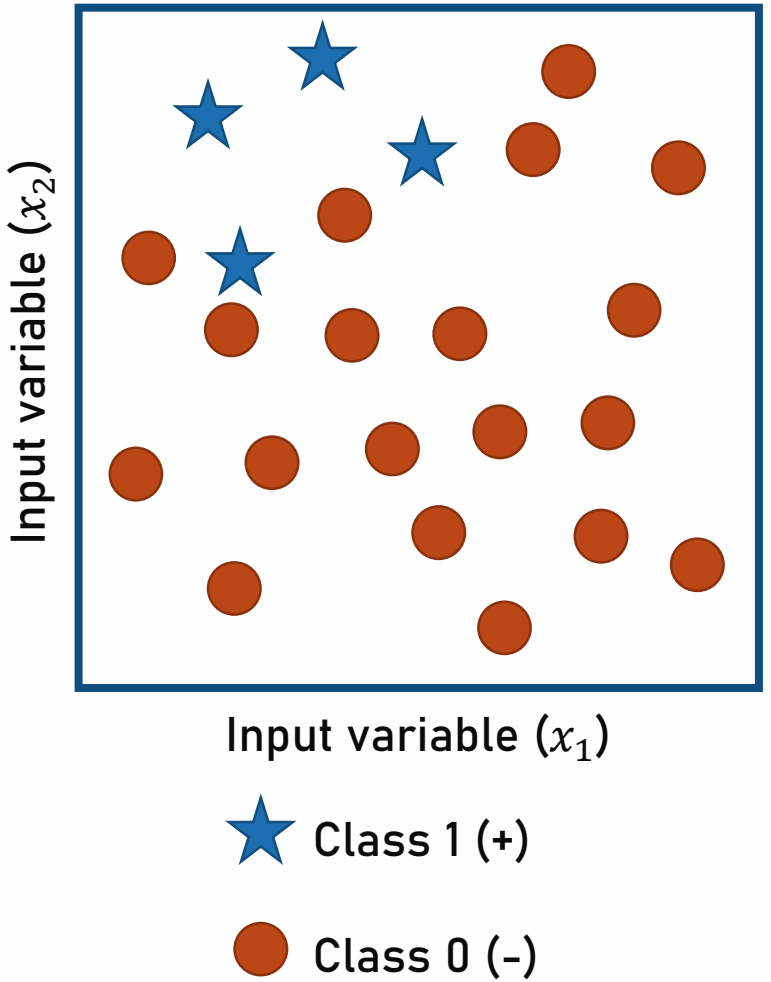
		Pred. (+)	Pred. (-)
Real	(+)	0	30
	(-)	0	70

TPR=0.00
TNR=1.00

Plots for evaluating the classification model

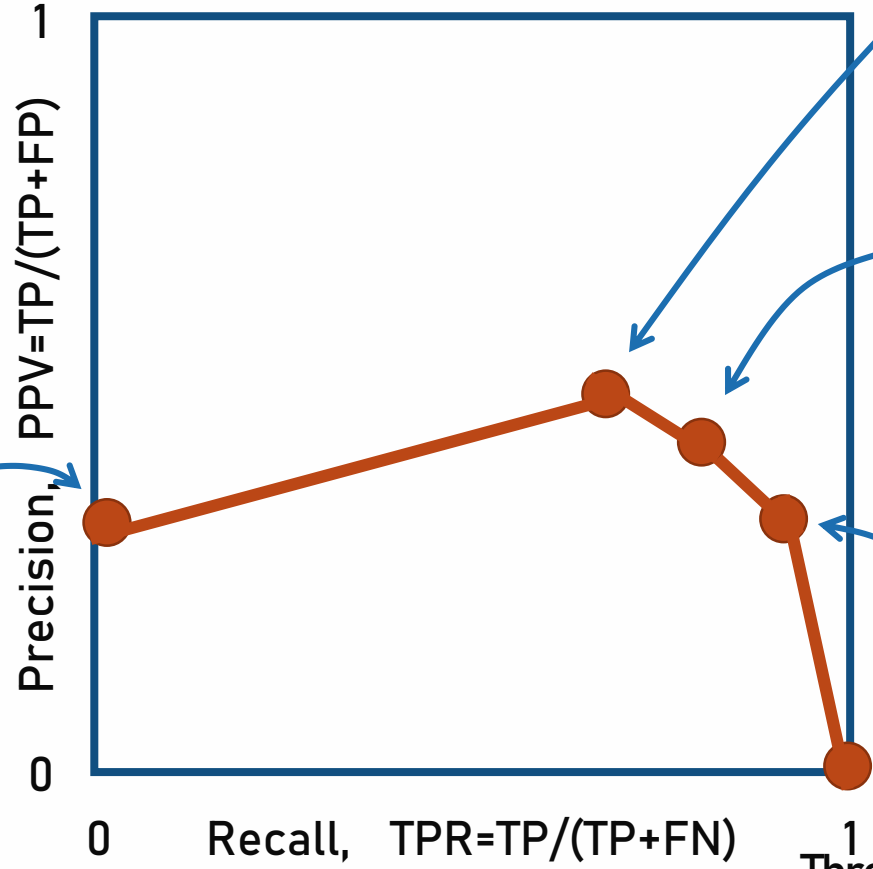
The problem of an unbalanced data set

- An unbalanced data set is one where the class sizes are significantly different.
- Classifiers tend to recognize objects belonging to the more numerous class better.
- In the case of an unbalanced data set, the ROC curve and AUC may not reflect the true performance of the classifier.
- In such a case, the precision-recall (PR) curve is used.
- The PR curve and the area under the PR curve are especially useful when the minority class is more important than the majority class and the classifier should recognize the minority class particularly well.



Plots for evaluating the classification model

PR curve: an example



Threshold = 1.00

		Pred. (+)	Pred. (−)
Real	(+)	0	30
	(−)	0	70

TPR=0.00
PPV= -

Threshold = 0.95

		Pred. (+)	Pred. (−)
Real	(+)	1	29
	(−)	2	68

TPR=0.03
PPV=0.33

Threshold = 0.75

		Pred. (+)	Pred. (−)
Real	(+)	20	10
	(−)	20	50

TPR=0.67
PPV=0.50

Threshold = 0.50

		Pred. (+)	Pred. (−)
Real	(+)	24	6
	(−)	32	38

TPR=0.80
PPV=0.43

Threshold = 0.25

		Pred. (+)	Pred. (−)
Real	(+)	28	2
	(−)	55	15

TPR=0.93
PPV=0.34

Threshold = 0.00

		Pred. (+)	Pred. (−)
Real	(+)	30	0
	(−)	70	0

TPR=1.00
PPV=0.00

Thank you for your attention!