Reference card

Accuracy

Model accuracy metrics

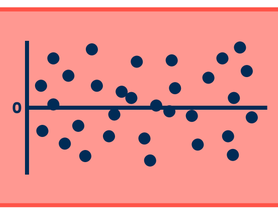
Model accuracy refers to how well a model is able to **accurately predict values**.

Residuals

A residual is the **difference between the observed value and the predicted value** for a data point.

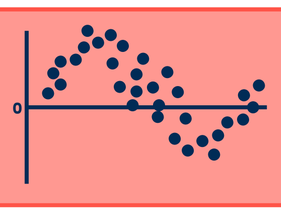
What we want:

Residuals are randomly scattered around zero, with no discernible pattern.



What we don't want:

Clear patterns or trends in the residuals, suggesting bias or shortcomings in the model.



Mean absolute error

The **average** of the **absolute values of the residuals**.

$$MAE = \sum_{i=1}^{n} \frac{|y_i - x_i|}{n}$$

y_i = predicted value
x_i = true value
n = number of data points

Mean squared error

The **average** of the **squared residuals**.

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

y_i = predicted value
x_i = true value
n = number of data points

Root mean squared error

The **square root** of the **average** of the **squared residuals**, i.e. the **square root of MSE**.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - x_i)^2}{n}}$$

y_i = predicted value
x_i = true value
n = number of data points

Model accuracy challenges

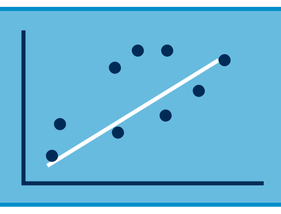
The **challenges** encountered in achieving **precise** and **reliable predictions**.

Bias

The **error** introduced by the **model's assumptions or simplifications**, causing it to consistently miss the mark when making predictions.

A model with **high bias oversimplifies the underlying relationships** in the data and **fails to capture the true patterns**.

High bias leads to **underfitting**.



Variance

The model's **sensitivity to fluctuations in the training data**. It measures **how much the predictions of the model vary** when trained on different subsets of the data.

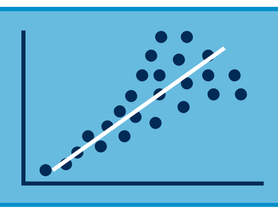
A model with **high variance overfits the training data** by memorising the training data **instead of learning the underlying patterns**.

High variance leads to **overfitting**.

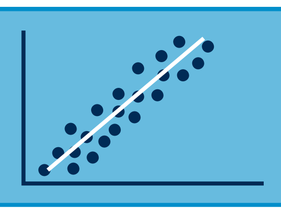
Heteroskedasticity

The **variability of the error terms are not constant** across the range of independent variables. Instead, we want **homoskedasticity** which means the variability of the error terms are constant.

Heteroskedasticity is present if our residuals have a fan or cone shape.



Homoskedasticity is present if the spread of the residuals are relatively constant.



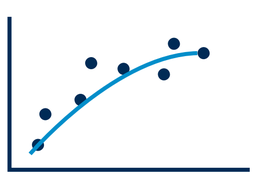
Poly-n trend line

A poly-n trend line is an extension of a linear equation that **includes polynomial terms to increase flexibility and better fit the data**.

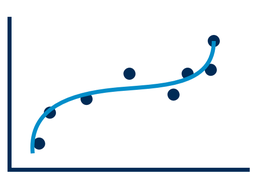
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_n x^n$$

x = independent variables
n = the powers
β = the coefficients

n = 2



n = 3



n = 4