DATA SCIENCE
and
BUSINESS INTELLIGENT

Author: Egi Safitri

Meeting 18

Learning Objective

Peserta mempelajari cara mengevaluasi Hasil Pemodelan dari:

A. Regression

B. Classification

E. Clustering

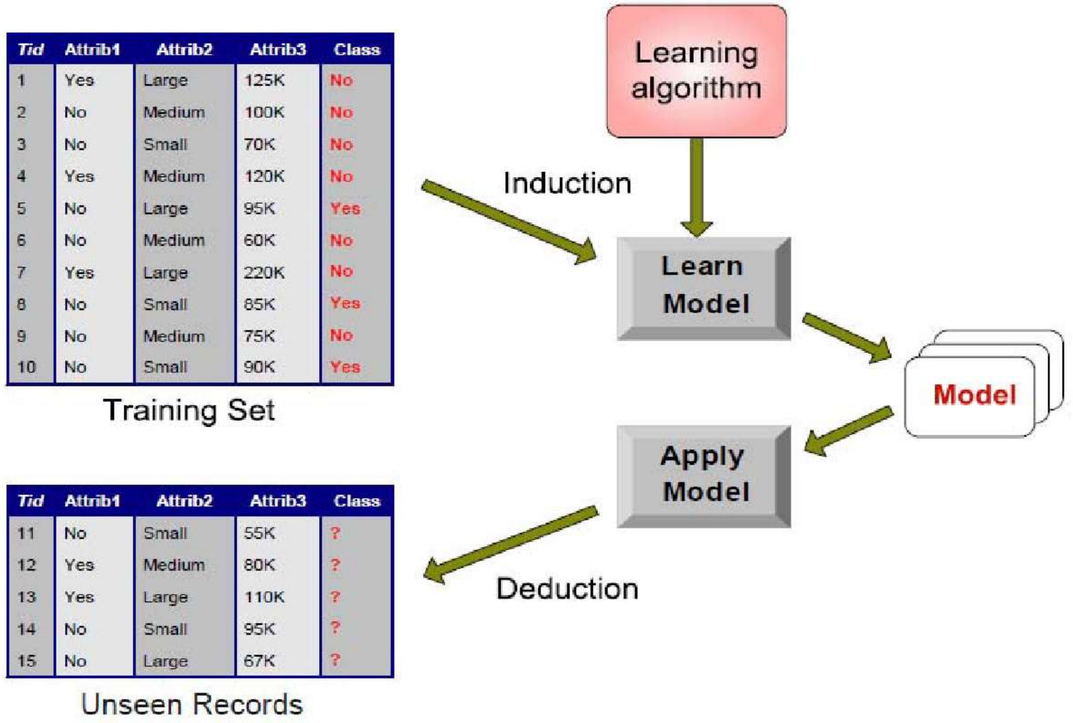
Menggunakan Rapid Miner.

Mengevaluasi Hasil Pemodelan

Model Evaluation

Central Question:

How good is a model at predicting the dependent variable?



9.1 Methods for Model Evaluation

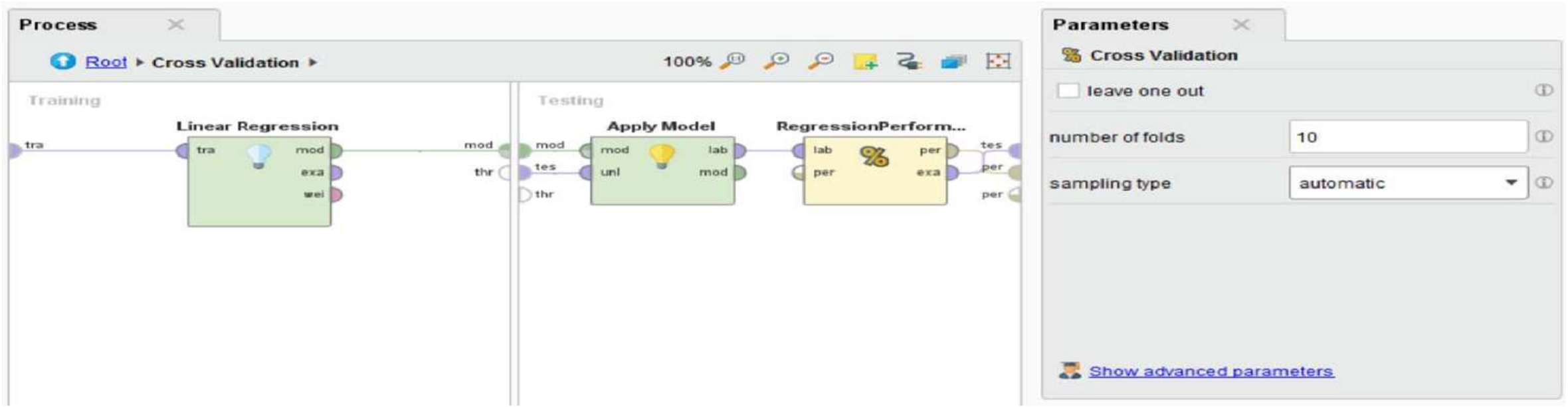
- How to obtain reliable estimates?

9.2 Metrics for Model Evaluation

- How to measure the performance of a regression model?

Methods for Model Evaluation

- The same considerations apply as for classification
- X-Validation: 10-fold (90% for training, 10% for testing in each iteration)
- Split Validation: 80% random share for training, 20% for testing
- Estimating performance metrics in RapidMiner
- X-Validation Operator + Regression Performance Operator



Metrics for Model Evaluation

Recap: Which metrics did we use for classification?

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\text{Correct predictions}}{\text{All predictions}}$$

		PREDICTED CLASS	
		Class= Yes	Class= No
ACTUAL CLASS	Class= Yes	TP 25	FN 4
	Class= No	FP 6	TN 15

$$\text{Acc} = \frac{25 + 15}{25 + 15 + 6 + 4} = 0.80$$

Metrics for Regression Model Evaluation

- Mean Absolute Error (*MAE*) computes the average deviation between predicted value p_i and the actual value r_i

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i|$$

- Mean Square Error (*MSE*) is similar to *MAE*, but places more emphasis on larger deviation

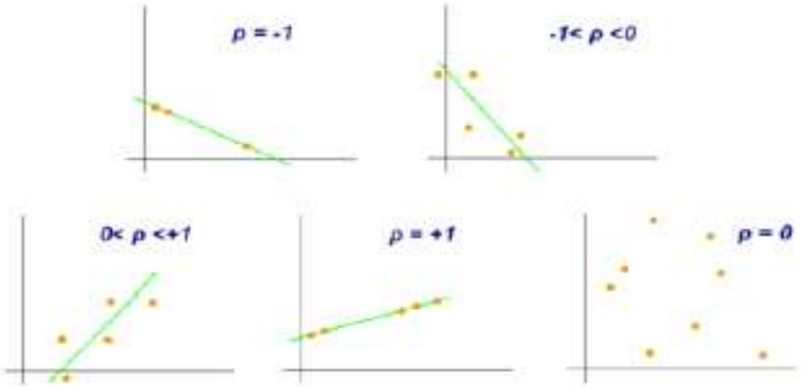
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2$$

- Root Mean Square Error (*RMSE*) has similar scale as *MAE* and places more emphasis on larger deviation

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

Metrics for Regression Model Evaluation

- Pearson's Correlation Coefficient (PCC)
  - Scores well if
    - high actual values get high predictions
    - low actual values get low predictions



$$\mathrm{PCC} = \frac{\sum_{\text{all examples}} (pred - \overline{pred}) \times (act - \overline{act})}{\sqrt{\sum_{\text{all examples}} (pred - \overline{pred})^2} \times \sqrt{\sum_{\text{all examples}} (act - \overline{act})^2}}$$

- R Squared: Coefficient of Determination
  - measures the part of the variation in the dependent variable y that is predictable from the explanatory variables X.

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

  - $R^2 = 1$: Perfect model as y can be completely explained from X.
  - called Squared Correlation in RapidMiner

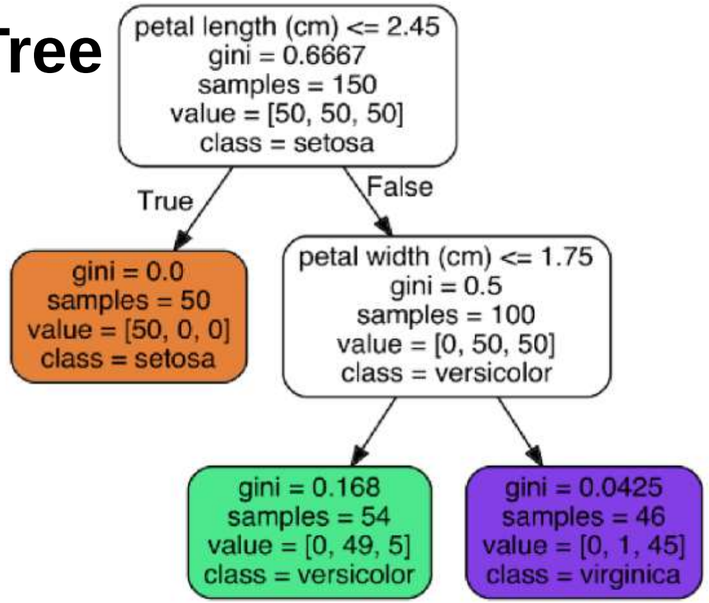
Regression Trees

 The basic idea of how to learn and apply decision trees can also be used for regression.

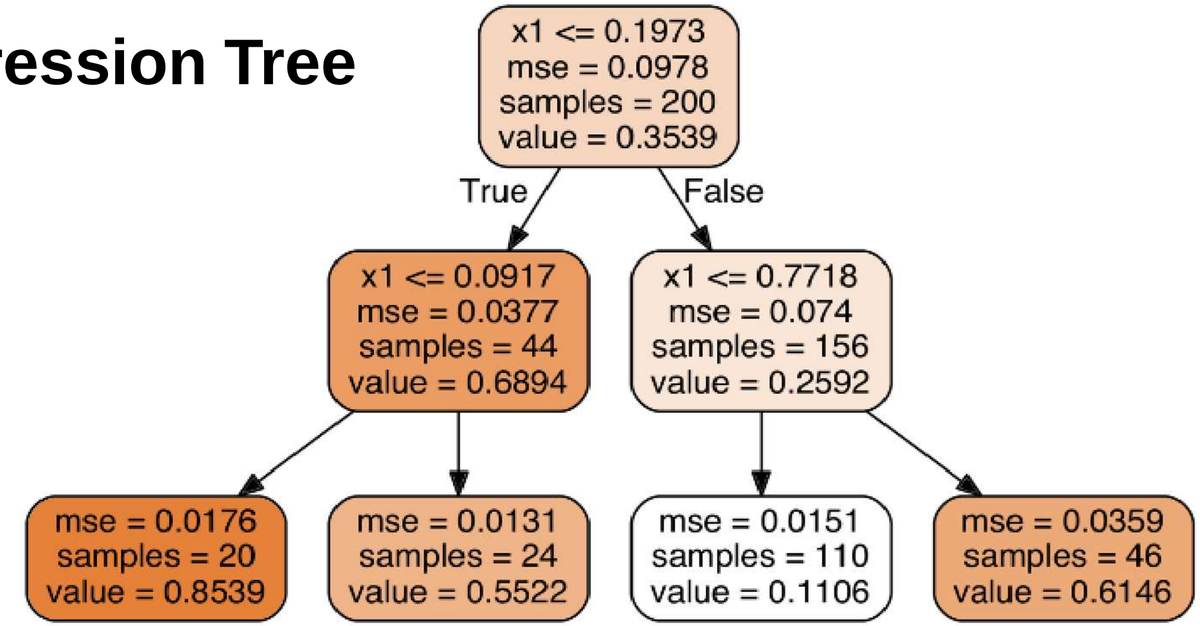
 Differences:

1. Splits are selected by maximizing the MSE reduction (not GINI or InfoGain)
2. Predict the average value of the trainings examples in a specific leaf.

Decision Tree



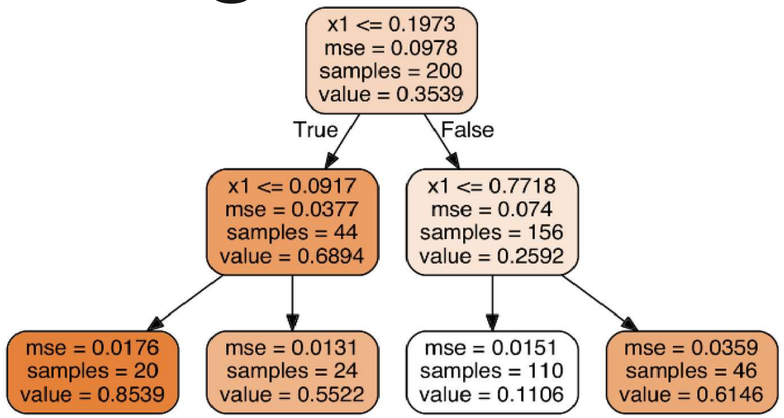
Regression Tree

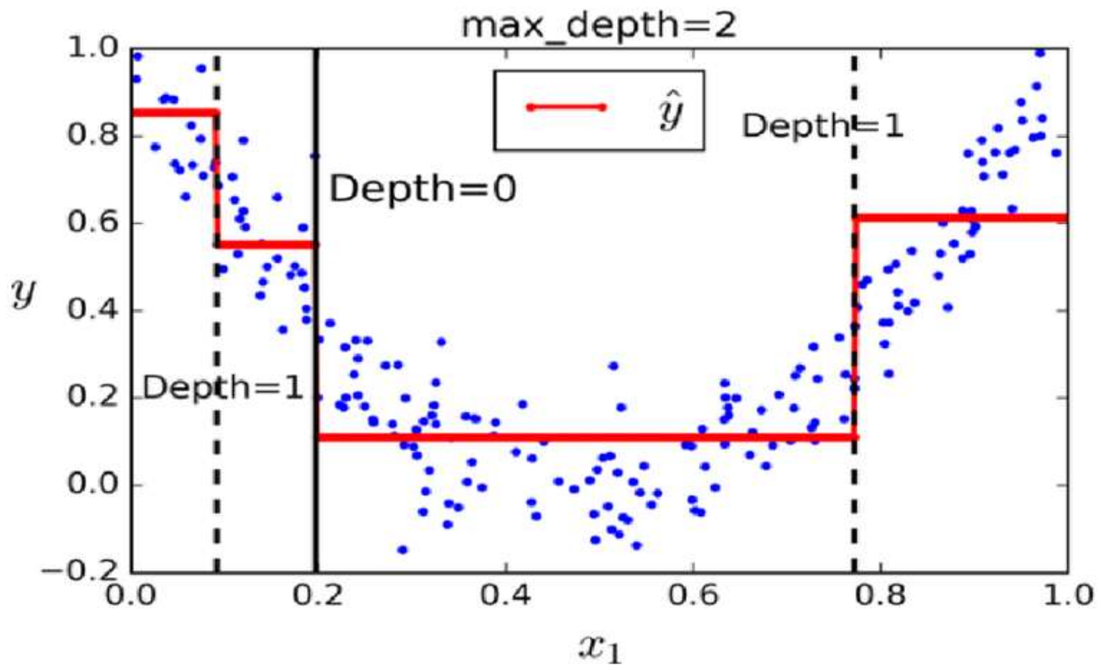


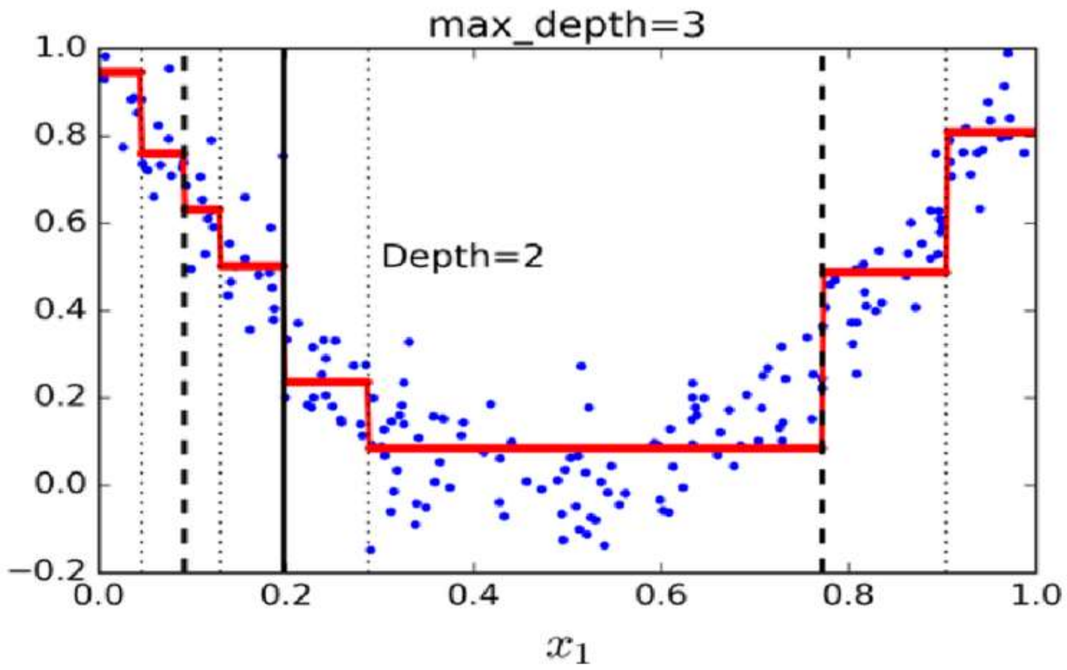
Regression Trees Fitting the Training Data

Pre-pruning parameters deterime how closely the tree fits the training data.

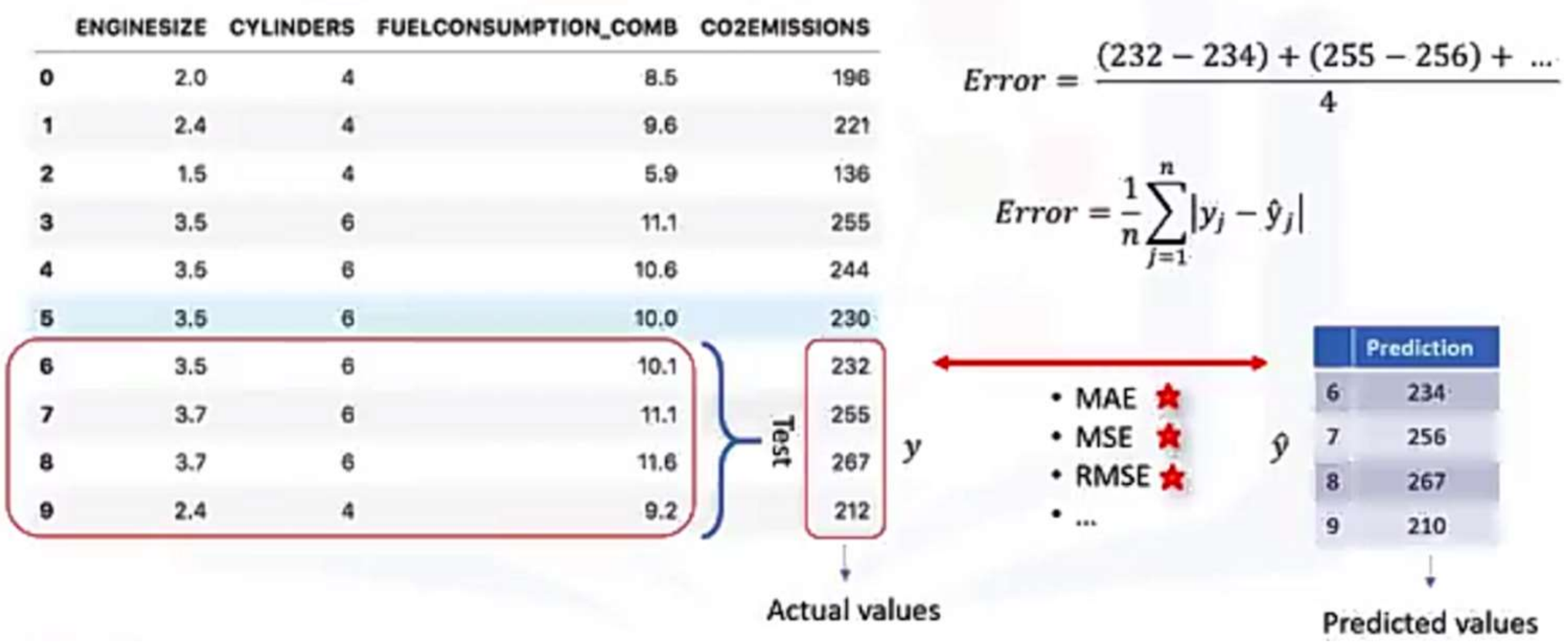
- e.g. max_depth parameter



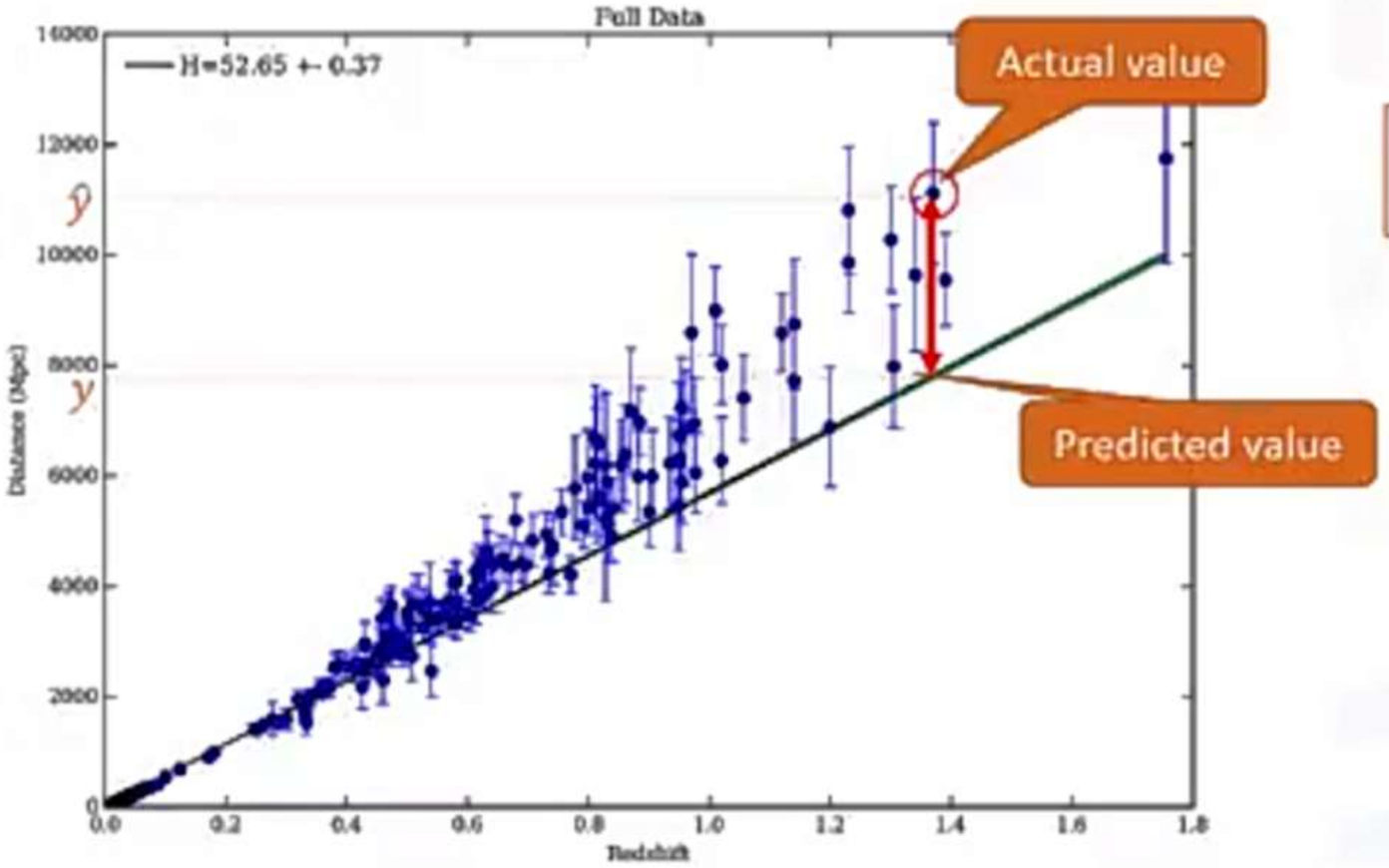




Error

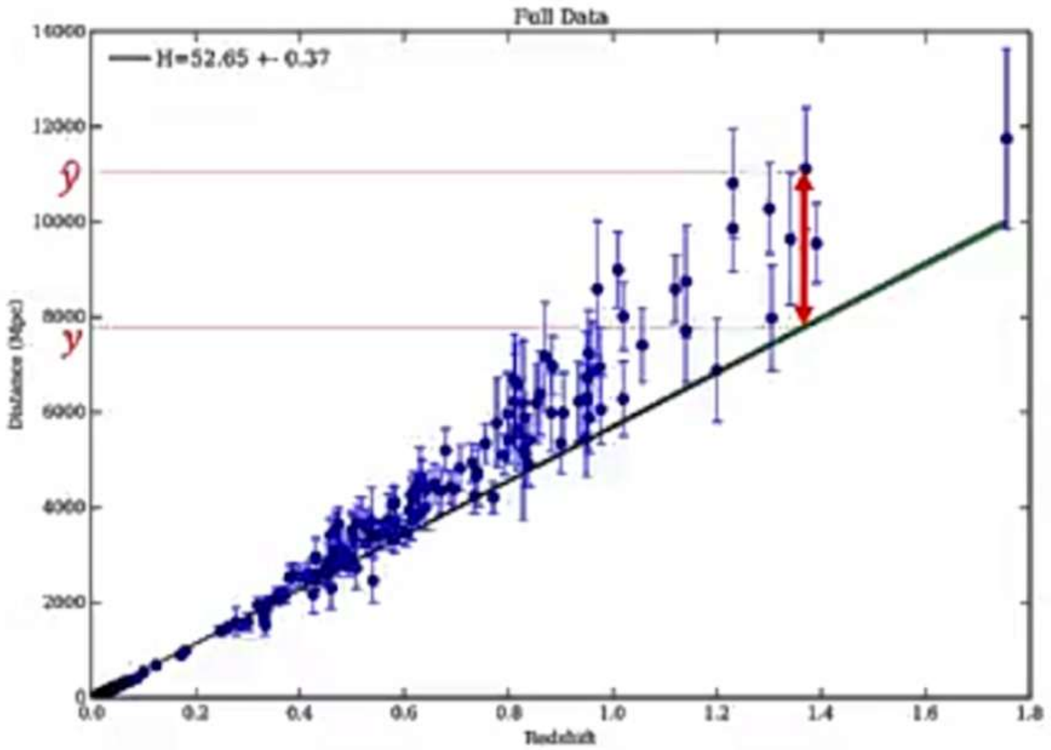

ENGINESIZE CYLINDERS FUELCONSUMPTION_COMB CO2EMISSIONS
0 2.0 4 8.5 196
1 2.4 4 9.6 221
2 1.5 4 5.9 136
3 3.5 6 11.1 255
4 3.5 6 10.6 244
5 3.5 6 10.0 230
6 3.5 6 10.1 232
7 3.7 6 11.1 255
8 3.7 6 11.6 267
9 2.4 4 9.2 212
Test
y
Actual values
$Error = \frac{(232 - 234) + (255 - 256) + \ldots}{4}$
$Error = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$
• MAE
• MSE
• RMSE
• ...
$\hat{y}$
Prediction
6 234
7 256
8 267
9 210
Predicted values

Error dari Model



Error: measure of how far the data is from the fitted regression line.

Error dari Model



$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

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

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

$$R^2 = 1 - RSE$$

Perbandingan

Regresi Linier (LR) vs DTR

- DTR mendukung non linearitas, di mana RL hanya mendukung solusi linier.
- Ketika ada sejumlah besar fitur dengan lebih sedikit kumpulan data (dengan noise rendah), regresi linier dapat mengungguli DTR/Random Forest Regression (RFR). Dalam kasus umum, DTR akan memiliki akurasi rata-rata yang lebih baik.
- Untuk variabel bebas kategorikal, DTR lebih baik daripada regresi linier.
- DTR menangani kolinearitas lebih baik daripada LR.

RL vs SVR

- SVR mendukung solusi linier dan non-linier menggunakan trik kernel.
- SVR menangani outlier lebih baik daripada RL.
- Keduanya berkinerja baik ketika data pelatihan lebih sedikit, dan ada banyak fitur.

DTR vs RFR

- RFR adalah kumpulan DT, suara (vote) mayoritas atau rata-rata dari forest dipilih sebagai keluaran yang diprediksi.
- RFR akan kurang rentan terhadap overfitting daripada DTR, dan memberikan solusi yang lebih general.
- RFR lebih robust dan akurat daripada pohon keputusan.


THANK
YOU