Machine Learning - Assignment 2

Evgeny Manturov

October 10, 2023

1 Task 1

a) The Gini index is allegedly computed for a single attribute, which is either of Customer ID, Gender, Car type, Shirt Size, etc. It cannot be computed for the overall collection of training examples, since it is unclear what type of mathematical operation is supposed to be used to combine all attributes into 1 index value.

b) For any CustomerID:

$$P(C_0) = \frac{0\ or\ 1}{1} = 0\ or\ 1$$
$$P(C_1) = \frac{1\ or\ 0}{1} = 1\ or\ 0$$
$$GINI = 1 - (1+0) or 1 - (0+1) = 0$$

c) For Gender:

$$P(C_0) = 6/10$$
$$P(C_1) = 4/10$$
$$GINI = 1 - (\frac{6}{10})^2 - (\frac{4}{10})^2 = 0.48$$

d) For Car type:
For Sports:

$$P(C_0) = 8/8$$
$$P(C_1) = 0/8$$
$$GINI(Sports) = 1 - (\frac{8}{8})^2 - (\frac{0}{8})^2 = 0$$

For Family:

$$P(C_0) = 1/4$$
$$P(C_1) = 3/4$$
$$GINI(Family) = 1 - (\frac{1}{4})^2 - (\frac{3}{4})^2 = 0.375$$

For Luxury:

$$P(C_0) = 1/8$$
$$P(C_1) = 7/8$$
$$GINI(Luxury) = 1 - (\frac{1}{8})^2 - (\frac{7}{8})^2 = 0.219$$

Weighed sum:

$$\frac{8}{20} \times 0 + \frac{4}{20} \times 0.375 + \frac{8}{20} \times 0.219 = 0.1626$$

e) For Shirt type:
For Small:

$$P(C_0) = 3/5$$
$$P(C_1) = 2/5$$
$$GINI(Small) = 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = 0.48$$

For Medium:

$$P(C_0) = 3/7$$
$$P(C_1) = 4/7$$
$$GINI(Small) = 1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = 0.5$$

For Large:

$$P(C_0) = 2/4$$
$$P(C_1) = 2/4$$
$$GINI(Small) = 1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5$$

For XL:

$$P(C_0) = 2/4$$
$$P(C_1) = 2/4$$
$$GINI(Small) = 1 - (\frac{1}{8})^2 - (\frac{7}{8})^2 = 0.5$$

Weighed Sum: GINI ≈ 0.5

f) Lower GINI = better, so Car type is the best here.

g) Customer ID is nominal - it does not describe real-world item properties. The model created with Customer ID as initial division value will be 100% overfitting-effected from the start - it will be 100% correct about the items from the test dataframe which are (precisely via Customer ID) in train dataframe, and only 50% (effectively random guess) effective about items with new Customer IDs.

2 Task 2

Python file will be included in the archive.

3 Task 3

The downloadable "wine" dataset has no column labels. RapidMiner's AutoModel functionality came in handy. RapidMiner has created an entire model lifecycle automatically upon choosing the column used as labels. Since the task implies using only decision tree, it has been used and the results (including accuracy) displayed by the RapinMiner GUI:

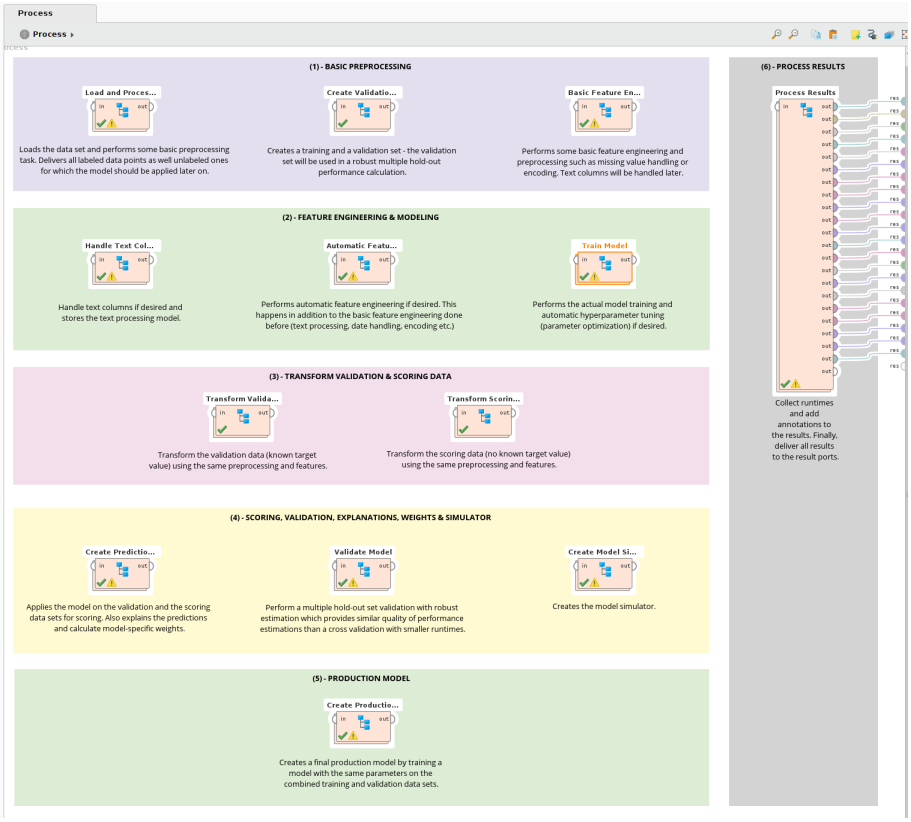

Figure 1: Process description by RapidMiner

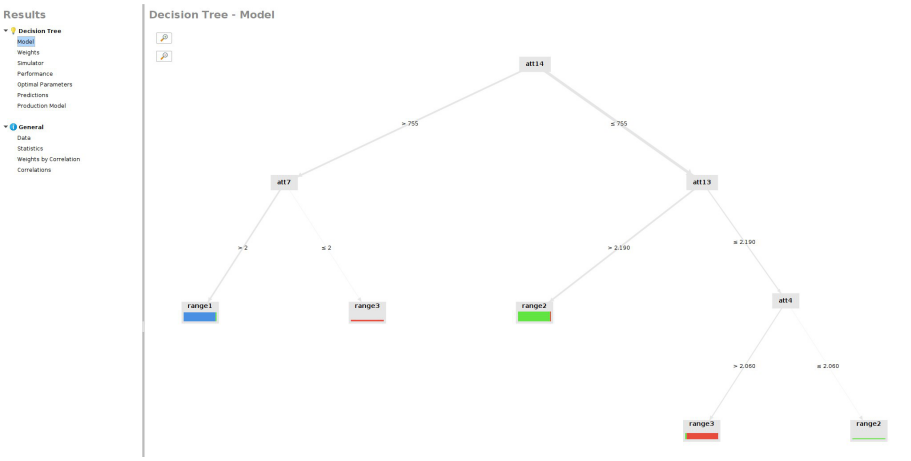

Figure 2: Tree drawn by RapidMiner

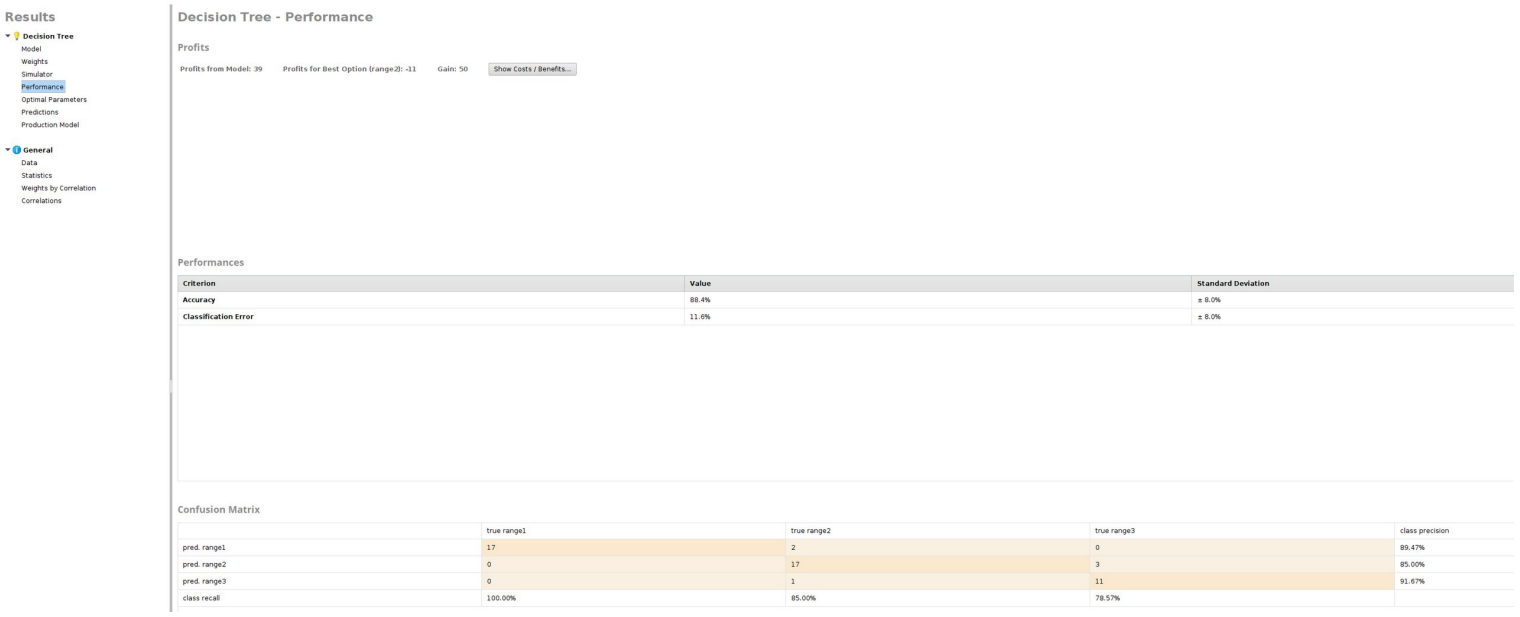

Figure 3: Accuracy by RapidMiner