Lab 02: Decision Tree with scikit-learn

# Decision Tree Classifier Evaluation

## Interpretation

### Interpreting the classification report

The classification report is of the form:

precision recall f1-score support

draw ***Pd Rd Fd Sd***

loss ***Pl Rl Fl Sl***

win ***Pw Rw Fw Sw***

accuracy ***Accu S***

macro avg ***Mp Mr Mf S***

weighted avg ***Wp Wr Wf S***

In which there are 22 calculated values as follows:

* ***Px***

The precision of the category ***x***, with ***x*** .

The value dictates the likelihood that the model is correct when predicting an observation as .

Is calculated by:

In which:

The precision value of category .

Number of true positives of category , that is, the number of observations that are of category and is correctly predicted as .

Number of false positives of category , that is, the number of observations that are NOT of category and is predicted as .

* ***Rx***

The precision of the category ***x***, with ***x*** .

The value dictates the sensitivity of the model to the category .

Is calculated by:

In which:

The recall value of category .

Number of true positives of category , that is, the number of observations that are of category and is correctly predicted as .

Number of false negatives of category , that is, the number of observations that are of category and is incorrectly predicted as NOT .

* ***Fx***

The f1-score of the category ***x***, with ***x*** .

The value dictates how “good” is the model at predicting an observation as .

Is calculated by:

In which:

The f1-score value of category .

The precision value of category .

The recall value of category .

* ***Sx***

The number of observations with category ***x***, with ***x*** .

* ***S***

The total number of observations used to perform the test.

* ***Accu***

The accuracy of the model.

The value dictates the likelihood that the model makes a correct prediction.

Is calculated by:

In which:

The accuracy value of the model.

A category.

Number of true positives of category , that is, the number of observations that are of category and is correctly predicted as .

Number of true negatives of category , that is, the number of observations that are NOT of category and is correctly predicted as NOT .

Number of false positive of category , that is, the number of observations that are NOT of category and is incorrectly predicted as .

Number of false negatives of category , that is, the number of observations that are of category and is incorrectly predicted as NOT .

* ***My***

The macro average of the metric ***y***, with ***y*** .

Is calculated by:

In which:

The macro average of the metric .

The category .

The value of the metric on the category .

* ***Wy***

The macro average of the metric ***y***, with ***y*** .

Is calculated by:

In which:

The macro average of the metric .

The category .

The value of the metric on the category .

The weight of the category .

### Interpreting the confusion matrix:

A sample of a confusion matrix:

Chart, treemap chart

Description automatically generated

The confusion matrix is an matrix, with , whose value is the number of observations. The rows of the matrix represent observations that are defined as a certain category according to the dataset, that is, ground truth. The columns represent observations that are predicted by the model. As such, the value at , or , is the number of observations that are of the category but is predicted as .

Note that in the confusion matrix, the diagonal values, values at indexes , represents correct positives predictions, while the other values represent false predictions.

## Results

### The 40/60 Split

The classification report:

precision recall f1-score support

draw 0.25 0.26 0.26 3870

loss 0.63 0.63 0.63 9981

win 0.84 0.84 0.84 26684

accuracy 0.73 40535

macro avg 0.57 0.57 0.57 40535

weighted avg 0.73 0.73 0.73 40535

The confusion matrix:

Chart, treemap chart

Description automatically generated

### The 60/40 Split

The classification report:

precision recall f1-score support

draw 0.28 0.30 0.29 2580

loss 0.67 0.66 0.67 6654

win 0.85 0.85 0.85 17789

accuracy 0.75 27023

macro avg 0.60 0.60 0.60 27023

weighted avg 0.76 0.75 0.75 27023

The confusion matrix:

Chart, treemap chart

Description automatically generated

### The 80/20 Split

The classification report:

precision recall f1-score support

draw 0.26 0.28 0.27 1290

loss 0.67 0.66 0.67 3327

win 0.85 0.85 0.85 8895

accuracy 0.75 13512

macro avg 0.59 0.60 0.60 13512

weighted avg 0.75 0.75 0.75 13512

The confusion matrix:

Chart, treemap chart

Description automatically generated

### The 90/10 Split

The classification report:

precision recall f1-score support

draw 0.28 0.28 0.28 645

loss 0.68 0.69 0.69 1664

win 0.86 0.86 0.86 4447

accuracy 0.76 6756

macro avg 0.61 0.61 0.61 6756

weighted avg 0.76 0.76 0.76 6756

The confusion matrix:

Chart, treemap chart

Description automatically generated

## Comment:

The metrics (precision, recall, and f1-score) are better as the proportion of the training set increases. This is as expected as the model performance are better when there are more training data.

For all train/test splitting proportion, the metrics (precision, recall, and f1-score) of the category “win” are higher than that of “loss”, which are higher than that of “draw”. This is due to the inconsistency in the proportion of observations for three classes: 65.83% for “win”, 24.62% for “loss”, and 9.55% for “draw”. The lack of consistency in the number of observations per category in the original dataset leads to the model perform better on categories with higher data, and worse for others.

# Accuracy of Decision Trees with Different Depths

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **max\_depth** | None | 2 | 3 | 4 | 5 | 6 | 7 |
| **Accuracy** | 0.748 | 0.658 | 0.664 | 0.676 | 0.689 | 0.693 | 0.703 |

The accuracy increases by the order of the max depth [ 2, 3, 4, 5, 6, 7, None ]. This is expected as an increasement in max depth would allow a tree with more branching, leading to more nodes to be considered. In other words, trees with higher depth would have a higher number of attributes (more data) being utilized, leading to higher accuracy.

Note that the configuration “max\_depth=None” defines a tree without depth limitation.