# Chapter 9: Evaluating and validating a decision tree model

## Introduction

Contents for this lesson include an introduction and a focus on aspects associated with evaluating and

validating a decision tree model including set up using Workflow blocks.

Model evaluation using statistics and charts will be introduced followed by a discussion of model validation: Statistical validation, structural validation and business validation.

This will be followed by a demonstration and then a summary.

Figure 1: Contents

A screenshot of a computer

Description automatically generated

Taking a structured approach to model assessment untangles the complexities associated with evaluating and validating a decision tree model.

Figure 2: Introduction

A screenshot of a computer

Description automatically generated

The first step is model evaluation. For a decision tree, ensuring only significant predictors are included,

the tree depth is adequate, node sizes are acceptable, and the tree is explainable and simple are complemented by statistics and charts to further determine model capability.

Once this is complete, the model should be validated. Model Validation applies the model to both

Partitions, generates statistics and charts and compares the model performance across partitions.

Model validation shares a lot of ground with model evaluation in that the same statistics and charts are used and takes three approaches:

Statistical validation, which compares model statistics across partitions, structural validation compares the tree structure and nodes sizes across partitions and business validation focuses on charts.

The focus of model validation is to ensure that the model performs consistently across partitions.

Figure 3: Scoring data and the Analyse Models bock

A screenshot of a computer

Description automatically generated

Two blocks are needed when evaluating and validating models. The Score block and the Analyse Models block. These are both located in the scoring group.

The score block is used to score a dataset and requires two inputs, the data to score and the model to score it with.

A decision tree model outputs propensity scores for each dependent variable category and these are used when generating statistics and charts with the Analyse Models block.

Figure 4: Model evaluation

A screenshot of a data sheet

Description automatically generated

Model evaluation focuses on the scored development partition. When a dataset is scored with a decision tree, two new variables are added containing model propensities for each dependent variable category.

These are used to determine which category of the dependent variable a observation Is predicted to fall into. A classification cut-off of 0.5 is used such that if the propensity of *bad* is equal to or in excess of 0.5, then the observation is classified as bad, otherwise it Is classified as *good*.

As a result of this a classification matrix and statistics can be generated to assess given insight into model performance.

Figure 5: Statistics

A screenshot of a computer screen

Description automatically generated

The Analyse Models block generates 8 statistic: Accuracy, Sensitivity, Specificity, Precision, NPV, F1, the C-Statistic and K-S test.

Accuracy relates to overall model accuracy, this sums the correctly predicted bads and goods and expresses the result as a proportion of the overall dataset size, here approximately .84, this means that 84% of observations are correctly predicted.

This statistic can be misleading as one category may be better precited that the other leading so other statistics are provided to give a more holistic approach.

Sensitivity focuses on the target category accuracy, here *bad*, this is simply as assessment of what proportion of the bads have been predicted as *bad*, here, at just over .5, or 50%.

Specificity focused on the other category accuracy, here *good*, with a value of approx. 95%. The overall accuracy of .84 is weighted by 95% accuracy of the good category and 50% accuracy of the bad category.

Precision focuses on the target category predicted outcome. For example, 2224 observations are predicted as *bad*, of that 1668 are *bad*. This gives a figure of .74332. This statistic can be interpreted as when *bad* is predicted, is it right approx. 75% of the time, meaning that when *bad* is predicted, It's generally correct.

NPV, is the same statistic for the other category and at approx. .85 means that when a observation is predicted as *good*, the model Is correct 85% of the time.

The F1 statistic can be used in conjunction with the Accuracy value - the Accuracy value may be high due to one category accuracy weighting the results. The F1 statistic can be interpreted as an overall average accuracy and here is .6 in comparison to the overall accuracy of .84.

The C-Statistic, also known as AUC or area under the curve is an industry standard statistic that is used when assessing models. The threshold varies across industries, but in general a value in excess of .75 indicates a model fit to deploy.

The K-S Test statistic measures the models’ ability to separate observations into the dependent variable categories and the higher the value the better, again, an industry standard is for a value in or around .6.

Most of the statistics have a range from 0 - 1 with the exception of the C-Statistic and K-S Test value which vary from 0.5 - 1 and 0 - 100 respectively.

Across all statistics, and in general, higher values are more desirable.

## Demonstration

So lets get on to a demonstration.

This demonstration uses a model developed in a previous lesson. There are three variables in the model. *capital\_gain*, *relationship* and *education\_num*.

Figure 12: Tree model

A computer screen shot of a chart

Description automatically generated

All leaf nodes, the point at which a prediction takes places, are of adequate size. All being at least 4% of the dataset size and the widest split has 5 nodes. There is the possibility to grow some segments further but as to whether there will be better isolation of the *bad w*ith an acceptable distribution is questionable, for example, consider leaf node number 7.

The node size is 13% of the overall dataset size with 1838 observations, characteristics for membership of this node are missing *capital\_gain*, *relationship* = *married* and *eduction\_num* greater than 7 and less than or = 9.

There are 533 *bads*, accounting for 29% of the node. Selecting this node and choosing Optimal Split by Entropy Variance splits on the variable *occupation* with 5 splits. As can be seen, most observations, 7% are concentrated in one node.

The generated split looks acceptable but with a variable such as *occupation*, which may have lots of categories, category merging may not be common sensical.

Validation will highlight whether this split is acceptable or not but for the time being it is retained.

Figure 13: Adding more splits

A computer screen shot of a chart

Description automatically generated

There is a questionable segment, node number 3, as can be seen almost all observations in this node are *bad*, again, validation will lead to understanding whether it can be retained or must be merged with another node.

One final node to focus on is node number 5, this has 53% of observations and only 5% of these, 341 are in the *bad* category. It would be great to split this further but given the small % of *bad*, additional splitting may provide no further benefit. To investigate, the node is selected and Optimal Split by Entropy Variance clicked.

The variable selected is *education* with 8 splits. The splits are further refined such that three are created that reflect an order. The resulting splits are acceptable with adequate distributions.

Figure 14: Final tree

A computer screen shot of a chart

Description automatically generated

An attempt could be made to grow the tree further as some of the *education\_num* splits are adequately sized with adequate distributions but here model building will stop at this point.

Onto validation. Notice the asterisk on the Decision Tree, clicking CTRL+S saves tree settings and the asterisk disappears.

Statistical and business validation can be assessed using the Analyse Models block, this is found in the Scoring group. This block requires a scored dataset, so the first thing to do is score the development partition with the decision tree model using the Score block, also found in the Scoring group.

Figure 15: Adding a Score block

A screenshot of a computer

Description automatically generated

Notice the configuration status messages when the block is dragged onto the canvas: Block requires one dataset input and No input connection providing a model.

The development partition and model are both connected to the Score block and as a result of the block

being included in Auto-Run, a scored dataset is automatically output. Here, its name is changed to *dev\_scored*.

Opening the scored dataset with the Data Profiler, it can be seen that two new columns have been added *P\_bad* and *P\_good*, these variables contain the probability of *bad* and *good* for each scored observation*.*

Clicking the data tab and splitting the screen with the decision tree, it is easy to understand how scores

Were arrived at.

Figure 16: Understanding scores

A computer screen shot of a computer

Description automatically generated

For example, the values for the first observation were generated as a result of it being a member of node 23.

This node contains 190 observations, so there will be another 189 observations with the same values as they are also contained in the same leaf node.

Notice that the tree has 19 leaf nodes, therefore there are 19 distinct *P\_Bad* values and 19 distinct *P\_good* values.

Now that the development partition is scored, the Analyze models block can be added to the Workflow and the scored development partition connected. The block configuration dialog contains two pages: Analysis Type and Variable Selection.

The Type of analysis drop-down provides two options: Classification and Regression. As the dependent variable in this instance is categorical, classification is appropriately selected.

The variable selection page allows selection of the variables to analyse in the connected dataset. Presently there is only one dataset connected. From the True class dropdown, the dependent variable is selected, here *DV*.

From the Truth category drop-down. The category of interest is selected, here, *bad* and from the Predicted probability drop-down the corresponding variable that contains the probability scores for the selected category chosen, here *P\_bad.* Once complete, OK is clicked and the Analyse Models report accessed.

Figure 17: Model Analysis Report

A screenshot of a computer screen

Description automatically generated

## Summary

This lesson focused on using statistics and charts to evaluate and validate a decision tree model.

The model was first evaluated using statistics and charts. Model validation looked at applying the model to both partitions and assessing the consistency and stability across results and three approaches were presented:

Statistical validation which focused on comparing statistics across partitions, structural validation, which looked at comparing the tree structure across partitions and business validation focusing on charts.

This was then followed by a demonstration using Altair Analytics Workbench capabilities.