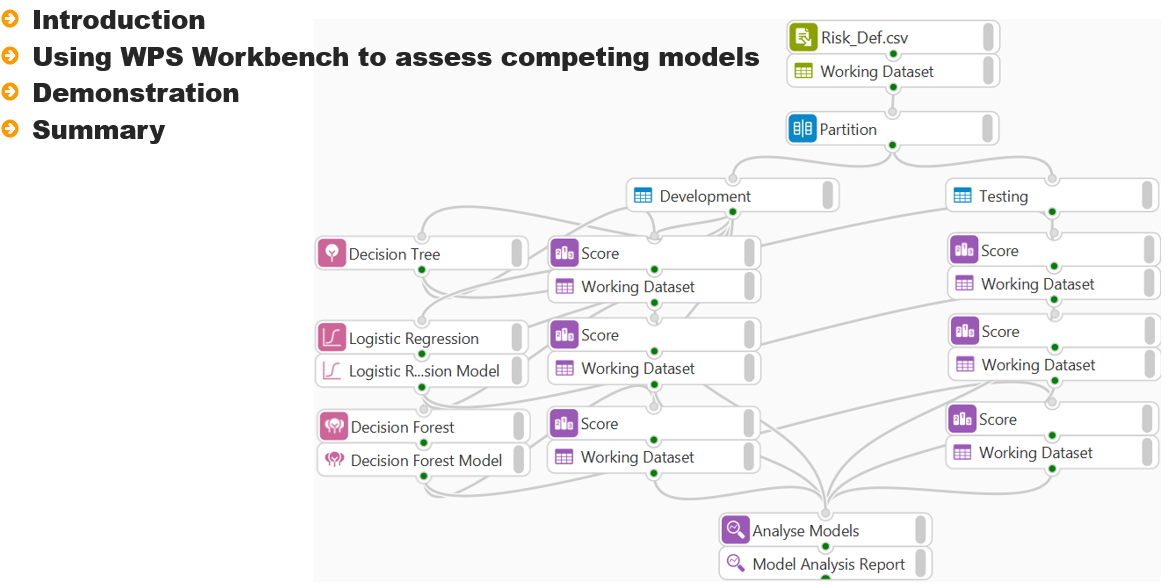
# Chapter 10: Challenging the model

## Introduction

This lesson includes an introduction and focuses on using Altair Analytics Workbench Workflow modelling

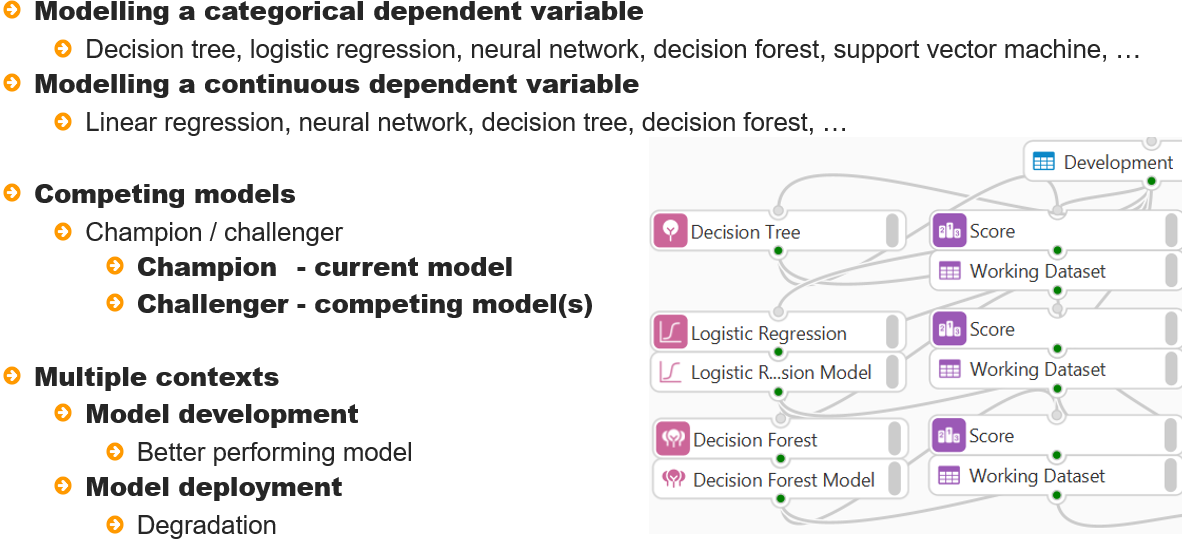
capabilities to challenge a model prior to demonstration followed by a summary.

Figure 1: Contents



There are many ways to model a dependent variable, whether it is continuous or categorical. Once a model has been developed, other models can be assessed to determine whether they outperform the current model and can takes its places as the model of choice.

Figure 2: Introduction



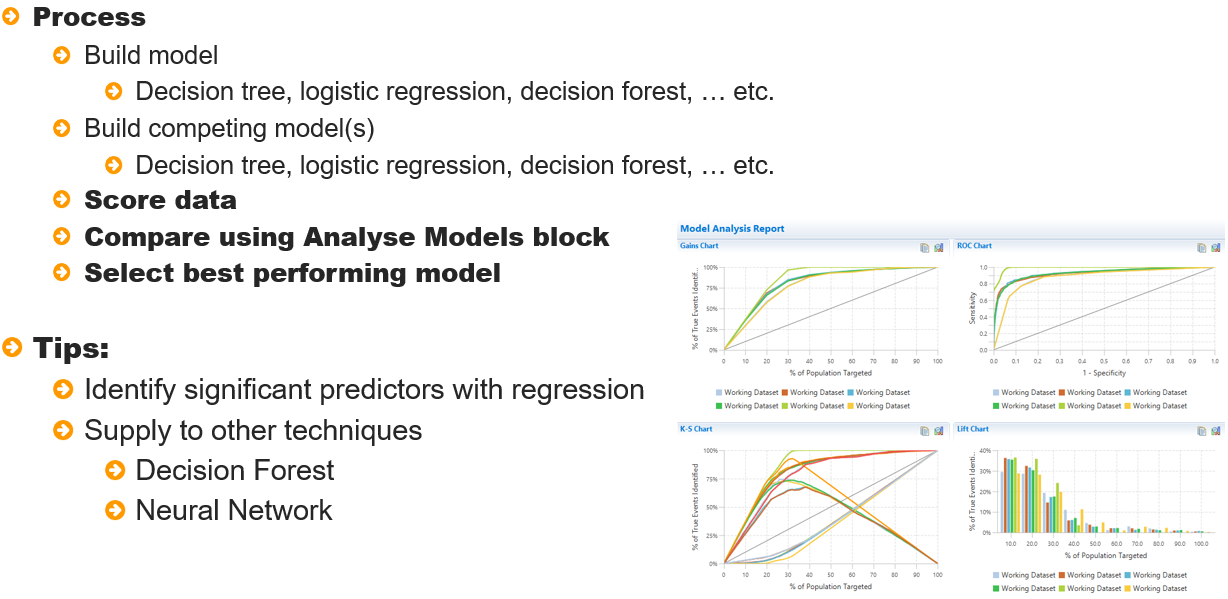
This process is referred to as champion/challenger. The champion is the currently accepted model and the challenger is the competing model or models. The context of where champion / challenger can be applied is twofold.

Firstly, model development - multiple models can be developed and the best performing model selected

Champion challenger can also happen when models have been deployed and are in production.

The current production model is the champion, at the point it degrades such that it no longer predicts well, a model rebuild may be necessary and one or more challenger models can be assessed.

Figure 3: Champion/challenger process



The process to challenge a model is straightforward and regardless of whether an established model

Is being challenged or multiple models to predict the same dependent variable are being assessed, the

Steps are identical.

Assuming there is interest in modelling a binary dependent - the first step is to build a model, once complete other models are built, data is scored, and results can be compared using the Analyse Models block. Simply select the best performing model.

Some tips in relation to developing other model types - as a decision forest and neural network may take considerable time to run, the process can be made more efficient by using regression to identify significant predictors to supply to these techniques.

The demonstration for this lesson will focus on a decision tree model previously developed and challenge it with two competing models, a logistic regression that uses a stepwise procedure to identify significant predictors and a decision forest, this will use the significant predictors identified with logistic regression to build a model.

## Demonstration

So onto a demonstration.

This demonstration uses a decision tree model developed in a previous lesson. The model contains five variables *capital\_gain*, *relationship*, *education,* *education\_num* and *occupation.*

This is the current champion model. Two additional models will be developed to compete with this model, a logistic regression and a decision forest. Logistic regression will be used to assess all variables and a stepwise method will be used to select and include only significant predictors.

Variables identified here will be used to model the same dependent variable using a decision forest and all models will be assessed using the Analyse Models block.

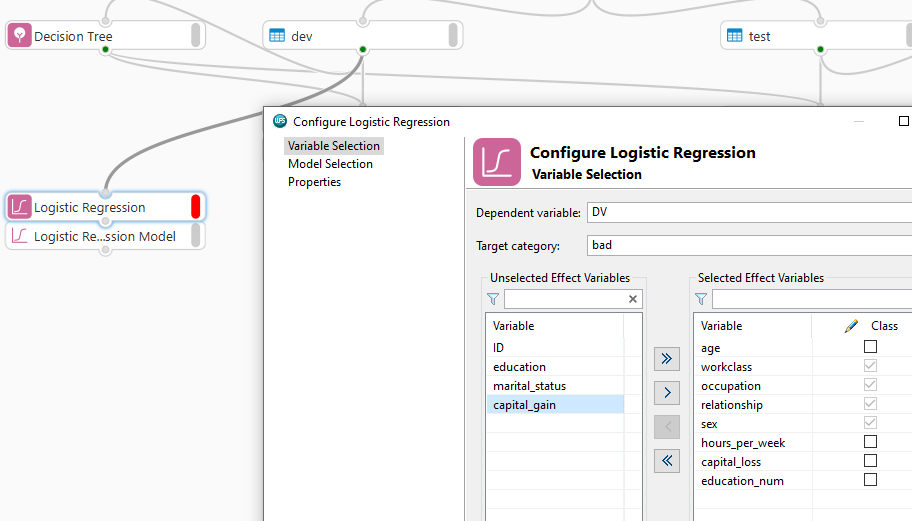
Viewing the data prior to proceeding shows that there are 13 variables. The variable *DV* will be used as the dependent variable and a selection of remaining variables used as independent variables.

As *capital\_gain* has a high proportion of missing values it will not be included as a possible predictor. Another variable that will not be included are *ID*. *education\_num* and *education* record education, as *education\_num* is a numeric continuous variable, it is preferred.

The variables *relationship* and *marital\_status* record similar characteristics and here *relationship* will be used.

To begin, a logistic regression block is dragged from the Model Training group to the Workflow canvas and the development partition connected.

Figure 4: Adding a Logistic Regression modelling block



The block has three pages, Variable selection, Model Selection and Properties. The Variable Selection page allows selection of the dependent and independent variables.

From the dependent variable dropdown *DV* is chosen and *bad* selected from the event dropdown. All variables are moved into the selected effect variables area and *capital\_gain*, *education*, *ID* and *marital\_status* are removed.

Note that some variables have been correctly identified as class variables, however *education\_num* has also been selected as a class variable, as it contains integer values, deselecting ensures it will be used as a continuous predictor.

There is interest in only including significant predictors in the model. An option to ensure all variables are assessed and only significant predictors retained is available from the model selection page.

The method drop down provides five options, the option none, will build a model with all independent variables supplied. The other four methods include and retain variables only if they meet specific significance thresholds, these are available and can be modified once an appropriate method has been selected.

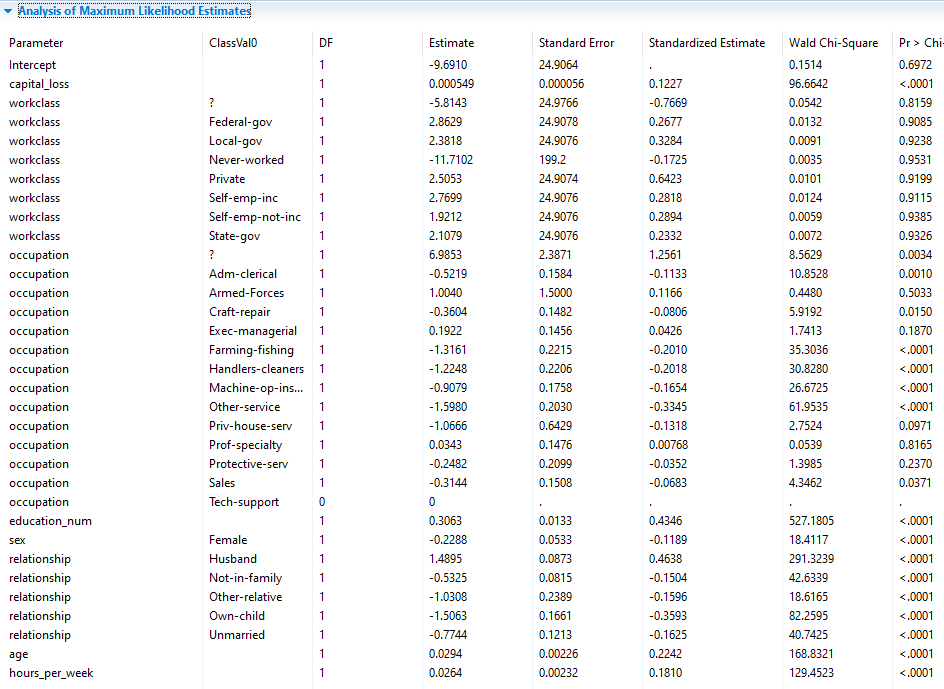
The option selected here is Stepwise, this allows only significant predictors into the model. Once a new variable is added to the model, variables already present are re-assessed to ensure their continued significance.

Clicking OK runs the model and once complete double-clicking the resulting Logistic Regression Model block reveals that output is presented across two tabs: Logistic Regression Results and Scoring Code.

The logistic regression results provide six expandable sections including Model Information, Model Fit Statistics, Testing Global Null Hypothesis: BETA = 0, Analysis of Maximum Likelihood Estimates, Odds Ratio Estimates and Association of Predicted Probabilities and Observed Responses.

A discussion of the specifics of these areas is not warranted here, however from either the Analysis of Maximum Likelihood Estimates or the Odds Ratio Estimates tables, the variables included in the model are visible.

Figure 5: Analysis of Maximum Likelihood Estimates



Eight variables were included: *capital\_loss*, *workclass*, *occupation*, *education\_num*, *sex*, *relationship*, *age* and *hours\_per\_week*.

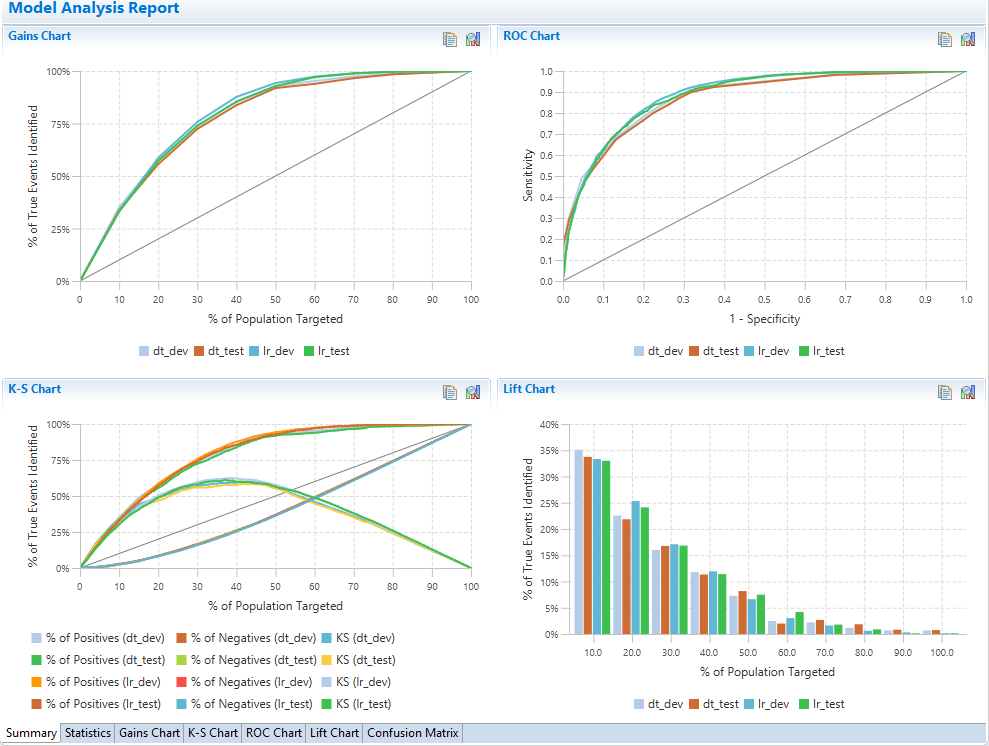
These variables are all significant predictors in the model and will be used when developing a decision forest. The next step is to score the development and testing partitions, here score blocks are added, connections made and data scored.

Prior to connecting to the Analyse Models block, the scored datasets are renamed to *lr\_dev* and *lr\_test* respectively so that they can be easily identified. The partitions scored with the decision tree are also renamed appropriately.

Once complete, the scored partitions are connected to the Analyse models block and configured. Here the same True class, Truth category and corresponding Predicted probability variable for both partitions are selected and OK clicked.

Opening the Model Analysis Report, it can be seen there are lines for both the decision tree and logistic regression applied to both partitions.

Figure 6: Summary

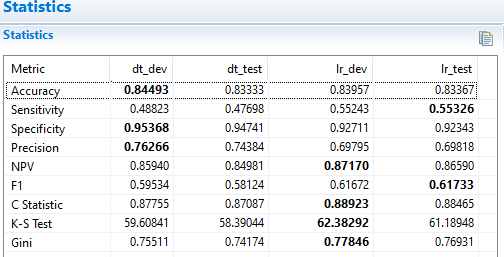


From the Summary its clear to see that the logistic regression is outperforming the decision tree. This is not surprising given that more variables were used to build the logistic regression model.

Notice that the curves are smooth for the logistic regression model in comparison to the decision tree, this is because an equation is used to generate scores for each observation based on the independent variable characteristics as opposed to grouping observations and assigning them the same score.

From the graphs the logistic regression is outperforming the decision tree. This is reflected in the statistics.

Figure 7: Statistics



Focusing on the development partition. The accuracy is slightly higher for the decision tree but the target category is not predicted as well as the logistic regression.

Also, the logistic regression gives a better average predictive accuracy as relayed by the F1 statistic. Other values such as the C-Statistic and K-S Test value also favour the logistic regression.

Both models perform consistently across partitions but at this point the logistic regression with eight variables would be the model of choice.

## Summary

This lesson introduced the concept of champion challenger, whereby competing models are assessed.

This was followed by a demonstration challenging a decision tree model with a logistic regression and a decision forest and the Analyse Models block was used to assess models and select a champion.