# Chapter 6: Modelling with Altair Analytics Workbench

## Introduction

This lesson consists of an introduction and progresses to a classification of modelling techniques before introducing those most commonly used and also discusses questions around which and when to use.

Modelling capabilities available from Altair Analytics Workbench will also be introduced prior to speaking about partitioning before a summary.

Figure 1: Contents

A white background with black text

Description automatically generated

Any modelling endeavour needs a focus or business problem to solve. For example, we need to attract more customers, or we're not offering the right products or we're overstocked for the season or any one of a number of issues experienced across and within industries and verticals.

Figure 2: Introduction

A screenshot of a computer

Description automatically generated

Once an issue has been identified it can be possibly be solved by modelling and there are many solution

available if that is the chosen route for example response modelling to increase customers, market basket analysis to offer the right products and forecasting for stock control, among others.

There are many modelling techniques and sometimes the wealth of capabilities can be overwhelming and regardless of the technique used all will predict something and output either a propensity, a category a group an offer a value or something of that nature.

Figure 3: Classifying techniques

A diagram of a company

Description automatically generated

In general modelling techniques can firstly be classified as supervised or unsupervised. Supervised techniques require a dependent variable, a target, whereas unsupervised techniques require no target variable.

Supervised techniques can be further classified in relation to whether they provide insight into the model workings such as the reasoning behind the predictions or the impact of each variable in the model and generate output or whether they simply generate an output.

Modelling techniques that provide some insight into how the model arrived at its decision or output include regressions, decision trees and scorecard. Modelling techniques outputting simply a result include neural networks and ensemble modelling techniques.

Unsupervised techniques can be further sub divided also as those that describe or group data and that find associations between elements in that data.

Modelling techniques that describe data include cluster analysis, segmentation and data reduction techniques such as principal component analysis and factor analysis. Those that can be classified as finding associations including Market basket analysis and association rules.

Figure 4: The dependent and independent variables

A screenshot of a computer

Description automatically generated

All supervised modelling techniques require an outcome, or dependent variable and predictors also called independent variables. For example, a logistic model will predict the propensity to default using demographic, financial and bureau details.

Unsupervised modelling techniques require inputs only, for example demographic details are used to identify groups of customers to promote products to in retail.

A special case of an unsupervised technique is market basket analysis and association rules where model inputs can act as predictors of an outcome or can be the outcome.

Regardless of the modelling technique a great deal of time is invested in data preparation and developing the dependent variable. This can be time consuming and requires consideration and should be related to a business objective with stakeholder approval

## Regression / Scorecards

Regression and scorecards are prime examples of supervised techniques that can both explain a model and classify data. They require both independent variables and a dependent variable. Linear regression is used to model a continuous dependent variable and logistic regression is used to model a categorical dependent variable.

These techniques can assess independent variables for inclusion and output an equation or a scorecard

providing the ability to quantify changes and understand the impact of each independent variable on the outcome.

 Figure 5: Regression/scorecards

A screenshot of a computer

Description automatically generated

Regression outputs an estimate of impact of each independent variable on the dependent variable as an

Equation and by supplying input values, what if scenarios can be evaluated. Scorecards output scores for characteristics, these can be summed, and if a threshold is reached a decision can be made. Given this it is easy to see how scores are arrived at and how they can be changed.

## Decision tree

Decision trees also fall under the heading of being a supervised technique that can explain and classify data. Decision trees can be used to model a continuous or categorical dependent variable and independent variables can be of any type.

Figure 6: Decision tree

A screenshot of a computer

Description automatically generated

They output a propensity score, as logistic regression does, and so can be and is used in similar circumstances or in tandem. Decision Trees are popular as they are easy to use and interpret and have applications beyond modelling such as exploring a dataset and complementing other techniques

## Neural networks

Neural networks required independent variables and a dependent variable, which can be either

continuous or categorical, but they do not provide any insight into model predictions and only generate an outcome. For this reason neural networks are commonly referred to as a *black box* technique. They may outperform other modelling techniques but will not provide any reasoning behind its predictions

Figure 7: Neural network

A diagram of a network

Description automatically generated

## Ensemble methods

Ensemble methods can also be considered as *black box* as a result of combining results from lots of models. The most commonly used ensemble methods derive from decision tree methodologies and

include boosting, bagging and random forests in some form or other.

Figure 8: Ensemble methods

A screenshot of a computer

Description automatically generated

These methods build lots of models, usually in the 100s or 1000s. The prediction from each model is used to calculate the final outcome and this entanglement means that predictions, and therefore the model workings cannot be well understood.

## Clustering / PCA

Unsupervised methods such as cluster analysis and principal component analysis require inputs only and

describe data by either grouping observations or variables into identifiable segments or factors. Their use is widespread across industries and verticals to create marketable groups and segments.

Figure 9: Clustering/PCA

A screenshot of a computer

Description automatically generated

## Association rules

Another unsupervised technique that is commonly used is Market Basket Analysis and association rules.

These techniques look for associations in data whether between products, services, or progress through

pages on a website.

Figure 10: Association rules

A close-up of a diagram

Description automatically generated

Market basket analysis originated from the analysis of the contents of shoppers’ trollies and outputs

rules with antecedents and consequents; the left hand and the right hand of the rule.

For example, those purchasing bread and milk also purchase veg. Inputs can act as outputs So another finding may be that those purchasing veg and frozen foods also purchased milk. The output is easy to understand as it is in the form of rules but may can be generated.

## Technology and data

 New methods are constantly being presented to address and deal with a progress environment and the impact of technology and the explosion of data cannot be ignored.

Figure 11: Technology and data

A screenshot of a computer

Description automatically generated

Statistical methods have been joined by machine learning methods designed for large data that focus on outcomes whereas traditional statistical methods may struggle with large volumes of data but do provide a level of insight absent to some degree from machine learning methods.

Which technique to use and when?

The technique to use and the method to apply are all dependent on goals.

Whether there is a need to be very accurate or whether there is a need to understand and explain results may impact the technique chosen but bear in mind that techniques do not have to be used in isolation and one technique may be used to compliment another.

For example, a decision tree might be used to illustrate logistic regression odds ratio or to give some clarify behind the predictions of a neural network

Figure 12: Which technique to use and when

A close-up of a question mark

Description automatically generated

Altair Analytics Workbench modelling blocks

Altair Analytics Workbench provides an array of modelling techniques available from the Workflow via the

Model training group blocks or via code. Modelling blocks are drag drop with point and click

functionality and automatic scoring code generation.

Figure 13: Altair Analytics Workbench modelling blocks

A screenshot of a computer

Description automatically generated

Other modelling functionality can be accessed through the SAS, SQL, Python and R code blocks.

Prior to modelling

Prior to any modelling the data should be partitioned. This is a mechanism to ensure that models

are accurate and can be used in production. A dataset is split, generally into two partitions a development and a training partition, but more can be created if desired.

Figure 14: Prior to modelling

A screenshot of a computer

Description automatically generated

The testing partition is isolated and a model is built using the development partition. The model is then

applied to the testing partition to ensure consistency prior to production.

## Summary

This lesson discussed modelling techniques and presented a simple classification to easily distinguish

varying approaches.

Common modelling techniques and their characteristics were illustrated as well as introducing questions

relating to which technique to use when.

Altair Analytics Workbench Modelling capabilities were presented as well as discussing the need to partition

Data prior to modelling.