*Chapter Six notes* ***for Agri-food research***

Summary from page 187

*Summary Algorithms built from data can be used for classification and prediction in technological applications.*

Prediction is in contrast to causal/explanatory modelling.

Algorithms are more akin to black boxes – they make predictions but it is difficult to explain how the work for any particular prediction.

Classification is important in general, but on the fringes of Argifood research. For example, you might want to try and classify sub-species of manuka trees from different areas of Tasmania. You can collect samples from high/low altitude and moderate/cold conditions and propagate them all in Sandy Bay. They look the same, but you do a chemical analysis on them to obtain the levels of key essential oils. Can you lump the different plants into a few different groups with “similar” essential oils. Or are they all the same. This is an opportunity to use a classification algorithm: using the essential oil data you try to reduce its dimensions and “score” each plant. If the scores of the plants are “clustered” into you groups you look at the groups – do these correspond to where the samples come from? Do the colder mountain manuka plants appear to be fundamentally different (in terms of overall types of oil present?

**artificial intelligence (AI**): computer programs intended to perform a task normally associated with human abilities. Spiegelhalter, David. The Art of Statistics (Pelican Books) (p. 381).

**big data:** an increasingly anachronistic phrase sometimes characterized by four Vs: a huge Volume of data, a Variety of sources such as images, social media accounts or transactions, a high Velocity of acquisition, and possible lack of Veracity due to its routine collection. Spiegelhalter, David. The Art of Statistics (Pelican Books) (p. 382).

*It is important to guard against* ***over-fitting*** *an algorithm to training data, essentially fitting to noise rather than signal.*

**over-fitting:** building a statistical model that is over-adapted to training data, so that its predictive ability starts to decline. Spiegelhalter, David. The Art of Statistics (Pelican Books) (p. 396).

*Algorithms can be evaluated by the classification accuracy, their ability to discriminate between groups, and their overall predictive accuracy.*

Assessing predictions:

1. Past data. **In-sample prediction assessment**. This is the same data that you used to create the model – so beware overfitting.
2. Wait for more data. **Out-of-sample prediction assessment.**
3. Hold back some data from the modelling process – then try to predict itand rank models based on this subset of predictions. This is **pseudo-out-of-sample prediction error assessment**. Otherwise known as **cross-validation**.

Why is a model that predicts out-of-sample data well perhaps more attractive (to use in the future) than one that predicts in-sample?

Less risk of significant **over-fitting**.

***Complex algorithms may lack transparency, and it may be worth trading off some accuracy for comprehension****.*

Maybe the devil-you-know (devil might be simplifying assumptions that are not true) is more robust, trustworthy and useful for interventions.

*The use of algorithms and artificial intelligence presents many challenges, and* ***insights into both the power and limitations of machine-learning methods is vital****.*

Some of the main problems include:

1. Lack of robustness.
2. Lack of generalisability. What works here, doesn’t work over there. In agriculture: works in one season but the next, for example.
3. Bias can be imbedded. Where does the data on recidivism come from – those that have already convicted (so if there is a bias in the past data set, an algorithm will predict future criminality *using that bias*).
4. Lack of transparency. Explaining to others how a model works is important, even if only to convince others to use it/

This book [Spiegelhalter] emphasizes the classic statistical problems of small samples, systematic bias (in the statistical sense) and lack of generalizability to new situations. The list of challenges for algorithms shows that although having masses of data may reduce the concern about sample size, the other problems tend to get worse, and we are faced with the additional problem of explaining the reasoning of an algorithm.

Having bucketloads of data only increases the challenges in producing robust and responsible conclusions. A basic *humility* when building algorithms is crucial.