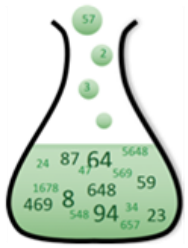
Department of Health

Random Forest Analysis: The prediction of fragile A&E systems

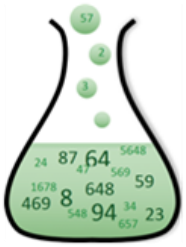
Edward Beake



Data Science Hub

Overview

- Demystifying Machine Learning
  - Introduction to machine learning
  - Bagged Ensemble, bootstrapping in machine learning.
  - From Decision Trees to Random Forest.
- Fragile A&E systems
  - Winter pressures in A&E units and fragile systems
- Worked Example
  - Using random forest in a Department of Health problem.

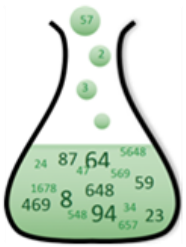


Data Science Hub

Machine Learning

Machine Learning

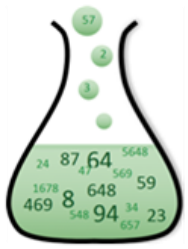
- Machine learning techniques make predictions by recognising patterns and trends in data.
- The performance of ML at a task increases with experience of the task.
- The rules in a ML model are not hard coded, they are learned from data.



Data Science Hub

Machine Learning

- The two main classes of ML techniques are supervised or unsupervised:
  - ➢ Supervised ML learns to make predictions from labelled data (data that has the 'answer' – includes the target variable).
  - ➢ Unsupervised ML learns to cluster unlabelled data.



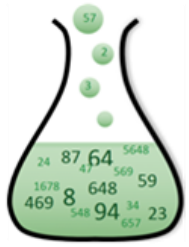
Data Science Hub

Machine Learning – Basic Terminology

- ML techniques build models based on input data, this is split into **training** and **test** sets.
  - training – to build the model
  - test – to evaluate the model
- Input data consists of **instances** (individuals in a population – the rows in a table) and **features** (information on those individuals – the columns of a table).

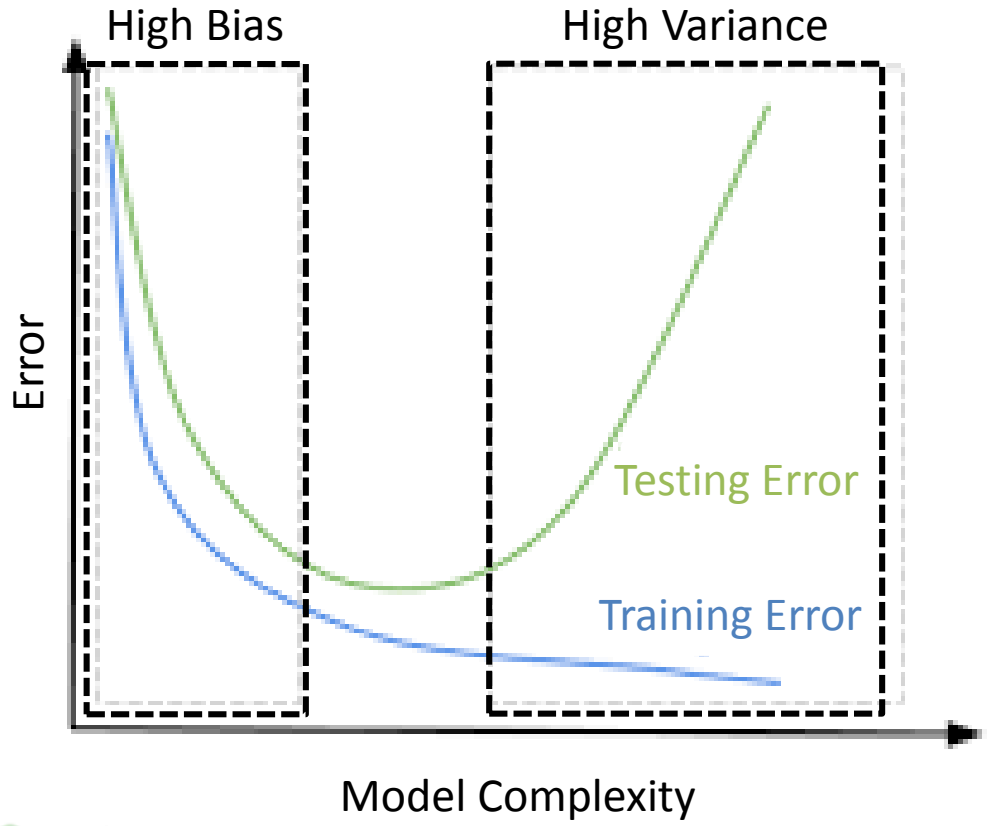
features | Target variable

	Name	Height	Weight	Gender
instances – training	Ed	5’8”	70	M
	Sarah	5’7”	65	F
	Kate	5’9”	68	F
	John	5’11”	80	M
	Adam	5’10”	75	M
	Chris	5’10”	73	M
	Sophie	5’7”	56	F
instances – test	Becky	5’6”	60	F
	Will	6’0”	90	M
	Jane	5’10”	65	F

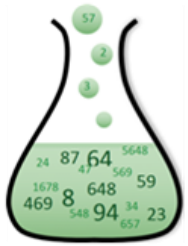


Data Science Hub

Bias/Variance Trade Of

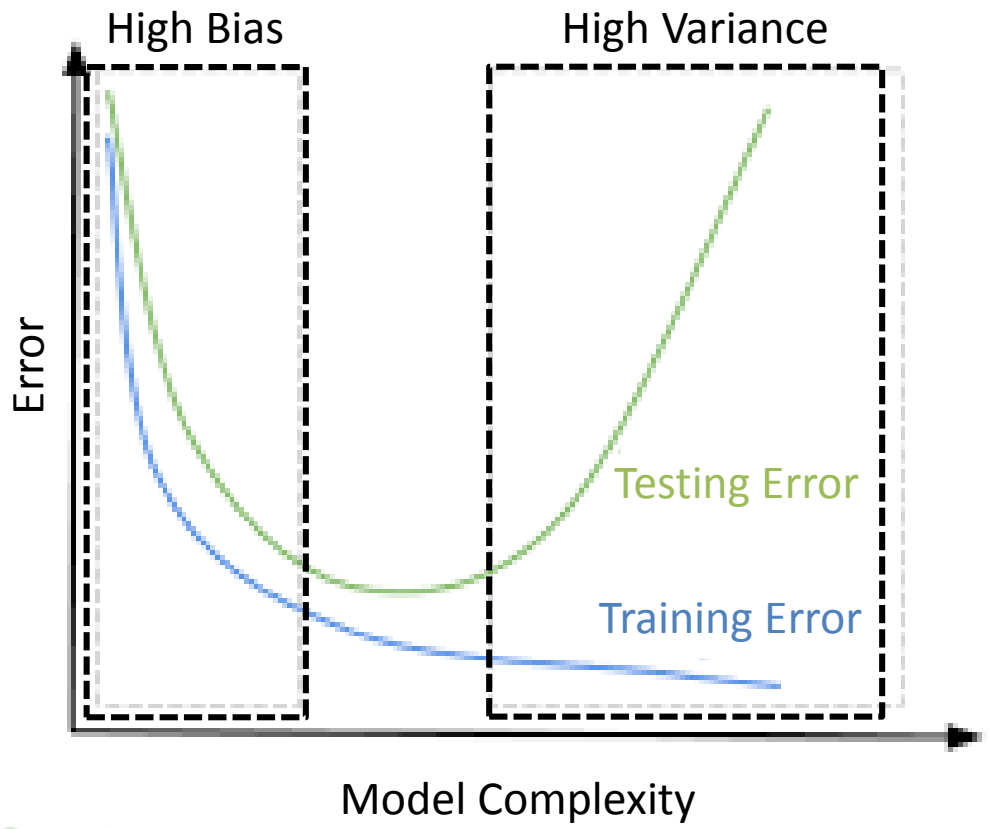


- A **low complexity** (eg. Linear) model is said to have a **high bias**.
- A **high complexity** (eg. quadratic) model is said to have a **high variance**.



Data Science Hub

Bias/Variance Trade Of

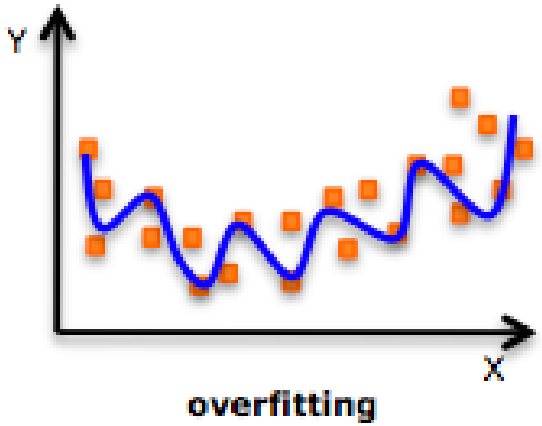


- A model with a high test error will **not** make good predictions on unseen data – it will **generalise poorly**.
- The correct bias/variance trade off can not be known *a priori*, model selection is a process of **trial and error**.

Data Science Hub

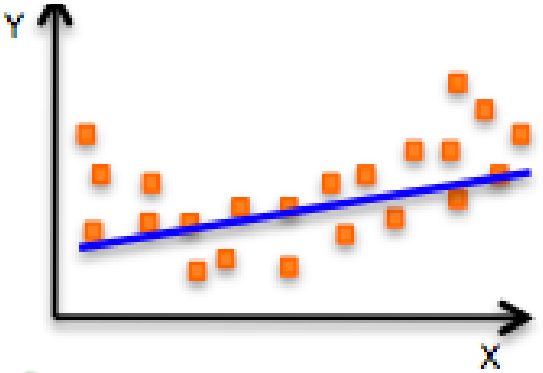
Underfitting/Overfitting

A Machine Learning model can overfit or underfit the data.



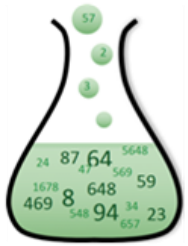
overfitting

- Overfitting – the model fits the training data ‘too well’ - the model fits the noise in the data, not the underlying trend.



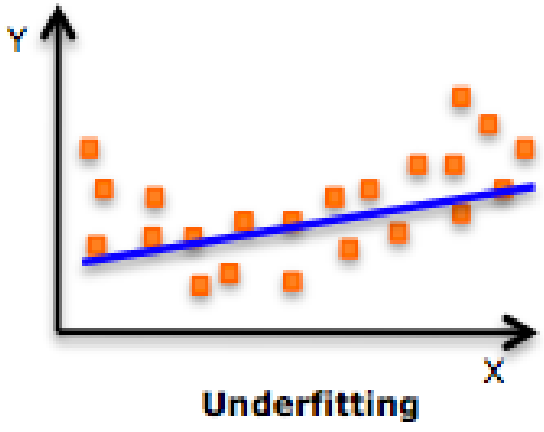
Underfitting

- Underfitting – the model does **not** fit the training data ‘well enough’ – the model does not fit the underlying trend.



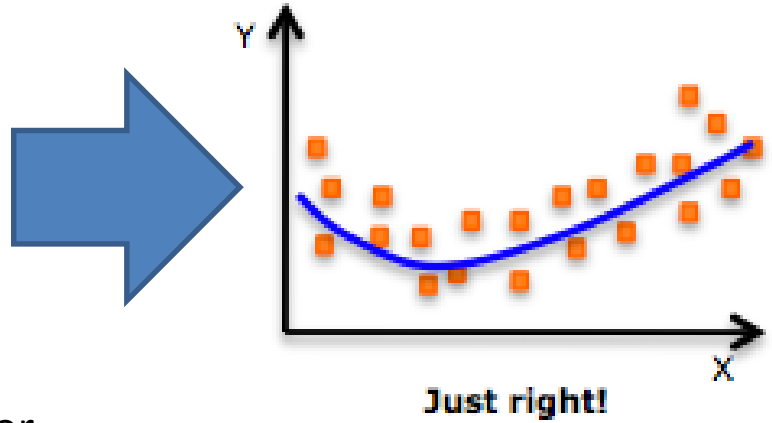
Data Science Hub

Underfitting/Overfitting

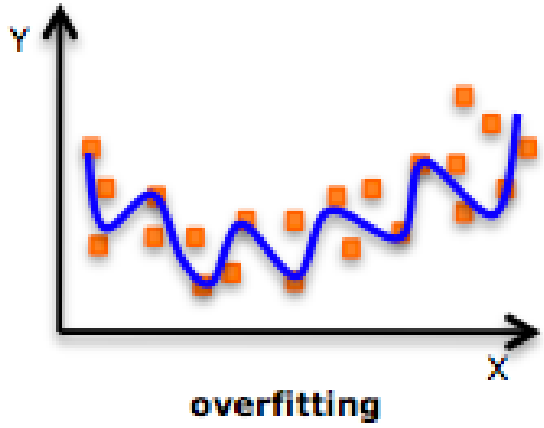


Underfitting

- High test and training error

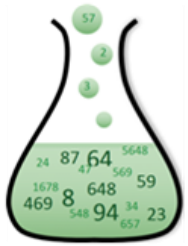


Just right!



overfitting

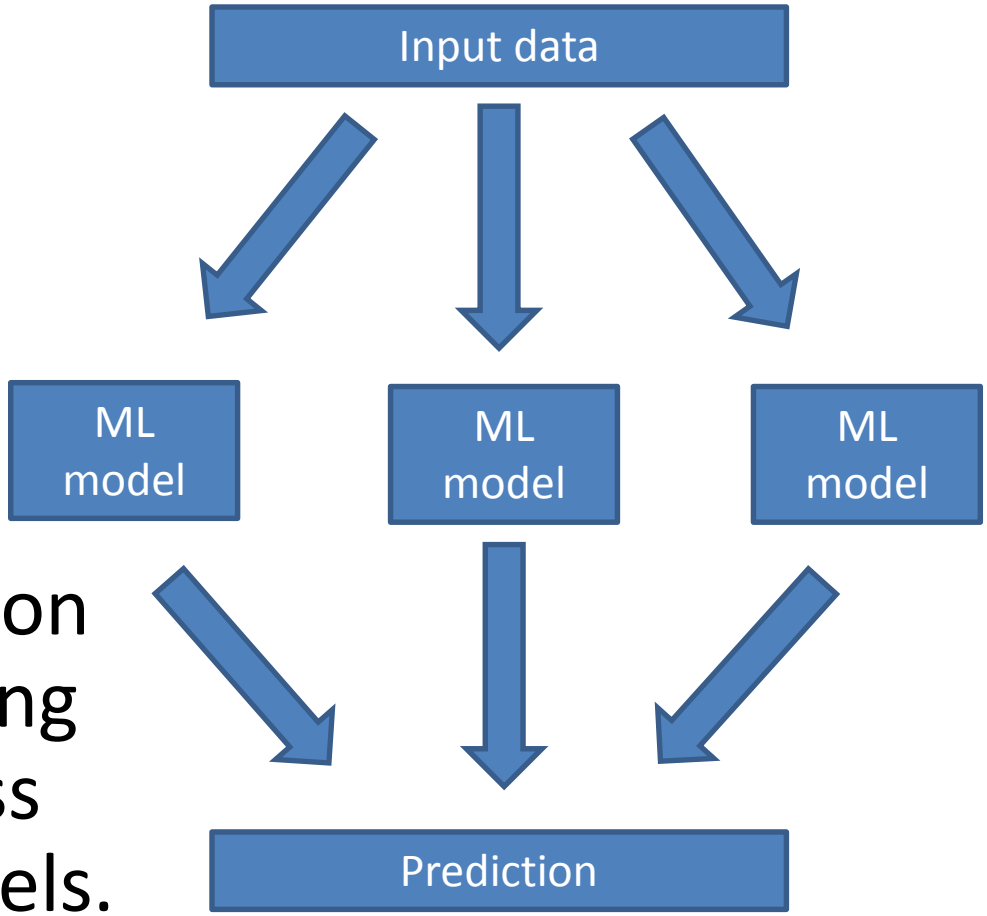
- High test error
- Low training error
- Generalises poorly

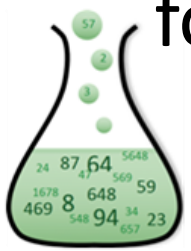


Data Science Hub

Bagged Ensemble

- Build multiple base models and then combines their predictions to give a final output.
- Combined either by averaging for regression models or by predicting the most popular class for classification models.



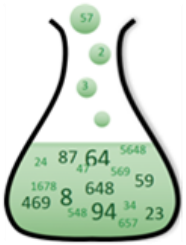


Data Science Hub

Bagged Ensemble

- By combining lots of different models, that may either be over or under fit, the final bagged ensemble model will ‘capture’ the underlying trends and patterns in the data.

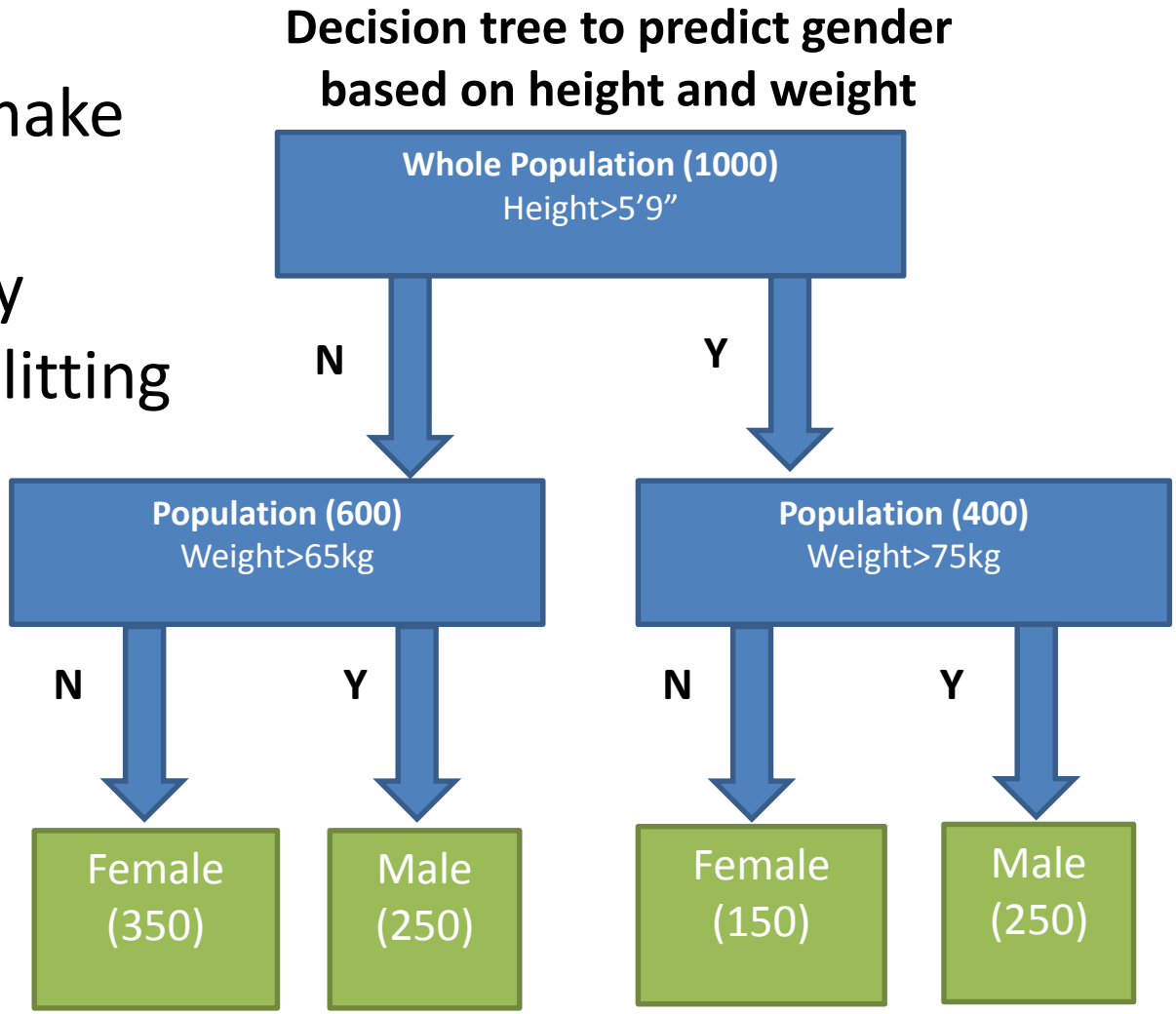
»The final prediction should be more accurate than any individual base model.



Data Science Hub

Decision Trees

- Decision trees make predictions or classifications by progressively splitting the population

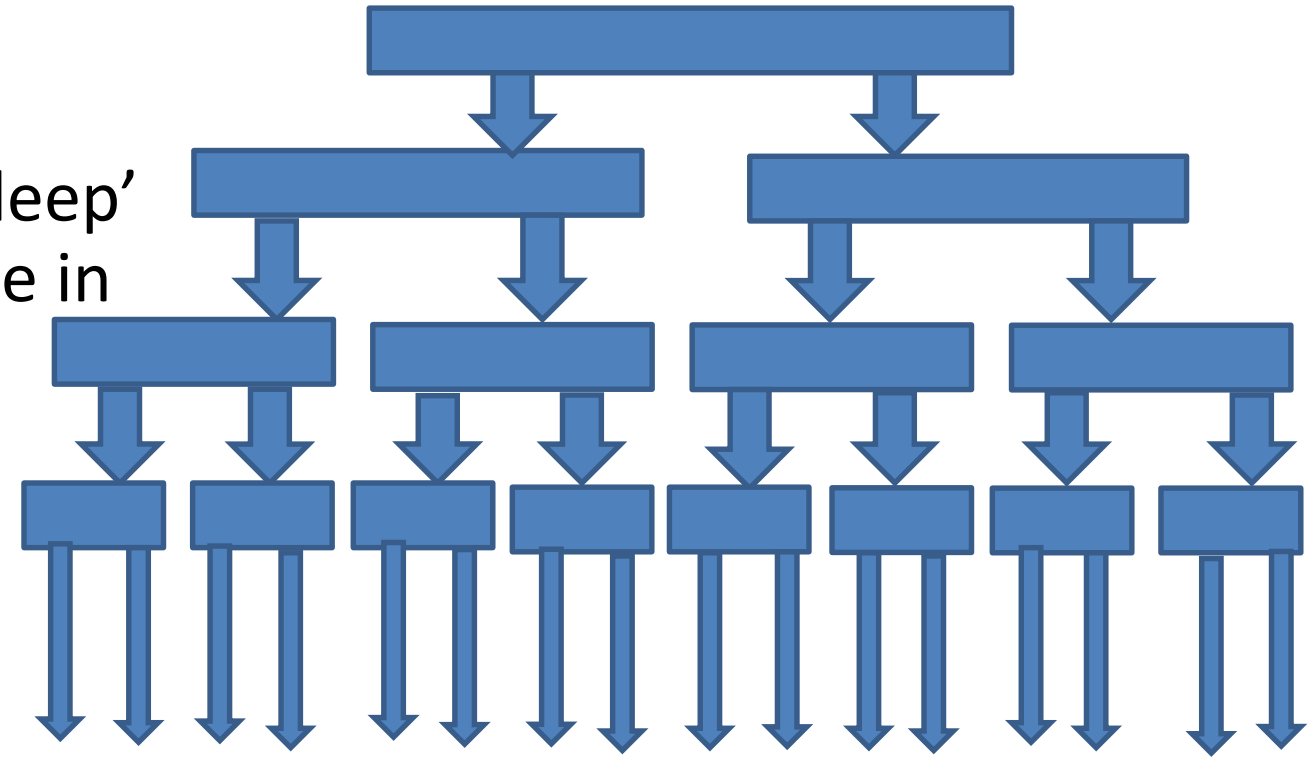


Data Science Hub

Decision Trees

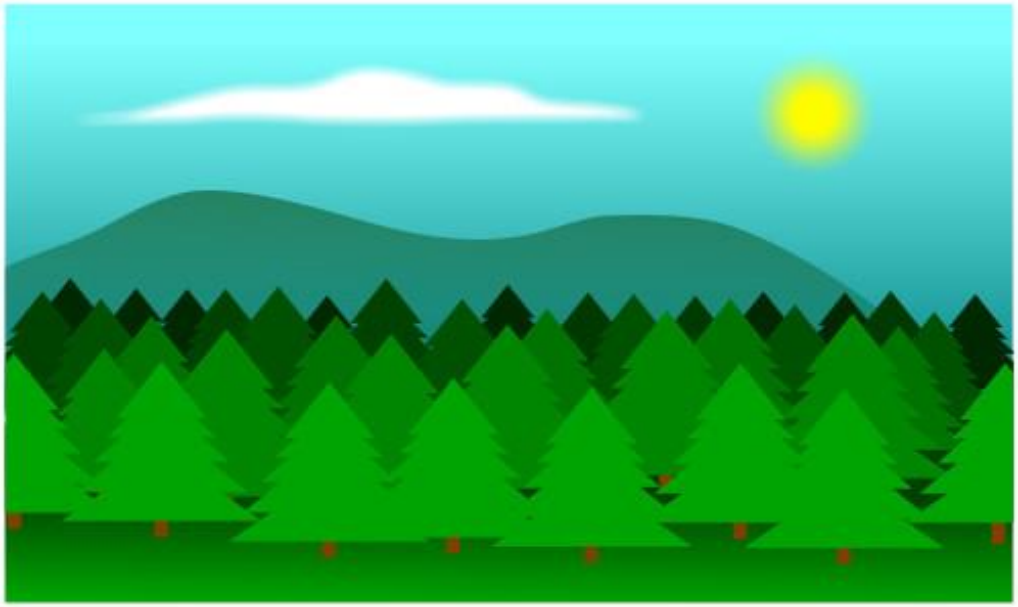
Decision trees are prone to overfitting, can become very ‘deep’ and fit the noise in the input data.

Over fit Decision tree



Data Science Hub

Random Forest



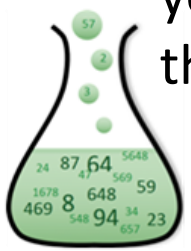
Random Forest

In appreciation of the random forest model;

The Random Forest is my shepherd; I shall not want.
He makes me watch the mean squared error decrease rapidly.
He leads me beside classification problems.
He restores my soul.
He leads me in paths of the power of ensembles
for his name's sake.

Even though I walk through the valley of the curse of dimensionality,
I will fear no overfitting,
for you are with me;
your bootstrap and your randomness,
they comfort me.

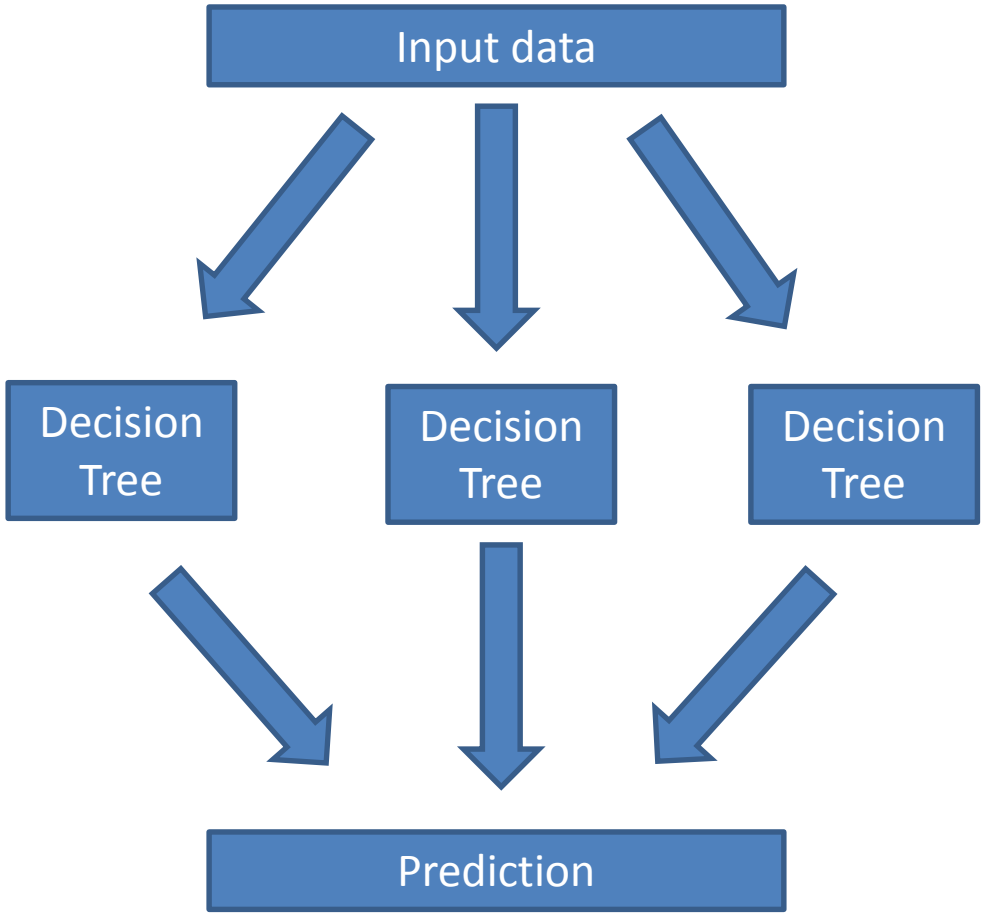
Source: R-Bloggers, *Oct 2017*

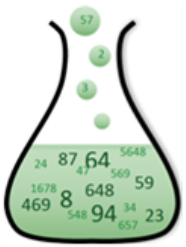


Data Science Hub

Random Forest

Random Forest is a bagged ensemble model based on the base learner of decision trees.

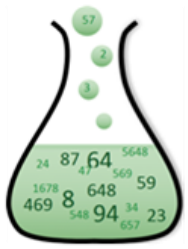




Data Science Hub

Random Forest

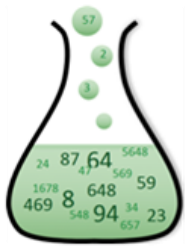
Each individual decision tree is very ‘deep’ (many nodes) and so overfit, but the final random forest should generalise well as it does not rely on the predictions from any one overfit model.



Data Science Hub

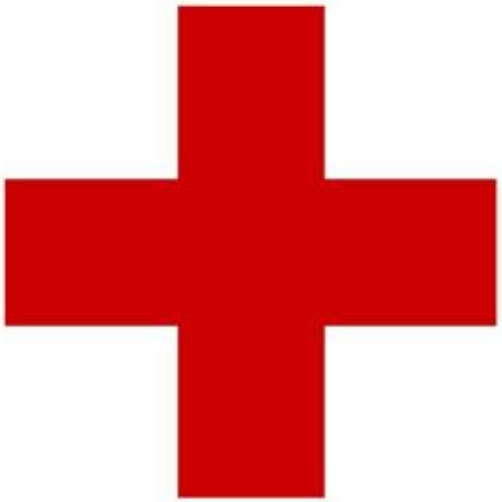
Random Forest – the randomness

Each individual decision tree is different as only a random subset of features are considered at each split node, random forest averages between a wide range of decisions from very different trees.



Data Science Hub

Fragile A&E systems



Fragile A&E Systems

News · UK News · NHS

Doctors warn of A&E crisis as hospitals influx of winter patients

Total of 45 hospitals have had to close A&E departments and divert ambulances to other hospital due to a surge in demand

SHARES COMMENTS

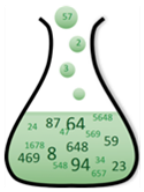
Enter your e-mail for our politics newsletter Subscribe

- A&E performance drops during winter.
- The performance of some A&E units may drop significantly more than others, these are termed as fragile A&E units.
- For planning purposes it would be advantageous to be able to predict these units in advance.

Data Science Hub

Random Forest to predict fragile A&E systems - workflow

Input data	Model	Output
Winter Sitrep data on A&E units and winter performance of A&E units (taken from 4 winters 12-13 to 15-16)	Random forest (built in the R language using the Randomforest and Carat packages)	Predicted A&E winter performance (the percentage of patients seen within 4 hours)



Data Science Hub

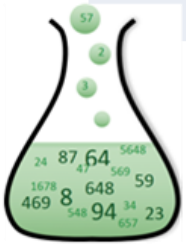
Random Forest to predict fragile A&E systems - prediction

Input data	Model	Output
Winter Sitrep data on A&E units and winter performance of A&E units (taken from last winter, 16-17)	Random forest (model trained on data from the last 4 winters)	Predicted A&E winter performance (the percentage of patients predicted to be seen within 4 hours over winter 17-18)

Data Science Hub

Predictions

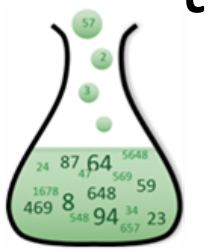
Trust	Performance (%)
Shrewsbury And Telford Hospital NHS Trust	79.2
Medway NHS Foundation Trust	79.3
Gloucestershire Hospitals NHS Foundation Trust	79.4
The Princess Alexandra Hospital NHS Trust	79.7
Pennine Acute Hospitals NHS Trust	80.3
Kettering General Hospital NHS Foundation Trust	80.4
West Hertfordshire Hospitals NHS Trust	80.7
Portsmouth Hospitals NHS Trust	80.8
University Hospitals Of North Midlands NHS Trust	81.0
Stockport NHS Foundation Trust	81.0



Data Science Hub

Results

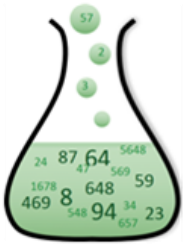
- The trusts with the lowest performance as predicted by random forest were all categorised as fragile by NHS E\NHS I, but some were in category 4 (most fragile) and some were in category 3 (somewhat fragile).
- These trusts will be monitored over the coming winter to assess any value added by the random forest methodology.



Data Science Hub

Conclusions

- Random forest may be (relatively) simply applied to most prediction (regression or classification) problems.
- The use of random forest may give new insights into your problem and add value to your predictions.



Data Science Hub