Institute
and Faculty
of Actuaries

Machine Learning for Actuaries

Steven Perkins
Hazel Davis

26 October 2018

Agenda

I. Data Science / Machine Learning Background

II. Case Study Introduction

III. Data Processing and Visualisation

IV. Model Building

V. Model Validation

VI. Reporting

VII. Final Thoughts

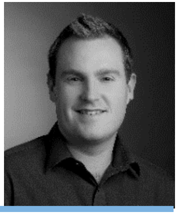
Aim: *Provide an overview of a data science process*

Institute and Faculty of Actuaries

Speakers

Hazel Davis

Actuary in personal lines pricing with operational research background



Steven Perkins

Actuary with data science and general insurance experience as well as PhD in machine learning

Members of Modelling, Analytics and Insight from Data (MAID) working group, working on applying new techniques to traditional actuarial areas

MAID now replaced with data science member interest group

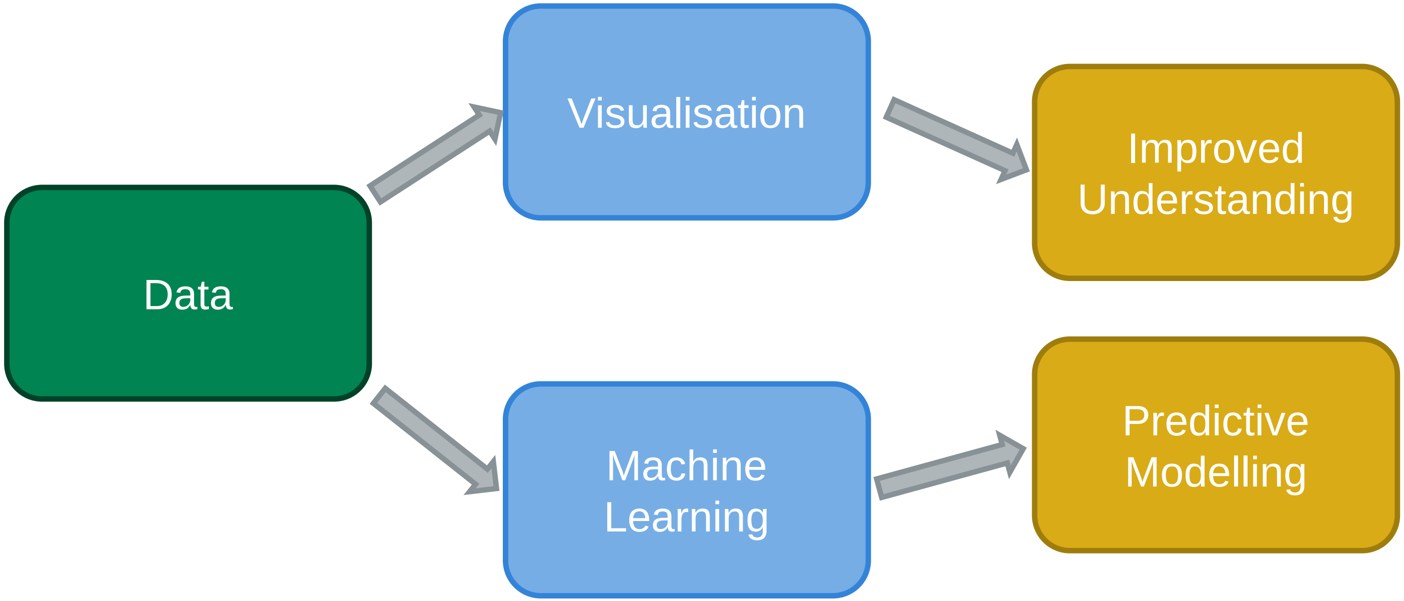
Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

Machine Learning Overview

26 October 2018

Data Science Overview



Institute and Faculty of Actuaries

Data Science Benefits to Actuaries

Improved Data Quality

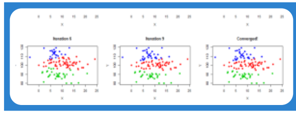
- A key driver for companies to improve data capture and storage

New Data Sources

- Opportunities for actuaries to explore alternative data sources

Speed of Analysis

- Machine learning models can generally be fitted and validated quickly

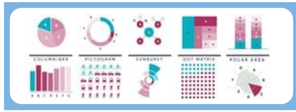


New Modelling Techniques

- Alternative modelling approaches allows different perspectives to be gained on data

New Approaches to Problems

- Wider variety of models quickly - select the best model technique for a given problem



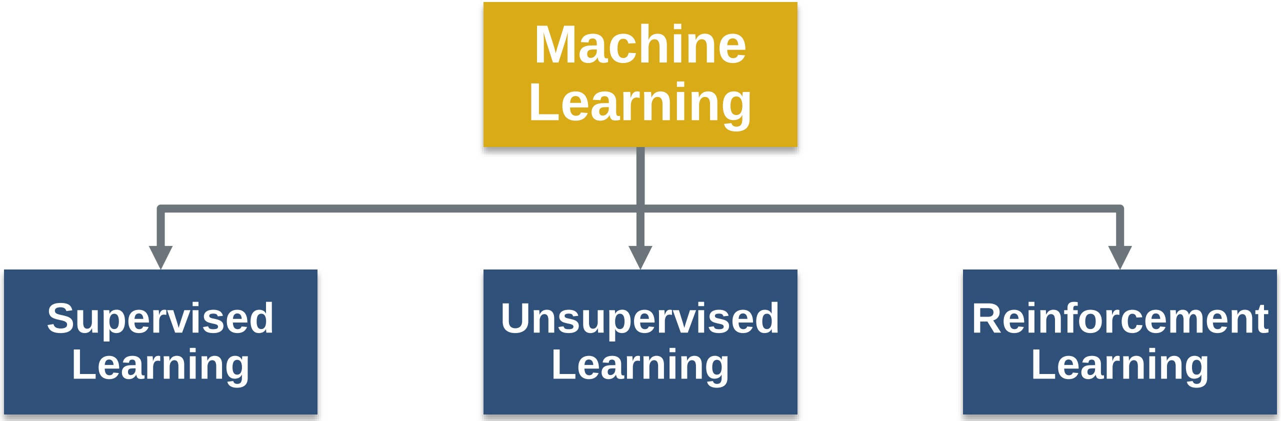
Improved Data Visualisations

- Stunning visualisations of data which can itself provide new perspectives on a task

Institute and Faculty of Actuaries

Machine Learning

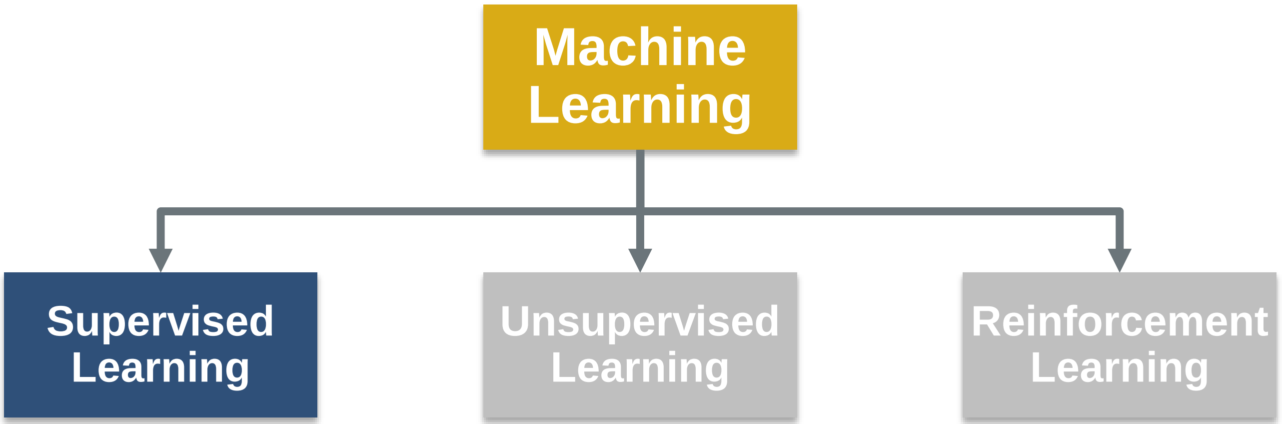
Wikipedia: *"Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers."*



Institute and Faculty of Actuaries

Machine Learning

Wikipedia: *"Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers."*



Institute and Faculty of Actuaries

Supervised Learning

Training Data

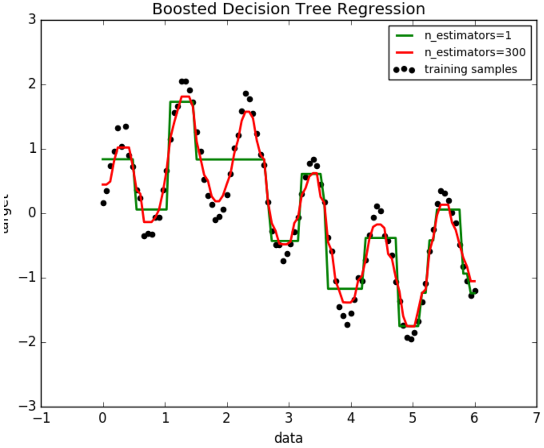
- Input Variables
- Target Variable

Goal: Predict the target variable using the input variables

Example

- Target Variable in Historical Data: Policy Fraud Indicator (Yes/No)
- Input Variables: Age, Employment Status, Occupation, Email Address, Location, IP address, etc.

Goal: Use historical fraud cases to predict future fraud cases



Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

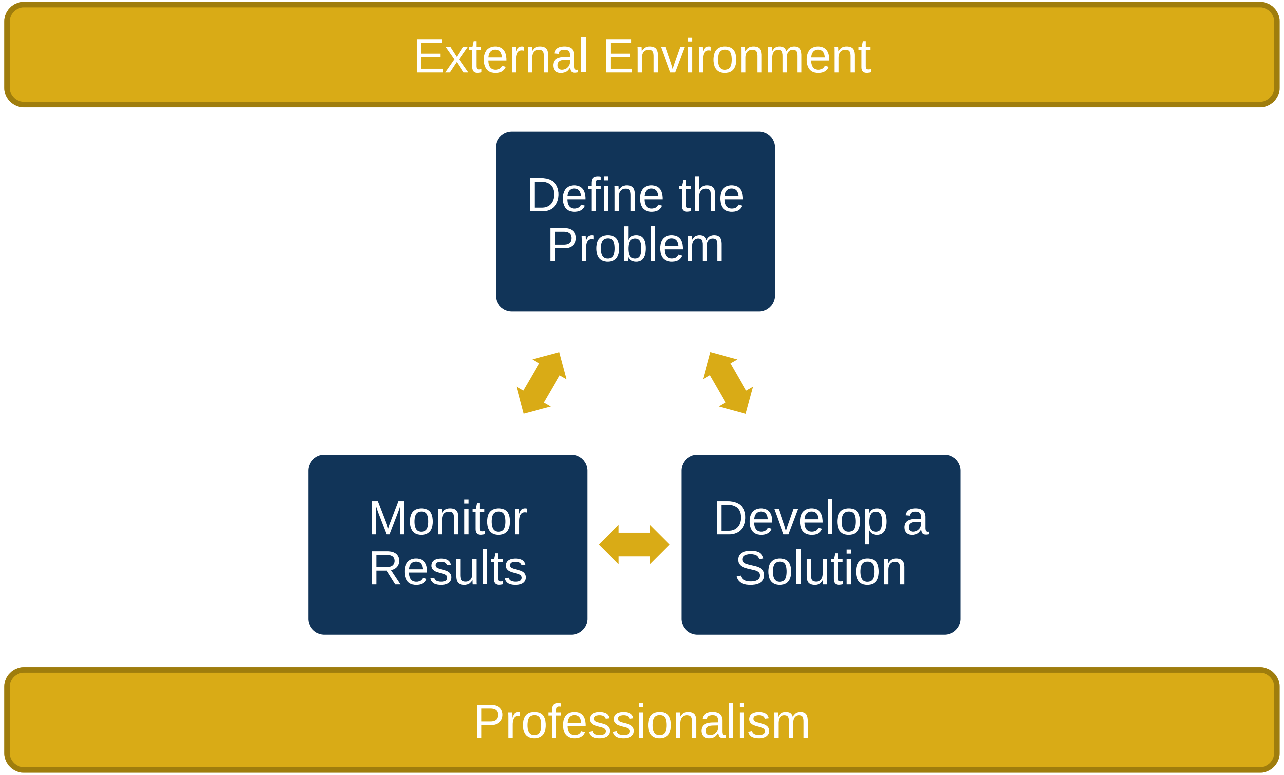
Case Study

26 October 2018

What will the death rate be in the UK ‘next year’?

Institute and Faculty of Actuaries

Actuarial Control Cycle



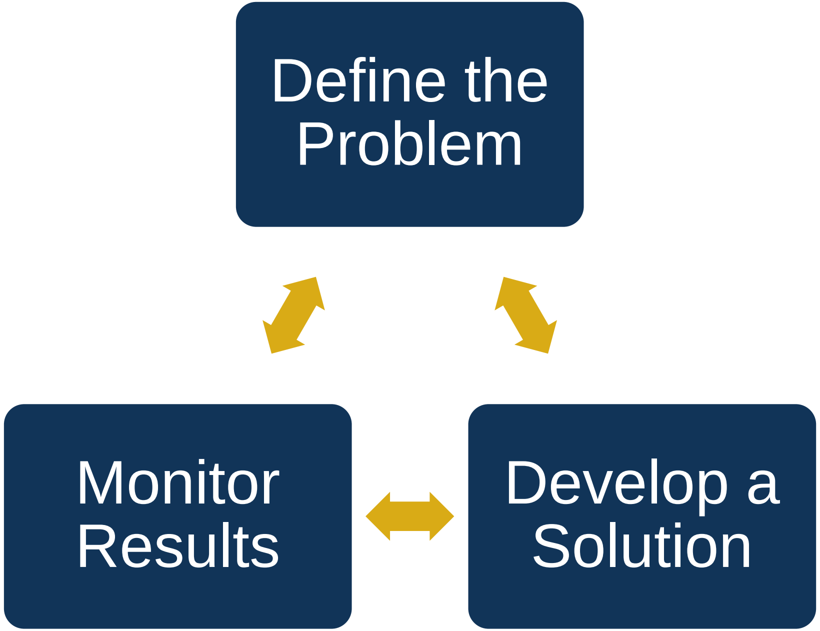
Institute and Faculty of Actuaries

Data Science Process

1. Problem Specification
2. Data Collection
3. Data Processing and Visualisation
4. Model Building
5. Model Validation
6. Reporting
7. Monitoring

Institute and Faculty of Actuaries

ACC vs Typical Machine Learning Process



1. Problem Specification
2. Data Collection
3. Data Processing and Visualisation
4. Model Building
5. Model Validation
6. Reporting
7. Monitoring

Institute and Faculty of Actuaries

Data Science Process

1. Problem Specification
2. Data Collection
3. Data Processing and Visualisation
4. Model Building
5. Model Validation
6. Reporting
7. Monitoring

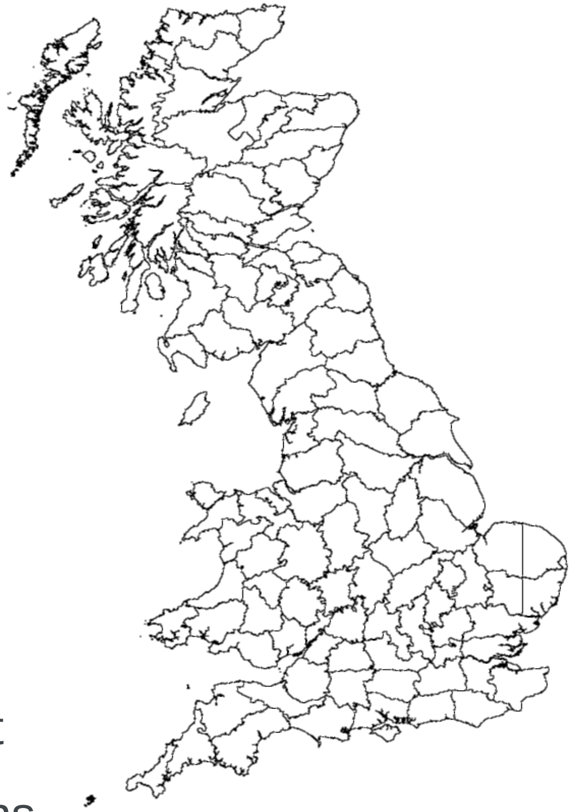
Institute and Faculty of Actuaries

Problem Specification

- Predict UK 2016 death rate

- Supervised learning problem
- Trained on past UK death rate data
- Using publicly available geographic data

- Build a range of models to see which perform best
- Engineer new input variables to improve predictions



Institute and Faculty of Actuaries

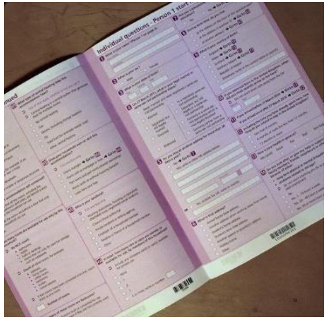
Data Science Process

1. Problem Specification
2. Data Collection
3. Data Processing and Visualisation
4. Model Building
5. Model Validation
6. Reporting
7. Monitoring

Institute and Faculty of Actuaries

Data Collection

- Target variable: UK death rate by region 2012-2016 ONS
- Explanatory variables from Census 2011
  - Population age profiles
  - Population density
  - Average wages and working hours
  - Pension income
- No personal data used – GDPR compliant!



This Photo by Unknown Author is licensed under CC BY-NC-SA

```
# 2011 Census Age Distribution by Area
ageProfile <- fread("Data/Age Distribution.csv", stringsAsFactors = TRUE)
```

Institute and Faculty of Actuaries

Data Cleaning

- Scotland & NI figures in different format, so revised scope to England & Wales
- Data checks
  - male deaths + female deaths = total deaths etc
  - 2014 deaths similar to 2015 deaths etc
  - Missing entries
  - Formats

```
# Recalculate Total Population
df$totalPopRecalculate <- df$populationMaleThousands + df$populationFemaleThousands

# Check for population differences
df$popCheck <- df$totalPopRecalculate - df$populationAllThousands
summary(as.factor(abs(df$popCheck)))
```

Data Science Process

1. Problem Specification
2. Data Collection
3. **Data Processing and Visualisation**
4. Model Building
5. Model Validation
6. Reporting
7. Monitoring

Institute and Faculty of Actuaries

Data Processing

- Potential explanatory variables come from different sources
- Join datasets using one or more variables to define the link between the datasets
- This is one of the higher risk areas of data manipulation
- Particularly problematic when datasets do not have a clear linking key

```
# Add statistics year and unique key to each table
for (i in 2012:2016) {

  # Create year index
  eval(parse(text = paste("deathRates",i,"$Year <- ", i, sep = "")))

  # Create unique key
  eval(parse(text = paste("uniqueKey <- paste(paste(deathRates",i,"$AreaCodes, sep = \"\"), paste(deathRates",i,"$Year, sep = \"\"), sep = \"_\")", sep = "")))
```

```
df <- merge(df, hoursAndPay, by.x = "uniqueKey", by.y = "uniqueKey", all.x = T, all.y = F)
```

Institute and Faculty of Actuaries

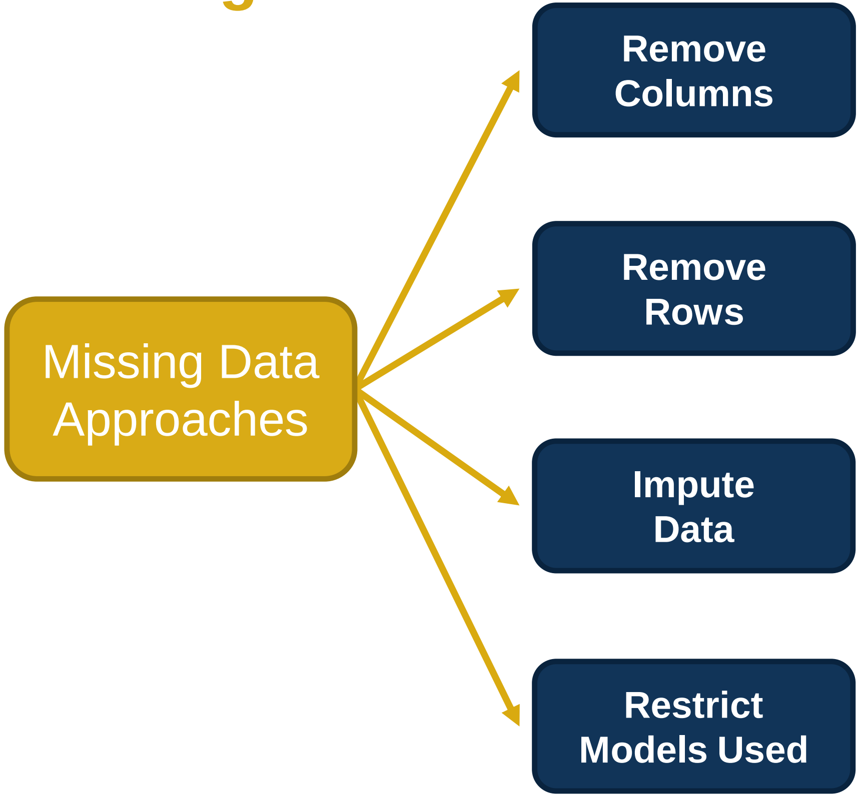
Data Processing

- Understanding area keys essential for joining data
- Join and transform files to create one dataset
  - By area code and year
  - Target variable plus all potential input variables

uniqueKey	AreaCodes	AreaName.x	populationAllThousands	populationMaleThousands	populationFemaleThousands	deathsAll	crudeDeathRate	Year
E06000001_2012	E06000001	Hartlepool	92.2	44.9	47.4	905	9.8	2012
E06000001_2013	E06000001	Hartlepool	92.7	45.2	47.4	923	10.0	2013
E06000001_2014	E06000001	Hartlepool	92.6	45.2	47.4	972	10.5	2014
E06000001_2015	E06000001	Hartlepool	93.0	45.0	47.0	1067	11.5	2015
E06000001_2016	E06000001	Hartlepool	92.8	45.3	47.5	990	10.7	2016
E06000002_2012	E06000002	Middlesbrough	138.7	68.0	70.7	1394	10.0	2012
E06000002_2013	E06000002	Middlesbrough	138.9	68.2	70.7	1335	9.6	2013

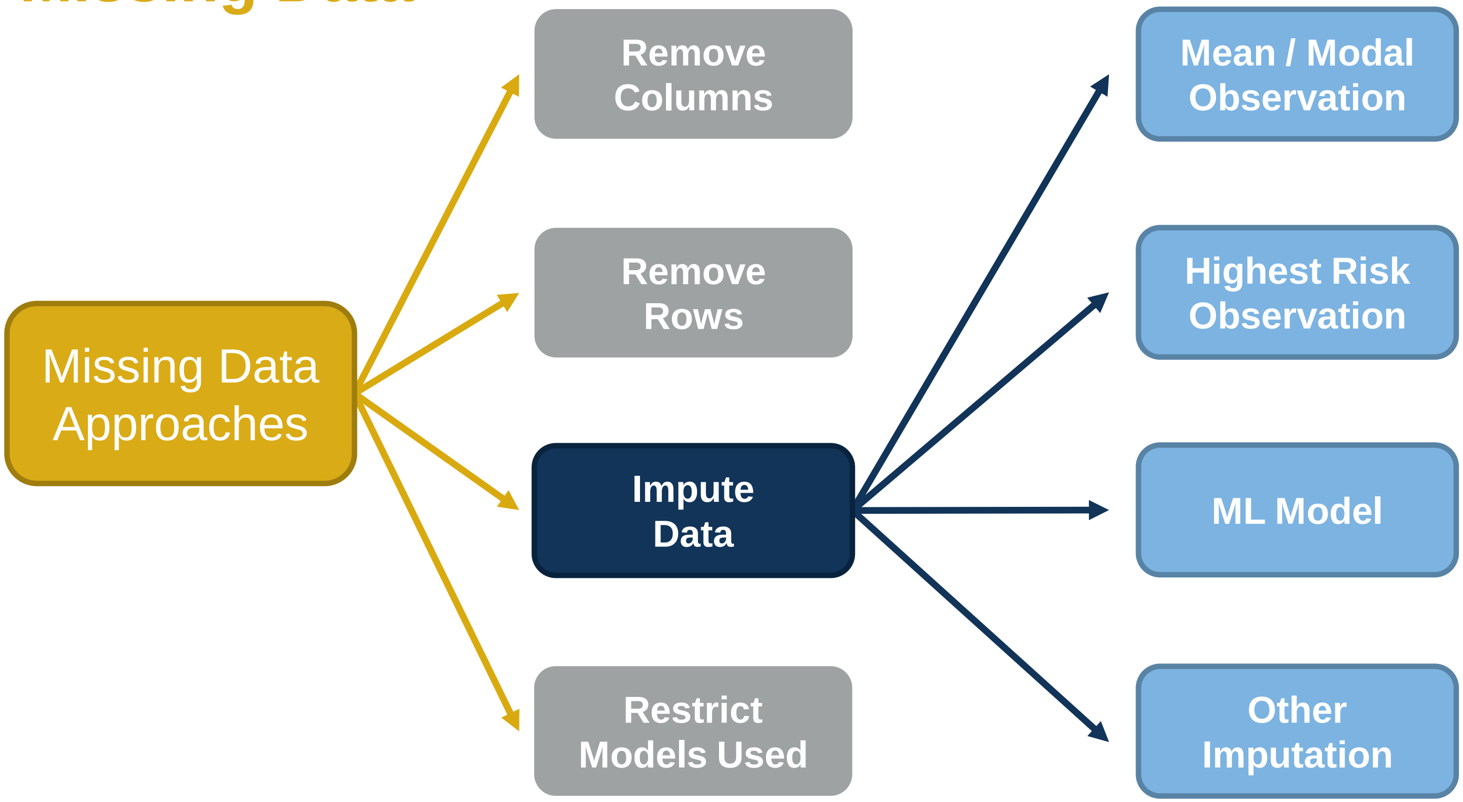
Institute and Faculty of Actuaries

Missing Data



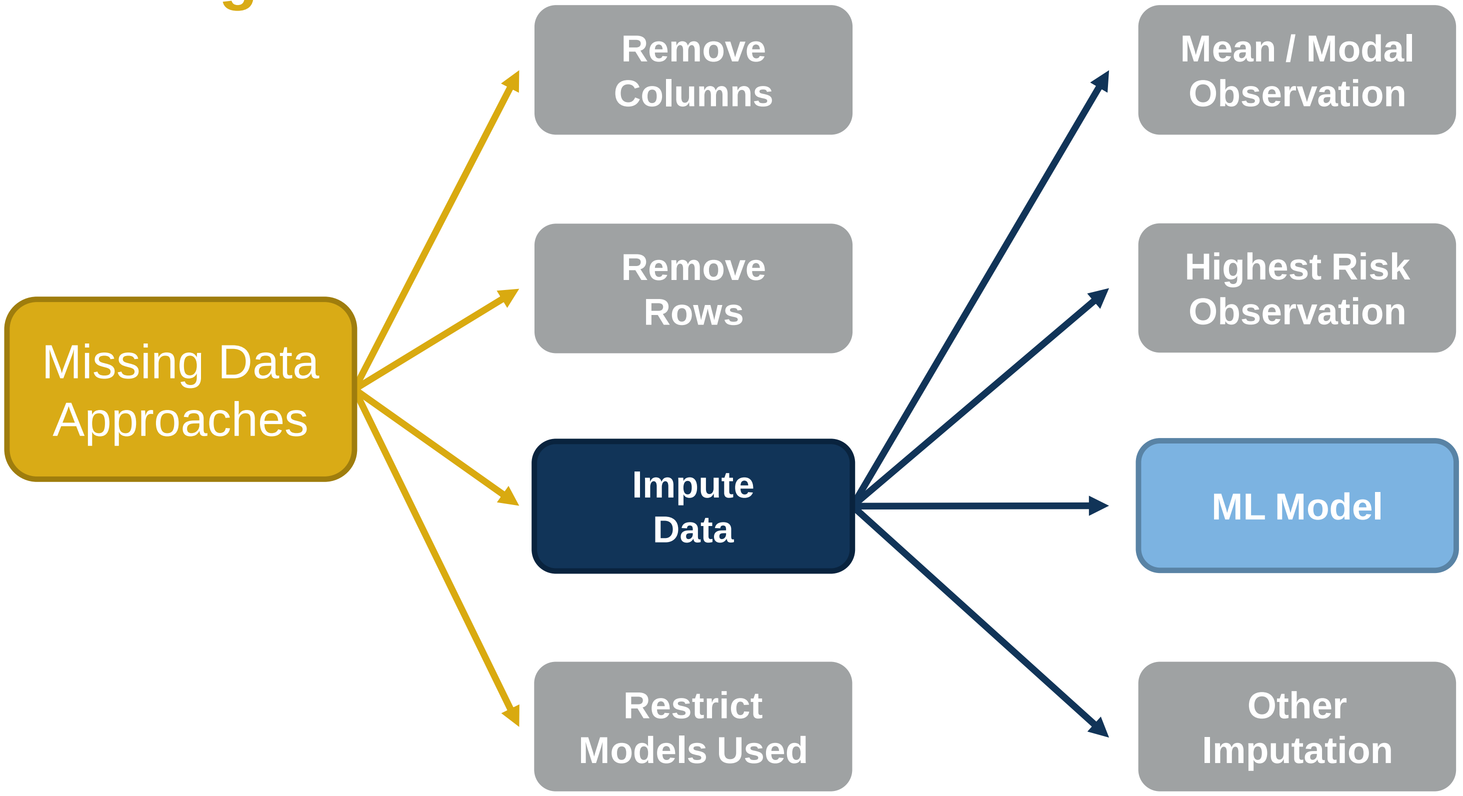
Institute and Faculty of Actuaries

Missing Data



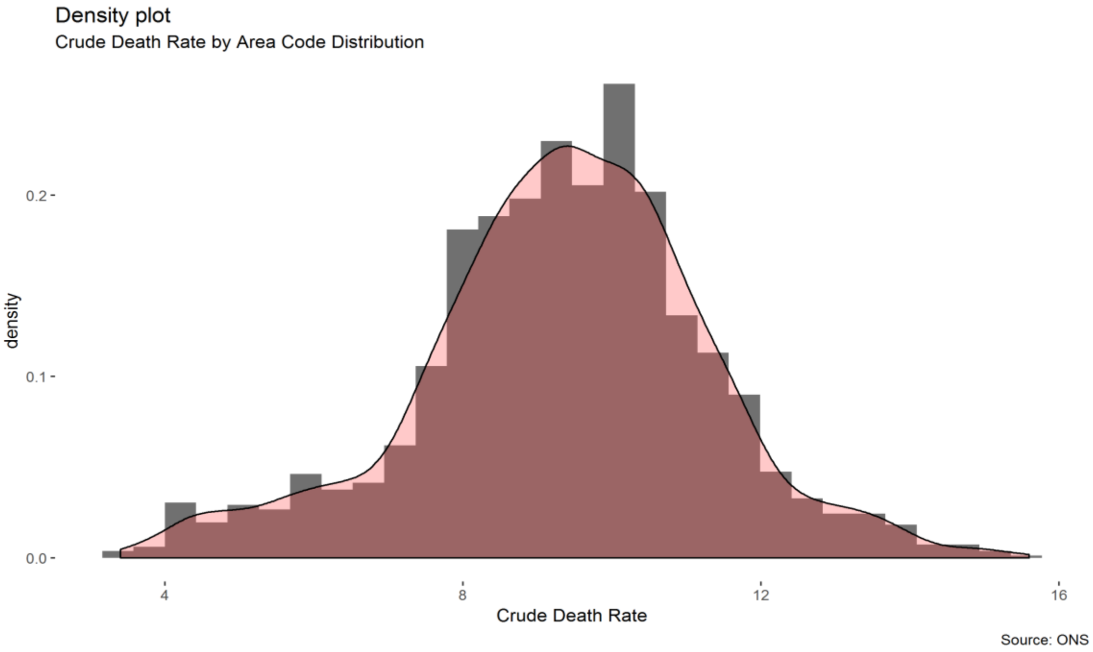
Institute and Faculty of Actuaries

Missing Data



Institute and Faculty of Actuaries

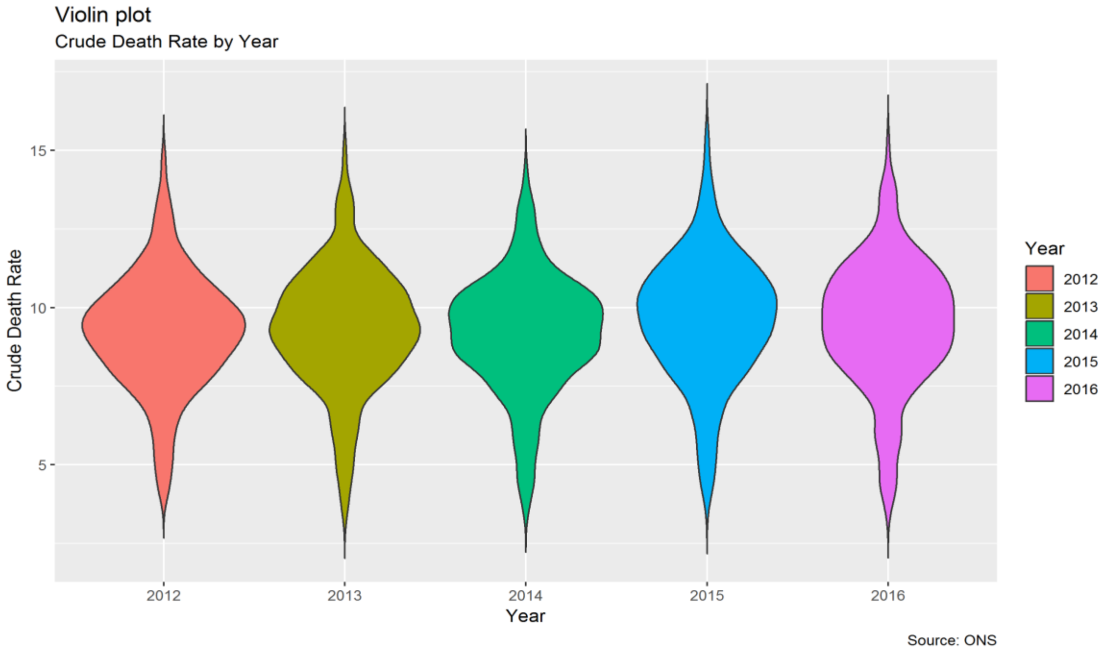
Initial Data Visualisations (1 of 4)



- Death rates – relatively normally distributed

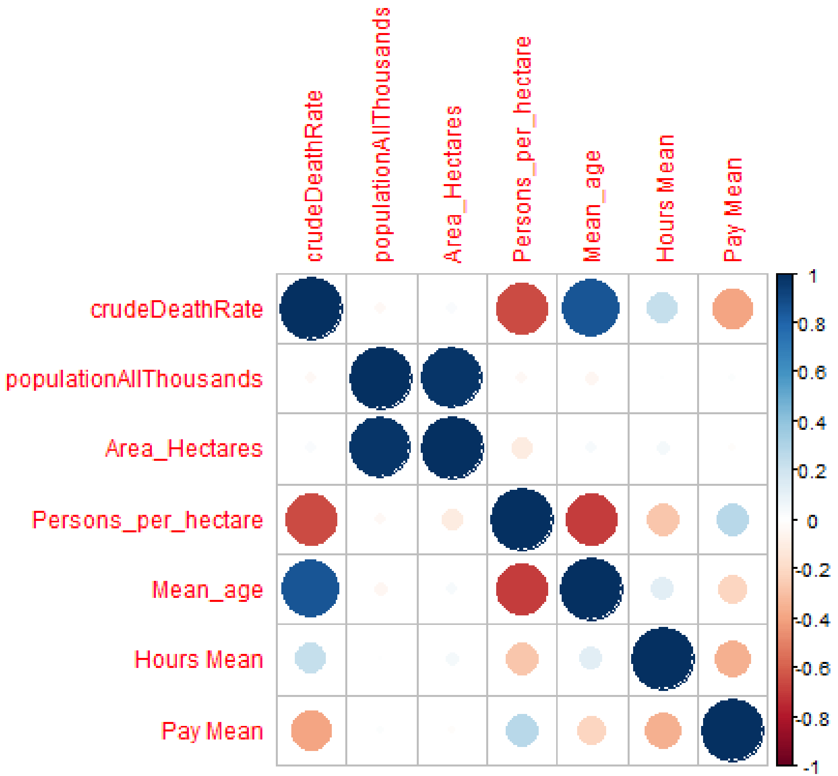
Institute and Faculty of Actuaries

Initial Data Visualisations (2 of 4)



- Death rates by year - relatively consistent distribution by year

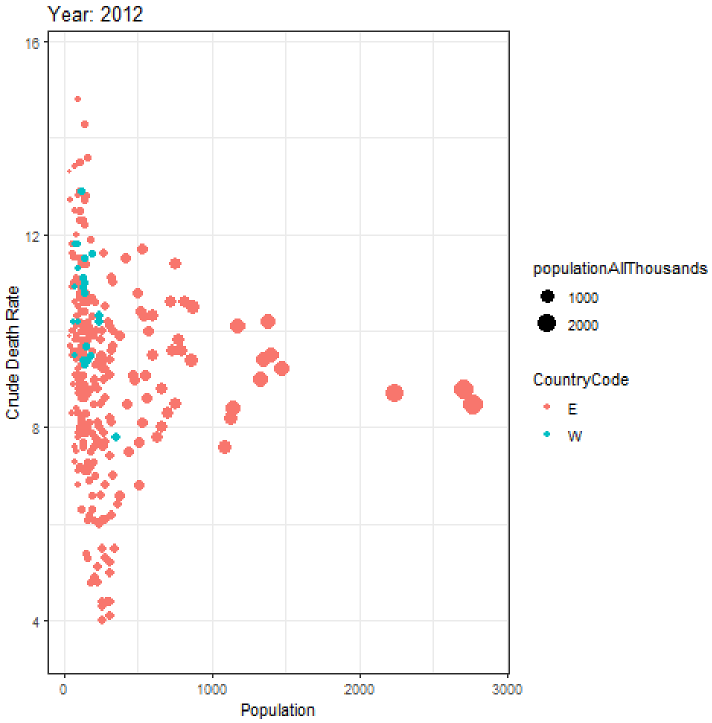
Initial Data Visualisations (3 of 4)



- Strong correlation between age and death rate
- Weak correlation between hours worked and death rate
- Strong negative correlation between population density and death rate
- Moderate negative correlation between average pay and death rate

Institute and Faculty of Actuaries

Initial Data Visualisations (4 of 4)

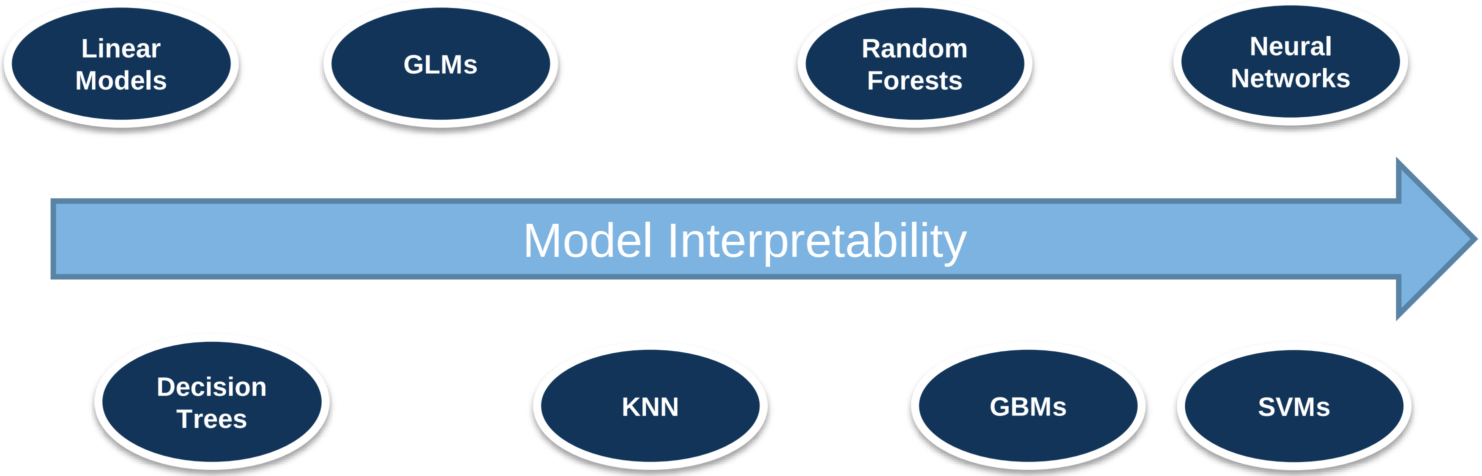


Data Science Process

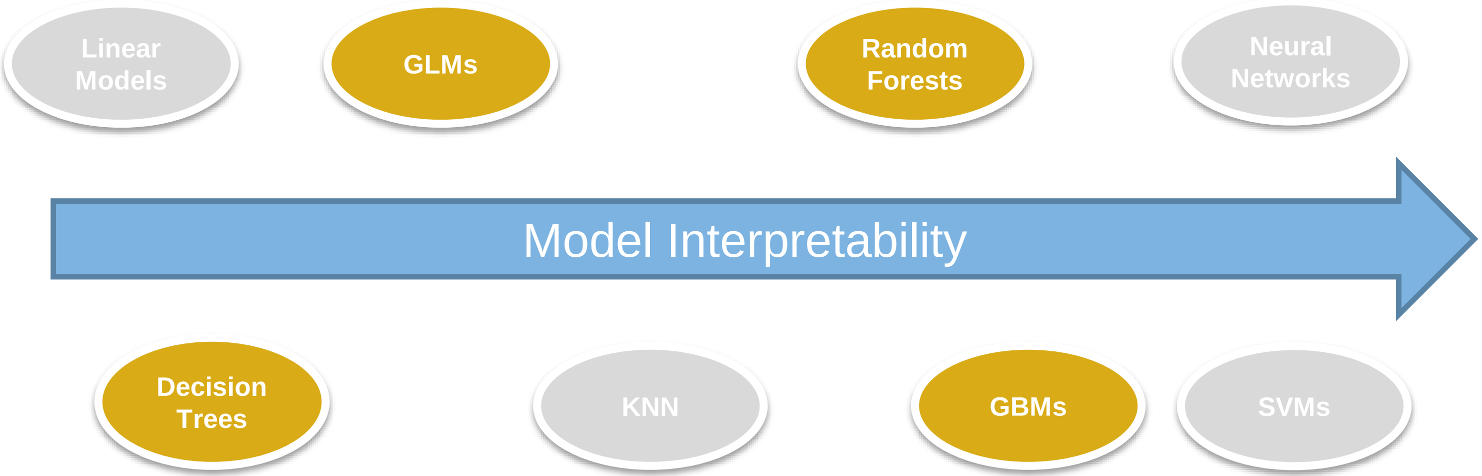
1. Problem Specification
2. Data Collection
3. Data Processing and Visualisation
4. **Model Building**
5. Model Validation
6. Reporting
7. Monitoring

Institute and Faculty of Actuaries

Modelling – Model Types



Modelling – Model Types



Institute and Faculty of Actuaries

Modelling – Example Model Building Code

```
modelFmla <- as.formula("crudeDeathRate ~ .")

interaction.depth <- 20
n.minobsinnode <- 5
shrinkage <- 0.02811323

set.seed(7645)

gbmModel <- gbm(
  modelFmla,
  dfTrain,
  distribution = "gaussian",
  n.trees = 250,
  cv.folds = 4,
  n.cores = 2,
  interaction.depth = interaction.depth,
  n.minobsinnode = n.minobsinnode,
  shrinkage = shrinkage
)

saveRDS(gbmModel, "R Models/gbmModel1.rds")
```

Define relationship between target variable and predictors

Select hyper-parameters

Select starting point for random numbers used in model fitting process – vital for reproducibility

Build the actual model

Save the Model!

Institute and Faculty of Actuaries

Modelling – Initial Model Performance

- Initial performance is assessed based on root mean squared error (RMSE) of the holdout data (2016 death rates)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(actual_i - predicted_i)^2}$$

- Baseline model: prior year death rate is best estimate prediction for the current year

Model	Prior Year	Tree	Pruned Tree	LASSO	GBM	Random Forest
RMSE	**0.451**	0.664	0.663	0.778	0.465	0.654
Median Absolute Error	**0.300**	0.420	0.419	0.511	0.297	0.450
RMSE vs Prior Year Model	**0.000**	+0.213	+0.212	+0.327	+0.014	+0.203

Institute and Faculty of Actuaries

Modelling – Feature Engineering

Initial Fields
(by Area Code & Year)

- Country
- Population
- Population Density
- Age Distribution
- Mean & Median Pay
- Mean & Median Hours
- Death Rate (target var)

Engineered Fields

- Death Rate Prior Year
- Variance of Death Rate
- Coefficient of Variation of Death Rate

Institute and Faculty of Actuaries

Modelling – Updated Model Performance

Pre Feature Engineering

Model	Prior Year	Tree	Pruned Tree	LASSO	GBM	Random Forest
RMSE	**0.451**	0.664	0.663	0.778	0.465	0.654
Median Absolute Error	**0.300**	0.420	0.419	0.511	0.297	0.450
RMSE vs Prior Year Model	**0.000**	+0.213	+0.212	+0.327	+0.014	+0.203

Post Feature Engineering

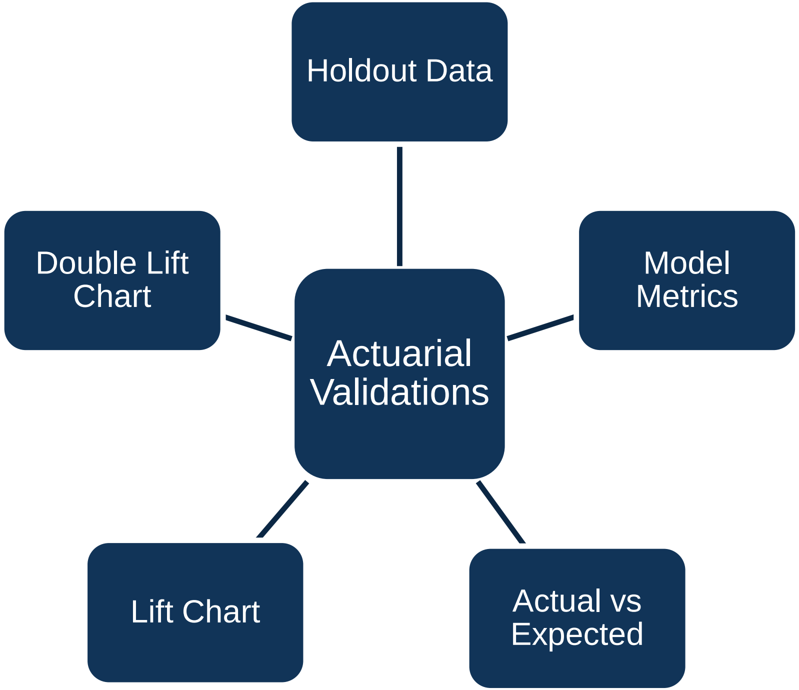
Model	Prior Year	Tree	Pruned Tree	LASSO	GBM	Random Forest
RMSE	0.451	0.540	0.540	**0.426**	0.471	0.438
Median Absolute Error	0.300	0.344	0.353	**0.277**	0.302	0.281
RMSE vs Prior Year Model	0.000	+0.089	+0.089	**-0.025**	+0.020	-0.013

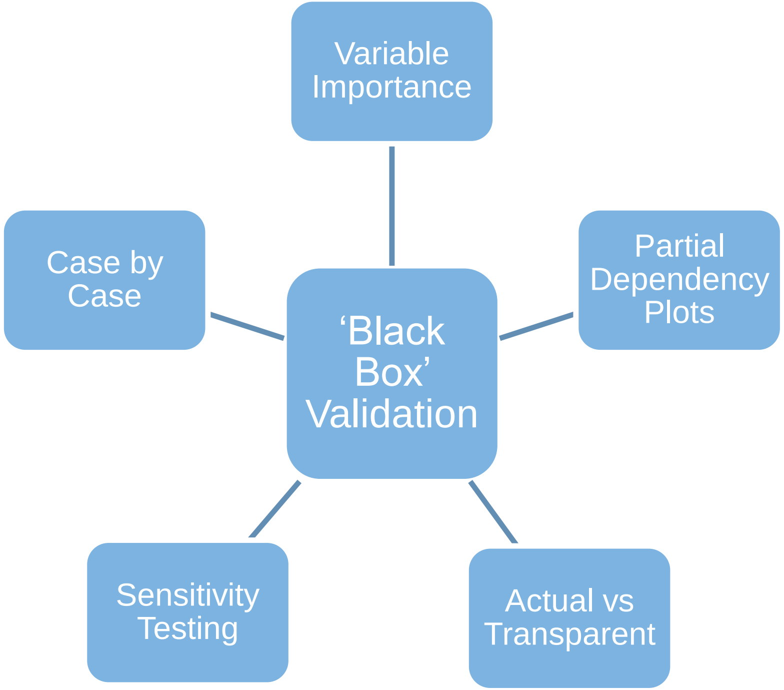
Data Science Process

1. Problem Specification
2. Data Collection
3. Data Processing and Visualisation
4. Model Building
5. **Model Validation**
6. Reporting
7. Monitoring

Institute and Faculty of Actuaries

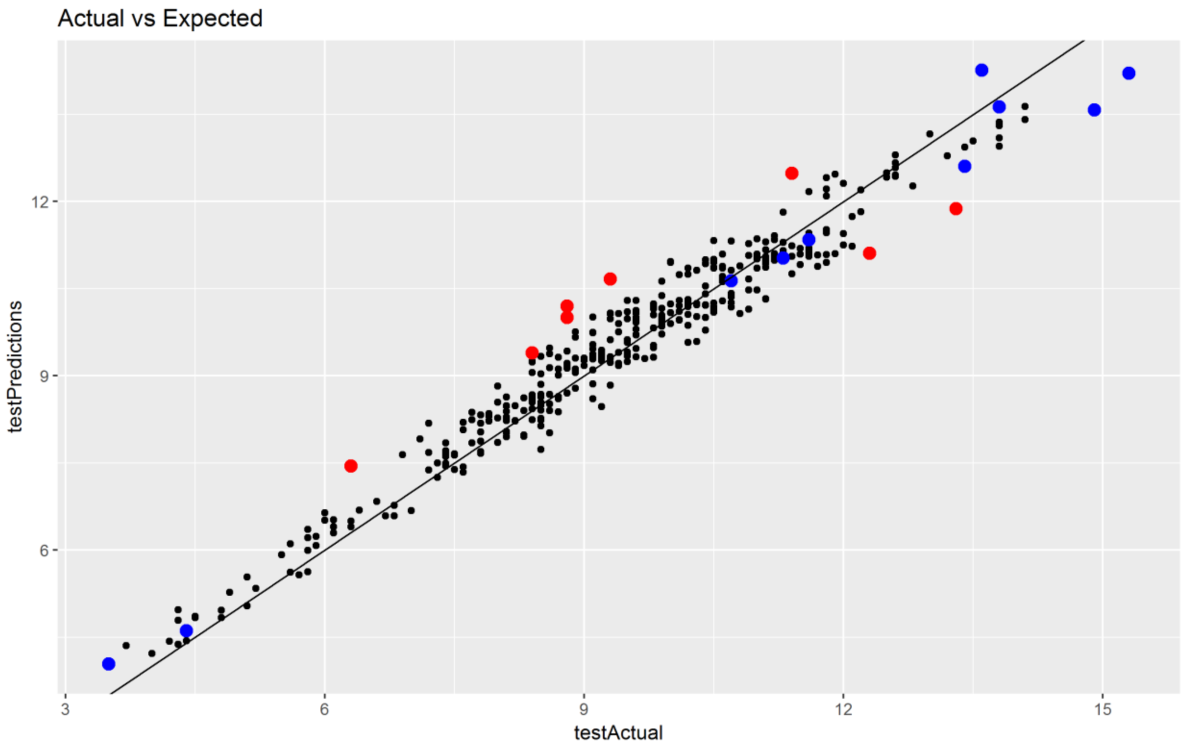
Validation Concepts





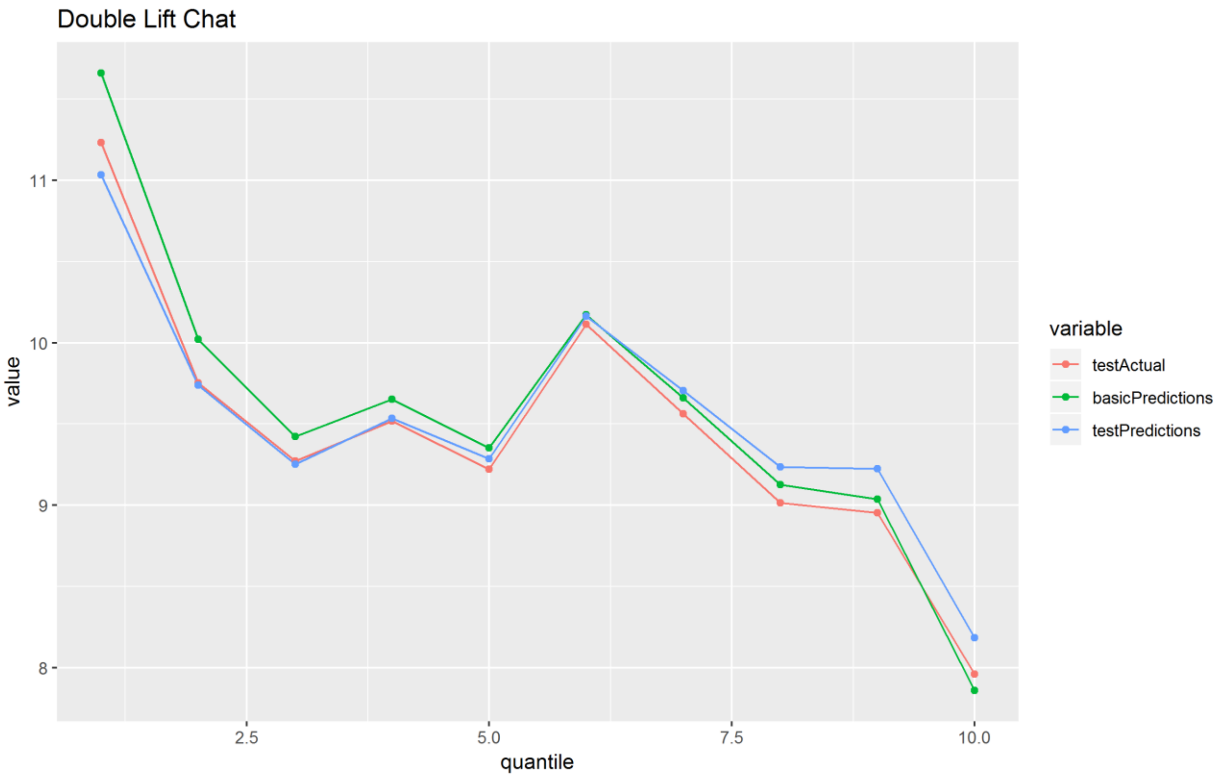
Institute and Faculty of Actuaries

Actual vs Expected



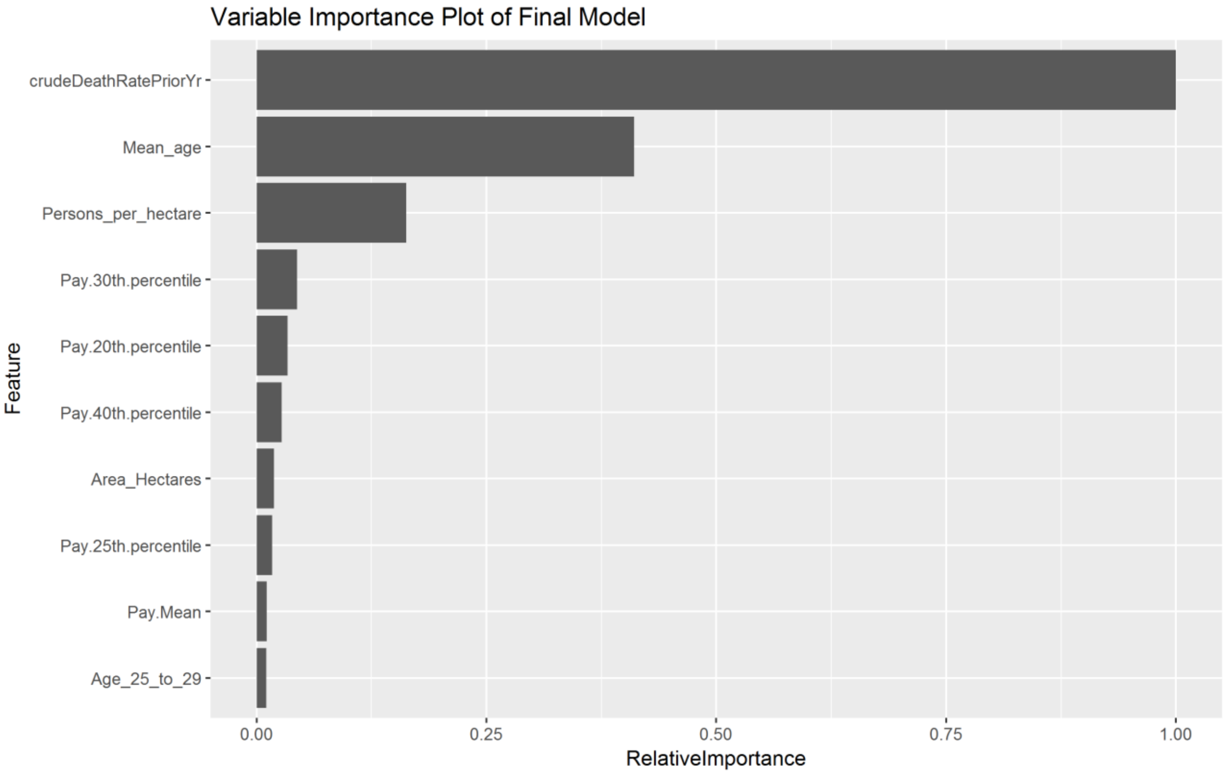
Institute and Faculty of Actuaries

Double Lift Chart

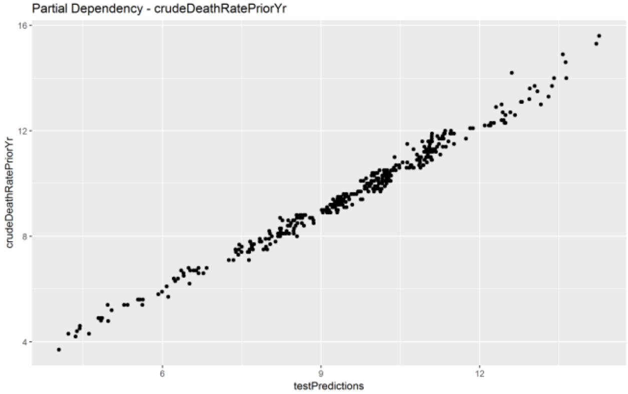


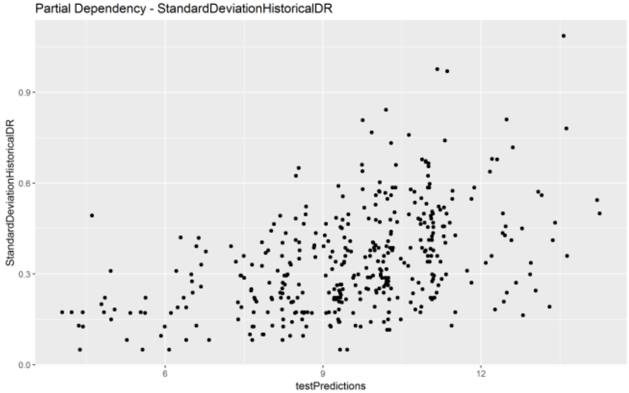
Institute and Faculty of Actuaries

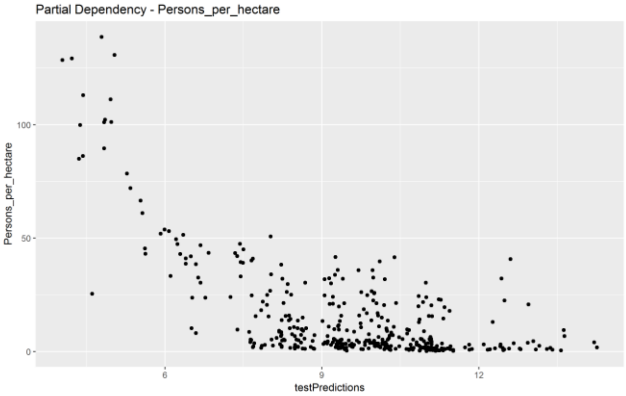
Variable Importance

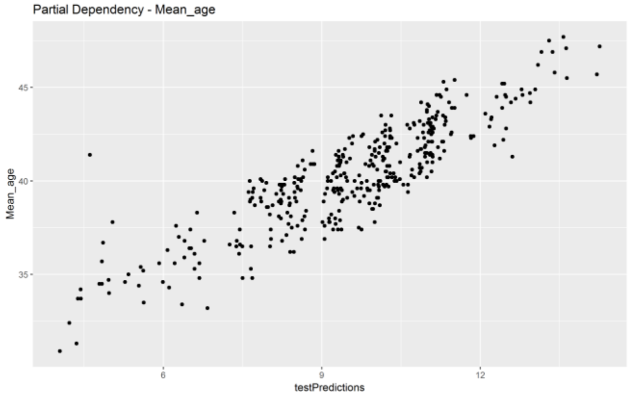


Partial Dependency Plots



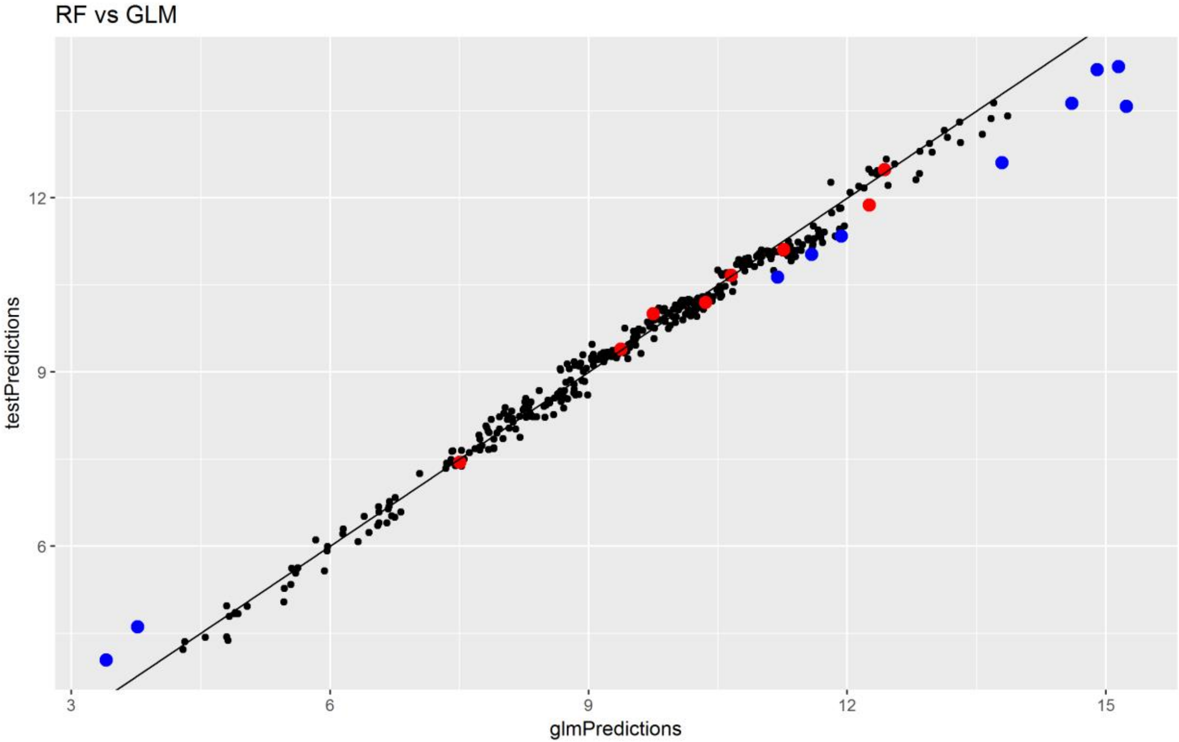






Institute and Faculty of Actuaries

Actual vs 'Transparent' Model



Case-by-Case Review

Institute and Faculty of Actuaries

Data Science Process

1. Problem Specification
2. Data Collection
3. Data Processing and Visualisation
4. Model Building
5. Model Validation
6. Reporting
7. Monitoring

Institute and Faculty of Actuaries

Results Communication

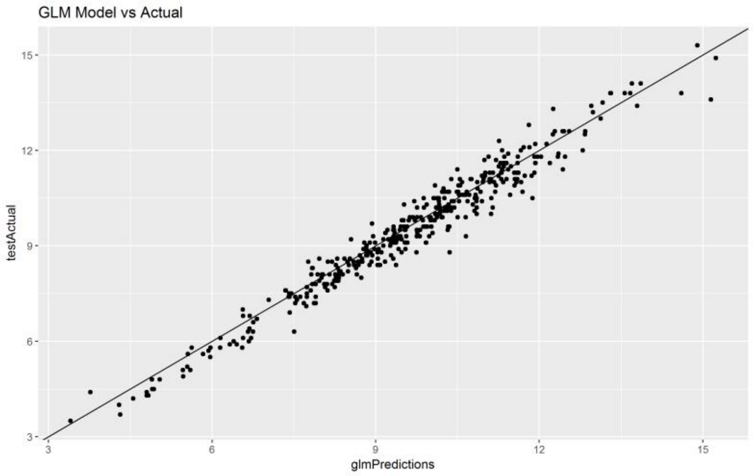
- Normal TASs apply
  - Data source, checks and controls
  - Assumptions
  - Model approach and testing
  - Results with limitations and uncertainty

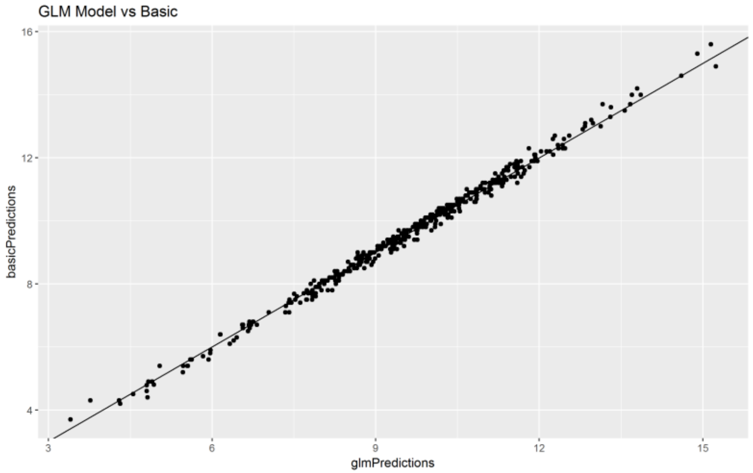
- Tailor communications to audience
  - Avoid jargon
  - High level or detailed results as appropriate / possible

Institute and Faculty of Actuaries

Death Rates Conclusion

- Best model: LASSO regression
- Compared to basic model:
  - 5.5% reduction in RMSE
  - 7.5% reduction in Median Absolute Error
- Insight gained into correlations:
  - For example: high density areas predicted to have low death rates





Death Rates Conclusion

So… is this a good model?

Death Rates Conclusion

So… is this a good model?

Maybe…

Death Rates Conclusion

So… is this a good model?

Maybe…

Context 1: Life insurer who is managing exposure to risk and hence relies on the best possible understanding of UK death rates

Institute and Faculty of Actuaries

Death Rates Conclusion

So… is this a good model?

Maybe…

Context 1: Life insurer who is managing exposure to risk and hence relies on the best possible understanding of UK death rates

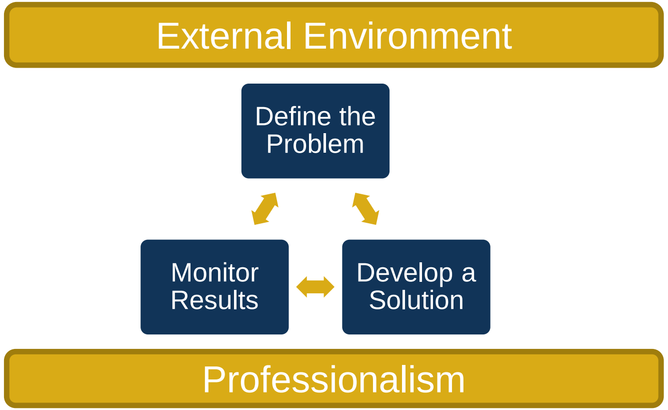
Context 2: Small life insurer who is considering improving capabilities of their pricing system to allow machine learning models to be used

Institute and Faculty of Actuaries

Death Rates Conclusion

So… is this a good model?

Maybe…



Context 1: Life insurer who is managing exposure to risk and hence relies on the best possible understanding of UK death rates

Context 2: Small life insurer who is considering improving capabilities of their pricing system to allow machine learning models to be used

Always remember the business context!

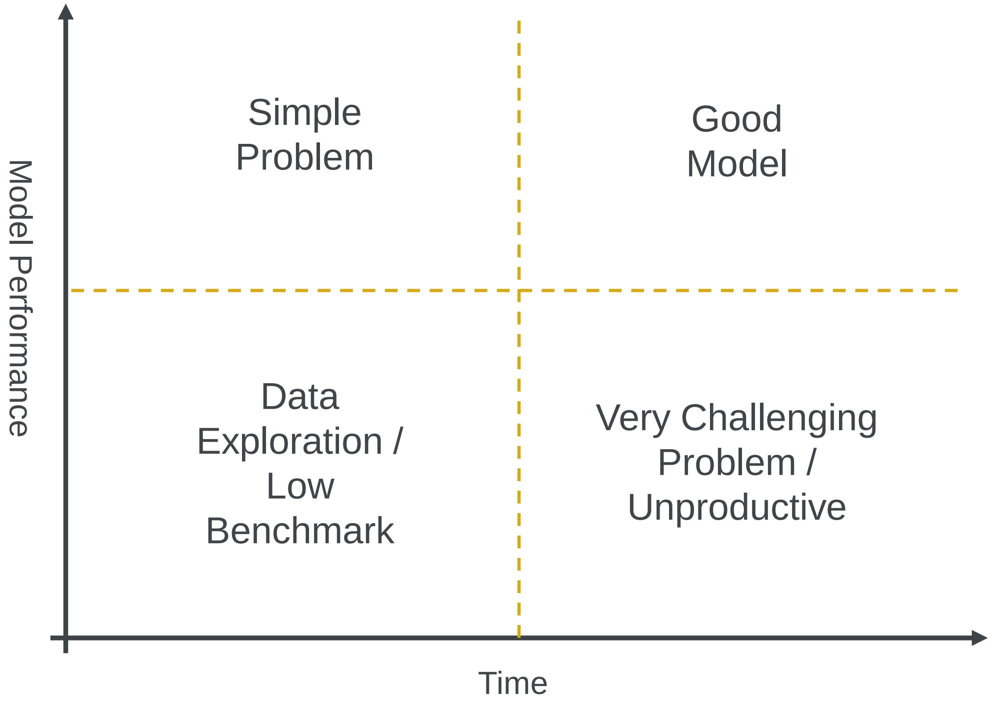
Institute and Faculty of Actuaries

Institute
and Faculty
of Actuaries

Final Thoughts

26 October 2018

Identifying Projects



Data Science Risks

Macro Risks

- Widening inequality as a result of automation
- Not enough junior staff being trained
- New staff unfamiliar with ‘the basics’
- Increased risk of data breaches – GDPR

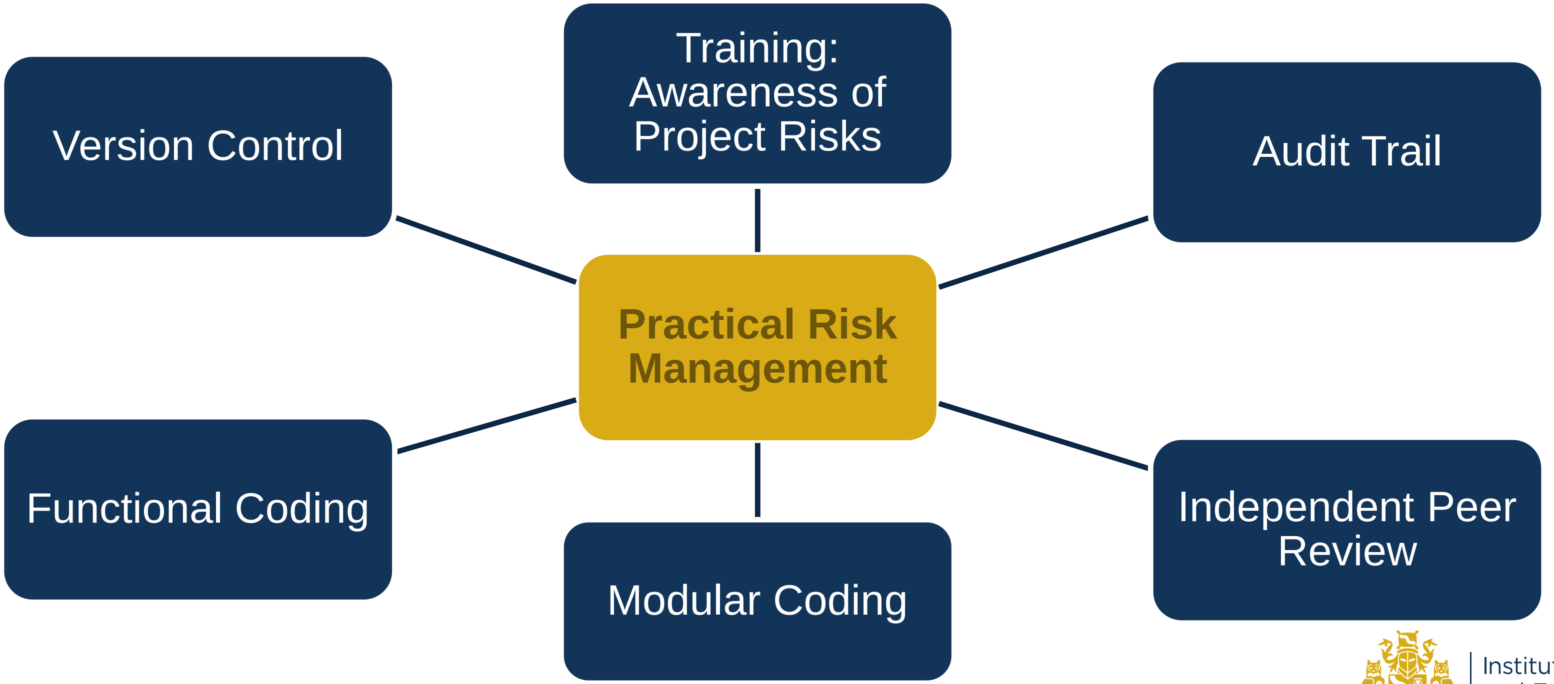
Micro Risks

- Building models which are poorly understood
- Actuarial models built by individuals with little / no actuarial knowledge
- Using incorrect, inappropriate or otherwise flawed data
- Actuaries reviewing coded models vs spreadsheet
- Models appraised out of context

Algorithmic Risks

Institute and Faculty of Actuaries

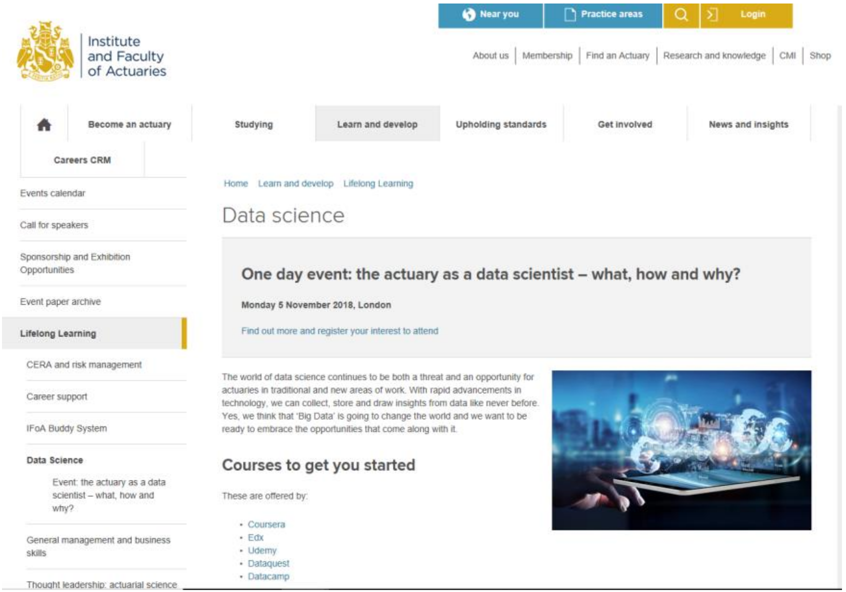
Data Science Risk Management



Institute and Faculty of Actuaries

Upskilling

- Practise!
- Lots of code and examples online
- ‘Point and click’ software available
- Data science member interest group
- IFoA lifelong learning area



Institute and Faculty of Actuaries

Questions

Comments

The views expressed in this presentation are those of invited contributors and not necessarily those of the IFoA. The IFoA do not endorse any of the views stated, nor any claims or representations made in this [publication/presentation] and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this presentation.

The information and expressions of opinion contained in this publication are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this presentation be reproduced without the written permission of the IFoA.

Institute and Faculty of Actuaries