Institute
and Faculty
of Actuaries

# Practical Data Science for Actuarial Tasks

A practical example of data science considerations

by Modelling, Analytics and Insights in Data working party – New approaches to current actuarial work

Steven Perkins, Hazel Davis ,Valerie du Preez

January 2020

**Disclaimer**

The views expressed in this publication are those of invited contributors and not necessarily those of the Institute and Faculty of Actuaries. The Institute and Faculty of Actuaries do not endorse any of the views stated, nor any claims or representations made in this publication 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 publication. 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 publication be reproduced without the written permission of the Institute and Faculty of Actuaries.

# Abstract

The increased presence of data science in financial services will mean that many actuaries will have some level of familiarity with the basic concepts behind machine learning. However, it remains a challenge for actuaries to integrate these new techniques into their work. This may be as a result of:

- Difficulties obtaining and working with large datasets;
- Challenges learning the technical skills required.
- Challenges to validating and communicating results from models due to their 'black box' nature;

This paper aims to address elements of these issues by firstly providing commentary around the key stages of the data science process and the considerations which should be applied by actuaries. This process will then be illustrated via a detailed worked example which attempts to predict future death rates by region for England and Wales using open source data.

Model validation is a key element of actuarial work and may be one of the areas in which standard data science approaches may fall short of the regulatory requirements which actuaries often have to adhere too. Even in instances where models are not being heavily regulated it is still often the case that actuaries will have to justify their models / conclusions to key stakeholders who will generally require a level of justification for the approaches taken.

The case study in this report will provide details of the steps taken to validate the final data science model produced.

*All opinions presented in this report are those of the authors and do not represent the views of the Institute and Faculty of Actuaries (IFoA) or any other persons or organisation.*

# Contents

# 1 Background

Data science is a growing field of interest for actuaries with some of the potential benefits being:

- Improved data quality and quantity;
- The ability to use new data sources;
- The improved speed of analysis;
- Improved model accuracy via new modelling techniques;
- Improved data visualisations;
- The ability to approach problems in new ways.

In previous work produced by the Institute and Faculty of Actuaries' Modelling Analytics and Insights from Data (MAID) working group we have highlighted that the new approaches to modelling offered by adopting a data science workflow closely resemble the modelling stages which are used by an actuary as part of the actuarial control cycle (Bellis, 2006).

Building on these observations this paper aims to demonstrate some of the advantages above in a practical case study whilst also setting out some of the key considerations of a data science workflow. In particular how to frame a problem, collect data, build models and then validate and communicate the results in an appropriately robust manner.

Chapter 2 provides some background on the key stages of problem specification, model development and model deployment within a project. Whilst there are many considerations and approaches which can be taken to delivering a robust data science workflow, the aim of this chapter is to review some of the key considerations which should be applied at each stage of the modelling process.

Chapter 3 is then designed to provide an overview of how these steps can be applied in practice. This overview is provide through an end-to-end data science workflow which includes defining a problem and then proceeding to develop a machine learning model which can be used to make predictions against future observations.

# 2 The Data Science Process

## 2.1 Problem Background

The Actuarial Control Cycle (ACC) begins with 'Specifying the problem', which is done within the context of the general commercial and economic environment. Similarly, consideration of the problem background and the business environment is critical to the success of a data science exercise, in order to identify the desired outcomes and understand the factors to be analysed. To truly understand the underlying problem being investigated, the actuary should ask:

- Why is the problem being investigated?
- What is the business reason for tackling the problem?
- What is the objective and what outputs must be delivered?
    - E.g. is a predictive model required or just a summary of findings?
    - Should the focus be on enhancing the process, predictive power, speed of analysis, interpretability of results or a blend of the above?

The answers to these questions will guide the way the problem is specified, what solution is developed, and what outcomes are monitored. For example, a regulatory requirement may have strictly specified outputs. A known weakness in a current model may steer the solution to a particular new model type or new data source. If the business value of tackling the problem is expected to be substantial this could potentially justify a larger project with more required resources, whilst an urgent issue might require a quick solution that precludes certain approaches. The problem itself may be an initial piece of analysis to inform priorities or the direction of future work.

As in most projects, the desired timescales for both the modelling work and the ultimate delivery of an implemented system will influence the choice of approach. The resources available and constraints imposed, whether human, IT, or data availability will need consideration. If the task will need to be repeated or updated frequently then allowing for this in the design upfront will save a lot of time later on. Understanding the end user and any ongoing support requirements for the developed solution can help achieve success with the project.

The stakeholders for the problem under investigation will also influence the solution. For example, the technical knowledge of the audience will affect how important the interpretability of results is. Knowing who is engaged (or not) in the project can also help make sure the solution is politically acceptable to the business.

Machine learning techniques can be applied to a wide variety of business areas. The ACC is also extremely transferable. This means that actuaries are well placed to use their skills to help provide answers to a variety of business problems that may not seem immediately 'actuarial' e.g. propensity analysis based on customer behaviour or identification of suitable products to meet customers' needs. Communication of the findings should be tailored to the audience. This becomes particularly important where the models are being used in areas outside of those traditionally actuarial, where stakeholders may not be used to analyses of this nature.

The rise in popularity of machine learning techniques means that they can become a 'solution in search of a problem' – stakeholders may want to be seen to be using the latest tools without fully considering whether the problem is suitable for machine learning. If insufficient or poor quality data is available, then machine learning projects would be inappropriate even if stakeholders wish to employ these techniques.

The clearer the objectives are the more likely it is that a suitable model can be built which will deliver useful answers. Understanding what level of accuracy is required can help select the right data, appropriate techniques and specify the problem.

Understanding modelling limitations, dealing with uncertainty and communicating this with the results are also vital to delivering a successful actuarial project, and this is particularly true of a machine learning project. This is considered in more detail in the following sections.

Finally, the ethical considerations around the proposed task should also be reviewed to ensure that stakeholders and society are not adversely impacted by the project. Whilst a full discussion on the ethics surrounding data science projects is beyond the scope of this paper it is important to note that the ethical significance of data science and the implications for industries and the wider public are constantly evolving. It is important to keep up with the challenging associated ethical issues, which has led to the partnership between the Royal Statistical Society (RSS) Data Science Section and the Institute and Faculty of Actuaries (IFoA) on the practical and ethical implications of data science (Institute and Faculty of Actuaries News, 2018) (IFoA and RSS, 2019).

An ethical charter is being developed and this guidance focuses on principles of data ethics and discusses ways of considering these within data science work and we would encourage actuaries to consider such principles when reviewing a proposed data science task.

## 2.2 Problem Specification

One of the challenges actuaries and data scientists will often face is the ability to translate a qualitative description of a particular problem into a suitable format to be tackled by modelling approaches. This can be challenging for a number of reasons and there can often be more than one valid approach.

At first, actuaries may be inclined to fall back on traditional, standard, established approaches to solving a problem, and for good reason: a reliance on established methods can often provide reassurance to key stakeholders and utilising established models can often increase delivery speeds, at least until new approaches have been learned. However, by taking a step back and considering the process behind the problem a deeper understanding can often be gained, and potentially better models can be produced (Loser, June 2018). In the context of utilising machine learning the following key questions need to be considered.

### *Data*

A key consideration in all stages of development of a machine learning model is understanding the data available. Considering the data at the earliest stage, potentially before any significant modelling work has been performed, can help shape the overall project by:

- Identifying suitable sources of existing data;
- Considering the additional data which might be available;
- Setting up new data capture capabilities for current and / or future tasks;
- Determining where data quantity and/or quality is sufficient to justify further detailed modelling work.

Understanding the data available for a task can be performed in three fundamental ways:

- A top-down approach involves creating a wish list of data items to solve a problem before attempting to identify the potential sources for each data item.
- A bottom-up approach involves reviewing known data sources to identify known data fields which are suitable for the task.
- A blend of a top-down and bottom-up approach.

In the early stages of a project it is important to determine whether the data available is sufficient to carry out the task to ensure that time isn’t wasted on a project which is not viable.

### *Model Structure*

Supervised learning

Many machine learning tasks involves a scenario where the user has a number of input variables (or ‘features’) as well as a target variable which needs to be modelled. Tasks of this nature naturally fall into the category of supervised learning and the problem can be approached in a relatively standardised manner. The key elements to establish will be:

- What target variable is most appropriate to model?
- Will the problem by classification or regression?

Sometimes these questions will have clear answers: for example a cancellation propensity model would have a target variable indicating whether or not each policy in the dataset had a cancellation. This would naturally fit the definition of a binary classification task, where each policy is ultimately predicted to be in one of two states. A similarly well-defined task may be predicting the life expectancy of a pension scheme member – in this instance the task would become a regression problem because the outcome would be a continuous variable reflecting the predicted life expectancy.

In certain cases, particularly where the problem is more loosely specified, there may be more than one possible definition of an appropriate target variable which could be used. It will then be vital to consider which target variable best suits the problem. However, the speed of applying and running various machine learning models may mean that such models can be built using a range of target variables and then assessed against a unified set of performance metrics at the end.

Unsupervised learning

Whilst the above supervised tasks are likely to be the most common problem actuaries face there can be instances where there are is no target variable or where there are limited historical examples of the target variable. When there are no examples of a target variable the task is likely to be aiming to group observations based on their similarity to one-another. In this instance the task is likely to be a candidate for unsupervised learning.

Cases where there are limited observations with known outcomes (labelled data) but which form part of a longer dataset with many unlabelled data examples may be candidates for semi-supervised learning methods. These utilise the target variable where this is available but then look to also gain an understanding of the problem by also utilising the unlabelled data.

Finally, whilst all these approaches are presented as separate methods for well-defined problems it can also be the case that methods can be nested to enhance the overall output. For example, an unsupervised learning model could be used to create a grouping which may serve as an input variable to a later supervised learning model. The general approach of combining the output from a range of machine learning models is widely referred to as 'ensembling'.

### *Implementation*

Another key consideration is the context and delivery method for any output from the task. If the machine learning model being produced will be built and maintained locally by an actuarial / data science team there may be few limitations around the methods and data which can be used. However, in many cases models will need to integrate within the wider business systems and hence it is important to understand the capabilities of those systems. Machine learning allows complex models to be built quickly but if the systems utilising those models cannot cope with particular model structures then those models will not be suitable outputs. Similarly, there may be a requirement from key stakeholders for a minimum level of transparency for the final actuarial models produced which may preclude certain model formats.

Nevertheless, in these cases less complex machine learning approaches may still be sufficient and the more complex approaches can serve as a benchmark, allowing a quantification of the reduction in performance a less complex model observes.

### *Modelling detail*

The time spent on a modelling task can vary widely. Factors such as the experience of the practitioner will clearly be important. However, it is also possible for a task to be significantly extended by simply spending longer on the data collection to model validation stages set out in Sections 2.3-2.6. Additional time spent in these areas can often improve the final model but the marginal gains made may not justify the time spent performing the task and a quantification of this marginal benefit is often not known at the outset of a project.

It will therefore be important to determine the level of detail which is required for a modelling task as this is likely to impact all future stages of the modelling process. This will be impacted by:

- The task requirements from the project stakeholder;
- Deadlines for output;
- The materiality of the model to the company.

Considering the ultimate purpose for which models are being created will often naturally dictate the extent of the modelling work which should be performed.

### *Regulation*

The final note to consider as part of the specification of the problem is any regulations which will affect the model. This may restrict the data which can be used for modelling purposes (eg. GDPR (Information Commissioners Office, 2018) or gender directive (Financial Conduct Authority, 2012)). Similarly, there may be transparency requirements around the final models produced as well as other legislation impacting the model. Whilst this will not always be the case it is important that planning a modelling task includes an understanding of the wider environment and key stakeholders which the output of the project will be impacting.

## 2.3 Data Collection

Data is the pre-requisite for most modelling endeavours, even more so when utilising machine learning techniques. The quantity and quality of the data used to train and test the model underpins its robustness and the veracity of its output. Exploring what data is available and acquiring the data are often the bulk of the work.

> *"It is a capital mistake to theorise before one has data. Insensibly, one begins to twist facts to suit theories instead of theories to suit facts."* – Arthur Conan Doyle (1891) *A Scandal in Bohemia*

Considerations around data collection, in the context of developing a machine learning model, are discussed under the following broad categories. These are high-level heuristics to consider; specific considerations depend on the nature of the problem the project aims to solve.

- Data quantity – is sufficient data available for training and validating the model?
- Data quality – is the data reliable?
- Data acquisition – what legitimate sources of data are available?

### 2.3.1 Data quantity

Machine learning techniques did not emerge overnight; they are a culmination of decades of advances in statistical methods and computer science. Its popularity in recent years is in part driven by the explosion of data – this provides a two-fold impetus to the adoption of machine learning techniques.

Firstly, the so-called "big data" necessitates the use of advanced techniques. Big data is characterised by the following:

- Volume – the scale and magnitude of data, enabled by inexpensive digital storage;
- Velocity – the speed at which data is generated, enabled by internet-enabled data generating devices e.g. mobile devices, wearables, and sensors;
- Variety – the source and type of data, enabled by the digitisation of information e.g. text, images, audio, and video.

Secondly, the increased quantity of reliable training data has boosted the performance of machine learning models.

At the earliest stages of a project, the key consideration is whether there is sufficient data to train *and* test the model, specifically:

- The training and testing samples should be sufficiently large and representative; and
- The input variables or "features" of the model should have sufficient explanatory power.

Machine learning projects are iterative by nature; an assessment of the above will also be carried out at the later stages of the project (see Sections 2.4, 2.5 and 2.6 for details of these assessments).

#### 2.3.1.1 Sample size

What constitutes a "sufficient" sample size is project specific. However, there are a number of rules-of-thumb to use as guidance.

##### Type of problem

The minimum sample size required for a multiclass classification (a classification with more than two distinct classes present in the target variable) problem is higher than a binary classification problem (a classification problem with exactly two distinct classes present in the target variable) or a regression problem (see Section 2.2 for a definition of these problems). This is such that the samples are representative of each class or category. If there are not enough representative samples, a possibility is to reframe the problem as a binary classification problem. Alternatively, over/under-sampling techniques could be used to achieve a representative sample size for each class (see Section 2.5).

##### Type of machine learning algorithm

Algorithms that aim to capture complex non-linear relationships between the input variables and target variables generally requires more training samples. To get an estimate of the sample size required, publications from practitioners (in the industry or in academia), who have utilised the algorithm(s) under consideration, is a good source of reference. At a later stage of the project, the impact of sample size on model performance can be quantitatively assessed (see Sections 2.5 and 2.6) by utilising a "learning curve".

##### Number of input variables

The minimum sample size required is often directly related to the breadth and complexity of input variables (or features) being considered. Typically a modelling problem which only has a small number of lower complexity input variables will require fewer historical observations than the same problem with a high number input variables and / or highly complex input variables. This is generally referred to as the curse of dimensionality. In the instances where a dataset has a high number of input variables a data scientist will initially aim to apply dimensionality reduction techniques to ensure the breadth of the data is better suited to the number of historical observations available for modelling.

#### Statistically-significant sample size

When estimating the mean of a distribution using an observed sample the Central Limit Theorem can be used to define the margin of error of the estimator as well as determining the minimum number of observations to produce a sufficiently narrow confidence interval (Billingsley, 1995).

Due to the complexity of a typical machine learning task the requirement is often to estimate a number of mean outcomes for a whole range of different cohorts within the data. Therefore when using data science approaches practitioners typically adopt a standard training-holdout split rule (e.g. 80:20 rule of training data to holdout data). Using a subset of your data for independent validation can reduce the bias within the models developed, however it does potentially require additional data to be collected or for the model to be developed using a smaller training sample which will increase the variance of the model outputs.

Whilst the above points provide a key consideration when selecting a modelling approach there is no clear, quantitative guide for when and where machine learning approaches will outperform traditional approaches, or for the specific sample size required for certain problems.

#### 2.3.1.2 Input variables

At the early stage of the project, domain knowledge of the problem at hand is required to ascertain the input variables that would be useful. Input variables could come in the form of structured or unstructured data. For example, in the case of an interest-rate forecasting model, text from central banks could be incorporated as input variables in addition to structured data such as inflation, amount of public sector debt, GDP growth rate etc.

It is worth noting that a greater number of input variables does not necessarily lead to improved model performance. Indeed, in an enviable but double-edged sword case of “too many” input variables, feature engineering is required (see Sections 2.5) to reduce the number of input variables.

At the modelling stage, the explanatory power of the input variables is quantitatively assessed and is often used as one of the criteria in choosing the final model.

### *2.3.2 Data quality*

In addition to being sufficient, data has to be reliable and fit for purpose. Data quality is characterised by the following aspects.

#### *Accuracy*

The accuracy or ‘ground truth’ of the target variables used in supervised learning is critical if the model is to perform well in the future, or the model will learn to predict incorrectly (as it will learn to reproduce inaccurate outputs). The accuracy of the explanatory variables is desirable, but if these are inaccurate in a consistent manner, then they may still be predictive. However, any change in consistency (either an improvement or degradation) could then cause a systematic bias in model outputs until the model is retrained.

#### *Completeness*

Missing data is undesirable, but does not necessarily render a field useless. Some model forms handle missing values better than others. It is often worth flagging missing variables and checking how they are treated. In some models you may expect to see the data with a missing value appear to behave as a weighted average of the rest of the data – indicating the cohort is made up of a typical mix of the possible known values, and they can be treated as such. In some data sets you may wish to fill in the missing values, and you might take the modal value (assume you are correct most often), a mean (assume correct on average). If the reason for the value being missing can be identified e.g. it is an older subset of the data before a question was asked, then this may determine the best way to treat the subgroup. Alternatively, it may be possible to build a model to predict what the true value of the missing field is from the other known information about that data point.

#### Timeliness

Understanding the latency of the data can help inform the usefulness of a model built with it. For example, better data may be available with a 6 month lag, but if the conditions are moving quickly you might build a more useful model on 1 month old data.

#### Consistency

If past data is used to predict the future then consistency is important, as described earlier. It could be that a change in the format of the data could break the model, or it could be that fields that look the same but that are being used differently over time mean that less predictive value can be derived from the variable.

### *2.3.3* Data acquisition

The data will need to suit the project, so useful data sources will vary depending on the objectives. For example, a requirement to deploy a model for use in the future will mean the data sources used for modelling not only need to provide useful historical information but will need to be accessed on an ongoing basis, potentially in real time. Understanding what is implementable will allow effort to be focussed in useful areas.

Data sources may be internal or external. The response variable and the many potential independent variables may well come from in house data, however as data science and machine learning techniques open up new ways to use data, even familiar sources may require new strategies for collecting data.   e.g. text mining the comments on claims forms, using metadata, wider sharing of data between departments that have had less interaction historically.

External data could be provided by a third party and used in a contractual manner. This may mean that the frequency of data updates is agreed, fields are well defined and the data quality is understood and guaranteed.

External data may be publicly available and provided freely for use by anyone, for example the UK government publishes data from central government, local authorities and public bodies to help people build products and services at data.gov.uk and the Office of National Statistics can be a rich source of public data, similarly free for use.

Alternatively external data may be obtained by the analyst from publicly available sources, but gathered in a way that is less conventional and potentially illegitimate. For example, web scraping is one way to obtain data that is available freely on the internet, but care should be taken that it is within terms of use of websites. Companies may have policies prohibiting such practices to avoid reputational damage, and actuaries should consider whether it is professional to proceed if it could be deemed inappropriate or illegal.

Big data also encompasses a greater variety of data types including images, audio and videos – these will need a collection strategy that suits these files as well as storage capability to store the potentially larger file sizes than might be required for more traditional data items.

If a model is to be used in the future then the ongoing availability of the relevant data inputs, and how to implement the model so that it performs as designed must be considered.

If a model that is built on historical data from a range of sources is required to make future predictions, then the range of sources must continue to be available in the future, and the model will need connecting to these data sources in an efficient manner. Depending on the nature of the model, this could require for example, real time linking to external databases that are updated automatically by some third party, or it could be that the original data collection process will need repeating in order to refresh the data to keep it up to date.

The analyst may therefore wish to limit the modelling to consider only data sources that are known to be implementable, or it may be that the scope of the project will include IT infrastructure type work to take care of future operations. The effort needed to keep the model functioning once built should not be overlooked.

Other key data considerations include:

### *Accessibility*

The effort needed to implement each data source may vary and be deemed as an appropriate investment or not depending on the predictive power of factors from that source. These may not be known until the modelling is complete. One advantage of machine learning is that the models can be built quickly, so they could be rebuilt without some factors if implementation issues are discovered later – however, if this could be pre-empted then some wasted effort can be avoided.

### *Quality*

If a data source is known to be of good quality and consistency, and known that it will exist into the future, then the effort needed to implement it is likely to be worthwhile. If the data source latency is good that may be valuable for future modelling, but if the latency is poor then monitoring the quality of your model will be important to check that it does not deteriorate.

### *Reusability*

If you find a source of data that is predictive for one model, e.g. claims costs for one product, can you scale this to use the same data on different products with little extra effort in obtaining the data?

*Scalability*

If you have tested a model on a small subset of data, will the factors be available for the full dataset and can they be used on the bigger batch without e.g. file size/storage space becoming prohibitive.

*Flexibility*

You might start with simple factors from an external data source but if there is potential to enhance the breadth of data used in modelling via powerful feature engineering then it may be a more valuable data source in the future.

*Automation*

If collecting the data is a very manual process then errors may be more likely to creep in and keeping the model working in the future on up to date data is going to require a lot of ongoing effort. The more the data collection processes can be automated, the easier it will be to repeat to keep data up to date or rebuild models in the future. Some processes e.g. web scraping may be automatable but this may break down quickly as website designs change. If a data source is in a consistent location with consistent data names and formats than automating the data collection and processing will be easier.

*Security*

If testing the data source involves personal data – either because the data itself is personal in nature, or because personal data such as DOB and postcode is used to match customers to data sources, then what data is stored, where and how, e.g. who has access to what, what type of encryption is needed, and how long it can be kept for, must be considered in light of GDPR and data security considerations. Moreover, deploying the model might mean the real time sending of personal data between different sources then the security is going to be paramount. People with appropriate expertise should be consulted to set up the data extraction and implementation.

*Compliance*

Just because you can access a range of factors may not mean that you should use them. Consulting with appropriate risk and compliance or legal professionals may be appropriate before considering using certain data sources.

## 2.4 Data Cleaning

### 2.4.1 Data Joining

One of the key risk areas in data processing arises when two datasets are being joined. Even when there is a clear linking key between the datasets this can often be an area where inexperienced analysts make mistakes. Where no clear linking key exists, or there are potential formatting differences between the linking keys in two datasets then the risks of failed data joins further increases.

When joining datasets an actuary should check that the data before and after the data join is as expected. Data joins which are only partially successful may result in new cases of missing data which may need to be investigated and handled alongside any other missing data instances within the dataset.

### 2.4.2 Data Pre-Processing

When building any model the process of reviewing, cleaning and adjusting the input data is often vital to the success of the end model. When building machine learning models this stage in the workflow remains equally, if not more, important than when building models using traditional actuarial approaches. In particular, machine learning models can often be more sensitive to data choices which may not have been explicitly been made by actuaries using more traditional modelling techniques.

*Variable type*

One modelling choice which may be unfamiliar to actuaries used to traditional modelling tools and approaches is the choice of variable type which is required for machine learning models. In particular, data is generally input into models as either a numerical variable (ie. a real number) or a factor variable (a variable with a finite, discrete number of possible different values). Whilst in certain cases it may seem obvious the data input choices often does affect the underlying model and sometimes choices exist where they might not initially appear to.

For example, three colour options (red, cyan, and yellow) would appear to naturally fit the definition of a categorical variable, however these colours could also each be converted to a three dimensional numerical vector using their values from the standard RGB colour model. These values could then be entered into a model as a numerical input rather than a factor variable.

*Factor grouping*

Once an actuary has decided to input a particular variable as a factor, further choices are often present as the various factor levels. But for the previous example the user may subjectively decide to group the ‘cyan’ and ‘blue’ factor levels together based on their known similarity and hence only use three distinct factor levels for modelling. Alternatively they might decide to retain the original four factor levels. Whilst elements of analysis can be used to support the approaches taken there is often a significant degree of subjectivity involved in these decisions.

*Numerical caps and floors*

Whilst numerical fields may appear to be easier to work with initially, modelling decisions also need to be taken for these input variables. In particular, the values of high or low outliers may need to be truncated to avoid the potential issue where bias is introduced into the model from extreme observations. Similarly, if future numerical inputs are not capped then there is a risk that the model developed will extrapolate predictions to new cases which may lie outside of the original modelling data.

*Missing data*

Missing data remains a key consideration for machine learning. Whilst certain machine learning models are capable of processing observations with missing data, many are not. It can also be important for the actuary to understand the causes of missing data and, in particular, whether missing data is likely to occur again in the future. The approaches which can be taken to handle missing data are discussed in Section 2.3.2

### *Feature engineering*

Feature engineering is the process of translating data fields available in a raw dataset into new, additional fields for use in model building. Feature engineering can be carried out using a manual or automated approach, with often the key considerations being any time constraints and whether the engineered new input variables (or features) need to be interpretable by a human. For even a moderate sized dataset there are often a huge variety of feature engineering steps which could be taken and hence this is often an area of the data science workflow which can increase model performance but also significantly increase the time taken to produce the model.

### *Data Partitioning*

Finally, the data scientist will need to partition data so that an independent holdout dataset can be used to validate the performance of the model. However, the selection of the model validation dataset can often be a process which can introduce bias into the model building steps.

Key considerations are whether an ‘out-of-time’ holdout sample should be taken (meaning the most recent period of data would be kept independent of the model building process to validate the performance), a random sample (where random observations are withheld from the model building process) or a more complex / bespoke selection approach is applied to the validation data.

It may also be appropriate to have more than one holdout dataset. For example it is common to use a training dataset and an independent validation dataset to tune the parameters of models, whilst a separate holdout (or ‘test) dataset is often used to determine which of the different model types is most appropriate. However, different approaches are possible, for example if a cross validation approach is taken (see (Panlilio, Canagaretna, Perkins, du Preez, & Lim, 2018))

Finally, it will be important to ensure that the holdout dataset remains large enough that the conclusions drawn are robust.

## 2.5 Modelling

When creating a data science workflow machine learning will typically be used to create a selection of predictive models for consideration. A key advantage machine learning methods have over traditional modelling approaches is the speed and ease with which models can be built and evaluated. Whilst the building and tuning of machine learning models is generally automated, the choice of models can often remain the choice of the modeller. An actuary using machine learning approaches will need to select appropriate model architectures by considering the key features of the models being developed, some of which are discussed below.

### *Model Implementation*

Arguably the most important consideration in the model building processes is a consideration around which models can ultimately be implemented in practice. Typically if a machine learning model is being used to produce new real time predictions (eg. to deliver insurance premium quotes to new customers) then the final model will need to be hosted in a suitable IT system. This system may have constraints over the scope of models which can be utilised.

### *A priori modelling beliefs*

If an actuary has a strong belief that the underlying processes which produced the data has a specific format then this should be reflected in the models being developed. For example, two key types of supervised learning models are linear models and non-linear models. A full discussion of the differences of these model types is outside the scope of this paper, however, the different types of model will interact with the input data differently and, as such, may be able to reflect the underlying processes in a more or less accurate way.

#### *Model complexity and build time*

As a general rule, the more complex a model is the greater the computational time required to develop it. Some of the most complex models for complex tasks, such as image processing and speech recognition, are deep learning models, but typically these models require greater volumes of data and can take longer to train. Therefore if there are time or data volume constraints to consider then these may impact the types of models which should be considered for a particular task.

#### *Parameter tuning time*

One of the key reasons that machine learning approaches can develop models more quickly than traditional approaches is that the model parameter tuning process can be fully automated rather than requiring manual approaches. However, time constraints will often impact the extent of parameter tuning which is performed, which a modeller having to balance the improvements in model performance against the additional time it takes for the algorithm to tune parameters.

#### *Model transparency*

An increasing focus is being given to the transparency of the models produced using machine learning approaches. There may also be regulatory restrictions around the types of models which can be used for certain purposes, such as insurance pricing. The level of transparency often depends on the types of models being used, with basic decision trees and linear models often being easier to interpret for an end user compared to models such as neural networks. However, whichever model architecture is selected it is important that appropriate validation of the models is considered, and this can help increase the transparency of even more complex machine learning models.

## 2.6 Validation

Model validation remains a key component of the data science workflow, just as it is in a traditional actuarial model building process. A basic data science model validation will involve selecting a single performance metric. The typical performance metrics used are designed to produce a goodness-of-fit assessment of a machine learning model by looking at the actual and expected performance of the models produced. This assessment should be carried out on fully independent holdout data which has not been used as part of the model building process. This then provides a quantification of how a model generalises to a new dataset. This approach can be applied to a full range of models developed and then this can be used to rank all the models developed on a leaderboard to select the most appropriate model for a particular dataset.

Whilst the above processes provides an outline of the basic steps for model validation, it fails to answer key questions such as:

- Are there alternative performance metrics which could / should be considered as part of the model selection process?
- Are there specific input variables or combinations of input variables where the model predictions are particularly accurate or uncertain?
- Is there a less complex / more transparent model which performs as well or almost as well as the most accurate model?

- Does the new model outperform any existing model?
- Does the most accurate machine learning model produce the desired business impact?

To answer some of these questions it is generally appropriate to perform further detailed models validations, some of which are described below. This does not represent an exhaustive list of model validation approaches, but it does highlight some of the key model components an actuary should review.

### *Actual vs Expected*

One of the key model validations for both traditional and machine learning models is an actual vs expected plot and the related residual plots. Plotting the actual results against the predicted result from the machine learning model allows an actuary to understand if there are any areas of systematic bias within the models being built. For examples, does the model have a tendency to predict values which are too high for cases where the actual observations are low. Where systematic patterns are identified in the residuals this may prompt a review of the models under consideration, with a model which is less accurate overall perhaps being more preferable than a more accurate model if there are significant systematic patterns in the residuals.

### *Lift curves and gains chart*

Lift curves and gains charts are similar methods to visually represent the performance of a model. For example, for a binary classification problem a cumulative gains chart all observations will be ordered based on their predicted probability of a positive response using the selected model. The cumulative gains chart will then plot the cumulative proportion of actual positive responses received for increasing proportions of the ordered observations. This can allow simple observations such as "the lowest risk X% of the population contains Y% of the positive responses". A model would be deemed better if the lift chart or cumulative gains chart has a steeper gradient.

It is however important to note that lift curves and gains charts do have limitations. For example, a model can have a good visual performance from their gains chart because the ordering of observations is accurate, but the underlying probabilities assigned to a positive response may not be appropriate at an observational level.

### *Variable Importance*

A variable importance plot shows the relative reliance a machine learning model places on specific variables. These should be reviewed so that the modeller can understand how many of the input variables are key to the final predictions and whether or not the variables being used are in line with the expectations. For example, if there was a strong correlation between a particular input variable and the target variable in the underlying data then we should not be surprised to see this specific input variable appearing high in the model variable importance. However, variables that are only weakly correlated with the target variable but which appear high in the variable importance plot may prompt further investigation.

### *Partial Dependency Plots*

Partial dependency plots include the values or factor levels of a specific input variable on the x-axis and the predicted value on the y-axis. These plots can be produced for all input variables to then review if there are systematic patterns in the predictions based on each specific input variable.

Where a systematic pattern in the predictions is observed this should be compared against the actuals to ensure that the prediction patterns is consistent with the underlying data being used.

### *Actual vs 'Transparent'*

Where model transparency can be an issue (eg. for a neural network) it can often be insightful to also create a basic 'transparent' model which can be easily interpreted by the modeller. Plotting the predictions from the selected model against the predictions from the transparent model allows a modeller to:

- Identify cases where the selected model and the transparent model are similar, to gain comfort over the appropriateness of these predictions.
- Identify and review cases where there are large differences between the two approaches. This then allows the user to test the reliability of the selected model in these outlier cases to validate whether it is making more or less appropriate predictions compared to the transparent model.

### *Business validation*

Finally, the technical model validations above will help determine the most appropriate model from a statistical perspective. However, it is important then to consider the commercial relevance of the selected model. In particular, where possible the model should be tested against an independent dataset to understand what the impact would be on other business KPIs so that the key stakeholders can see the impact of using the model on the business. Where the new model is replacing an existing model then there may also need to be a review of the expected changes in outputs under the old model and the new model to understand the impact of the new model on a case-by-case basis.

Overall, whilst the leaderboard approach described at the start of this section may be an appropriate method for narrowing down the final models to be reviewed, the final model selection will often rely on some of these more detailed model validations to ensure that the final model meets the end users requirements.

## 2.7 Reporting

An effective actuary needs more than technical actuarial and modelling skills in their toolkit. The efforts of all of the work that has gone into developing the models, projections, analyses or estimates will only be effective if the findings lead to good decisions, actionable insights and appropriate solutions. This would depend on whether the actuary has understood the question being investigated and also whether the findings are communicated in an effective way to the end user so that the advice is interpreted appropriately.

Communicating the findings of the machine learning case study in an appropriate manner, keeping in mind the aims of the exercise as well as the end user (or reader) is critical. Ensuring the concluding message is clear, and the limitations of the findings explained, would enable relevant and appropriate use of the findings.

The requirements set out above are largely the same for all actuarial work and therefore the reporting requirements around machine learning models remains similar to the reporting requirements actuaries face around other models. In particular, we would still expect actuaries to produce and report findings in a manner which is compliant with their usual professional requirements, such as the UK Technical Actuarial Standards. This includes elements such as clear communications of model results and limitations.

There are however some specific additional considerations which may form an appropriate part of future reporting of results from machine learning models.

### *Model selection*

A traditional actuarial approach will require an actuary to create a single model, with the results or the model itself then presented to the end user. A robust data science workflow will allow an actuary to not just consider a range of model types from the conceptual perspective, but actually develop and test a range of modes. There is therefore an additional reporting consideration for the actuary to outline how and why a particular model was deemed to be the most appropriate for the specific task.

### *Assumptions and judgements*

Traditionally reporting on model assumptions would concern the disclosure of key parameter choices made by the actuary when they created their model. When building machine learning models the actuary doesn’t tune / select individual parameters in the same manner. Instead the key assumptions and judgements are more likely to reside in the selection and processing of the data, the range of modelling techniques tested and the extent of parameter tuning applied. It is therefore important the judgements around these key areas are also appropriately communicated to the end users, in line with TAS requirements.

### *Modelling tools*

Given the breadth of tools available for data science work it may become more of a requirement for actuaries to disclose the tools they are using to develop machine learning models. This is because, for example, the benefits and risks which models create may vary depending on the specific tools used to develop the selected models.

### *Testing approaches*

Where machine learning models are being developed and deployed in programming languages rather than more traditional software (eg. Microsoft Excel) it may be appropriate for the actuary to outline the testing regime which has been applied to validate the model performance and the quality of the coding which has been used to develop the model.

## 2.8 Monitoring

In previous work the MAID working party has highlighted that whether data science or traditional methods are used the actuarial control cycle remains highly relevant to the model building process (Panlilio, Canagaretna, Perkins, du Preez, & Lim, 2018). A vital part of the actuarial control cycle is the monitoring section as this ensure that a model performs in line with expectations and remains suitable for future applications.

Monitoring approaches for machine learning models remain similar to traditional actuarial approaches. Broadly speaking the monitoring of a model can be broken down into two main sections:

A. Initial monitoring
B. Ongoing monitoring

The remainder of this section will consider these phases of model monitoring. Many of the considerations will be familiar to actuaries. However, there are potentially additional considerations which can arise when monitoring machine learning models which are highlighted below.

### *Initial Monitoring*

Initial monitoring of a model is often concerned with ensuring that a model is performing in line with a handful of key performance metrics (KPIs). In certain cases, such as insurance applications, it can be particularly challenging to monitor initial model performance due to delays in receiving accurate feedback. For example, an actuary would ideally assess a new risk pricing model by understanding how well it predicts claims costs, but there could be a significant delay in reporting and settlement of claims which may mean that other metrics need to be used initially to understand performance of these models.

Another key consideration in the initial phases would be identifying any potential issues which might be generated by the model. In particular are there any errors or anomalies being produced which would need to be monitored on a case-by-case basis.

Similar to model validation, this sort of case-by-case monitoring can often be more challenging when using machine learning methods. Machine learning models are designed to identify patterns and correlations within data on an aggregate basis but it can often be difficult, if not impossible, to understand the reasoning behind any single output. A related point being that many machine learning models are non-linear models. This contrasts with traditional linear techniques, such as GLMs. A key challenge behind non-linear models is that a small shift in a single input field can sometimes generate a significant movement in the final output – a phenomenon which is generally avoided with linear methods.

### *Ongoing monitoring*

Ongoing or longer-term monitoring is carried out for two purposes: firstly to ensure that the model continues to perform in line with initial expectations and, secondly to monitor model aging and, in particular, identify when a model needs to be re-calibrated or completely overhauled. As part of this modelling it will be important to set out the triggers for future model updates and for complete re-modelling work to be performed.

Actuaries have also become increasingly comfortable building and maintaining basic regression models and GLMs. For these models it is straightforward to re-calibrate the outputs by making manual adjustments in a transparent manner to alter the calculation process. Whilst certain machine learning models retain this level of transparency, many more powerful techniques do not have this advantage and therefore it will often be more challenging to make minor subjective adjustments to the model itself.

However, one of the potential advantages to utilising machine learning techniques is the speed at which models can be built and this can also impact decisions to re-model. Using traditional approaches may require a substantial process to update a model and hence minor adjustments may instead be preferable. However, with machine learning models it may be preferable and just as efficient to perform a full remodel. It is then straightforward to place the updated model into a direct head-to-head with the existing model to see which performs best on independent data.

### *Machine Learning for Monitoring*

The final note of this section is to highlight that utilising machine learning for initial predictive modelling may only be one benefit. Another option for future developments is to utilise machine learning to monitor other models. For example, a particular machine learning model may be put live with a suite of KPIs being monitored. Rather than an actuary directly monitoring these KPIs a separate machine learning model could be designed to detect anomalies in the reported KPIs. The idea being that a wider range of KPIs can potentially be monitored than might be the case manually, with push notifications being generated when models perform outside of their expected levels, prompting a potential human intervention. This would be a more advanced application of machine learning within the actuarial environment, with potentially significant time savings.

### *Dynamic Models*

The final point to note in this section is that certain machine learning models can be ‘dynamic’. In this case ‘dynamic’ models mean that they can respond to the environment around them and potentially change future outputs based on the environment. A real world example of this is the Alpha Go computer programme (Silver, et al., 2016) which continues to learn and evolve each time it plays a new game. This approach can have advantages when trying to manage the model aging process but can provide additional challenges for both initial and ongoing monitoring. In particular, historical model predictions may no longer be relevant to benchmark current model output (on a case-by-case or an aggregate basis). In this instance, it would therefore be important for an actuary to monitor and attempt to understand where (and ideally why) a model is changing.

# 3 Case Study

In this chapter we provide an illustrative summary of the key stages of the data science workflow in a specific case study. Note that the context of this problem means that no formal reporting or monitoring was required and hence these sections have been omitted. We do however note that these remain important elements in data science process as set out in Section 2.7 and 2.8 respectively.

## 3.1 Problem Background

Mortality rates in the UK have been falling for a number of years (Wong-Fupuy & Haberman, 2004) and the mortality tables which are produced by actuaries are heavily used in both the pensions and life insurance practice areas. Actuaries will typically create mortality tables which use an aggregation of national data [is this true?]. However, there is also a recognition that mortality rates vary considerably across the UK and therefore an accurate model which estimates the regional variations of UK mortality could be used to assist life insurance companies in their pricing and reserving exercises.

## 3.2 Problem Specification

The aim in this case study is to understand the regional variations in UK death rate and ultimately to build a machine learning model which can predict the death rate by region in England and Wales in future years. This problem was selected for three main reasons:

1. There is a considerable amount of open source data which can be used to model death rates in the UK;
2. The problem is potentially relevant to actuarial work in both life insurance and pensions;
3. The modelling work and information provided does not provide any commercially sensitive material.

Specifically, the aim will be to build a machine learning model which can accurately predict the death rate in future years for each of the regions of England and Wales using mortality data from the previous years. The regions in England and Wales will be the UK administrative areas, as defined by the UK Office of National Statistics (ONS).

This problem naturally fits into the class of supervised learning problems because a training set will exist which includes historical observations of the target variable (crude death rate) as well as predictor variables (also known as 'features' or 'variables'). The crude death rate is simply defined as the number of deaths per thousand people in the region.

## 3.3 Data Collection

A number of potential datasets and data sources were considered for this problem, however for ease of access and processing the final data used was all sourced from the ONS. The sourced data includes:

- Death statistics, including a crude death rate, by administrative area for 2012 to 2016;
- A summary of hours worked by administrative area, including mean hours worked and percentiles;
- A summary of annual earnings by administrative area, including mean pay and percentiles;
- A summary of population density by administrative area;
- A summary of the age distribution by administrative area from the 2011 UK census.

The 2016 data will be used as an independent dataset to validate the model built using the 2012-2015 data. The choice of data to use in this case study was largely based on a few key considerations:

- Relevance to the underlying problem;
- Ability to link the datasets in a robust and meaningful manner;
- Availability.

## 3.4 Data Cleaning

The cleaning of the sourced data was carried out in two phases. Initially the excel data files provided by the ONS were manually reviewed and condensed into a structured tabular format which could easily be imported and processed using suitable software. It is relatively common for some degree of manual review to be required for new datasets to ensure there is an understanding of the data fields being used.

Following the manual review, data was imported into a suitable environment for pre-processing, cleaning and joining the separate datasets. For this project the data science platform R was used for all data processing and modelling. The data processing took three main stages.

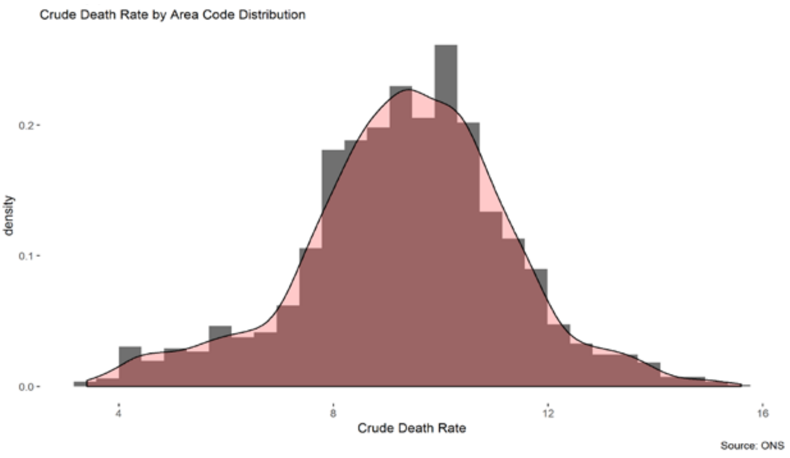
### *Data Joining*

The raw data used was contained in ten separate data tables. The first challenge was to join these tables to create an overall table to use for modelling. To create a robust data join between data tables a suitable linking key needs to be created. In this instance two possible linking keys were used:

- The first key was the UK administrative area code to use when linking data sets by region;
- The second key was the concatenation of UK administrative area code and year, to use when linking data sets by region and year.

The second linking key was also defined as the unique key for every record in the dataset. Joining datasets is one of the higher risk areas of data manipulation and therefore basic reconciliations and checks were performed to ensure the join has performed as expected. Note that not all of the datasets had information for every year of analysis. Where this has been the case, for example with average pay data, the raw data has been joined to every year of the death rate data, and hence there is an implicit assumption that the average pay in a region has not changed over the period being modelled.

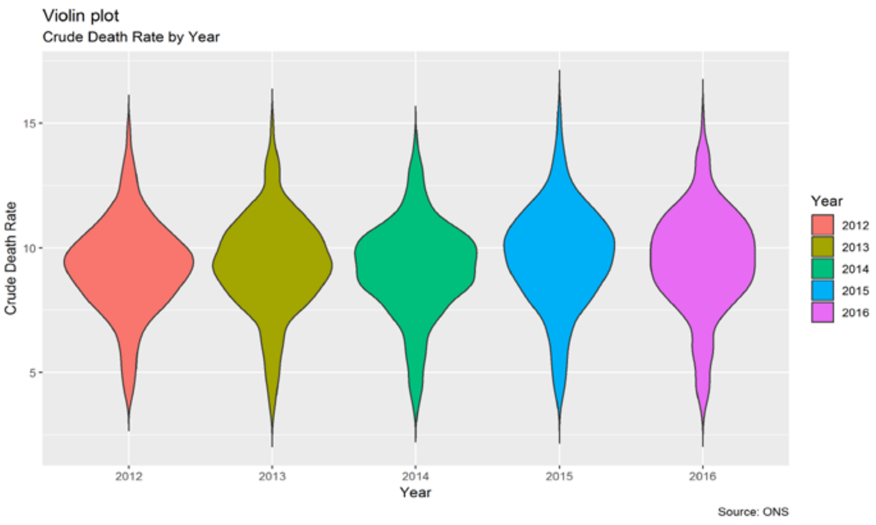
**Figure 1**



### *Data Cleaning*

Once the data was joined the data had to be cleaned. The primary data cleaning which was performed in the pre-processing stage was to review missing data. This is because many machine learning algorithms are not designed to work with missing data and therefore an appropriate strategy is needed to handle these cases. In total the dataset used had 1,956 observations and the missing data was:

**Figure 2**



- 390 cases where 'pay' data is missing;
- 45 cases where 'age' data is missing;
- 5 cases where 'hours' data is missing.

Common options for dealing with missing data include:

- Removing columns with missing data;
- Removing rows with missing data;
- Imputing missing data;

- Restricting machine learning models used to only those which can handle missing data.

In this case various strategies were applied to impute (populate) the missing data using the observations with non-missing data to automatically determine appropriate values for the cases with missing inputs.

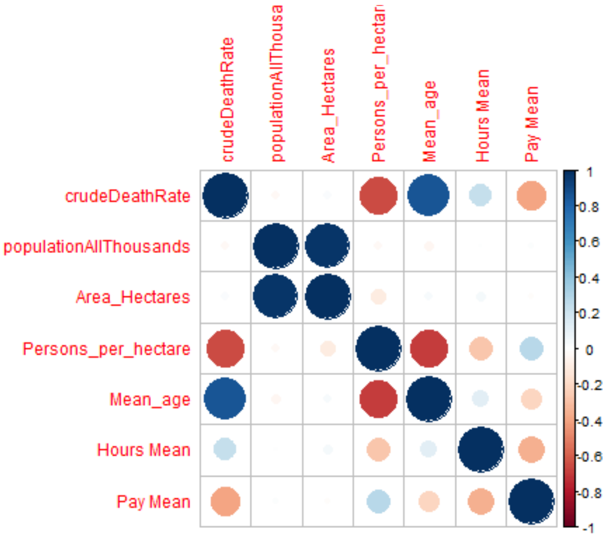
### *Data Review and Checking*

Both before and after processing data it is important to perform a basic review of the dataset. This can be via summary tables as well as graphical reviews. The first data item to review was the target variable. Figure 1 shows the crude death rate distribution for all the different observations by area and year. The key observations from this plot is that there is a reasonable spread of crude death rates with a relatively normally distributed target variable.

Figure 2 once again shows the distribution of the crude death rate by area, however the violin plot also splits out the death rates by year. This demonstrates that the death rate distribution is relatively similar across all years. The bulk of areas have a crude death rate between 5 and 12 deaths per thousand people. There are however some subtle differences by year, with 2015 and 2016 showing a longer upper tail, suggesting that there may have been a slight increase in the extreme crude death rates observed. Further details plots by region confirm that the changes in crude death rate from one year to the next are generally minor on a region-by-region basis.

Having reviewed the target variable similar plots and distributional summaries can be produced for all variables to enhance the understanding of the data. The next stage in reviewing the data is to understand any correlations between the data fields. Figure 3 shows a correlogram for the most significant features in the data. The correlogram shows that there is a strong positive correlation between the average age and the crude death rate by region. Broadly speaking this implies that areas with higher average ages experience higher death rates, which aligns with our understanding of mortality.

Figure 3



There is a weak positive correlation between the average hours worked and the crude death rate. Broadly speaking this implies that areas where people work longer hours on average experience higher mortality.

There is a strong negative correlation between population density (Persons_per_hectare) and the crude death rate. Broadly speaking this implies that areas with higher densities have lower death rates and vice versa. Further investigation of the correlogram suggests that this may be due to a further correlated effect, because the average age in a region is inversely correlated to the population density, broadly implying that younger people live in more densely populated areas.

Finally, there is a moderate negative correlation between average pay and death rate. Broadly speaking this implies that areas with higher average pay experience a lower death rate.

## 3.5 Modelling

The aim of this case study is to predict the 2016 crude death rate by region for England and Wales using basic population statistics. In many cases an actuary would be building a model with the aim of improving an existing model and therefore they would have a benchmark model to challenge. In this case no such model existed but to create an appropriate benchmark a simple model was created which predicted the crude death rate in a given year to be the crude death rate from the prior year. This was selected as the benchmark model.

A series of machine learning models were built and tuned to test whether they could outperform the existing models. The model structures which were considered were: decision trees, gradient boosted machines (GBM), random forests and LASSO regression models. Note that these models represent a small subset of the machine learning models which could have been used, but provide a basic set of models to test for this problem. The proposed candidate model for each of the four classes of machine learning models was selected after extensive tuning of the hyper-parameters of the model. These hyper-parameters provide the controls which define exactly how the models learn from the data provided. The final proposed model for each of the four model classes was the one which performed the best during this model tuning phase.

Each of the four candidate models was built and validated using historical death rate data from 2012 to 2015 with each model then being compared against the independent death rate data in 2016. The initial metrics used to assess the performance of these models were the root mean squared error (RMSE) and the mean absolute error (MAE) defined as:

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

$$MAE = \sum_{i=1}^{n}\left|actual_i \; - predicted_i\right|$$

The aim will be to find the model which minimises both the RMSE and the MAE. The table below shows the results for the machine learning models created as well as the baseline mode which uses the prior year crude death rate as an estimate. However, the table shows that some of the machine learning models outperform the basic model, with both the random forest and the LASSO regression model having a lower RMSE and MAE against the independent death rate data from 2016.

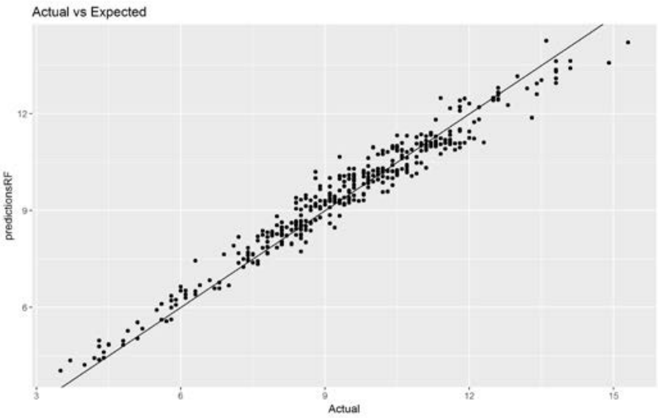
| Model | Prior Year | Tree | LASSO | GBM | Random Forest |
|---|---|---|---|---|---|
| RMSE | 0.451 | 0.540 | **0.426** | 0.471 | 0.438 |
| MAE | 0.300 | 0.353 | **0.277** | 0.302 | 0.281 |
| RMSE vs Prior Year Model | 0.000 | +0.089 | **-0.025** | +0.020 | -0.013 |

## 3.6 Validation

These model metrics in the table above suggest that a machine learning model has been identified which will improve both the basic model metrics. In certain cases, this would be sufficient to accept these models. However, typically an actuary will need to have a greater understanding of their model and hence more detailed validations should be performed. In the rest of this section we have elected to validate the random forest model shown in Section 3.5. We note that this model did not perform as well as the LASSO regression but the validations in this section are designed to highlight methods of understanding machine learning models which are perceived as less transparent, which is why we have focused on a complex tree-based model for our validations.

### *Actual vs Expected*

**Figure 4**

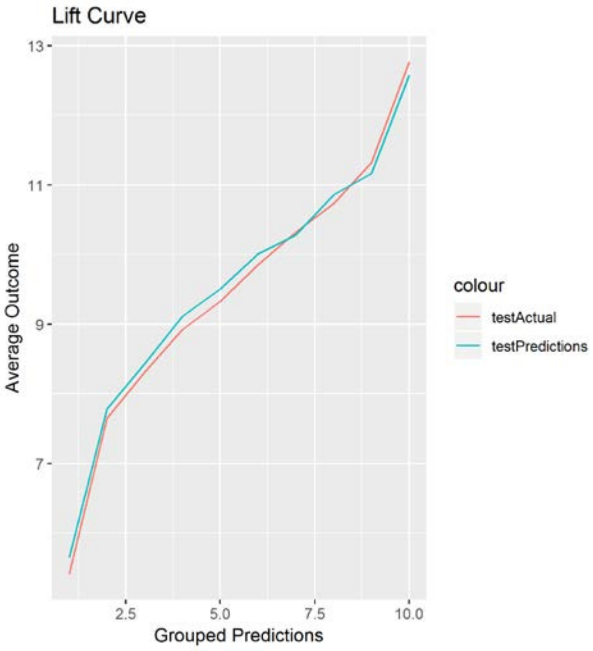


The first validation performed is to consider the actual vs expected plot for the model. Figure 4 shows that the selected random forest model predictions against the actual outcomes for the 2016 data. This demonstrates that the model performs relatively well against the actual outcomes with no obvious outliers being identified. The area with the biggest uncertainty is at the two tails with the model having a slight tendency to under estimate the values where the actual observations are higher (eg. a crude death rate greater than 12) and to slightly over estimate the values where the actual observations are lower (eg a crude death rate less than 6). A residual plot would confirm these observations but has been omitted from this summary.

### *Grouped Actual vs Expected*

In addition to an actual vs expected plot, a variation of this can be produced to help assess the relative performance of the model. In Figure 5 the x-axis contains ten broadly equal size groupings of the ordered test predictions and the y-axis shows the average observed / actual outcome in each group. The plot shows that the actual observations and random forest predictions on the independent 2016 data are relatively closely aligned, supporting the actual vs expected analysis and further demonstrating that this is a relatively strong model.

**Figure 5**



### *Variable Importance*

The lift chart and actual vs expected analysis confirm that the model is generally making predictions which appear to be appropriate. However, typically an actuary will want to have a deeper understanding of *why* a model is making certain predictions. In a traditional GLM framework this will involve reviewing the coefficients of the model, however in a machine learning framework simple coefficients are only available for a subset of models. Therefore it is important to consider alternative approaches to understanding how individual features are impacting predictions.

Figure 6

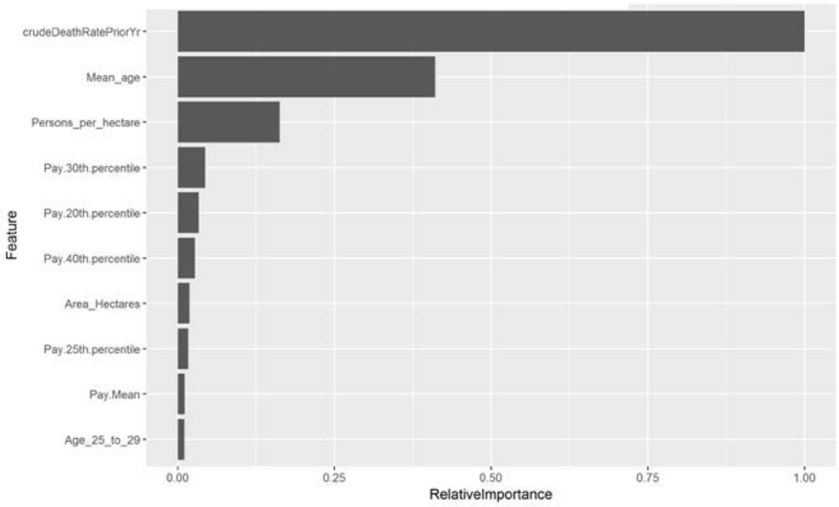


Figure 6 shows a variable importance plot for the random forest. Variable importance is typically available for machine learning model and it demonstrates the relative significance of each input variable on the performance of the model. Figure 6 shows that the most important feature in the model is the observed prior year crude death rate, which would match our prior expectation that this input should be very significant given the changes in crude death rate from one year to the next are generally minor on a region-by-region basis. The other two important features are the average age and the population density features. Linking the variable importance plots with the initial correlation analysis confirms that both of these variables have a strong positive or negative correlation with the target variable, and hence the model has reflected these associations. This confirms that our model is primarily using the three features with the strongest associations with the target variable to make future predictions.

### *Partial Dependency Plots*

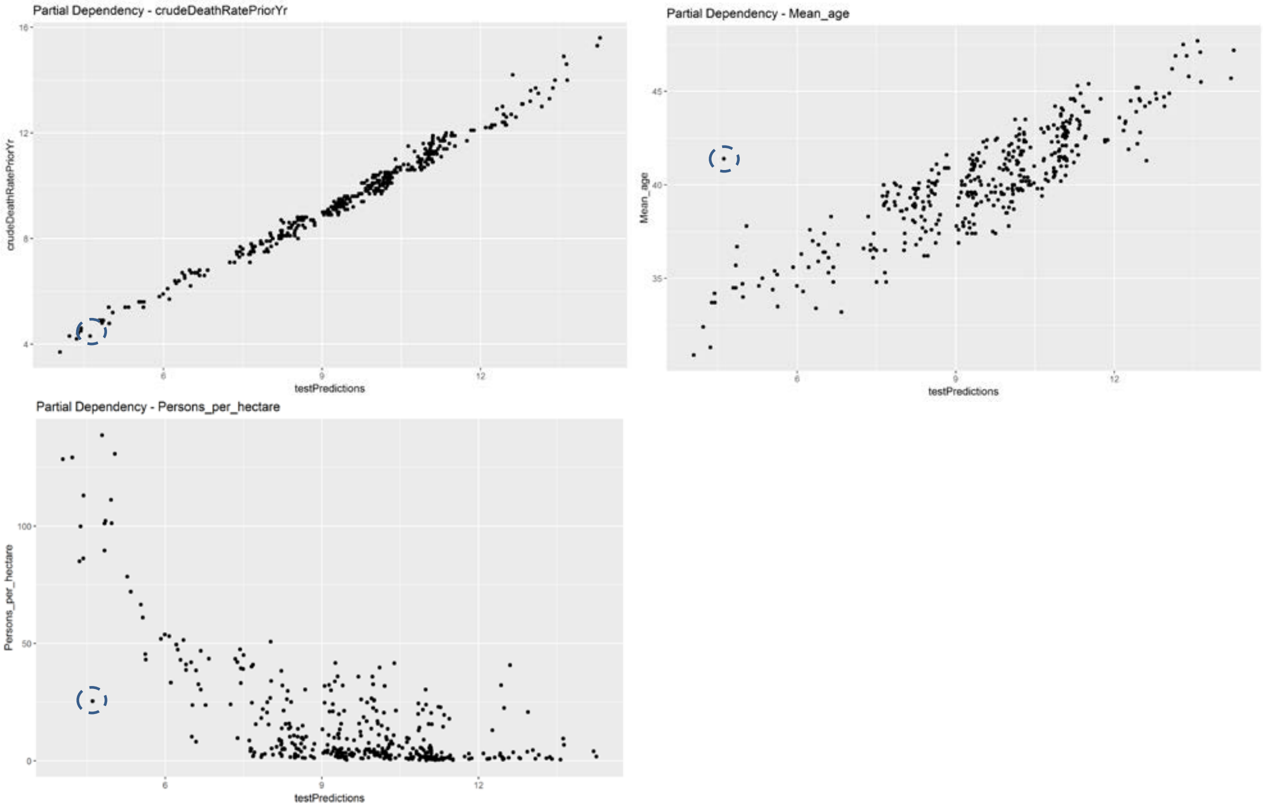
Variable importance plots show *which* features impact the model predictions, however they do not show *how* these features impact model performance. Partial dependency plots can be used for this purpose to understand predictions on a factor-by-factor basis. These partial dependency plots can take many formats.

Figure 7 shows three examples of partial dependency plots for the most important variables in the model, which have the variable values on the y-axis and the model predictions associated with those observations plotted along the x-axis. In this case the key observations are that, as expected the crude death rate predictions tend to be very closely related to the crude death rate in the prior year (top left graph).

The correlation between population density and crude death rate (bottom left graph) appears to be non-linear, with high density areas generally having a relatively low crude death rate but once the population density is below 50 persons per hectare there is a significantly weaker trend being identified.

There is also a strong positive correlation between the average age of an area and the crude death rate predictions (top right graph), however there is one notable outlier which has a predicted crude death rate of around 4.5 but a relatively high average age between 41 and 42. This observation is highlighted in all three partial dependency plots to show that despite having a high average age and low population density the model has predicted a low crude death rate, close to the observation in the prior year. Whilst we could investigate this observation further, the three partial dependency plots demonstrate that the model is performing in a sensible manner, even if this case is an outlier in certain aspects of the data. This review demonstrates that the model is performing in a sensible manner for the three most important variables. Similar analysis can be performed for all input features to determine more broadly if the model is performing in line with our expectations.

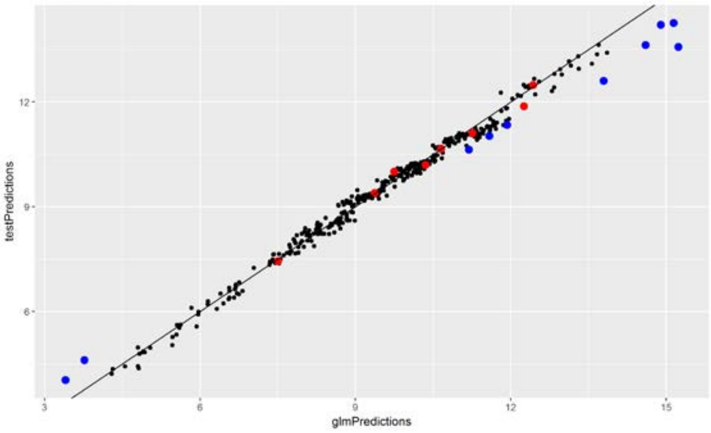
Figure 6



### *Actual vs 'Transparent'*

One specific concern introduced by machine learning is that these models produce predictions that are less transparent compared to traditional actual models, such as generalised linear models (GLMs). An approach which can be taken to improve the interpretability of machine learning models is to compare the predictions against a 'transparent' model which is easily interpretable. Figure 8 shows the comparison of predictions for the random forest model (y-axis) against a 'transparent' model, in this case a GLM (x-axis). This plot shows that the predictions from each model are relatively similar in all cases.

Figure 7



Highlighted in red are the outlier predictions from the random forest model compared to the actual observations. What we see is that the GLM and the random forest both predict very similar values for these observations, suggesting that the cases are simply ones where it is harder for a statistical model to make predictions, and the residual error is therefore similar for both models.

The outliers highlighted in blue are the observations where there is the largest difference between the random forest and the 'transparent' model. This highlights the key differences between the two modelling approaches and helps to focus the review of the random forest. Figure 8 shows that the main differences between the two models appear at the extreme ends, which is likely to be as a result of fewer observations in this part of the distribution.

It is important that an actuary is aware of the limitations of the model rather than relying solely on high-level metrics to determine whether a model is accurate. Therefore where it is the case that different models have different predictions then specific observations should be reviewed on a case-by-case basis to determine which (if any) of the two models is most appropriate to use for these extreme predictions and whether or not any manual intervention should be highlighted for these cases.

# 4 Conclusions

In this paper we have set out a high level overview of the key stages of a data science workflow. In Chapter 2 we have provided an outline of the key considerations an actuary or data scientist should consider as part of this process, though we note that many further levels of analysis and discussion could be applied to these stages. We have then provided an illustrative example of a data science workflow via our case study in Chapter 3. This case study has been designed to be easily replicated using open source data, which can be found via the sources in the appendix.

Overall this paper demonstrates that it is possible for actuaries to build and validate machine learning models in a robust manner. Whilst many of the traditional techniques for obtaining and processing data remain valid, the volume of new data available creates additional opportunities to use data in creative ways. However, it is important to ensure that appropriate controls are put in place to ensure that the end models created are robust. Similarly, many of the model validation approaches traditionally used by actuaries remain appropriate, however, new concepts, such as variable importance plots become important new tools for gaining a deep understanding of model performance.

Finally, we believe that as with all actuarial modelling the extent of the checking, documentation and validation applied should be proportionate to the problem being solved. In certain cases it may be appropriate to apply minimal model validation where in other cases a detailed governance framework is appropriate.

# 5 References

Bellis, C. (2006). Actuarial Control Cycle. *Encyclopedia of Actuarial Science*.

Billingsley, P. (1995). *Probability & Measure.* New York : Wiley.

*Financial Conduct Authority*. (2012). Retrieved from https://www.fca.org.uk/publication/archive/fsa-gender-directive.pdf

IFoA and RSS. (2019). Retrieved from https://www.actuaries.org.uk/system/files/field/document/An%20Ethical%20Charter%20for%20Date%20Science%20WEB%20FINAL.PDF

*Information Commissioners Office*. (2018). Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/

*Institute and Faculty of Actuaries News*. (2018, May 10). Retrieved from https://www.actuaries.org.uk/news-and-insights/news/rss-and-ifoa-partner-data-science

Loser, F. (June 2018). Machine learning vs actuarial methods in claim prediction. *31st International Congress of Actuaries.* Berlin.

Panlilio, A., Canagaretna, B., Perkins, S., du Preez, V., & Lim, Z. (2018). Practical Application of Machine Learning Within Actuarial Work. *Institute and Faculty of Actuaries*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., . . . Sutsk. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, pages 484–489. Retrieved from https://deepmind.com/research/alphago/

Wong-Fupuy, C., & Haberman, S. (2004). Projecting Mortality Trends. *North American Actuarial Journal*, 56-83.

# 6 Appendix: Data Sources

| Data Description | Data Use | Data Source |
|---|---|---|
| Annual summary of earnings, place of residence by Local Authority | Predictor Variable | https://beta.ons.gov.uk/datasets/ashe-table-8-earnings/editions/time-series/versions/1 |
| Death Statistics by Area Code (2012) | Target variable (training data) | https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/deathsregisteredbyareaofusualresidenceenglandandwales |
| Death Statistics by Area Code (2013) | Target variable (training data) | https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/deathsregisteredbyareaofusualresidenceenglandandwales |
| Death Statistics by Area Code (2014) | Target variable (training data) | https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/deathsregisteredbyareaofusualresidenceenglandandwales |
| Death Statistics by Area Code (2015) | Target variable (training data) | https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/deathsregisteredbyareaofusualresidenceenglandandwales |
| Death Statistics by Area Code (2016) | Target variable (hold-out data) | https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/deathsregisteredbyareaofusualresidenceenglandandwales |
| Average pension by area | Predictor Variable | https://www.gov.uk/government/statistics/income-and-tax-by-borough-and-district-or-unitary-authority-2010-to-2011 |
| 2011 Census Age Distribution by Area | Predictor Variable | https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/2011censuskeystatisticsforlocalauthoritiesinenglandandwales/r21ewrttableks102ewladv1_tcm77-290566.xls |

| | | |
|---|---|---|
| Population Density by Area | Predictor Variable | https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/2011censuspopulationestimatesbyfiveyearagebandsandhouseholdestimatesforlocalauthoritiesintheunitedkingdom/r12ukrttablep04ukv2_tcm77-304141.xls |

Institute
and Faculty
of Actuaries

**Beijing**

14F China World Office 1 · 1 Jianwai Avenue · Beijing · China 100004
Tel: +86 (10) 6535 0248

**Edinburgh**

Level 2 · Exchange Crescent · 7 Conference Square · Edinburgh · EH3 8RA
Tel: +44 (0) 131 240 1300 · Fax: +44 (0) 131 240 1313

**Hong Kong**

1803 Tower One · Lippo Centre · 89 Queensway · Hong Kong
Tel: +852 2147 9418

**London (registered office)**

7th Floor · Holborn Gate · 326-330 High Holborn · London · WC1V 7PP
Tel: +44 (0) 20 7632 2100 · Fax: +44 (0) 20 7632 2111

**Oxford**

1st Floor · Park Central · 40/41 Park End Street · Oxford · OX1 1JD
Tel: +44 (0) 1865 268 200 · Fax: +44 (0) 1865 268 211

**Singapore**

163 Tras Street · #07-05 Lian Huat Building · Singapore 079024
Tel: +65 6717 2955

www.actuaries.org.uk