Institute
and Faculty
of Actuaries

Practical Application of Machine Learning Within Actuarial Work

by Modelling, Analytics and Insights in Data working party

Alex Panlilio; Ben Canagaretna; Steven Perkins; Valerie du Preez; Zhixin Lim

30 January 2018

Abstract

Machine learning techniques are increasingly being adopted across the financial sector. Workstream 2 sets out to explore the use of these techniques in existing actuarial practice areas.

In Section 1, a clear objective is outlined. We consider the various practise areas and highlight potential applications of machine learning techniques. In Section 2, machine learning concepts are introduced and explored at a high level. Parallels between the machine learning project cycle and the actuarial control cycle are drawn to highlight the similarities.

In Section 3, four case studies, showcasing the applications of machine learning techniques, are introduced (and detailed in the appendix), including:

1. Utilising unstructured data in forecasting interest rates
2. Pricing of marine hull
3. Supervised learning in exposure management
4. Mortality experience analysis

In Section 4, an overview is provided for some of the programming platforms used. The list is by no means exhaustive and lays the foundation as a starting point for actuaries to use. In Section 5, an overall conclusion is drawn and a number of lessons learnt provided. The conclusions drawn from the case studies were in most cases inconclusive although we have gained enough intellectual property to be confident of the merits of these techniques within our respective areas. We have identified the limitations and potential improvement to our work.

Contents

1 Working Party Overview

1.1 Working Party Aims

To investigate whether the application of machine learning techniques can improve the models and/or assessments we use within traditional actuarial practice areas i.e. how would traditional actuarial practice areas benefit from data science; and in particular machine learning; techniques?

Mission Statement from Terms of Reference

To identify key actuarial function areas and processes which have scope to be improved by the implementation of mathematical modelling, predictive analytic tools and data science. Once a list of applicable areas is established, the workstream aims to explore and use these new methods and techniques to produce possible solutions to improve these areas. This will be summarised into a report for the Institute and Faculty of Actuaries ("IFOA") Modelling, Analytics and Insights from Data (MAID) Steering Committee.

1.2 Background to working party

The working group started by identifying broad categories of models used within the traditional actuarial practice areas. From this, we identified areas, which could potentially benefit from the application of machine learning techniques. This is not an exhaustive list.

TRADITIONAL ACTUARIAL PRACTICE AREAS	General Insurance	Pensions	Life, Health & Care	Investment
Pricing	√		√	
Product Design / Propensity Customer Behaviour	√		√	
Reserving	√			
Capital Modelling	√		√	
Exposure Management	√			
Scheme Valuation		√		
Surplus Distribution			√	
Strategic / Tactical Asset Allocation				√
Asset & Liability Management / Hedging				√
Claims Management	√	√	√	√
Data Cleansing (Table 5)	√	√	√	√
External Data Sources (Table 5)	√	√	√	√

Table 1: Potential application of machine learning techniques

GENERAL INSURANCE	METHODS TO EXPLORE
Pricing	• **Supervised Learning**: decision tree, forests and penalised regression • **Unsupervised Learning**: using a non-linear approach • **Deep Learning** and high level decision making • **Experience monitoring** with a larger dataset
Product Design / Propensity Customer Behaviour	• **Big Data** on consumer information • **Sentiment Analysis** using external sources and social media
Reserving	• **Different cohorts** Making projections more predictive; claim predicting pattern could vary • Explore **supervised learning (penalised regression)** • **Experience monitoring** with a larger database
Capital Modelling	1) **Network / Graph Modelling**- looking at driving dependencies rather than correlation assumptions 2) Strategically flexible, more **decision aid** based model on environment 3) Portfolio / Reinsurance optimisation – **genetic algorithms**
Exposure Management	• Build predictive models based on **weather patterns** • (See Table 5. **Data Cleansing**)

Table 2: General Insurance

PENSION	METHODS TO EXPLORE
Scheme Valuation	• More granular individual information from alternative data sources e.g. social media • More sophisticated longevity model • Tailoring investment strategy to individual circumstances

Table 3: Pensions

INVESTMENT	METHODS TO EXPLORE
Strategic / Tactical Asset Allocation	• Utilising alternative data e.g. text-heavy data, social media feeds, satellite images etc. • Improvements to Mean-Variance Portfolio Optimisation
Asset & Liability Management / Hedging	• More granular data for asset/liability modelling • Enhanced market risk monitoring

Table 4: Investment

LIFE, HEALTH AND CARE	METHODS TO EXPLORE
Pricing	• **Supervised Learning**: decision tree, forests and penalised regression • **Unsupervised Learning**: using a non-linear approach • **Deep Learning** and high level decision making • **Experience Monitoring** with a larger database
Capital Modelling	1) **Network / Graph Modelling**- looking at driving dependencies rather than correlation assumptions 2) Strategically flexible, more **decision aid** based model on environment 3) Portfolio / Reinsurance optimisation – **genetic algorithms**
Surplus Distribution	• More granular individual information from social media sites • More sophisticated longevity model

Table 5: Life, Health and Care

ALL PRACTICE AREAS	METHODS TO EXPLORE
Data Cleansing	1) **Reducing errors** i.e. data validation 2) **Filling in gaps** i.e. missing latitude and longitudes 3) **Increasing sample size** using Machine Learning extrapolation 4) Web scraping, **word search** / **natural language** analysis
External Data Sources	• Quandl / Dun and Brad Street / Bloomberg / social media feeds / credit agency
Feedback Loop / Actuarial Control Cycle	• Year on year to keep track of outputs

Table 6: All areas

Following the review above, the group was then subdivided into separate working groups in order to explore the following case studies:

- General Insurance – pricing techniques using experience data

- Exposure Management – data cleansing
- Mortality – analysing suicide rates within a defined population
- Investments –interest rate forecasting

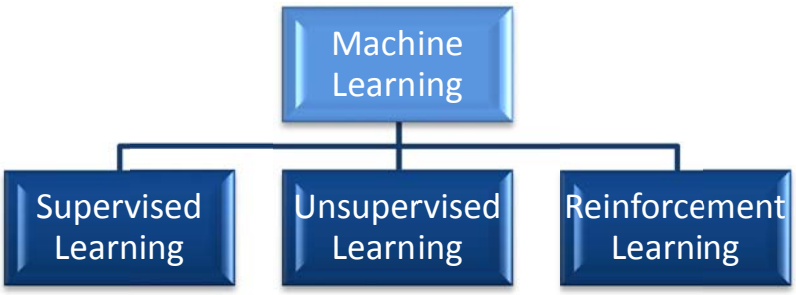
This report summarises the work of the group and is a starting point for further research and discussion on the topic.

2 Machine Learning Overview

Improvements in computational power has given rise to the use of machine learning techniques in a wide variety of areas, including finance, driverless cars, image detection and speech recognition among others. In a world of high volume and varied datasets, machine learning techniques are an essential toolkit to provide actionable insights from the data.

2.1 What is "Machine Learning"?

The term "machine learning" is an overarching term which covers a wide variety of techniques overlapping areas of mathematics, statistics and computer science. Machine learning is often referred to as artificial intelligence (AI), with the two words (often incorrectly) becoming synonymous. This is because many high profile, cutting edge, AI use cases utilise machine learning techniques to create intelligent systems. Broadly speaking, machine learning algorithms can be divided into three classes depending on the type of problem they are applied to.



Supervised learning

Still by far the most common application of machine learning are instances of "supervised learning". A historical set of training data is used to create a model which explains the underlying correlations within the data. Critically, the training data has both input variables (often referred to as "features") as well as a target variable (often referred to as the "response variable"). Supervised learning algorithms will use the input variables to attempt to identify the target variable. Once a supervised learning model has been built and validated, it can be used to make predictions for future datasets where the input variables are known and the user would like to estimate the unknown target variable. Most actuaries will be familiar with the modelling process described above, with many traditional actuarial tasks naturally fitting into the supervised learning framework.

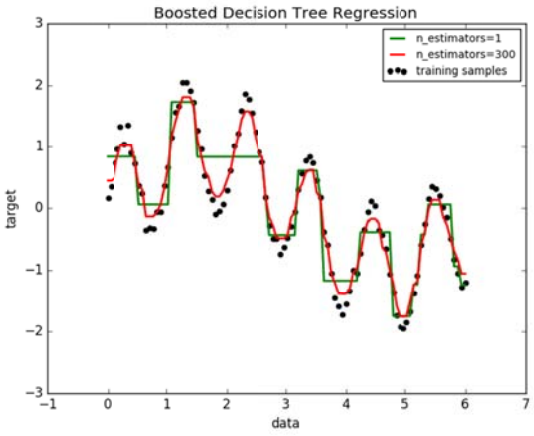


Figure 1: Example of supervised learning (decision tree regression)

Supervised learning includes three categories of algorithms:

- Binary classification;
- Multiclass classification;
- Regression.

Binary classification tasks are those where the response variable is categorical in nature with exactly two classes (e.g. fraud *vs.* not fraud).

Multiclass classification tasks are those where the response variable can be separated into a finite number of specific "classes" (e.g. class A, B or C). Binary classification is therefore a special instance of multiclass classification where there are only two classes.

Finally, regression tasks are those where the response variable is numerical and continuous in nature (e.g. predicting claims cost).

Some examples of supervised learning algorithms used in practice include:

- Decision trees;
- Random forests;
- Gradient boosted machines;
- Generalised linear models;
- Support vector machines;
- K-nearest neighbour;
- Neural networks.

The above list of techniques is far from exhaustive, but provides an introductory set of techniques for supervised learning. These algorithms can be applied to both classification and regression tasks, though the actual parameters used within the models will need to be adjusted accordingly.

As highlighted, many actuarial modelling projects naturally fall into the category of supervised learning, with tasks such as insurance contract pricing or pension scheme valuation naturally fitting into this framework. The difference between actuarial and machine learning approaches to such tasks is often relatively small, as discussed further below. This makes supervised learning tasks a natural place for actuaries to initially explore machine learning techniques.

Unsupervised learning

Unsupervised learning covers a variety of techniques which have been designed to solve distinctly different types of problems. Similar to supervised learning, a historical set of training data is used to create a model which explains the data. The training data contains input variables, however for unsupervised learning tasks there is no response variable. This therefore means that unsupervised learning does not rely on previously known or labelled data to be trained. Instead it takes the input data as it is, then infers patterns and structures from it "blindly".

Initially, the concept of unsupervised learning can feel counterintuitive to non-statisticians as it may be believed that nothing meaningful can be learned without some form of outcome data (e.g. a response variable). However, unsupervised learning often allows users to gain a deeper understanding of their data, even if it is not obvious at the outset quite what will be learned. By far the most common implementation of unsupervised learning is cluster analysis, which is also often used for dimensionality reduction and anomaly detection.

Common unsupervised learning algorithms used in practice include:

- K-means clustering;
- K-nearest neighbour;
- Hierarchical clustering
- Principal component analysis;
- Support vector machines;
- Neural networks.

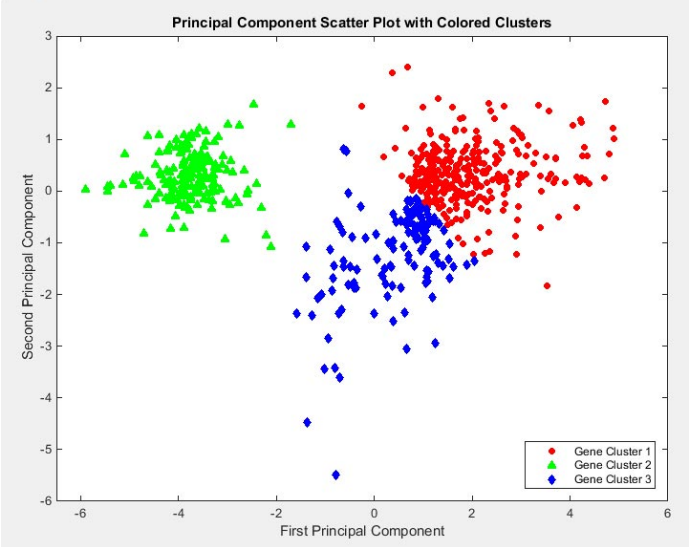


Figure 2: Example of cluster analysis

Again, the above list of techniques is far from exhaustive, but provides an introductory set of techniques for unsupervised learning. It will immediately be obvious that certain techniques appear on both the supervised and unsupervised learning lists. This is because these algorithms are particularly flexible in how they can be implemented, allowing the user to apply them in a variety of ways.

There are perhaps fewer instances where unsupervised learning can be applied within actuarial work. However, cases such as image recognition, text analysis and speech recognition may increasingly become useful areas to

apply unsupervised learning. A more traditional actuarial task might involve cohort creation, with individuals perhaps grouped based on their year of birth. Unsupervised learning might provide a potential alternative to traditional cohort creation, creating more homogeneous risks and hence improving modelling.

Reinforcement learning

Perhaps the most complex area of machine learning is reinforcement learning. Unlike supervised and unsupervised learning, reinforcement learning does not rely on a historical dataset to build a model. Instead the model is created and updated dynamically. A reinforcement learning algorithm will provide a predicted response, but once it is shown the actual outcome it will incorporate this new information into the model to improve the next prediction it makes. Over time the predictions should improve as the algorithm learns more about the environment it operates in. Convergence of a reinforcement learning algorithm to an optimal solution is not guaranteed and improving the theory covering the convergence of these is an active area of research. However, one classical example of a problem which can be solved using reinforcement learning is a Markov decision process. There may be relatively fewer immediate applications of reinforcement learning in actuarial work, but this may well change over time as statistical methods improve.

2.2 Other Machine Learning Concepts

Model Validation

A common issue within statistical modelling is over-fitting a model. It is therefore important to validate a predictive model to ensure the performance generalises to new cases. Within machine learning and actuarial work a standard way to scrutinise and improve models is to use model validation in order to assess whether a predictive model retains a sufficient level of performance against new data. Under a standard validation approach, a holdout sample is taken prior to any modelling work. Models are built using a "training" dataset which is independent of this holdout "test" data, with this "unseen" dataset being used to test the predictive power of the model. Models are typically scored using an appropriate performance metric(s) (decided at the outset); the model achieving the highest score on this independent dataset will usually be selected to be used in practice.

Cross-validation is an alternative model validation technique for assessing how the results of a statistical analysis will generalize to an independent set of data. It is mainly used in settings where the goal is prediction and the user wants to estimate how accurately a predictive model will perform in practice.

Cross-validation is used to find out the predictability of the model by splitting a dataset into k-parts (often referred to as 'folds'). k-1 parts are used for the actual training of the model and the remaining single part is used as an independent dataset for validating the performance of the model. This approach iterates until all combinations of the k partitions have been tested. In this manner k models are produced and scored against (different) independent datasets. The final model might then be a simple average of the k models produced, as shown in Figure 3. Without an explicit holdout dataset, this allows models to benefit from utilising all the available data whilst benefiting from validation of performance on k datasets which are independent of the actual model building process.

An extension of this is to this use of cross validation can be in tuning model hyper-parameters. In this instance the data is split into a training and test (or 'holdout') dataset, but the training set is then split into k-parts. Models are built with a variety of hyper-parameters once again using k-1 elements of the training set. The best hyper-parameters can be selected based on their performance on the k^{th} ('validation') part of the

Figure 4: Cross-validation using 5 folds of 16% as well as a 20% holdout

training set. This allows a user to identify what appears to be the 'optimal' (or at least best performing) set of model parameters to go forward and be scored against the test data. Utilising cross validation for the hyper-parameter tuning ensures that the test dataset remains truly independent of the model fitting process and therefore will give an unbiased view on the performance of the model against other candidate models. This approach is illustrated in Figure 4.

Deep Learning

An area of machine learning gaining particular notoriety is deep learning. Deep learning has become a recognisable term for many due to its being the catalyst for many recent high-profile advancements in artificial intelligence. It is fundamentally an advanced application of artificial neural networks (ANNs), with an example of a basic feed-forward ANN shown in Figure 5. Much like other modelling techniques, such as generalised linear models, in a supervised learning context a neural network takes inputs (shown in blue) and uses these to predict an outcome (shown in red). The neural network does this by transforming the inputs via a hidden layer (shown in green). The actual model parameters are the weights that are placed on each linking edge of the model (the black lines).

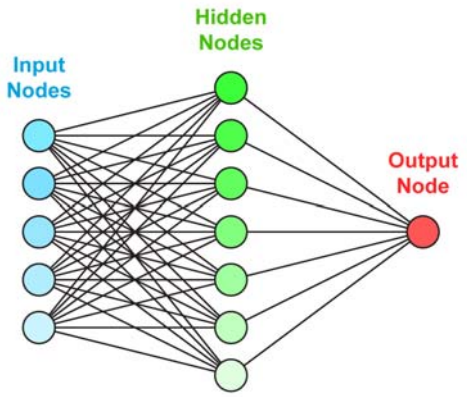


Figure 5: Example of a basic neural network

As an example, Figure 5 shows a relatively basic feed-forward ANN with 5 input nodes. Each input node will take a value which for ease could be assumed to be either 0 or 1. The value of the first node in the hidden layer will be determined by the values of these 5 input nodes with the relative weight placed on each input learned as part of the model fitting process. Other nodes in the hidden layer will be similarly defined, though they will naturally have different weightings on each of the input nodes. In much the same way the value of the output node will then be determined by the values of the 7 nodes in the hidden layer, with the relative weight placed on each hidden node again being determined through the model training. Inputting a new observation containing values of the 5 input nodes will lead to an output prediction for the value of the target variable. This is calculated by passing the inputs through the network using the learned weights from the training process to firstly calculate the values at the nodes in the hidden layer and then use these to calculate the value of the output node.

The above is a very high-level description of a basic ANN, but more complex neural network architectures have become popular more recently due to their empirical performance for a variety of challenging tasks. However, this has all been made possible through the ability to design complex model structures which reflect the nature of these tasks. Whilst in certain cases the underlying mathematics behind these complex model structures has been known about for a while it is the increase in data volumes along with the improved computing capabilities (including GPU processing) which has resulted in an increase in the usage of ANNs.

Deep learning itself can be defined in a number of ways but, broadly speaking, it covers the extension of the above basic model architecture to cover any case where there is more than one layer of hidden nodes, or multiple layers of neural networks. As such, these ANNs make predictions based on processing repeat layers of signals, imitating the learning process via neural pathways in the human brain.

2.3 Machine Learning or Data Science?

Two terms which are often used interchangeably are "machine learning" and "data science", and therefore it is important that actuaries are clear around the distinction in these two terms. Machine learning covers a suite of statistical techniques and algorithms used for modelling data. Data science is a broader term which includes all methods, processes, and approaches to extract insights from data. Therefore, data science will include areas such as:

- Data collection;
- Data cleaning;
- Data engineering;
- Data visualisations;
- Application of the scientific method;
- Advanced programming.

Data science also includes knowledge of machine learning to determine which approaches are best suited to particular tasks. What should immediately become apparent though is that many of the above skills are also common with those already used by actuaries as discussed further below.

2.4 Data Science vs Traditional Actuarial Approaches

As already noted above, many aspects of data science are very similar to those already used by actuaries. Figure 6 below gives an overview of a typical data science project lifecycle.

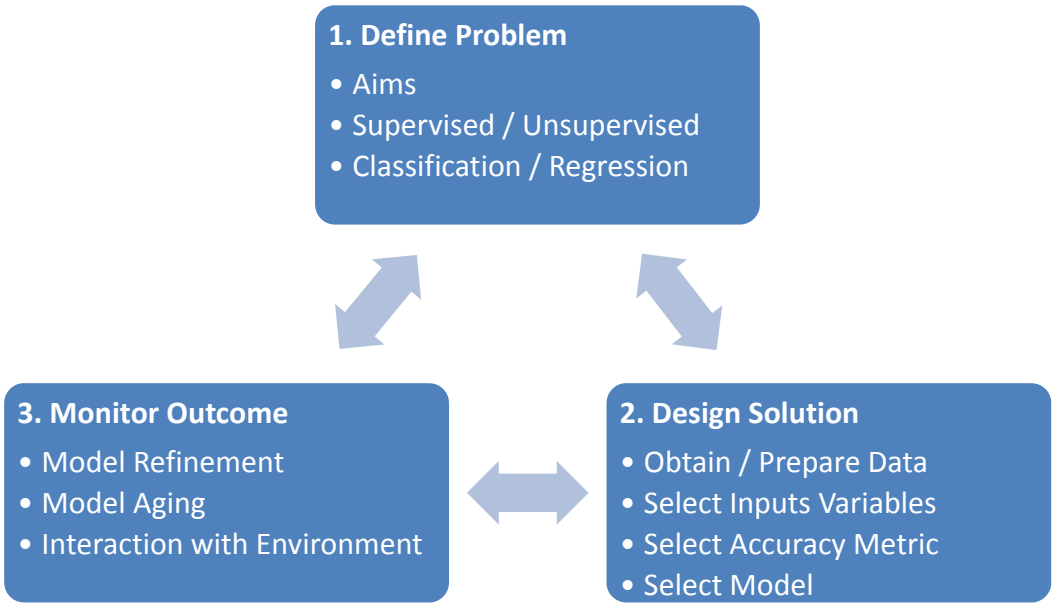


Figure 6: An overview of a typical data science project lifecycle

When shown in this format, it becomes clear that there is a significant overlap between the phases data scientists go through when developing a model and the actuarial control cycle. The main differences between actuarial and data science approaches occur in phase two, when a solution is being designed. Actuaries will typically use their domain knowledge to select an appropriate model format before spending a reasonable amount of time making choices of parameters which are sensible and justifiable for the purpose of the model. In this way, actuaries typically spend a large amount of time minimising the parameter error within their models. By contrast, data scientists may look to test a wider array of potential models, taking advantage of the speed with which machine learning can fit suitable models. This means that a data science approach can focus on reducing the model error initially, by testing a variety of model forms, with parameter tuning coming later and perhaps receiving less scrutiny than under an actuarial approach.

The other key difference is that actuaries will often build models which are financial in nature, utilising their existing domain knowledge. Data scientists tend to tackle a wide variety of modelling tasks (financial and non-financial) and therefore are often more reliant on gaining an understanding of the domain specific elements of a task from other domain specialists. Other key differences arise due to the following:

- The statistical techniques used by actuaries and data scientists often differ;
- The approaches taken to validate assumptions may differ;
- The approaches used for variable selection may differ;
- The approaches used to assess the performance of a model may differ.

However, despite these differences, what should be clear is that data science and actuarial modelling approaches have much in common. This leaves the actuarial profession well placed to utilise these new techniques within the scope of their existing work.

2.5 How Actuarial Tasks May Benefit from Machine Learning

The exponential increase in data generation, capture and storage along with improved computer power is likely to benefit actuaries in two primary ways. Firstly, improved data and computational capabilities is likely to mean that traditional actuarial tasks can be tackled with increasingly sophisticated approaches. The second opportunity arises because many actuaries will have the necessary skills to capitalise on new opportunities which arise to expand the profession into new areas.

This paper focuses on the first of these benefits, demonstrating them via a series of case studies. The key benefits of applying machine learning to actuarial tasks can broadly be split into six categories.

Improved Data Quality

As more data is created and storage becomes more cost effective, there is an increase in the opportunity to improve the quality of data which businesses are using. Similarly, as competitors start utilising better data, businesses not attempting to do the same may become left behind. These effects are allowing companies to improve the quality of their data going forward, and this can only benefit actuaries as higher quality data will produce better models regardless of the techniques being employed.

New Data Sources

Modelling by actuaries has typically taken a relatively standard format. Model inputs tend to be numeric or categorical fields which are used to predict a numerical outcome. However, machine learning potentially opens up opportunities for actuaries to explore alternative data sources. For example, text fields could be explored to understand key themes and images could potentially be incorporated into predictive models.

Speed of Analysis

As many data scientists and actuaries will agree, much time taken to produce a model is used to gather, clean and manipulate data. These tasks will largely be similar, regardless of the methods used. However, once a modelling data set has been produced, machine learning can be beneficial. Models can generally be fitted and validated in a short space of time, allowing tasks to be completed quickly.

New Modelling Techniques

Utilising alternative modelling approaches, such as unsupervised learning, allows different perspectives to be gained on data. Techniques such as anomaly detection or time series modelling can potentially produce a stronger predictive power for certain problems, improving the performance of actuarial models.

New Approaches to Problems

Actuaries typically use relatively standard modelling approaches to tackle a variety of problems. However, in all cases, models will suffer from model error and therefore being able to produce a wider variety of models in a short space of time will allow actuaries to better select the appropriate modelling approach for a given problem.

Improved Data Visualisations

With new modelling techniques and new software for machine learning, users have an increasing power to produce stunning visualisations of data which can itself provide new perspectives on a task. One popular example would be a word cloud, which can be used to highlight the most common words within text. However, many other data visualisation methods also exist and are regularly used by data scientists.

The case studies in the following section seek to demonstrate some of these benefits within a variety of actuarial tasks.

3 Executive Summary of Case Studies

3.1 Utilising Unstructured Data in Forecasting Interest Rates

Interest rate forecasting is of importance in various actuarial practice areas, including investment, asset liability management (ALM), insurance liabilities valuation, and capital modelling.

The case study describes a model which "reads" and provides sentiment analysis on central bank communications. Central banks like the Bank of England (BoE) exert vast influence on the level of interest rate via monetary policies. The tone or sentiment in central banks communications sets an expectation in the market.

Supervised machine learning techniques are used to train an ensemble model that classifies BoE communications in a fully automated and scalable way. The result of the sentiment analysis is used in an interest-rate forecasting model. Given the inherent uncertainty in making forecasts, the interest-rate forecasting model provides a range of feasible outcomes.

The case study employs R, an open-source programming language.

3.2 Pricing of Marine Hull

This report analyses how supervised machine learning can be used as a pricing tool to predict future aggregate claim costs for a given type of vessel. The results showed that a 'Generalised Linear Model Blender' was the most accurate learning algorithm for aggregate claims.

Of the inputs used, the sum insured of a given vessel was clearly the most influential feature in predicting the expected total claim value. The trained model was also able to find some interesting patterns (such as year of build) which would help identify the key rating factors that make up a better or worse risk, although the accuracy could be improved with more data.

It was felt that derivation of annual vessel-mileage from the latitude/longitude data together with a multiclass approach would improve the sophistication of the analysis. However, it was concluded that the concepts and logic could be applied to a number of other case studies where a large enough volume of credible exposure and claim data is available.

3.3 Supervised learning in Exposure Management

This report analyses how supervised machine learning can be used as a data cleansing tool to predict missing fields (i.e. 'year built' and 'stories') within a property exposure dataset. The results showed that an 'eXtreme Gradient Boosted Tree Regressor' was the most accurate learning algorithm for both 'year built' and 'stories'. The total insured value (TIV) or property value was the most influential feature in predicting the number of stories of a building. The latitude and longitude of a property proved to have the greatest influence in predicting the year a building was built. Other key findings were that the 'stories' model had an accuracy or (Poisson deviance) error of 1 story. The 'year built' model had an accuracy or root mean squared error (RMSE) of 11.51 years. This report finds that machine learning is beneficial in finding definite patterns and predictions from trained data to complete blank unfilled data.

The study is limited to predicting continuous variables rather than multiple classes. Some features that have a more direct influence on pricing could not be modelled i.e. construction and occupancy codes as multiple classes.

3.4 Mortality experience analysis

By linking external data sources such as mood index, consumer confidence index and Dow Jones index to US mortality data (1980-2014) the investigation looked at whether machine learning techniques could be used to identify improved patterns and/or links with external drivers of mortality. The output shows, at a high level, that:

1. A possible correlation between the change in consumer confidence within the US and the change in suicides.
2. Dow Jones Index could not predict number of suicides and changes in suicides which was a null hypothesis.
3. We did not have enough data points of Mood Index to conclude regarding its relationship with counts of suicides in this analysis. (Null hypothesis could not be rejected).

4 Programming languages and Applications

One of the key questions for an actuary looking to utilise machine learning is the platform they use to utilise these new modelling techniques. This ultimate decision is likely to depend on a number of factors, but the primary consideration will be the prior programming expertise required for each option. Other natural considerations might include:

- The cost of each approach;
- The time it will take to learn how to use the software / programming language;
- The range of models which can be built;
- The nature of the output which can be produced;
- The stability of the operating system;
- The existing systems which are used by the company / actuary.

Those with little or no programming experience will probably want to start in a point-and-click environment, which are provided by a variety of third party software solutions. Those who have more experience with computer programming may be better placed to utilise open source programming languages. In the sections below, we provide a brief outline of the two key open source languages typically utilised by data scientists, R and Python, as well as a broader discussion around other software solutions which actuaries may choose to use instead.

4.1 R

R is an open source programming language developed by statisticians with the purpose of providing a diverse and high quality open source statistical analysis language. It was initially developed in the mid 1990's and was almost exclusively used in academia until more recently when the growth of data science has increased its user base.

Since R's purpose is to provide a tool for statisticians to use, it comes as no surprise that this is the most widely used programming language by data scientists. The key strength of R is the wider ecosystem, with an enormous variety of cutting-edge packages, which can make implementation of complex models amount to a single line of code. R has also been designed with data visualisations in mind and there are a number of packages available which provide excellent data visualisations.

The key limitation with R is the speed of processing. R was designed to allow sophisticated statistical analysis to be conducted more easily. However, the speed of this analysis was less of a concern when R was initially being created. In addition, the learning curve with R can be relatively steep, especially for those without a programming background.

Overall, the benefits of using R outweigh the limitations and this has resulted in it become the most widely used data science language.

4.2 Python

Python is a widely used object-oriented programming language which is known for its code readability. Unlike R, Python is a general-purpose programming language. This means that whilst it is known to be a powerful tool for data scientists, it is also a language which is more widely used by developers for tasks such as web design. This can have advantages in the variety of model building options and if a model needs to be translated into a production environment.

However, Python also has some limitations, most notably the array of packages available for data science. Python remains behind R in both the number and depth of packages available, but the options in Python has improved significantly in more recent years. This can mean that certain elements of data analysis might require more complex code in Python or a higher level of background knowledge to achieve the same result in R.

Overall, Python provides a strong alternative to R. The easy integration with the wider developer community is a key strength. However, there is potentially a steeper learning curve for those with little or no programming experience.

4.3 Other Proprietary Software

Whilst R and python represent the two core programming languages used by data scientists, an actuary who does not have experience with either of these pieces of software should not be discouraged from building machine learning models. With the rapid growth of machine learning techniques and increasing computer power, there are a number of third party providers who are looking to give individuals easy access to the power of machine learning. In many cases, a user can build powerful machine learning models with little programming knowledge by utilising the simple user interfaces which have been developed for this purpose.

Alongside the relatively low barrier to entry, the key benefit to propriety software is the speed at which complex models can be built. This can also be a weakness, as software can be limited in both the range of techniques which can be implemented and the calibration of the implementation itself.

Platforms such as Microsoft Azure or Amazon Web Service provide cloud-based platforms with a point-and-click machine learning environment, underpinned by large international companies. These have the benefit of scalability and flexibility over implementation solutions, with an associated higher cost. Alternatively, there are many other offline services provided by large companies, such as IBM and SAS or smaller companies. For example, some of the case studies in this paper (i.e. cases 2,3 and 4) were conducted using third-party software created by DataRobot ('the Software').

The correct platform for analysis will depend on a number of factors, but it is relatively straightforward for an actuary to get started in an open source software such as R or Python, or to trial other propriety software to understand the relative benefits of each approach.

4.4 Implementation Considerations

The final note in this section is around the implementation of machine learning models. Many machine learning models can be seen as exploratory in nature, with the key aim being an improved understanding of the underlying data. However, in other circumstances it will be important to be able to utilise the machine learning models created and, as with the various platforms for analysis, there is a wide variety of approaches which can be taken to implement machine learning algorithms within a business.

The actual systems required for the implementation of machine learning algorithms will depend on a number of factors such as the current infrastructure, the specific machine learning technique and the particular environment under which the algorithm will be implemented. Further discussion is beyond the scope of this paper but is likely to be a key consideration for end users. This means consideration of the implementation options should be taken early in a project and is a decision which may impact the modelling approaches which actuaries can take.

5 Conclusions and recommendations

Machine learning techniques and processes conform to the fundamental principles of actuarial science. The case studies (details are available in the appendix) have demonstrated that actuaries from all backgrounds can utilise these techniques either from first principles using programming languages such as R or Python or by purchasing software applications. Whilst the conclusions drawn from the case studies are in most cases inconclusive we have gained enough intellectual property to be confident of the merits of these techniques within our respective areas. We believe that dedicated time and effort spent on using these techniques will improve the quality of our analysis.

Other observations and lessons learnt:

- **Data is key**– Machine learning and AI techniques are essential toolkits in a world of high volume and varied datasets. However, acquiring and processing large datasets can be challenging. In our case studies, finding and preparing the relevant data for analysis takes up the bulk of our work. While less interesting, being able to work with data is a prerequisite.
- **Putting it into practice** – There is a significant amount of learning resources available on data science. The best way to learn is to start with a simple case study to understand the principles around data processing, model training, and model validation. With the fundamentals fully grasped, one can move progressively to more complicated and interesting projects.
- **Domain knowledge** – Advanced machine learning and AI techniques are easily accessible via open source software. The only barrier of entry is domain expertise. To utilise these techniques, understanding the problem, knowing which data is relevant, and being able to connect the dots are essential. This is an area actuaries could add value.

We intend to develop the four case studies within this paper to include a broader set of data where appropriate to explore further some of the themes we have uncovered. For example, on the marine hull pricing case study, we would like to derive mileage per vessel as a rating factor. In addition, we would like to add new case studies in the capital modelling and risk management areas.

We strongly believe these new methods will be critical for the profession in remaining relevant in a fast changing world where technology is disrupting all areas of the financial sector. The actuarial profession has a long history of innovation and proposing new methods of analysis; we see this as just another challenge.

We feel that more research is required, supported by a financial investment, and this should come from both the companies actuaries represent in conjunction with the IFoA. Finally, we would encourage actuaries at all levels to develop an understanding of machine learning techniques and how it can be utilised within their current and future roles.

6 Appendix – Case Studies

6.1 Utilising Unstructured Data in Forecasting Interest Rates

6.1.1 Background

Interest rate forecasting is of importance in various actuarial practice areas, including but not limited to:

- Investment e.g. the level of risk-free interest rate is an anchor on the return of a wide array of asset classes.
- Asset Liability Management (ALM) e.g. how much interest rate risk to hedge and when to hedge?
- Life and pension liabilities valuation e.g. valuation with reference to the interest rate curve.
- Capital modelling e.g. 1-in-200 interest rate risk scenario over 1 year.

Central banks like the Bank of England (BoE) exert vast influence on the level of interest rate via monetary policies. The BoE pursues monetary policies to achieve a balance of price stability (defined by an inflation target) and economic growth. These policies are usually implemented through setting the base interest rate, and through quantitative easing (QE) where central banks create new money electronically to buy financial assets such as gilts and corporate bonds.

The BoE's Monetary Policy Committee (MPC) meets regularly[1]. Its decision and the associated meeting minutes to increase, maintain, or decrease the base interest rate and the amount of QE is made publicly available. In addition, the BoE publishes various communications (e.g. speeches, inflation reports, press releases, and other forms of forward guidance).

6.1.2 Problem definition

BoE communications contain informational content that could be used to predict monetary policies (and hence, the level of interest rate) ahead of the MPC meetings. Due to the unstructured and nuanced nature of text data, machine learning techniques are well-suited to perform sentiment analysis on BoE's communications. "Sentiment" is defined in central banks' parlance – hawkish, neutral, or dovish. A hawkish tone denotes a tightening of monetary policy i.e. an increase in the base interest rate and a reduction in QE, while a dovish tone suggests the opposite.

A sentiment analysis model is trained on the MPC minutes (dating from June 1997 to February 2017). Once trained and validated, the output of the sentiment analysis model on forward-looking BoE communications (e.g. speeches, inflation reports, press releases etc.) is used as one of the features (i.e. independent variables in statistics-speak) in the interest rate-forecasting models. Details of the methodology are provided in the following sections.

6.1.3 Results and benefits

As shown in the table below, incorporating text data provides a marginal improvement in the predictive power of the interest rate-forecasting models. The performance of the model is back-tested (from June 1997 to January 2017) and measured using mean absolute error (MAE) i.e. the average absolute difference between the forecast and actual interest rates, and directional accuracy i.e. % of correct up/down movement predictions.

Interest rate-forecasting model	Including text data	Excluding text data	Improvement
Bayesian Structural Time Series (BSTS)	22 bps	22 bps	0 bps
Random Forest (RF)	30 bps	40 bps	10 bps

Table 7: Mean absolute error (1-month forecast)

[1] MPC meetings were held monthly before September 2016; they now take place 8 times a year

Interest rate-forecasting model	Including text data	Excluding text data	Improvement
Bayesian Structural Time Series (BSTS)	52%	50%	2%
Random Forest (RF)	54%	48%	6%

Table 8: Directional accuracy (1-month forecast)

Traditionally, text data has been quantified using rule-based methodologies e.g. counting the number of words considered to be "positive" or "negative". This has several disadvantages that can be addressed using machine learning techniques in the following ways:

- Speed of analysis/implementation – the traditional method requires classifying "positive" and "negative" words/terminologies specific to central banks. This is done automatically over iterations of "learnings" via machine learning techniques;
- Complex relationship modelling– the traditional method requires defining rules to quantify text data. Machine learning techniques are able to model complex relationships (not necessarily linear) between the text data and its relationship to monetary policy outcomes;
- Self-learning – machine learning techniques adapts to new data, continuously updating its lexicon of central banks-specific words and their relationship to monetary policy outcomes;

6.1.4 Framework and methodology

A summary of the model framework is visualised below; this is generalised and can be applied in developing other machine learning based models. In contrast, the methodology described in this case study is only one of many approaches in building a sentiment analysis model. Where relevant, an alternative methodology is highlighted for further investigation.

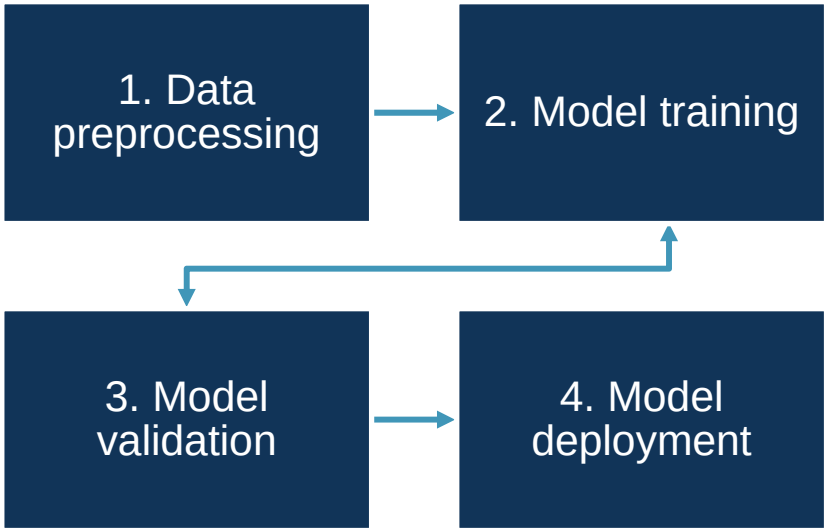


Figure 7: Model framework

In this case study, sentiment analysis is cast as a multiclass classification problem, and supervised learning techniques are used to train the sentiment analysis model. The MPC meeting minutes (dating from June 1997 to February 2017) provide a convenient corpus (i.e. a collection of text) for training the model. A classification label – hawkish (1), neutral (0), or dovish(-1) – is assigned to each minutes based on the decision to increase/maintain/reduce the base rate and the QE programme.

An alternative methodology, not in scope of this paper, is to cast sentiment analysis as a regression problem; in lieu of classification labels, the impact on interest rates can be assigned to the minutes. There are other factors

driving interest rates, hence domain expertise is required to isolate the effect of the release of the MPC meeting minutes.

Once trained and validated, the model can be used to classify other forward-looking BoE communications (e.g. speeches, inflation reports, press releases etc.) to predict monetary policies ahead of the MPC meetings. The prediction is then used as one of the features in an interest rate-forecasting model.

6.1.4.1 Text data pre-processing

The first step is to build up a lexicon or vocabulary using the MPC meeting minutes. The meeting minutes are in the form of PDF files and require pre-processing to create a "bag-of-words" (BoW) representation. Examples of pre-processing[2] include:

- Breaking the document into individual words (known as Tokenization)
- Removing "stop words" (e.g. "the", "and", "but" etc.), numbers, punctuations etc.
- Stemming i.e. reducing a word to its root form (e.g. "growth" to "grow")

The pre-processed text (before stemming) is visualised below. As expected, the words "growth" and "inflation" feature prominently in the MPC minutes, reflecting the balancing act between achieving economic growth and price stability.

Figure 8: Visualisation of BoE minutes (before stemming)

Under the BoW representation, the frequency of each word in the text is used as a feature for training the model. Frequency alone is generally a poor indication of the importance of a word. The weighting scheme used for the BoW representation is "term frequency-inverse document frequency" (tf-idf) which gives more weight to words appearing in fewer documents in the corpus. However, BoW disregards the order and context of the words. A popular alternative, which addresses these limitations, is the "paragraph vector"[3] representation.

6.1.4.2 Model training

The MPC meeting minutes are partitioned into training data and testing data. 70% of the meeting minutes, selected randomly, are used as the training data while the remaining is used to test the resulting model. The training data is in the form of the BoW representation of the MPC minutes (i.e. the feature matrix) and the associated classification label – hawkish (1), neutral (0), or dovish (-1).

[2] David Bholat, Stephen Hansen, Pedro Santos, Cheryl Schonhardt-Bailey (2015), *"Text mining for central banks"*

[3] Quoc Le, Tomas Mikolov (2014), *"Distributed Representations of Sentences and Documents"*

To allow for model uncertainty and to avoid over-reliance on a single model, multiple supervised learning techniques are used to build an ensemble model (i.e. a blend of multiple models). An overview of two of these machine learning techniques – Artificial Neural Network and Random Forest – is provided below.

- **Artificial Neural Network**[4] loosely mimics the structure of a biological neural network. The BoW representation of the MPC minutes forms the inputs; they are given weights and passed through layers of "neurons", non-linear functions which are activated when a threshold is reached, arriving at the classification output. The "learning" process (i.e. optimisation of the weights and thresholds) repeats until the training data classification error[5] is minimised.

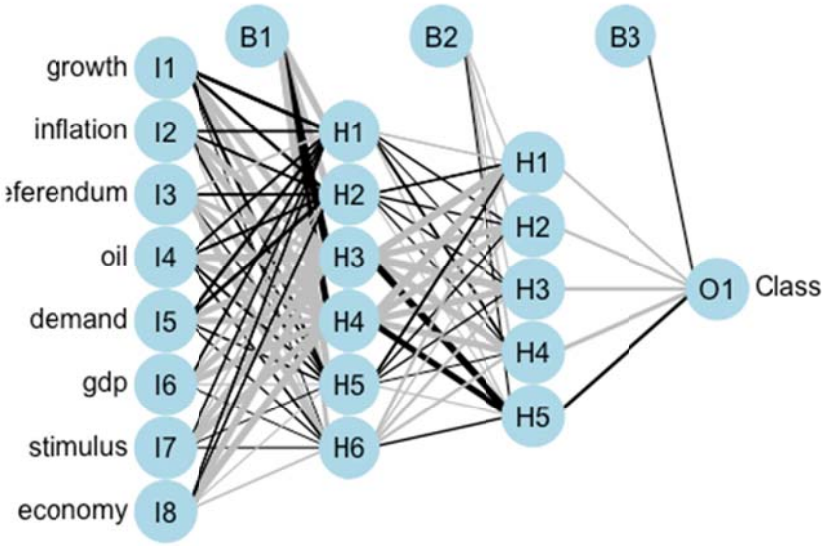


Figure 9: Visualisation of a shallow two-layer neural network

- **Random Forest**[6] addresses the tendency of decision trees to over fit the training data (which results in high in-sample accuracy but poor out-of-sample accuracy). It operates by constructing a multitude of decision trees (an example is visualised below), each different from the other due to a random selection of training data and features in the training data. The result of each decision tree is then aggregated; this involves having each decision tree vote with equal weight, with the majority vote taken as the output of the model.

[4]Michael A. Nielsen (2015), *"Neural Networks and Deep Learning"*
[5]In practice, the mean squared error (MSE) is minimised
[6]Leo Breiman (2001), *"Random Forests"*

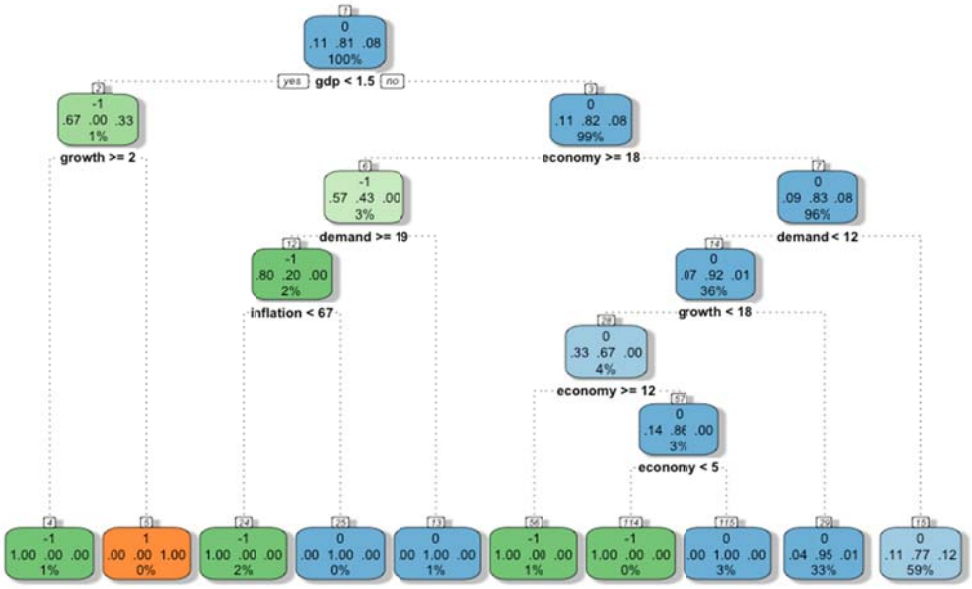


Figure 10: Example of a decision tree

The sentiment analysis model is an ensemble of separate models constructed using various machine learning techniques. The output of each model is aggregated in the same manner as in the "bagging" method used by a random forest described above. Alternatively, "stacking" can be used; this involves training a learning algorithm to combine the output of each model.

6.1.4.3 Model testing/validation

The remaining 30% MPC meeting minutes are used to test the sentiment analysis model. The out-of-sample performance metrics used include:

- Accuracy – the proportion of correct classifications out of all the test samples
- Precision – the proportion of correct classification out of test samples with a given classification (i.e. true positives)

Model training is repeated by tuning the hyper-parameters (e.g. in the case of the ANN, the number of hidden neuron layers) and/or re-selecting the training data until the out-of-sample performance is satisfactory. In practice, model tuning is more art than science. Progress is being made in the data science and analytics community to automate this process.

6.1.4.4 Deployment

Once trained and validated, the sentiment analysis model can be used to classify other forward-looking BoE communications (e.g. speeches, inflation reports, press releases etc.) to provide a prediction of BoE monetary policy ahead of the next MPC meeting. The prediction is fed into an interest rate-forecasting model, which has been trained using publicly available datasets obtained from the Bank of England (BoE), the Office of National Statistics (ONS), and the UK Debt Management Office (DMO).

The basis of the interest rate-forecasting model is Bayesian Structural Time Series (BSTS)[7], a supervised machine learning technique designed to work with time series data, and Random Forest (RF). For the purpose of this case study, the model is trained to forecast the 10-year spot rate. As part of model training, the QE programme, BoE sentiment, and factors which influence monetary policies (e.g. unemployment rate and inflation) are highlighted by BSTS and RF to have explanatory power.

The model is back-tested to measure its out-of-sample performance. This is achieved by repeating model training and forecasting across time, and comparing the forecast (e.g. mean of the conditional distribution of interest rate) to the actual interest rate. The performance of the model is measured using the following metrics.

[7] Steven L. Scott, Hal R. Varian (2014), *"Bayesian Variable Selection for Nowcasting Economic Time Series"*

- Mean absolute error (MAE) i.e. the average absolute difference between the forecast and actual interest rates
- Directional accuracy i.e. % of correct up/down movement predictions

It is also common practice to measure the performance of the model relative to a base model in order to gauge the improvement, if any, gained in forecasting accuracy.

Given the inherent uncertainty in making forecasts, in addition to a point prediction, the model also provides a range of feasible outcomes.

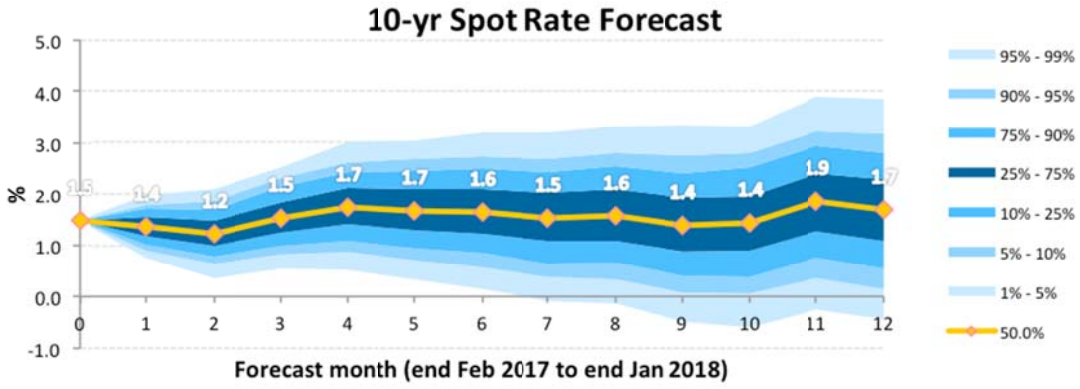


Figure11: 10-year spot rate forecast fan chart

6.2 Pricing of Marine Hull

6.2.1 Background

The project aims to identify the benefits of machine learning over traditional methods and highlight the challenges faced. Typically, the current approach is to apply a base rate to an exposure value (e.g. sum insured or gross tonnage) with some adjustment to reflect some of the nuances of the risk. It is felt that such a simplistic approach can lead to cross-subsidies within the pricing.

A key difficulty in modelling this area has been to obtain a sufficient volume of credible data, in particular having claims paired with its corresponding exposure data using an appropriate unique identifier. Generally companies do not store claim data at a sufficiently granular level in order to do this. However, with the help of a third party, historic claim values have been collected along with the vessel's IMO number. This allows each claim to be married up with its corresponding rating factors, and for the data analysis to be performed.

By benchmarking the findings against the current insurers' base rates we aimed to identify certain parts of the book which may be over- or under-priced.

6.2.2 Problem Definition

Approximately 18,000 exposure points were available for the purpose of this analysis. Each data point was defined as a unique policy (i.e. IMO/Year combination) and covered the period between 2010 and 2013. Each of the 1,162 non-zero claims were then allocated to one or more data point based on the IMO number and the following logic:

- If the claim occurred prior to the inception of the policy then it was defined as a historic claim and was used as a rating factor.
- If the claim occurred during the policy period then it was used as the predicted value.
- If the claim occurred after the policy period then it was not used.

Two different approaches were then considered:

a. Model the aggregate claim value

b. Model frequency and severity separately, combining the results to give the aggregate claim value

It was hoped that by looking at frequency and severity independently, greater insight could be gleaned into what were the key drivers into claims.

6.2.3 Results

Frequency / Severity approach

Using the 'Poisson Deviance' as the accuracy metric, the Software determined that a GLM Blender was the most appropriate model to predict expected claim frequency. By "Blender", we refer to an ensemble of algorithms or sub-models that are pre-processed and fed into a final model. In this case, the data has been split and processed with 3 variants of an "eXtreme Gradient Boosted Trees Regressor" and finally combined again into a single model via a GLM (Tweedie Distribution). This is referenced in the table below where best fit is identified through the lowest validation & cross-validation figures.

Model Name	Sample Size	Validation	Cross Validation	Holdout
GLM Blender	70%	0.4107	0.4071	**0.3735**
ENET Blender	70%	0.4155	0.4109	0.3737
AVG Blender	70%	0.4168	0.4125	0.3735
Advanced GLM Blender	70%	0.4169	0.4108	0.3725
eXtreme Gradient Boosted Trees Regressor with Early Stopping (Poisson Loss) and Unsupervised Learning Features	70%	0.4169	0.4126	0.3776
eXtreme Gradient Boosted Trees Regressor with Early Stopping (Poisson Loss)	70%	0.4180	0.4179	0.3742
Advanced AVG Blender	70%	0.4203	0.4209	0.3765

Table 12: Top Algorithms for Predicting Frequency

For modelling severity, the 'Gamma Deviance' was used to rank each model. The Software scored an 'eXtreme Gradient Boosted Trees Regressor' as the best fitting model. The results of these were then combined to see how well the aggregate claim value was modelled.

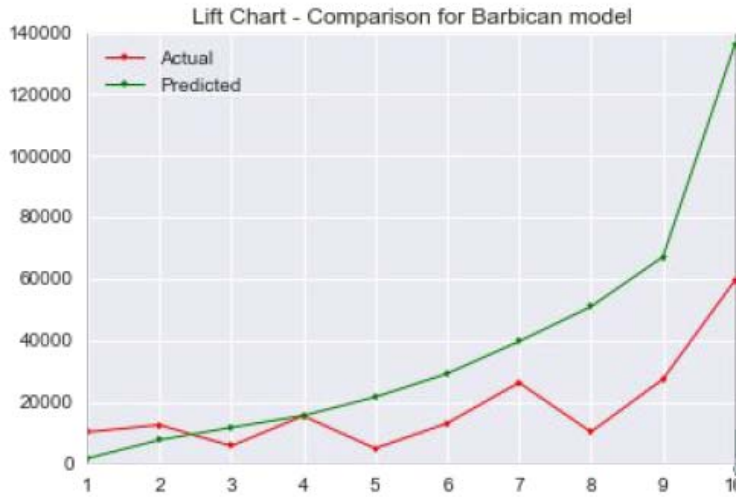


Figure 13: Lift chart comparing predicted versus actual values

The lift chart ranks the predicted value from lowest to highest, and buckets them into 10 parts.

For each risk within a given bucket, the average predicted value is compared to the average actual value for corresponding risks. For a well-fitting model, it is expected that the characteristics of the two lines on the graph should be similar. If this were the case, it would mean that the predictive model could distinguish between better and worse risks. If the two lines were also close together, then we would also have found a model which could, on average, accurately predict the total claim value.

As can be seen above, neither of these desired characteristics holds true. It could be due to the limited dataset introducing additional error when trying to break the problem into smaller components. However, the lift chart is not perfect but does potentially highlight higher risk in bands 7, 9 and particularly band 10. This could potentially be used for pricing even if the whole model wasn't adopted.

Modelling aggregate claim value

In contrast, modelling the aggregate claim value yielded a better result. Again a GLM Blender was the top-ranked result (Table 14), though this time the lift chart (Figure 15) showed that the predicted values were more similar to the actual underlying data.

Model Name	Sample Size	Validation	Cross Validation	Holdout
GLM Blender	64%	582.0261	598.2154	662.7883
Advanced GLM Blender	64%	594.2741	634.6337	629.1668
ENET Blender	64%	603.2168	639.5828	626.2795
eXtreme Gradient Boosted Trees Regressor with Early Stopping (Poisson Loss)	64%	603.9415	628.3817	645.6146
Advanced AVG Blender	64%	604.7846	657.4504	612.8320

Table 14: Top Algorithm for predicting aggregate claim values (GLM Blender)

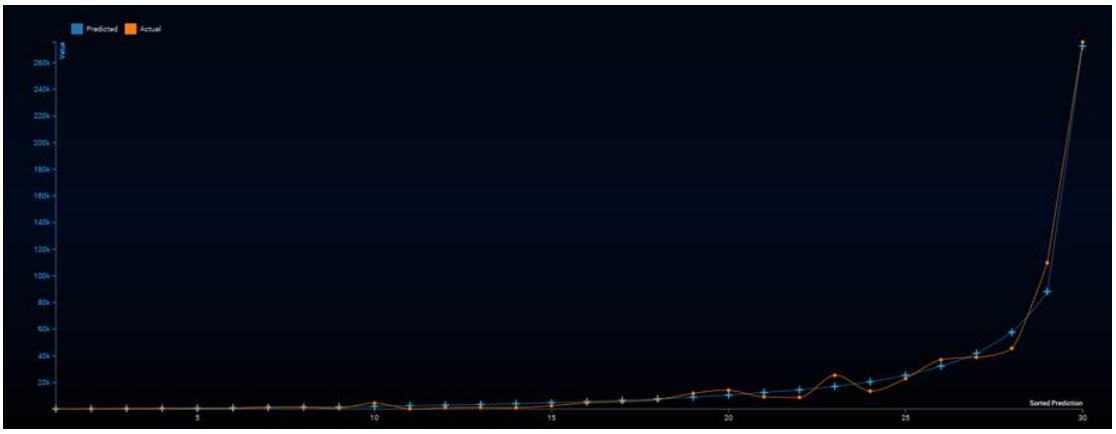


Figure 15: Lift Chart

Based upon this, it was decided that the second approach would be pursued to gain additional insights into the underlying book.

As seen in Figure 16, the Sum Insured and the vessel's physical attributes were the main driver of aggregate claim value. The size of any historic losses was also important, though surprisingly the number of claims scored low. Again this could be driven by the data set where the number of historic claims were limited.

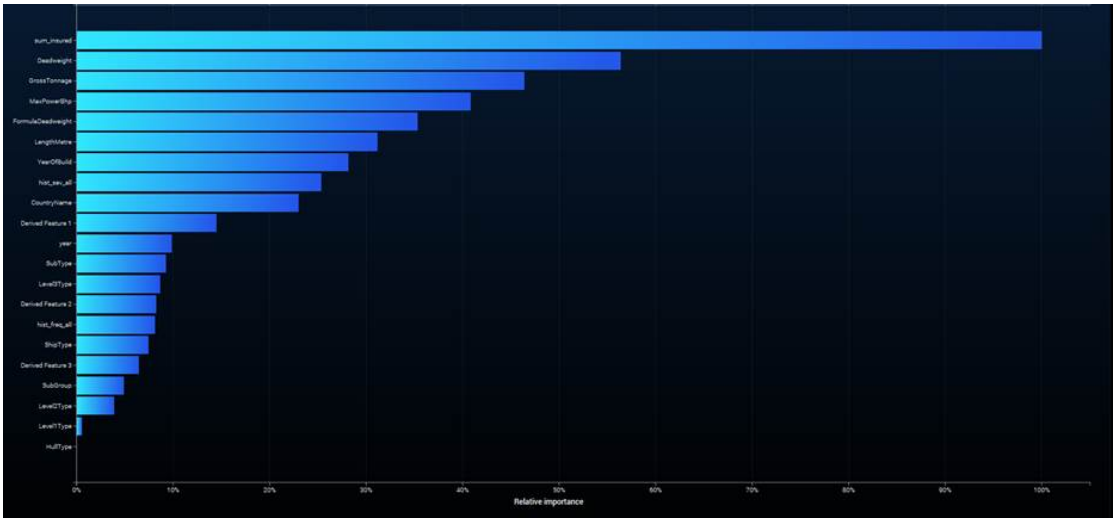


Figure 16: Chart showing key drivers for Aggregate Claim Value

Looking at each input in isolation, some interesting insights can be deduced:

- **Types of vessels**- seems not to have a big impact on predicted claim value, though Ro-Ro's and Cellular Container vessels perform better than other types.
- **Gross Tonnage**- higher value increases the expected claim value. However, the modelling suggests that it reaches a cap - once a vessel is above circa 10,000 tonnes the expected claim values stays relatively constant. Not unexpectedly, a similar theme emerges when looking at Formula Deadweight.
- **Year of Build**- indicates that old vessels (pre 1980) and newer ships (post 2000) are predicted to have lower claim values than those in between. This could be due to newer safety features on board newer vessels, and older boats being used less so less likely to be exposed to a loss (i.e. Year being used as a proxy for mileage, which was not available for this analysis).
- **Country of origin**- seems to have an influence, though this could be more to do with the subsequent routes vessels must take. The yellow dots below plot the expected claim value as the Country changes whilst the blue bars indicate the volume of data points within each section.

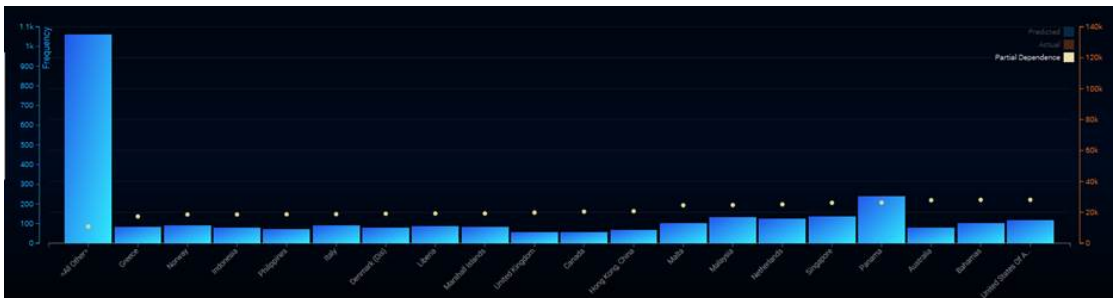


Figure 17: Model X-Ray for Aggregate Claim Value

a. Comparison against Market data

Each vessel within the dataset was run through a standard market pricing engine to derive an expected loss cost. This was aggregated by Year of Account and compared against the predicted values being produced from the Software engine. Disappointly, the average values were some way off. Of the 1,162 claims being used for the analysis, only a small percentage were sizeable claims (excess of 500k) whereas the base rates being used within the market probably assume a higher number of large claims.

6.2.4 Limitations and Scope

These concepts and logic can be applied to any class of business, though is limited in its effectiveness by the quality of data available. This analysis was based upon 4 years of exposure data, and one would expect to get more stable results if a larger dataset was used.

Rating factors such as mileage travelled, or the average location of the ship (using latitude and longitude) were not available in sufficient quantity to be used within this exercise. Having this positional and voyage level information would likely act as a proxy for how and where the vessels are used which would be expected to influence the frequency and severity of loss(es).

In addition, there is scope to link this data to external providers of financial information. For example, this could then be used to connect the ship owner's financial history, or the general economic outlook, to the claim data to understand if macro or micro-economic factors are important. Meteorological data would also be expected to give additional information into frequency and severity of losses. The Software could then be used to effectively process this data and find useful drivers of loss.

6.2.5 Conclusion

The Software has provided a quick iterative process for modelling, retraining and thus refining the model. Further work and data required to draw any firm conclusions on an improved basis for pricing risk. Finally, it is important to remain open minded on what the data is suggesting rather than reverting back the traditional views that maybe suggested by the traditional methods.

6.3 Supervised learning in Exposure Management

6.3.1 Background

The project aims to identify the benefits of machine learning over traditional methods and highlight the challenges faced. Exposure management departments at insurance companies will receive building and property exposure to catastrophes such as earthquakes and hurricanes. They model these perils in a catastrophe software from AIR called Touchstone. This model is made up of several modules for example a vulnerability module and a financial module. The latter module calculates our losses based on the property data which we supply to it. These attributes may include what the building is used for ('occupancy 'codes) and what the building is made of ('construction' codes e.g. reinforced concrete). The problem arises where the data we receive lacks this information. In this project, we aim to predict the fields 'year built' and number of 'stories'.

6.3.2 Problem Definition

The problem is split into two separate projects, one for 'year built' and one for 'stories.' Each is a regression problem aiming to calculate the dependent variable or *predictor*. The dataset is the same for both projects. It is a sample of 400,000 records of property data taken from the insurers' exposure database. Where 'year built' and 'stories' are known these are used as the training set. Part of the training set is partitioned and used to test on. It is validated against the actual value, thus the predictability of the model can be measured. Finally, an appropriate accuracy metric is selected for the model. Root mean squared error (RMSE) is used for year built and Poisson Deviance for stories (as the data for stories is highly skewed with most buildings being only one story high).

RMSE is a common accuracy metric. One calculates the residual

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

These are squared, averaged and square-rooted to omit negative residuals.

Results – Building Stories

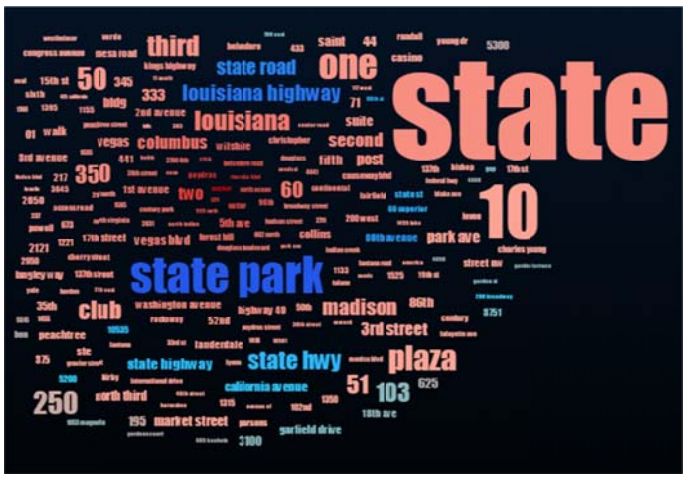


Figure 18: Stories Word Cloud

The algorithm selected for the model was an 'eXtreme Gradient Boosted Tree Regressor' with a (Poisson deviance) error of approx. 1 story. The data was heavily skewed with most properties being one story high and affected by a few tall skyscrapers. The value of the property was the most influential feature relating to building height. A word cloud generated from address text showed that if the first line of address contained 'state park' it was likely to be a taller building (Figure18).

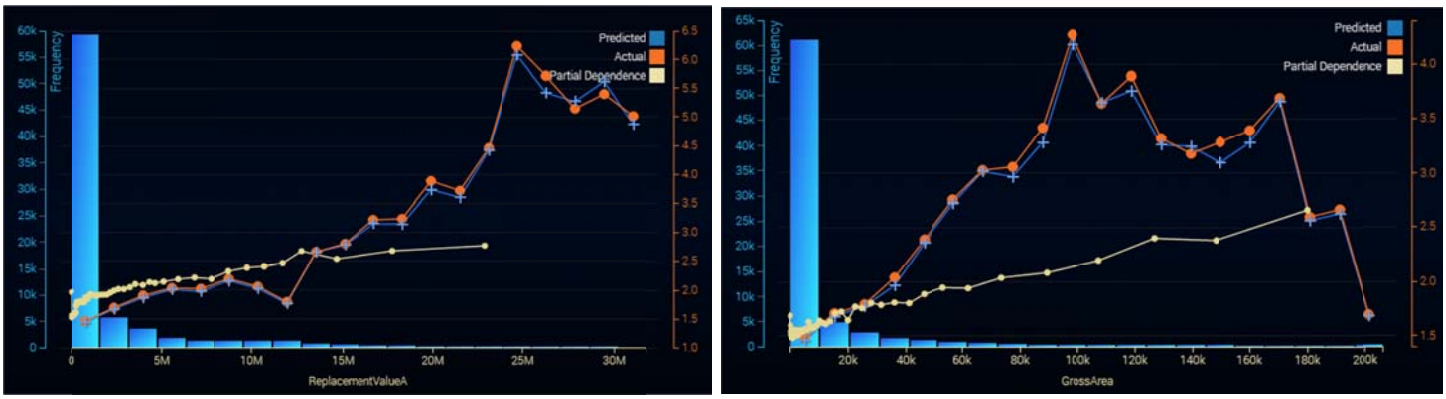


Figure 19: Model X-Ray for Stories: a) Replacement Value A and b) Gross Area

Figure19 shows that 'Replacement Value A' also known as building value and 'Gross Area' are related to the number of stories of a building. This can be seen by the close relation between actual and predicted values (blue and red lines accordingly) and by the increasing gradient of partial dependence (yellow line). It can be concluded that as buildings increase in value and/or plot area, the number of stories are likely to increase as well.

Results - Year Built

The same algorithm type was selected as above (see Table 19 for the top algorithms) with a Root Mean Squared Error (RMSE) of 11.51 years. Address fields which contained 'Burden Avenue' were much older and buildings containing 'San Juan' were more recently built. Longitude and latitude were the most accurate features that could predict when a building was built.

This algorithm 'eXtreme Gradient Boosted Tree Regressor' (also known as XGBoost) is a boosted tree that is pushed to its computational limits in terms of scalability, portability and accuracy. As a more complex model the XGBoost algorithm enhances standard tree regressors by:

- **Reducing the risk of over fitting**, making better decisions per branch by setting a score to each leaf.
- **Creates a tree ensemble** (trees within trees) it sums each of these tree groups together and chooses the best trees within the model.
- **Regularisation and sorting techniques** optimises the tree.
- **'Pruning' techniques** are used to not add neighbouring branches that are of lower score than the current leaf.

As a boosted tree, and with all supervised models, we define an objective function and optimise over it as part of the model fitting process.

Model Name	Sample Size	Validation	Cross Validation	Holdout
eXtreme Gradient Boosted Trees Regressor with Early Stopping	64%	11.5185	n/a	n/a
Advanced GLM Blender	64%	14.1973	n/a	n/a
GLM Blender	64%	14.7232	n/a	n/a
Advanced AVG Blender	64%	14.7626	n/a	n/a

Table 20: Top Algorithms for Predicting Year Built

6.3.3 Exposure Management

These concepts and logic can be applied to any kind of data quality problem where some data is missing but could be filled in by comparing with known non-missing data. This of course will vary on the type of dataset used and how good the available training data is.

Currently the Software does not support supervised learning for multiple classes. Some features that have a more direct influence on pricing could not be modelled as multiple classes (i.e. construction and occupancy codes). Thus the two continuous variables year built and stories were selected. Figure 21 shows that there is a relationship that 'year built' can be predicted from 'occupancy code'. The Software can show these relationships, but is not as yet able to predict categorical occupancy codes.

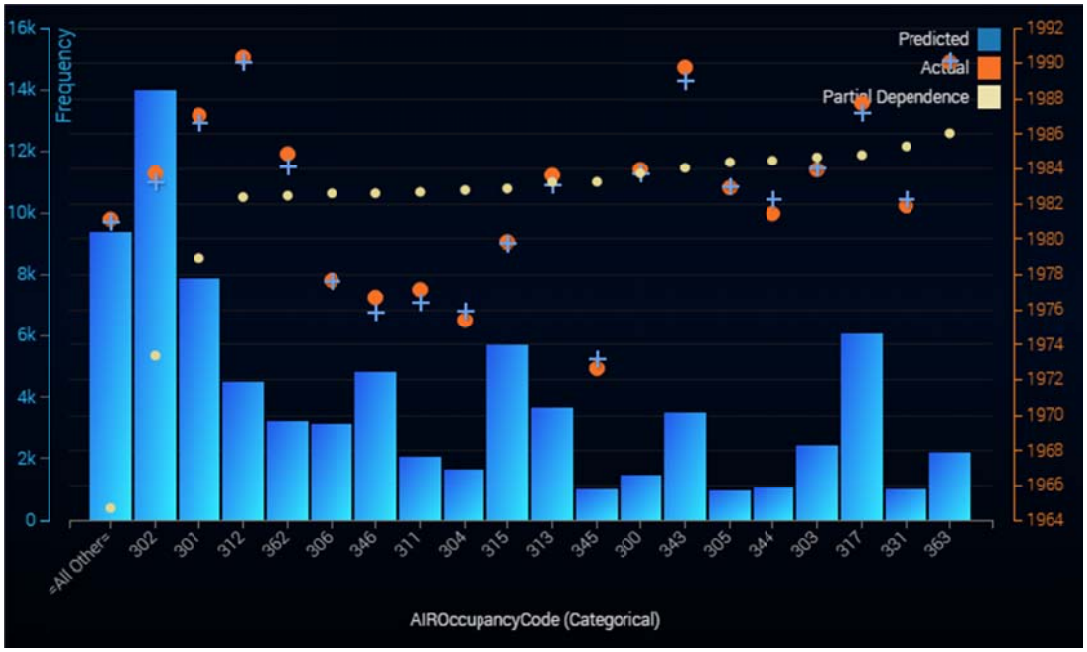


Figure 21: Model X-Ray for Year Built Drill Down on Occupancy code

There is scope for improvement by increasing accuracy of the model with more input data. The error metrics of 11.51 years for year built and 1 story could be reduced. By removing unimportant features, noise can be reduced and a more accurate model could be produced. Due to time constraints, it was not possible to see the financial impact of using machine learning. For a future investigation, it would be beneficial to see if using machine learning to predict fields was better than leaving these fields blank thereby increasing the accuracy of expected loss and pricing.

Within the property class in exposure management, latitude and longitude is normally the minimum requirement for address information that goes into a catastrophe model. Latitude and longitude could be further investigated i.e. if this data is removed from the overall dataset could other city features be observed? One could concentrate a weaker model on factors which relate different areas, such as age of city, within cities with (latitude and longitude) data and cities without. In principle, this highlights the importance of selecting the right input features that go into the model to be generalised.

6.3.4 Conclusion

The Software has provided a quick iterative process for modelling, retraining and thus refining the model. As such, there was a positive correlation between building value / gross area and the stories of a building. The most accurate algorithms were found, with 'eXtreme Gradient Boosted Tree Regressor' coming on top. Word clouds show that 'state park' was a key driver in accurately predicting a building with greater number of stories. For business impact it would be more useful to model occupancy and construction codes. However, this project has found very useful patterns despite not predicting multi-class variables. It should, therefore, prove to be effective in finding key relationships from data once the Software is able to predict multi-class features.

6.4 Mortality experience analysis

6.4.1 Background

This investigation aims to use supervised machine learning techniques to perform a component of experience analysis on death statistics. Experience analysis is fundamental work performed by the life insurer. Consequently improving the speed and the accuracy of analysis will add much value to their line of business.

6.4.2 Preliminary Analysis

The aim of our preliminary analysis was to look for patterns in 2014 US death records. We wanted to understand which models would best fit our set of standalone data on death statistics. We specified the following metrics:

- Outcome variable (target): "Age of Death"
- Accuracy Metric: "Root Mean Square Error" (RMSE)
- Feature List: including death circumstances and personal data (such as educational level)

It is important to note that within this work full population data was not available (i.e. individuals that did not die). As expected, results from initial findings on a standalone set of data showed marital status and age-related features were the best predictors for when a person will die.

Following the fitting of various statistical models the best predictive models in respect of the RMSE on a test dataset, were:

- Ridge Regressor with Binned Numeric Features
- Decision Tree Regressor
- Gradient Boosted Tree Regressor (Least Squares Loss)

Although the analysis was relatively basic and limited by the fact that we didn't have population information, we were able to extract relationships from the data with no human intervention in a few minutes using Machine Learning and our chosen modelling platform.

The table below shows the output leader board, with all the models trained ranked from the lowest to the highest RMSE when comparing predictions vs. actual values on a 16% validation sample. The most accurate predictive model (i.e. lowest RMSE on the validation sample) were Ridge Regressors (in years).

The RMSE on the 16% validation set is 6.14 years for the Ridge Regressors. In order to assess the validity of the predictions, we could compare this number to the RMSE we could get with a benchmark "dummy" model predicting always the average Age. The RMSE of such a benchmark would be 18.49 years on this dataset. We can conclude the results regression model is able to predict age of death with a decent level of accuracy compared to an average.

Model Name	Sample Size	Validation	Cross Validation	Holdout
Ridge Regressor with Binned numeric features	50%	6.1411	6.4568	7.3729
Decision Tree Regressor	16%	6.2812	n/a	7.5382
Gradient Boosted Tree Regressor (Least Squares Loss)	16%	6.3001	n/a	7.4851
eXtreme Gradient Boosted Trees Regressor with Early Stopping (learning rate = 0.12)	16%	6.3570	n/a	7.5303

Table 22: List of most accurate models run by DataRobot Autopilot mode, based on RMSE Leaderboard for best predictive models

6.4.3 Further Analysis

The next stage of our investigation aimed to link new external data sources to this US mortality dataset in order to get new & additional insights.

We define the following specific terms:

- Basic data: Death statistics in the US[8]
- External data: Mood Index; Consumer Confidence Index; Dow Jones[9]
- Technology in the industry: A software platform based on machine learning techniques
- Risk: A component of the risk of death i.e. suicide
- Experience Analysis: Historic statistics of suicides

The second analysis looked into combining external data sources with current mortality records to see patterns with macro data and suicides. The goal was to find potential causation or correlation.

This is an example of a multivariate time series use case when the aim is to forecast the change in the number of suicides based on external factors like mood index; market data (Dow Jones) and confidence index. These types of economic variables (suicides, confidence etc.) tend to have a high autocorrelation, which means that the value on Month M is expected to be highly correlated with the value on Month M-1 or M+1.

We defined the following metrics:

- Target variable: Difference in counts of suicide between Month M+1 and Month M
- Accuracy Metric: "Root Mean Square Error" and R-squared
- Feature List: Aggregated features based on historical values of Mood Index, Consumer Confidence and Suicide Counts (Male/Female).

In addition to gender-based groupings; age bands could also be used as an additional grouping to perform more granular analyses in the future.

Partitioning used in the analysis:

a. Training / Validation / Holdout (out-of-time validation)

One strategy we applied to validate the models was to use data from 1980 to 2000 to train the model, data from years 2001 to 2008 as a validation set (to select the right model) and years 2009 and 2014 as the final holdout set (to assess the stability of our approach).

[8] Death statistics: www.kaggle.com/cdc/mortality. A detailed report on deaths in the US released by the Centre for Disease Control and Prevention. The original dataset consisted of detailed death records for each death in the US in 2014 including causes of death and the demographic background of the deceased.We supplemented this dataset with additional records obtained from www.cdc.gov for 1980-2013

[9] Mood Index: https://secure.psychsignal.com/mood-index - this is a daily mood index from the US derived by Psychsignal who provide trader mood data, analytics and indices based on twitter data & markets.This was available from 2011-2014

Consumer Confidence Index: https://data.oecd.org/leadind/consumer-confidence-index-cci.htm - this is a monthly index of consumer confidence for each country (including US). It is defined as an indicator designed to measure consumer confidence. This was available for all years of our analysis i.e. 1980 – 2014. The idea behind the Consumer Confidence Index (CCI) is that if consumers are optimistic, they tend to purchase more goods and services. This increase in spending inevitably stimulates the whole economy.

Market Data: https://finance.yahoo.com/quote/%5EDJI/history?p=%5EDJI - this is the Dow Jones index which could act as a proxy for the US stock market.

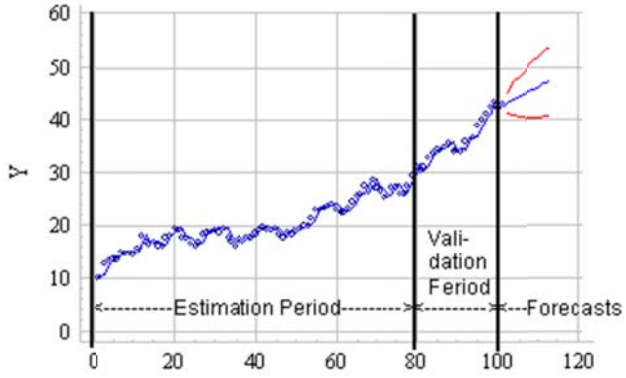


Figure 23: Partitioning approach
Source: https://people.duke.edu/~rnau/three.htm

b. Cross-validation with folds split by year

Another approach we used to estimate the accuracy (i.e. generalization error) of our models was to use Cross-Validation .In order to use cross-validation for time series, the dataset was partitioned by date using 5 folds of cross-validation, and the last two years of data were kept as a holdout set.

Each year between 1980 and 2014 was randomly assigned to a fold of cross-validation between 1 and 5. The final two years are the holdout set.

Pro: Reduce the variability of the generalization score (i.e. the RMSE expected on new data), because it is evaluated on a larger dataset.

Con: Using this technique for time series analysis can however introduce bias in the models since some parts of the data used for training the models can be posterior to where the model is evaluated. Thus, using aggregated historical features can introduce "target leakage" in the model.[10] Another strategy which wasn't used but could have considered is "back-testing".[11]

The evaluation metrics used were the root mean squared error ("RMSE") and the R-squared (also known as the "Coefficient of Determination"). These metrics were chosen because they are very suitable for outcome variables that are normally distributed (which is the case for suicide count changes).

6.4.4 Improving our Result: Iterations of our modelling approach

Changing the target variable to Change in Counts of suicides

Initially we approached the model by aiming to predict whether there was a link between the count of suicides and the external data. Following initial inspection this did not result in a clear pattern. Following further review and investigation we hypothesized that the change in the suicides would potentially be a better predictor.

Therefore our final analysis was based on whether there was a pattern/link between the change in suicides and the change in the external data used (i.e. confidence index).

Feature selection and iterating with the models

Several functionalities within the Software were used to iterate and improve our models:

1) **The Autopilot** is the standard way to apply models within the Software. It runs dozens of algorithms and pre-processing steps on the datasets and compare their accuracy. This was used to select a few algorithms suitable for this dataset.

2) **Feature Impact** allows us to rank the different predictors according to their influence on the model accuracy. The highest ranked variables have the highest impact on the model accuracy. It is derived from the Permutation Importance metric. Feature Impact was mainly used to select a suitable subset of predictors and simplify our models. The initial number of explanatory variables (aggregates of time

[10] http://machinelearningmastery.com/data-leakage-machine-learning/
[11] https://faculty.fuqua.duke.edu/~charvey/Research/Published_Papers/P120_Backtesting.PDF

series) being high, it is necessary to select a subset of them to improve the model performance and achieve interpretable results, in particular in presence of collinear features.[12]

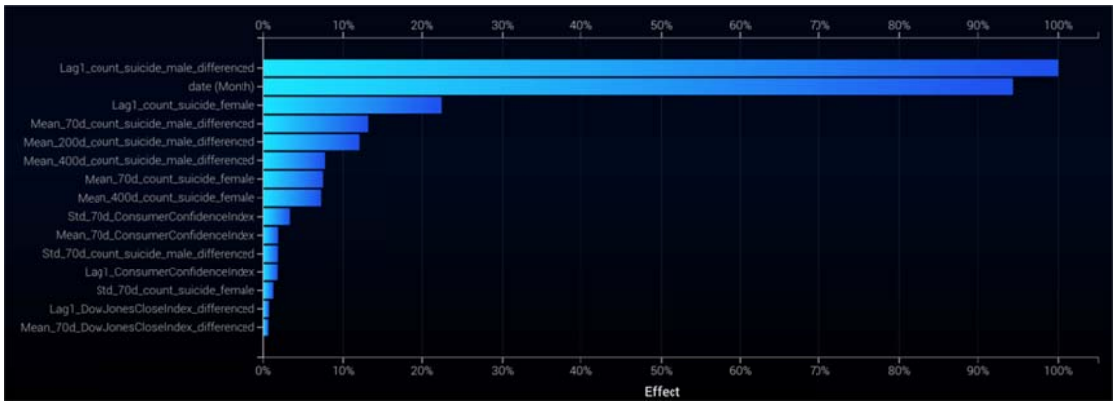


Figure 24: Feature impact Chart of an "ExtraTrees" classifier, showing the most important factors to predict counts of suicide.

Feature Impact (Figure 24) shows the most important factors in the model, with the most predictive variables having the highest value.

The most important factors are "Lag1_count_suicide_male_differenced", i.e. the lagged version of the difference in male suicides, followed by "date (Month)", i.e. the month of the year and then multiple variables named "Mean_***_count suicide_", which are averages of past values of suicide counts.

Definitions used:

- Seasonality: the month of the year has a high influence to predict changes of suicides
- Historical values (lags and moving averages) of counts of suicides and its changes
- Moving average of Consumer Confidence index

6.4.5 Results

Summary of Conclusions

Results showed Dow Jones could not predict number of suicides and changes in suicides (the null hypothesis). Similarly, there was not enough data points of Mood Index to conclude regarding its relationship with counts of suicides in this analysis (the null hypothesis could not be rejected). However, a possible relationship between the Consumer Confidence index and the number of suicides was identified, with increases in confidence potentially yielding a decrease in suicide count.

Non-linear dependence of Change in Confidence Index with Change of Suicides

[12] https://academic.oup.com/bioinformatics/article/26/10/1340/193348

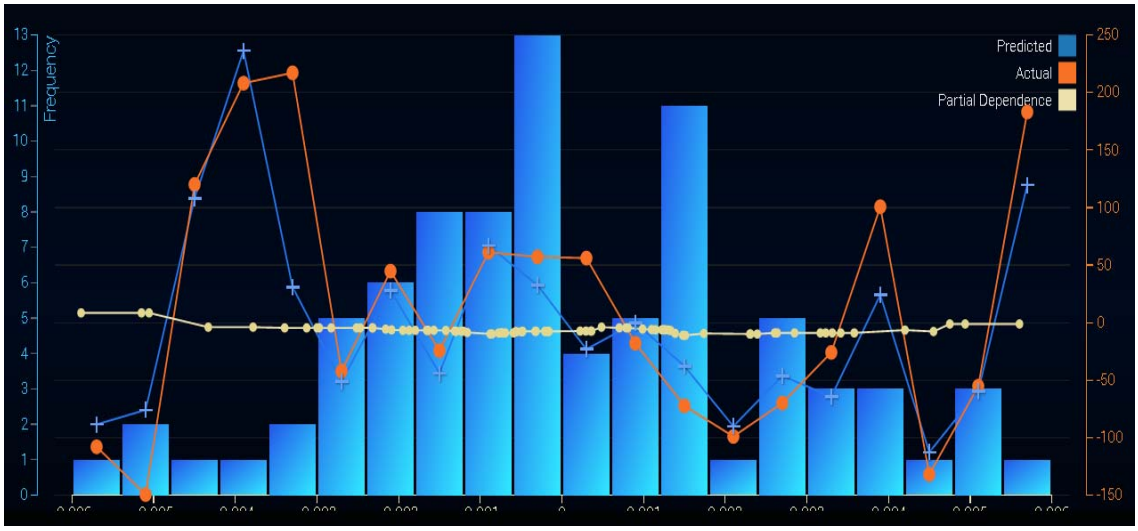


Figure 25: Model X-ray chart for Confidence Index change (X-axis) vs. Change of Male Suicides (Y-axis)

The graph (Figure25) shows a "Model X-ray" chart produced by the Software for an Extreme Gradient Boosting algorithm on the following predictor: "Change in Confidence index vs. previous Month". The decreasing trend outlined by the partial dependence in the above chart shows the model found that there could be a decrease of up to around 10 to 20 suicides when the Confidence Index is increasing, compared to when it is decreasing. This result was outlined after adjustment of other dependent factors in the model, like historical aggregates of counts of suicides and seasonality. This partial dependence chart is the equivalent of a dependency chart of a linear model, for which the slope of the curve is constant. It is widely used to diagnose non-linear models.[13]

Linear dependencies

Another way to look at the relationships in the data is to restrict ourselves to linear models built within the Software (e.g. Ridge Regression, Elastic-Net or pure Least-Square models). They have the benefit of summarizing the relationships in the data via the use of coefficients.

In our case, we built multiple linear models to predict the change in suicide counts in the following way:

$$Y_i = Suicide_{month\ i} - Suicide_{month\ i-1} = \alpha + \sum \beta * x_i + \epsilon_i$$

With Y_i being the change of suicide count for Month i, α the intercept, β the vector of coefficients of the model, x_i the feature vector on month i (for example change of Consumer Index, or value of Suicide in month i-1).

The Software trains several linear models on the dataset with different techniques, then ranks them according to their accuracy. In order to show the lack of relationship between DowJones and count of suicides, we have applied the following features:

1. Month (categorical variable indicating the month) to capture seasonality
2. DowJones monthly return for the past 2 months (i-1 and i-2)
3. Lagged Change of suicide (i.e. Yi-1)
4. Average of suicide change in the past 12 months.

With 3) and 4), we try to capture relationships related to the historical trend, 1) captures seasonality, and 2) expresses DowJones changes.

[13] "Elements of Statistical Learning" 10.13.2 [L. Breiman]

The Software automatically pre-processes input features into "derived features, for example categorical variables are processed into dummies (0/1 indicator variables), and numeric variables are standardized on the unit scale. The coefficient then indicates the effect of the feature on the average change of suicide count.

Model Name	Sample Size	Validation	Holdout
Ridge Regressor	67.57%	80.7288	111.7766
ENET Blender	67.57%	81.3537	114.3508
AVG Blender	67.57%	81.4141	115.2144
Linear Regression	67.57%	81.8049	117.1633
Advanced AVG Blender	67.57%	82.1225	117.4358

Table 25: Best Linear Model run by DataRobot autopilot with RMSE on Validation/Holdout set

After selecting the best linear model according to the RMSE (Ridge Regression), we observe the coefficients trained:

Derived feature	Coefficient
March	+161.3
January	+134.1
May	+56.2
July	+47.8
April	+26.1
August	+20.1
Standardized_Mean_2months_DowJonesCloseIndex_change	-0.9
June	-38.2
October	-42.3
Standardized_Mean_400d_count_suicide_male_differenced	-45.4
December	-67.9
Standardized_Lag1_count_suicide_male_differenced	-77.4
November	-78.7
February	-88.3
September	-100.8

Table 26: Coefficients for Ridge Regression Model

The above table outlines that the standardized change of Dow Jones in the past 2 months is likely to have a very low negative effect on the suicide count (close to 1 for 1 unit of standard deviation). However, seasonality has a strong effect (from -100 to +161 depending on the month), as well as the lagged values of the suicide count change.

This result is not sufficient to justify that Dow Jones returns have an effect on suicide counts.

6.4.6 Limitations and Future Improvements

All modelling work has limitations and specifically for this investigation, it was noted that there was no information to test against living population i.e. how well able to predict suicide or no suicide- commonly found in mortality models. Similarly, there wasn't enough Mood Index data to draw statistically significant conclusions from as Twitter only exists from 2011 onwards.

There is further scope to add new external data sources as they become available e.g. more historical twitter sentiment data (more than 4 years). Therefore the more data we have from Twitter, the easier it will be to create

a model and look for patterns. Other possible macroeconomic data sources could include seasonal climatic data, political news feeds, sociological data (race, ethnicity, sex) and population data (CMR).

A UK mortality investigation could be a reasonable next step as well as looking into other causes of death. Other features on customers could be considered as consumer confidence index is predictive, but does not necessarily depend on the underlying risk categories of the person e.g. employment level, education, location, etc.

Our case study resulted in an understanding of the concept of machine learning and how it could be applied to subsets of death data. It was not intended to be a detailed experience analysis into mortality. Further analysis of insurance specific mortality could be performed in the future.

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