Data Science and Machine Learning in Insurance

Insurers operate in an increasingly data-rich and algorithm-oriented world, where growing computational power allows machines to collect, transform and analyse data with even more efficiency. Data science and machine learning present an opportunity for actuaries to innovate traditional actuarial fields and embrace new techniques to improve business operations, governance processes and customer satisfaction. If insurers are to compete in this rapidly changing and challenging industry, an investment in data analytics is essential.

Data scientists and actuaries

Insurance companies are progressively expecting staff to add data science to their skill sets. Typically, the talent possessed by data scientists combines three qualities: coding, mathematics and statistics, and domain knowledge. Whilst programming allows data transformation and the creation of algorithms, the fundamentals of mathematics facilitate the use of data to develop models and predict future outcomes. Additionally, data scientists require a capacity to interpret actual phenomena and regulation to solve real world problems. Therefore, data science extends beyond the realms of machine learning and statistical analysis; it encompasses the entire spectrum of data processing. These competencies, together with technical actuarial expertise and knowledge of regulation, are becoming highly desirable in the recruitment of actuaries.

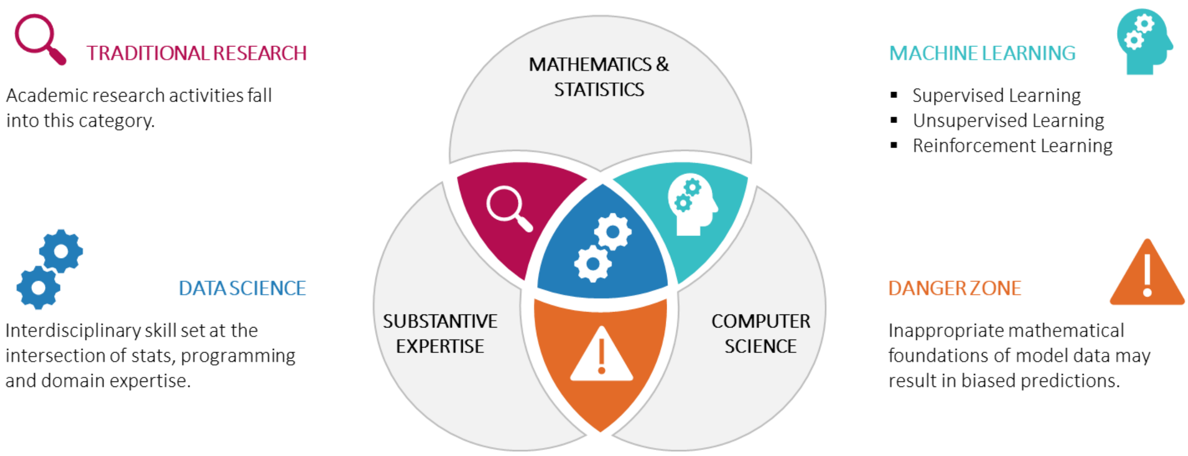

Figure 1 - Conway's Data Science diagram

Although data science and actuarial modelling have a lot in common - leaving the actuarial profession well placed to utilise new data analytics techniques - they diverge when it comes to how data scientists and actuaries operate in practice. The main difference between them arises when designing and implementing logical solutions. Actuaries typically use their domain knowledge to select appropriate models and then focus on calibrating parameters that are suitable to achieve the objective. In contrast, data scientists first spend time and effort testing numerous models, before estimating adequate model parameters. Furthermore, these areas differ as actuaries build models which are financial in nature, whereas data scientists often rely upon external experts to gain an understanding of domain specific elements. As a result, there are differences in the approach taken to validate assumptions, select variables and assess model suitability.

Big data: challenges and opportunities

Searching for patterns in large volumes of data is nothing new for the insurance sector. Nevertheless, this still poses a major challenge as unstructured and fluctuant data is not easy to analyse with traditional tools. In order to derive valuable business insights, big data requires innovative technology and methods to collect, store and process the large amounts of data gathered.

The growth of big data is encouraging insurance companies and regulators to define best practices around its use. It is common for big data literature to refer to the following '5 Vs' to describe the main challenges faced by organisations:

- Volume - data quantities have dramatically increased in recent years and will continue to do so. For example, IFRS 17 regulations are going to increase data requirements due to the need to group policies at more granular levels and upgrade data systems in order to run models efficiently.
- Variety - modern datasets can be generated from numerous sources which have diverse data structures. Historically, insurance companies have relied upon highly structured data sets stored in relational databases but with the explosion of telematics devices and social media, unstructured data has emerged providing insurers with additional information about their products and customers.
- Velocity - data does not only need to be acquired quickly, it also needs to be processed and analysed at a faster rate. Being able to understand and analyse data rapidly will help organisations speed up decision-making processes and maintain their position in a changing market.
- Veracity - with an increased volume and variety of data it is critical to ensure that data can be trusted and relied upon. To allow data analysts and actuaries to produce meaningful analyses and subsequently arrive at high-quality decisions, data must be reliable and properly understood.
- Value - business decisions resulting from big data analytics should lead to economic benefits and competitive advantages.

Figure 2 - The 5 Vs of Big Data

Given the sheer volume of data (and the increasing storage requirements) there also comes a need for external data management via cloud computing solutions. Clouds provide insurers with new platforms to manage rapidly growing data sets and improve the data storage process. Clouds evidently place new demands on IT departments, data analysts and the business functions they support, but can generate significant value across an insurance organisation:

- Underwriting - using larger pools of information to underwrite risks more effectively than competitors and optimise pricing strategies via more accurate predictive modelling techniques.

- Fraud detection - identifying policyholders with a higher likelihood of committing fraud in the underwriting process and monitoring the concealment or misrepresentation of applicant information. Big data can also be used in claims management to monitor social media for evidence of potential fraudulent behaviour.

- Claims management - implementing processes using experience and social media data to efficiently filter suspicious claims, shorten the claim cycle and reduce expenses.

- Reputation and customer satisfaction - unstructured information from social media can help insurers obtain policyholders' opinions on products and consequently, set a strategy to enhance customer retention.

From classical statistics to machine learning

For many decades, classical statistical methods have been applied by actuaries, but in a modernising industry, they are now becoming outdated. Generalised Linear Models (GLMs) for example, were historically used for pricing and reserving in the non-life insurance industry to determine how key variables (e.g. frequency and severity of claims) vary with rating factors. Over the last few years, GLMs have also gained ground in the life insurance sector, where they are regularly applied by actuaries to capture the most significant risk drivers and guide the calibration of decrement assumptions.

Yet GLMs have limitations of their own. They are parametric models which rely on the pre-determined underlying error distribution and link function. Moreover, they are not well suited to detecting interactions between variables and complex relationships. Limitations such as these can lead to poor goodness-of-fit and unreliable forecasts of future observations.

To overcome these limitations, and thanks to rapid technological advancements, machine learning (ML) is becoming more prevalent in the insurance sector. ML can create algorithms that use data inputs to recognise complex patterns, make strategic decisions and form informed predictions without explicit programming. Essentially, ML has the capacity to learn from historic experience data and make decisions without the need for human intervention.

This allows for the:

- Establishment of more complex relationships between features and outcomes than those in traditional models.

- Exploration of more subtle predictive features as well as detection of anomalies.

- Rapid adaptation to changes in data patterns and underlying business conditions.

ML algorithms are typically divided into three different classes depending on the type of problem they are to be applied to:

- Supervised learning - the goal is to predict the value of an outcome measure by using numerous input measures. The learning process is supervised due to the presence of an outcome variable that guides the learning process. Examples include regression, neural networks, and tree-based methods such as random forests.

- Unsupervised learning - there is no outcome measure, the goal is simply to describe the associations and patterns amongst a set of inputs. Examples include cluster analysis and Principal Component Analysis (PCA).

- Reinforcement learning - predicted outcomes are incorporated into the model to improve the next prediction. Over time, the predictions improve as the algorithm learns more about the environment it operates in and models are updated dynamically. Currently, it is not widely used in actuarial science, but this may change over time as statistical methods and computational power develop.

Penalised regressions (e.g. lasso, ridge and elastic net), which aim to limit the number of parameters, represent specific examples of supervised learning ML techniques which can address some of the drawbacks of GLMs. These models can reduce the variability of estimates at the cost of a negligible bias, by constraining and shrinking parameters.

Other ML techniques like decision trees, random forests and neural networks, are also making inroads into actuarial science. One area of application is in underwriting, where they can be used as a forecasting tool to categorise potential policyholders and decide whether to accept or reject on standard terms. Similarly, these techniques can be utilised in pricing and reserving - historical policyholder data, e.g. actual claim sizes and frequencies can be directed to optimise pricing strategies and estimate future losses.

Generally, ML models are either applied in their own right to formulate predictions, or together with other multivariate techniques e.g. clustering and PCA, to improve certain aspects of an analysis. Clustering and PCA are commonly used as a preliminary analysis technique to reduce dimensionality of data and eliminate redundant features. By reducing the dimensions of a training input set, the speed of training can increase and datasets can be streamlined to just a few dimensions, making it easier to perform data visualisations.

Cluster analysis can also be applied to optimise model point generation. Given time and processing power constraints it is often necessary to run actuarial models with grouped model points rather than full data runs. This approach automatically generates a user-defined quantity of model points, without manual intervention, through identifying groups of policies with similar distinctive and distinguishing features. In order to embed more sophisticated data analytics approaches and a wider range of ML methods, software houses have released new packages to satisfy a growing need.

Nonetheless, the development of ML within the insurance industry is still in its infancy. There are various reasons why insurers are hesitant to move away from classical statistical techniques and adopt ML methods. Firstly, linear models represent a simple and well-known statistical approach, and standardised software packages to implement such methods are widely available. Secondly, insurance companies have only recently started to build data analytics teams, therefore company-wide visions and strategies are still to be formed. Data science experts are often scattered across organisations, meaning expertise and activities are not optimally co-ordinated and organised. The onset of big data and advanced analytics also brings a need to invest in new technologies, provide specialist training and implement additional governance.

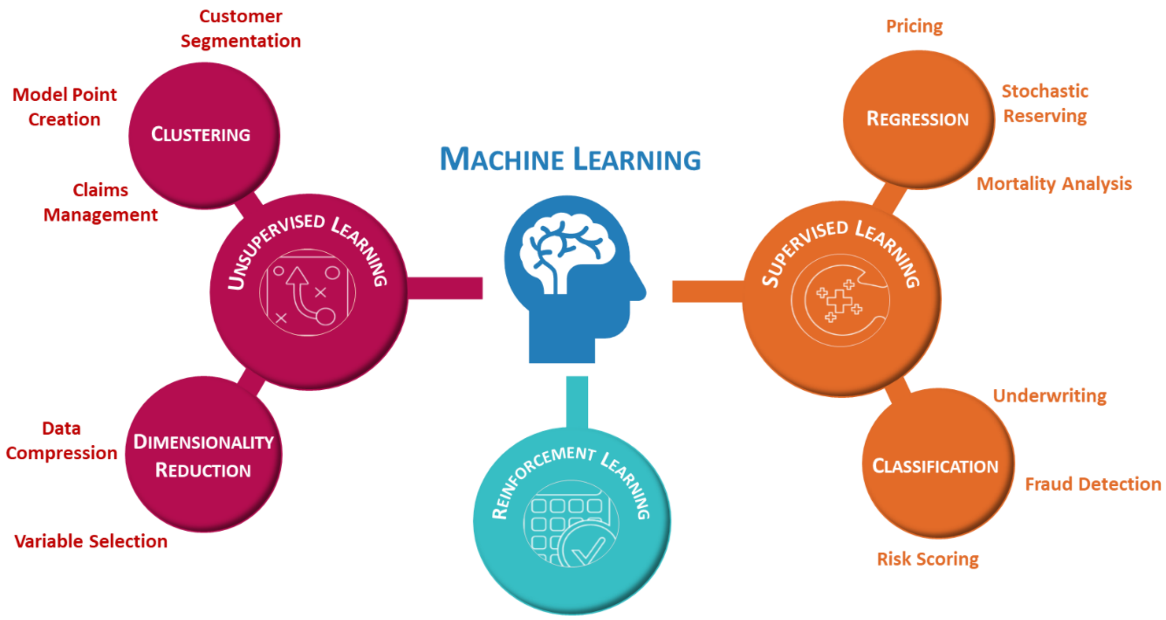

Figure 3 - Applications of ML in Insurance

Reaping the benefits

Actuarial expertise, combined with programming experience and knowledge of visualisation tools, can help insurers realise the benefits of systems refinement and enhanced customer engagement.

Data analytics and ML can add real value to the extraction of information and understanding of customers' risk profiles. Transforming systems to facilitate the discovery of data patterns can give a clear indication of future adverse events and potential costs.

Designing and executing cutting edge ML solutions will enable an insurance provider to improve efficiency and gain a competitive advantage. The scope of potential applications is vast and affects many different aspects of actuarial modelling, such as model point generation, pricing, reserving, claims management and reporting automation.

In addition, these innovations can support underwriting to precisely quantify risks and detect potential fraudulent claims by transforming systems to learn from data patterns, identify new scenarios, and react appropriately to automatically evaluate each scenario. This allows insurers to reduce time and costs where processes are largely manual, while supporting fast and efficient customer service. Predictive analytics can also cut out guess work and aid quick identification of key market segments and customer groupings, as well as predicting customer behaviour; helping insurers customise products and marketing to consumers' personal needs.

Selecting a provider that offers pragmatic solutions will add tangible value to the business, in particular:

- In-house research and innovation - the application of techniques to real-life problems are continually being investigated.
- Make it understandable for all parties - no blind trust in technique or the modeller.
- Comparison of approaches - add complexity only when needed.
- Multiple validations - get confidence around both accuracy and interpretation.
- Flexibility of development and implementation - make the solution viable for the business.

To learn how 4most can help transform your systems and make the most of the data you have available, please get in touch with our Client Partner, Chintan at chintan.patel@4-most.co.uk or Head of Data Science, Fab at fabrizio.russo@4-most.co.uk.