MILLIMAN BRIEFING NOTE

Use of internal data in insurance

Making better use of available internal data

Donal McGinley, FSAI
Bridget MacDonnell, FSAI, CERA
Eamon Comerford, FSAI, CERA

Milliman

The use of data science techniques enables the extraction of value from increasingly diverse sources of data. In a previous briefing note, we discussed the use of external data in insurance. This briefing note gives a high-level overview of how insurers can make better use of internal data to gain insight and drive competitive advantage. This briefing note is intended for readers who are relatively new to the area of data science in insurance.

The cost of storing data has been falling exponentially for decades, and many companies have started storing lots of potentially valuable internal data. Computing speeds have been increasing exponentially for decades, and data analysis software has been steadily improving, making it feasible for companies to analyse and gain insight from these larger data sets. A 2019 Milliman Data Science Survey[1] found that 50% of insurers surveyed are investing in harnessing more existing internal data and collecting more internal data going forward. Making the most of such internal datasets using the capabilities of data science techniques can allow insurers to gain greater insights into their business and target market and to gain a competitive advantage.

Data science is a term given to the broad array of activities used to gain insight and extract value from existing data sources, including techniques such as data analytics, predictive analytics, machine learning, data mining and artificial intelligence. It is certainly not a new discipline, but its widespread application to insurance is relatively new.

Analytics can be performed in any programming language. Traditionally, insurers have used a range of tools to analyse their data. For example, life insurers traditionally use Excel, SQL/Access, plus some specialised tools such as financial modelling software for life insurers and statistical analysis software for non-life insurers. However, insurers are increasingly turning to modern data science tools, such as Python and R, due to their strong computational capabilities. These powerful tools enable companies to use their internal data in new ways. They also allow companies to process larger quantities of data, and to better capture the links and correlations between the variables. They are often capable of handling unstructured data. Python and R both have many libraries providing state-of-the-art implementations, not just for classical statistics, easy data processing, machine learning and AI, but also for interfacing with other applications (e.g. Excel, SQL), building interactive web reports, and many kinds of data storage solutions. They also give access to tools that enable processing of big data, and tools which can compute extremely quickly on many CPUs or GPUs, either on-site or on cloud servers (known as 'parallel' or 'distributed' computing). They give users easy access to cutting-edge libraries and methodologies developed by leading companies in the data science world.

Data science tools such as R and Python enable companies to analyse data at all levels – policyholder level, product level and company level. They are fast enough to do calculations at a granular policyholder level[2]. They have powerful aggregation techniques to aggregate the policy-level results to product or company level. They can store the end results in traditional formats like Excel, CSV or SQL, or can save massive amounts of granular data to a server using big data storage tools. These calculations can be combined into a single automated process, enabling companies to run these powerful analyses every day or every week.

These data science tools also enable insurers to perform new types of analyses. For example, Python and R have cutting-edge machine learning and artificial intelligence libraries, and many insurers have started using these techniques in recent years. Insurers may find that these techniques work better than traditional techniques when analysing the vast amount of data which they now have.

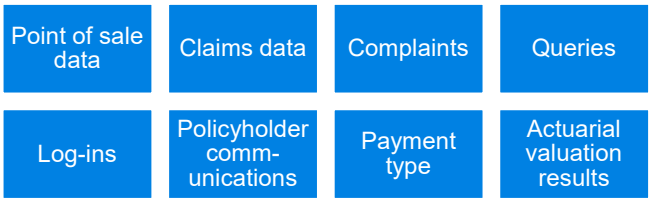
Source and Uses of Internal Data

Insurers have been using internal data and industry data to inform their business decisions and strategy for many years. However, the availability of new data science tools can enable greater analysis of existing internal data sources, which may not previously have been utilised due to system, time and computation constraints.

The following existing internal data sources can be used to perform new and advanced analyses:

[1] This survey focussed on Life and Health insurers

[2] Modern data science tools can do complex calculations on millions of individual policies per hour on a single CPU, and aggregate those results to higher levels



Data items such as complaints, online queries and phone queries are often logged and linked to the relevant policies. However, these data sources are not always analysed using sophisticated techniques. Enhanced information may also be collected by companies going forward now that such data can be utilised more effectively.

Understanding customer behaviour is of vital importance to companies. Using modern tools, insurers can analyse items of interest, such as how well the policy has performed; whether the policy performance is in line with policyholder expectations and the documentation sent to policyholders; analysis of policyholder transactions such as premium payments, switches, surrenders and lapses; and dynamic investigations into customer behaviour. When the analysis is done at a granular level, it can be aggregated up to higher levels, enabling the company to view the results at a high level for the entire company, at a granular level for any individual policyholder, or at an interim level for any particular group or cohort of policyholders.

One example that may be interesting is to investigate any links between the timing of complaints, issuance of account valuation statements or other policyholder activity and outcomes such as surrenders. Data regarding payment types such as frequency (e.g. monthly, annual) and transfer type (e.g. direct debit, standing order, cheque) may also be analysed.

Data science techniques can be used to amalgamate large volumes of data from many different internal sources. Companies may store raw data from many sources, from many different time periods, in a large data lake. The data lake may contain detailed snapshots of the portfolio at regular time periods (e.g. policy-level monthly valuation results), as well as other structured and unstructured data.

Data science tools can be used to process and extract the relevant data from the data lake. They can take data from different tables in the data lake and merge them into a single coherent structured table. For example, transaction data can be imported from an SQL server, CSV files and Excel files, and amalgamated into a single coherent data table. Actuaries can create automated processes to clean and check the data. Actuaries can spend time checking the audit trails and results summaries from these processes, and doing high-level sense checks on the results, rather than doing the process manually. They can then perform the analysis on this single clean reliable data source, rather than having to import, clean and check data at multiple stages during the analysis.

Policy-level results from traditional actuarial valuation models can be amalgamated with policy-level datafiles, and any other relevant sources of information, to create a single coherent, reliable policy-level data table which can be used in many different types of reports, analyses and investigations. For example, results can be better segmented by product level, new business during the year or quarter, age, gender or distribution channel. Analyses of movement can be conducted a policy level by comparing the current single data table to previous data tables. Detailed experience investigations, such as lapse, switch or premium elasticity investigations, can be performed at a granular policy level by comparing the current single data table to previous data tables (e.g. the policy-level monthly valuation files saved in the data lake). Historically it has been difficult to do such analyses at policy-level, as the data table may contain millions of rows, but modern data science tools can easily handle this volume of data[3]. Using a single set of granular, reliable data tables for many different tasks may improve efficiency and leave more time for checking, compared to having a separate manual process and dataset for each individual task.

Data Analytics Platforms

The choice of which platform to use to manage and analyse data can seem daunting. Microsoft packages like Excel and Access certainly still have their uses for smaller datasets, and can carry out once-off, simple analysis very effectively. The familiarity of these platforms can also help with collaboration between different stakeholders and the communication of results. However, to get more value from data, and to facilitate the repeatability and expansion of their initiatives, companies will have to look towards other platforms.

To interact with large databases, SQL has been the language of choice for many years. It is quick, can handle vast amounts of data, and can efficiently extract relevant information from the database.

However, other programming languages offer advantages over SQL for analysing and extracting value from the raw data. SAS is often used in the non-life insurance industry, but readily accessible, open-source languages such as Python and R have become the mainstream languages for data science and provide a wide choice of tools and techniques. R is often described as a statistical language written by statisticians for statisticians. It is a very popular

[3] A typical desktop computer with 32GB of RAM should be able to handle data tables of perhaps 1 billion numbers (e.g.10 million rows and 100 columns). Big data software or more powerful workstations may be required if the table is bigger than that.

tool for statistics and data analysis. In practice there does not appear to be much difference between their data analysis capabilities at present. Both languages are flexible and can respond quickly to any developments in the data science world. Both R and Python languages have active communities of developers, and we expect to see continuous improvements in their data analysis capabilities in the coming years.

When deciding what platform to use for data analytics, consideration should also be given to how the results will feed into your wider processes. Open-source languages can save results back into other formats such as SQL, or CSV. This allows analysis of the data to be carried out in Python or R, and then the results can be fed back into any existing long-term data store. However, there are clear risk management benefits in having a consistent end-to-end process, such as reduced chance of manual error, and faster end-to-end runtimes leaving more time for the results to be checked. Some open-source languages also have Excel add-ins that can be used to help bridge the gap between new and existing processes, aiding transition from one platform to the next.

Python and R have lots of interesting capabilities which insurers can use to augment and improve their processes. For example, both contain powerful visualisation libraries. The results can be presented in an interactive web-based dashboard, where the results are initially presented at a high level, and the interactive features such as dropdown menus and selection buttons can be used to show more detailed views of the results.

Ethical Considerations

Ethical questions can be raised when using internal or external data, including personal data, but through addressing these questions, they can inform the types of data that can be used, and the applications of the data. Different types of data will obviously come with different considerations. The purpose of the data, the subsequent actions taken as a result of the analysis performed, and the disclosure to stakeholders are all part of the conversation around ethics in data science in Insurance – putting the raw data in context.

'A Guide for Ethical Data Science' was jointly published by the Institute and Faculty of Actuaries and the Royal Statistical Society Data Science Section in October 2019. This is aimed at addressing the ethical and professional challenges of working in a data science setting, focussing on ethical themes such as seeking to enhance the value of data science for society, avoiding harm, and maintaining professional competence.

How can Milliman help?

Milliman can help you to make better use of your internal data, including advice on:

- Best practice frameworks for data science processes
- Improving existing data processes and procedures and converting models to use more efficient tools
- Identifying suitable tools and techniques for particular circumstances
- Implementing Data Science solutions
- Understanding the implications of results
- Training in data science tools and techniques (including Python and R)

For further information, please contact your usual Milliman consultant or those below.

Milliman

CONTACT

DUBLIN

Donal McGinley
donal.mcginley@milliman.com

Bridget MacDonnell
bridget.macdonnell@milliman.com

Eamon Comerford
eamon.comerford@milliman.com

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com