Use of external data in insurance

Complementing traditional data sources with external data

Cormac Gleeson, BAFS
Eamon Comerford, FSAI, CERA

Milliman

Data Science techniques are able to extract value from increasingly diverse sources of data. Here we discuss how the use of external data can complement companies' wider Data Science initiatives, and look at some of the challenges of integrating this data into existing systems.

Traditionally, insurers have relied most heavily on the data they have collected themselves and industry-specific data to inform their business decisions and strategy. However, as techniques to analyse data from different sources become more sophisticated, insurers can better understand the relationship between internal data and data from external sources, and can use this to extract value.

Data Science is gaining momentum within the insurance industry and techniques such as predictive analytics, machine learning, data mining and artificial intelligence are being used more and more by companies to extract value and insights from data. One of the main strengths of these advanced techniques is their ability to process larger quantities of data and to better capture the links and correlations between variables. Another strength of these techniques is their ability to better handle unstructured data compared to more traditional models, such as General Linear Models for example. However, a recent Milliman Data Science Survey[1] found that there has been limited use of external datasets in companies' Data Science initiatives. Such incorporation of external datasets into their current framework would allow insurers to get a more holistic picture of their business, and to make better use of the capabilities of Data Science techniques.

Where to Look

There is an abundance of data sources that companies can look to utilise, such as:

- Industry pooled data
- Economic data
- Public statistical data (e.g. CSO/ONS data)
- Census data and other Government body data
- Intergovernmental body data e.g. IMF, World Bank
- Credit data
- Customer surveys
- External fund performance data

- Market sales trends
- Social media

The appropriateness of each source will vary depending on the purpose of the data. Many of the sources listed above such as public statistics and census data, are freely available and can be found via a simple online search. Many economic data sources are also free, such as the performance of certain indices and interest rates, while others may need to be requested or paid for on platforms such as Bloomberg. Some of this data, such as census data, may only be updated at certain regular intervals, while others may be updated daily.

Uses of External Data

When looking at policyholder behaviour, certain macro factors such as unemployment rates, average incomes and levels of education can be included through census data and linked via address. This could lead to new insights. For example, the data may show that policyholders from areas with differing average unemployment levels exhibit different persistency behaviour.

Figure 1: Census Data used to illustrate that areas with high unemployment have a higher propensity to lapse
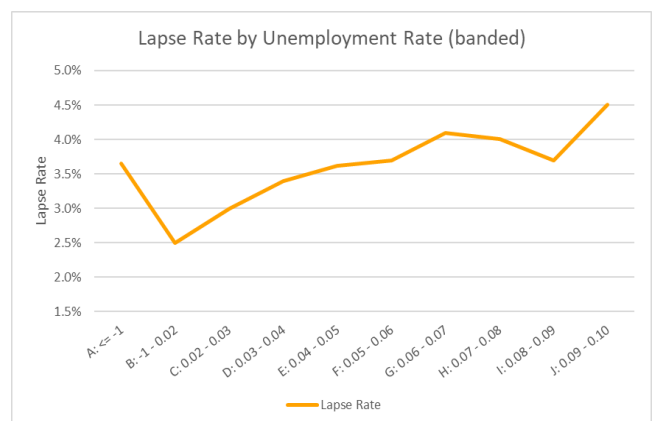
This type of census information can also prove useful when looking at inforce management and trying to better understand customers' use of policy options.

Combining standard economic data, such as Gross Domestic Product ("**GDP**"), stock market returns, interest rates and inflation with other, more specific internal sources could help better understand the underlying drivers on sales figures, market share, distribution methods, and customer engagement.

---

[1] This survey focussed on Life and Health insurers

Data Science techniques are capable of analysing the correlation and interaction between a large number of factors. This opens the door for a greater number of external data sources to be included in a company's data analysis. However, it is not limited to just being able to look at more and more data. There is considerable benefit in better understanding the interaction between internal data and larger macro factors. An example of this would be to look at the correlation between lapse rates or top-up rates on unit-linked policies with fund performance, and with the performance of appropriate indices. This can be used to analyse experience attributable to the performance of specific funds relative to the market itself, and not just analysing experience relative to the performance of the funds in isolation.

Companies will have to weigh up the costs and benefits of any Data Science initiative but using external data to complement internal data could yield significant benefits. For new or smaller companies who may not have large sets of data built up internally, using external data effectively could provide a practical, efficient method to augment their own data stores. For larger companies with existing databases, incorporating external data into analysis processes can provide an additional perspective and lead to a deeper understanding of policyholder behaviour.

Data Systems

There is an abundance of data available that could benefit insurance companies. With these data coming in varying formats, companies may need to adapt how they store and manage their own internal data in order to better integrate this. For example, a company might need to collect and store more granular address or post code information from policyholders in order to correctly map census characteristic data to them.

There is also the converse of this to consider – manipulating external data so that it is fully compatible with your own systems. This could range from more straightforward tasks such as changing date formats, to more involved processes like removing non-applicable data entries. Data "scrubbing" or data maintenance is not spoken about with the same veneration as predictive analytics or neural networks, but the results from advanced data analytics are only as reliable and useful as the quality of the data that underpins the processes. Thankfully, processes around data cleaning can largely be automated and streamlined once the initial framework has been developed.
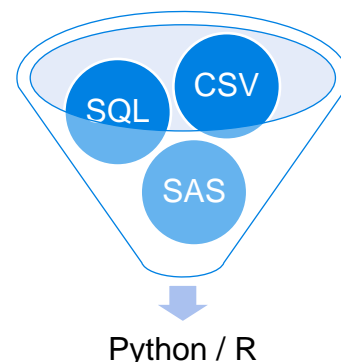
Data Analytics Platforms

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, both internal and external, 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 work with any database, and can store datasets efficiently using specific SQL servers.

However, other programming languages offer advantages over SQL for analysing and extracting value from the raw data. SAS is often used in the General Insurance industry, but free, open-source languages such as Python and R have become the mainstream languages for Data Science and therefore provide a wide choice of machine learning libraries. They can read data saved in different formats and from multiple sources, including SQL servers and SAS tables, and then create one cohesive dataset. This is vital for incorporating external data into existing company databases.

Figure 2: External data sources and formats can all be read in by Python/R

Python / R

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, particularly when external data is brought in alongside an internal process. 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.

Ethical Considerations

The use of personal data comes with certain ethical considerations. This is not limited to data collected by insurers themselves, but also to data from external sources, particularly as the availability and scope of such data

increases into areas such as social media data, personal finance and credit information, purchasing trends, etc.

Ethical questions can be raised when using external 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.

How can Milliman help?

Milliman can assist you with all aspects of your Big Data projects, and your Data Science needs in general, including advice on:

- Best practice frameworks for Data Science processes
- Collection and processing of data
- Identifying applications
- Identifying suitable tools and techniques for particular circumstances
- Implementing Data Science solutions
- Model development and validation
- Understanding the implications of results
- Constraints and practical challenges

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

Milliman

CONTACT

Eamon Comerford
eamon.comerford@milliman.com

Cormac Gleeson
cormac.gleeson@milliman.com