EBOOK

TWO SIGMA IQ

BEST STRATEGIES TO ENABLE DATA SCIENCE INVESTMENTS IN UNDERWRITING

INTRODUCTION

The explosion of data, recent advances in new technology, and increasingly competitive market conditions are driving interest by insurance companies in the application of data science. Insurance carriers understand the value of data and the insights it provides to underwriting. In addition to seeking better methods of harnessing their internal data, insurers are seeking unique new sources of third-party data to help gain a competitive advantage. The application of data science to the many parts of the insurance value chain has begun to prove worthwhile for some, but the full enterprise potential has yet to be realized as many projects are done in silos or as one-offs. Many data science projects are limited or fail because of issues with data quality and quantity, as well as challenges in defining the business goal.

Large amounts of data are needed to make the most of data science techniques. Large volumes of data and a wide variety of data sets help to produce more interesting insights and better results. Insurance carriers are interested in the potential insights gained from satellite imagery data, data from drones, telematics data, and other new types of available data sets. Such specific types of data can help underwriters better understand their customer base and the risks they are asked to assess. An increase in data volume leads to challenges in managing it all. Data is often trapped, becomes decentralized or is inaccessible. Historical data can lack consistency, be poorly defined and data quality often varies. Without an effective data management strategy, the success of machine learning and other data science techniques will be limited.

To make the most out of data science applications, insurers need to be thoughtful and develop a strategy for managing their data. Key elements of an effective strategy should include early alignment of data scientists and business user requirements; a commitment and plan to save all data; data provenance; centralized data access; and ensuring data quality. An effective data management strategy will lead to deeper and more accurate insights from machine learning and other applications and, in turn, better risk and overall business decision-making.

5 KEY ELEMENTS OF AN EFFECTIVE DATA MANAGEMENT STRATEGY

1 ENGAGE DATA SCIENTISTS IN THE PLANNING STAGE

Successfully applying data science techniques within an organization requires a solid strategy, as designing the most impactful and successful solutions requires much time and effort. By selecting and using the right sources, tools and methods, data scientists can enable underwriters to identify better risks in difficult pools and filter out the worst offenders in good classes of risk. Data scientists can also use machine learning techniques to automate and improve workflow challenges and improve operational efficiencies. But, to produce the best business results, a data scientist must thoroughly understand the problem he or she is trying to solve. The business goal should be clearly defined upfront, and discussed in detail by both the business executives and the data scientists.

For example, many insurance companies are becoming increasingly interested in using third-party data sources in the insurance underwriting process, but have been experiencing challenges in the application. Bringing new data sources into the underwriting workflow is more complicated than it seems and needs to be strategically planned. Most new external and even some internal data sets are rarely in useful formats. Allowing a data scientist to understand the intentional goal for the data prior to purchasing and ingesting particular third-party data sets will help to yield better results. A data scientist can ensure the best data is available in the appropriate format and advise as to the areas within the workflow where the insights gained can be applied for the best results. Applying their expertise later in the design or redesign process hinders the ability of the data scientist to make the most effective improvements and devise the best solution. Allowing business executives to collaborate with data scientists early in the planning stage of any data science-related solutions will produce the best results.

2 KEEP ALL DATA, NEVER DELETE DATA

Data science applications yield the best results from large amounts of data - the more data, the better. If an insurer wants to take advantage of machine learning applications, it is wise to save all of its data and never delete any of it. A data set that seems to have no use today may become useful later, especially since risk profiles and lines of business change over time. A particular risk attribute may seem unimportant today, but as risk landscapes shift and insurance coverages evolve, saved data sets may serve as great sources for new risk predictors and can be used to build new products and accompanying underwriting and pricing strategies. Saving all of the information involved in an underwriting decision can allow a carrier to understand what may have led to a claim. Having the ability to analyze the steps that were taken in underwriting specific risks can provide transparency as to any errors that may have been made in the process or help to better understand if any additional data and analysis could have been used in the underwriting process or risk analysis. These data sets will serve as the cornerstone for all future data science initiatives.

Unfortunately, under most current system design practices, data is not captured in the most usable format. Insurance companies have typically been focused on only capturing data for reporting purposes and not for modeling purposes. Typical policy administration systems don't capture records of change in enough granularity to support data science techniques. This is primarily due to the perceived lack of need, since qualified data scientists are scarce and the cost of storing data and converting it into usable formats has historically been high. However, advances in cloud technology and growing interest in data have changed both the demands and the cost structure.

Saving large quantities of data today is relatively inexpensive and simple due to the low cost and abundant and flexible capacity cloud storage solutions provide. The tools now available in the cloud allow insurers to scale up or down more cheaply and easily, and allow for economies of scale to work in an insurance carrier's favor. There is no need to host lots of infrastructure. With the growing interest in machine learning and artificial intelligence applications, the need for large volumes of data and the importance of creating a true system of record is more important than in the past. Having the ability to efficiently save historical internal and external third-party data will enable carriers to better apply data science applications to improve their underwriting process by allowing them to troubleshoot potential problems and errors and add additional insights and tools. Data is becoming an increasingly valuable asset and should be preserved.

3 ENSURE AND TRACK PROVENANCE OF DATA

In addition to having access to large volumes of data, insurers should track the provenance of their data. In other words, understanding and having a record of the where, what, and when of each data set or point is important when applying data science. Knowing the provenance of a data set helps to contextualize it against other data sets. It helps determine which data sets are most accurate and which might be the best fit for a particular solution. Seemingly equal or equivalent data sets, such as entity revenue, or roof condition or property condition on a specific insured can have different values. Data sets may come from different sources or come from the same source at different periods of time. Having a record of the data set's source, history and its transformations can help to contextualize which source of data is best or what the most logical approach of reconciling the data sets might be.

Understanding the provenance of data will allow an insurer to build a more confident version of a single source of truth. Knowing its source helps determine data validity and reliability - data from a trusted source is itself more trustworthy and reliable. It also allows for the identification of bias in data sets and internal data can be checked and validated against external data which will ultimately give an underwriter more confidence in the insights that result from the analysis. If data sources are known, a full audit trail will also be available for regulators and internal audit purposes. This audit trail will make it simple for carriers to identify errors as well as why and how decisions were made.

Creating a data catalog to store all of the sources of data is beneficial. Challenges do exist when recording data provenance and creating a data catalog. Data sets are unique, and some data sets are harder to track than others. Legacy core systems don't typically have data catalogs as adding tracking data to the already high volumes of data they contain can slow these systems down. In order to best capture data provenance, a modern cloud computing architecture with a data catalog is a must.

> "Investments in data collection and curation capabilities will be a key differentiating factor."
>
> ***Swiss Re Institute, No. 5/2020,*** Machine-intelligence in insurance: insights for end-to-end enterprise transformation

MAKE DATA EASILY ACCESSIBLE, ESPECIALLY FOR RESEARCH

Having large amounts of reliable and cataloged data is of greatest value when it is centralized and easily accessible. Data scientists, underwriters and other business users will best benefit from data and the insights it produces if it is accessible at the point of decision-making. The more easily an organization can access data, the better underwriters and data scientists can collaborate in designing products, go to market strategies, and risk management in a data driven manner. Having the ability to research and test available data is an important part of creating meaningful data science applications.

Easily available data in various formats, even overlapping data sets, are incredibly valuable for data science research. Raw data, processed data and open-source data files are useful formats. Data scientists find it useful to research and experiment with a variety of types of data sets. Having the ability to easily access a variety of data helps data scientists to identify patterns, perform and identify clustering, run models or algorithms and in general get a better understanding of the data and what solutions and insights might be drawn out of it.

Making data easily accessible to an experimental environment is also compelling. Having an underwriting system with a flexible testing environment that doesn't impact real data, but has easy access to historical data is incredibly valuable for testing and experimental purposes. For example, testing models, researching new signals in loss data, and experimenting with changes in rating and underwriting criteria provides for valuable experimental learning. Being able to access data to test proposed rating changes and determine their impact on a product and portfolio level allows an insurer to ultimately save time and increase certainty when bringing new products to market.

> Data powers all AI; the more data that inference engines can ingest, the faster and better they can learn, and the better their answers.
>
> ***Celent: Machine Learning in Insurance Fact from Fiction.***
> Donald Light

5 DETERMINE AND ENSURE DATA QUALITY

The majority of time spent on data modeling and data science is actually spent on data cleansing. As the old adage says, garbage in equals garbage out. Quality data science results depend on quality data, therefore all data sets need to be checked for accuracy, consistency, and gaps. The sooner data is prepped and cleansed the sooner it can be validated and used - and speed matters. Machine learning algorithms can be used to help identify quality issues and discrepancies between data set updates and overlapping third-party data sets can be used to validate each other. Tracking and ensuring data provenance and creating a single source of truth will also lead to better data veracity.

There are many reasons for data quality issues. Insurance companies have traditionally stored data in a variety of disparate systems, many of which are legacy systems. Having large amounts of information stuck in silos decreases quality. In addition curated data sets from third-parties can inadvertently contain discrepancies between data set updates, and vendors may switch formats and not inform the user. Some vendors use sub-vendors leaving room for more errors and/or gaps. Columns or rows could be lost, or data can even be incorrect. Insurers often have a lot of unstructured data (such as pdf files, images, website, email, IoT data etc.) that needs to be converted leaving more room for errors.

The more reliable a data set, the more actionable the insights that can be drawn. In the case of an insurer interested in applying data science to the underwriting process, more informed underwriting decisions can be made. Ensuring the highest of quality in your data sets will give you more accurate pricing, less premium leakage, better reserving and ultimately more efficient use of capital.

CONCLUSION

With the proliferation of data and the increased adoption of data science, managing data effectively is becoming more and more important. Having a successful data management strategy is critical for the successful application of advanced analytics such as machine learning. Aligning data scientists and business users in the planning stage and having large volumes of easily accessible, quality data that can be sourced and tracked are key elements of a data management strategy that best enables data science applications. An advanced integrated underwriting system built on a modern architectural platform can help both successfully manage data and apply the insights gained through the application of data science at the point of decision-making. Having the ability to capture an abundance of data and make it available through a modern data warehouse will provide insurers the ability to scenario-test, create new/better insurance products, and price risk more effectively.

HOW TWO SIGMA IQ CAN HELP

Having the right system/platform, one that can effectively access, utilize and apply a carrier's data is important in helping to implement a strong data management strategy. Many of today's core systems, however, are unable to effectively and efficiently access and utilize a carrier's data. According to Accenture, up to 75% of an insurer's data may be inaccessible to automated systems, a roadblock often referred to as "Dark Data". (Shining a Light on Dark Data, 2020)

Two Sigma IQ can help. In our early experience working with customers, we were met with the challenges of extracting data from existing legacy systems in order to be able to derive risk insights using data science techniques. This led us to develop our IQ Platform, an automated underwriting platform for the future. The IQ Platform integrates data science directly into the workflow in order to provide risk insights to the underwriter at the point of decision-making, all while running securely on our SOC 2 compliant cloud infrastructure. The results are improved accuracy, more consistency and greater efficiency within the entire underwriting process. The IQ Platform enhances organizational agility, accelerates time to market for new products, enables data-driven decision making for risk analysis, improves operational visibility and streamlines processes.

If you would like to learn more, please feel free to contact one of our client representatives.

John Paladino
john.paladino@twosigmaiq.com

TWO SIGMA IQ
www.twosigmaiq.com