Deep Science Publishing
https://doi.org/10.70593/978-93-49910-74-4_5

DeepScience
Open Access Books

Chapter 5: Data engineering pipelines in insurance analytics and actuarial modeling

5.1. Introduction to Data Engineering in Insurance

While actuaries apply mathematical and statistical approaches from risk theory to model and predict events in the insurance domain, data engineering is required to implement these models and predictions as pipelines that feed modern enterprise data strategies in insurance. In doing so, they produce services and products competent in supporting business and stakeholders' decisions on company, accounts, entities, exposure, financial forecasting, policy, provision, pricing, product, quota, reporting, sheets, investments, claims handling, strategy, and loss, among others. Data engineering in insurance, then, is responsible for the advancement and implementation of models aimed at risk control as services and products that build upon reputation. Models and predictions in insurance are normally conceived and created upon historical data, or time series; the task of operationalizing these services and products is devised, thereby, on the creation and maintenance of data pipelines and systems capable of collecting, storing, processing, and updating the volumes of historical data that is needed to ensure reliable model and prediction precision. This task ensures that data works reliably on the scales that are demanded by insurance products – responsible for transferring risk from policyholder to insurer – and processes – guaranteeing that lapse of time between payment and transfer of compensation for an insured event is, on average, the minimum possible, given financial and investment market conditions (Cernaianu & Corbos, 2019; Esposito et al., 2021; Henckaerts et al., 2022).

5.1.1. Understanding the Role of Data Engineering in Insurance

Insurance organizations have long relied on data repositories containing vast amounts of information collected over time to gain insights into complex decision-making problems. These data repositories are traditionally used in specialized departments for portfolio reporting and peer benchmarking. Recently, they have become the tool for controlling the underlying behavior and correcting any discrepancies. With an ever-evolving data environment and a regulatory need for controls, these data repositories now need to be the sole trust of record – organized, cleaned, and easily retrievable. This is where data engineering comes in.

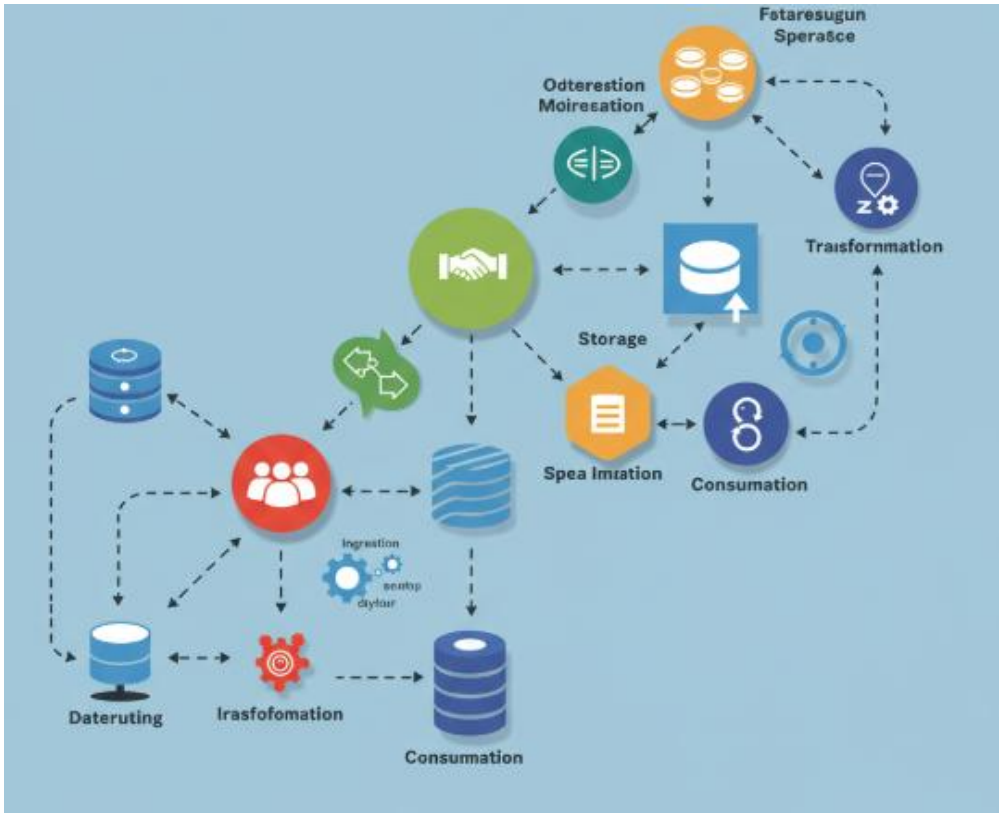

Fig 5 . 1 : Data Engineering Pipeline

Data engineering came about as an evolution of programming, data architecture, and data analysis best practices. With the growing demand for insights on business KPIs – be it for an insurance organization, bank, hedge fund, or tech giant, the value of a good data engineer has become paramount. Data engineers control the flow of data for ever-changing reporting and analytic needs. They build data pipelines which are the backbone of the automated creations of datasets for analysis each day, manage the servers on which the pipelines run, and are the custodians of master data management to ensure the trust

of record of original data used for reporting, analytics, and model development. In an age where an individual or organization can be categorized into numerous data points within financial services or the world over, analyst organizations whose core processes are based on the data engineering pipelines have the most powerful insights into decision-making problems (Verbelen et al., 2018; Wüthrich, 2020).

5.2. Overview of Insurance Analytics

Insurance Analytics (IA) is a discipline that deals with complex business problems that involve significant uncertainty and high associated costs. It uses mathematical methods, statistical techniques, and digital technologies, along with domain knowledge, to build predictive and prescriptive analytics to improve insurance products and decision-making in all business functions and at all levels of insurance. Predictive analytics has been well-established in the industry for two or three decades. This is considered "the first era" in IA. It has been combined with other related techniques, such as catastrophe modeling and financial modeling, to form a decision support system for the management of risk. There is a recent surge in the development and application of prescriptive analytics and what-if simulation techniques in IA. This is viewed as "the second era" in IA. In addition to the efforts to develop and apply descriptive and predictive analytics for data-driven explanations and risk forwarding, prescriptive analytics provides insurance leaders and other decision-makers with the tools to drive and change the future.

Many believe that IA is now at an inflection point because of its wider adoption and collaboration with technologies and core analytics in other domains. They see IA becoming a new unit of enterprise, one that connects data and information with all other enterprise processes. IA will allow broader, more consistent utilization of enterprise information and knowledge across all functions and levels of the organization. The formation of a new unit of security as another pole in enterprise technology suggests the importance of the insurance industry in the increasingly connected world. It suggests the potential of collaboration by insurance technology firms and software companies in R&D and the development of new cloud-enabled software tools and platforms to support analytical collaboration by actuaries and non-actuaries alike.

5.2.1. Key Trends in Insurance Analytics

At the core of insurance analytics lies the aspiration to optimize the outcomes of a portfolio of insurance policies by leveraging all available data. To balance risk mitigation with profitability, insurers aim to set adequate product premiums, discourage clients from filing unnecessary claims, indemnify customers reasonably, and, when needed, recover costs as swiftly as possible. It is therefore hardly a surprise that actuarial

modeling has originated at the forefront of insurance analytics. Indeed, modeling loss distributions has long been actuaries' primary task, generally using established statistical techniques from extreme-value or censored data theory. Once executed on an aggregate level via loss triangles and aggregated into pure premium curves, loss distribution modeling becomes pivotal for product pricing and provision. While this core tradition of insurance analytics shapes the field to this day, several structural trends are making the discipline as dynamic as any other analytics area. Many of these trends stem from rapidly advancing information technology, which generates milestones for the next generation of insurance products. In this tech-savvy data-centric narrative, recent take-up rates of fully mobile platforms, wearables, and autonomous devices further extend the digital interfaces through which insurance companies interact with, assess, and observe their clients' behaviors. Consequently, it has become progressively easier to design customized insurance products, to dynamically adjust product pricing to align with true risk levels or policyholder behavior and to provide timely assistance to the customer in case of a request for the payment of a claim. Moreover, the improved data feedback loops between insurers and insured have made it easier to influence desired client behavior, for example, through discounts on annual premiums conditional on effectively improving risk factors.

5.3. Actuarial Modeling Fundamentals

Actuarial models transform data into estimates to aid decision-making in insurance and reinsurance companies. Though estimates are often point estimates, they may take the form of distributions. A keen focus on data ensures that actuaries help decision-makers select from robust sets of alternatives based on developing data-driven insights. All models use the same basic building blocks: mechanisms that link inputs to outputs, inputs that feed into those mechanisms, and data for those inputs. Actuarial models may take many different forms, from multivariate regression equations that estimate points, expected values, or some other summary measure for particular classes of transactions, such as the expected loss from an auto accident, to models that derive input values from simulations. Modeling activities center on building relationships and approximations within and across the many basic building block components.

Though the components of models may differ across tasks and lines, key considerations are common. The concepts and mechanics underlying these components are well established, and the resulting models, whether they relate to pricing, reserving workers’ compensation claims, or any of many other areas, are similar. This set of considerations is important not just when building models, but also when using the output generated. The selection and quantification of model input variables, the identification of fiscal and

other variables that affect output, and the linking of those inputs to output through the model mechanism are at the heart of the estimates generated.

5.3.1. Key Components of Actuarial Modeling

To understand Data Engineering Pipelines and the domain context into which we situate them, we outline some basic components of Actuarial Modeling, covering the structure of a typical Actuarial Model and the interconnections between its components, as well as the actual business function of Risk Management which such models are intended to perform. Specific applications in Life and Non-Life Actuarial modeling disciplines utilize the overall structure in a more focused and tailored operation. It is generally done by performing inference about future uncertain events based on data from the historical past, or earlier experience. Actuarial Models essentially utilize hierarchies and other structured dependencies in Data behavior to perform such inferences and quantifications. Exposures to random risk, Critical Random Events, Random Event Models, Experience Datasets, and Inference Processes describe the anatomy of these Models. Specifically, risk quantification is directed towards forming an inference-conditioned best prediction, or Expectation, of certain specified Key Specific Model Outputs. These Key Outputs are specifically derived from Modeling Study Objective statements, which are a separate topic.

By its nature, Data Engineering Pipelines perform similar functions. The Data Engineering process is central to predictive modeling in statistics, econometrics, and Machine Learning, and thus provides the central idea around which our Modeling study Objectives are derived, as outlined in the current section. Warnings about data mining are well-known in statistical prediction, however in the context of Data pipelines, we offer certain additional observations and insights. We describe a Modular Information Theoretic Denoising Objective framework for these Pipelines and also review details of specific implementations thereof in the context of Actuarial and Risk Management.

5.4. Data Pipeline Architecture

Before data can be interpreted and used by business stakeholders and data scientists alike, meaning must be created out of low-level algorithms so that results of high-level business interest can be generated. This is the task of the data pipeline architecture, designed to take raw data in as input, follow a programmed sequence of data manipulations and transformation, and then produce useful data products such as visualizations, dashboards, or results of machine learning models that are leveraged for decision making. Insights derived from the analysis of data may go stale quickly. For instance, product recommendation engines must frequently update the affinity value to

avoid feedback loops that bias recommendations towards uninteresting products. Business actions like fraud detection based on behavioral analytics or compliance with anti-money laundering need to happen in real time, and decisions must be made on transactions and activities within tenths of a second and updated continuously if red-flagged. As such, the data pipeline architecture may be designed to operate in a batch processing or stream processing mode. In the batch processing mode, data is collected together and transformed 'in batches'. Often, this operates on low-level business summary data databases. By contrast, in stream processing mode, data products are created continuously. Often these products come off low-level use-case derived intermediary data products created via the batch processing mode. Before input data can be put through a data pipeline, this raw input data is usually ingested into core high-level business fact database tables. Core high-level fact tables hold business entities of interest at varying levels of aggregation. The intermediary data products play an important role as they are derived datasets used in the pipeline to produce other data products.

5.4.1. Batch Processing vs. Stream Processing

Many data systems operate in one of two common paradigms: batch and stream processing. The batch processing approach prepares the analytics on a large amount of data in one run over the entire dataset. On the contrary, stream processing makes computations continuously on data elements in real-time. Each approach has strengths and weaknesses and is more suitable for some applications than for others.

Batch processing is appealing for several reasons. For starters, batch processing's data load is already available in a serialized persistent format, generally on disk in some form of columnar file, optimized for scan/query performance, and compressed. Since the size of the dataset is very often in the terabyte range, usually a data lake is leveraged. The reason it can support such scale can be attributed to three design philosophies: it avoids the central bus bottleneck of having a single CPU, with only about a dozen compute cores, read and write to a small amount of memory, which in turn reads and writes to the one persistent storage device; it achieves economy of scale by using hundreds or thousands of processors to perform part of the work in parallel; it envisioned that parts of these huge datasets would be neglected, and it liberated the programmers from the burden of having to track the state of the program at each step through the dataset by treating the problem mathematically as one of application of closure to the problem space. In that way, by partitioning up the problem space into a very large number of units, bulk processing can hold its own against vectorizing the computations on much smaller and more manageable chunks of the data needed to be processed on the fly.

5.4.2. Data Ingestion Techniques

Data pipelines facilitate the extraction, transformation, and loading of data from one or multiple sources into a single target for additional processing and analysis intended for business decision-making. They encompass four functional components: data source; data ingestion tool; data transformation systems; and data storage, orchestration, or management systems. Presently, diverse cloud-based data storage infrastructures, data engineering platform solutions, and models are capable of managing any amount of data output from pipelines within taxation from a third-party cloud service provider. Cloud services pricing business models advocate for a pay-per-use approach, where the data is continuously stored on active storage types with increased costs, while most pipelines temporarily move cold data to lower-cost batch or object storage by pushing activities out of hours of operations.

Modern data pipelines accept data from a variety of data sources typically using batch, streaming, or micro-batching methods, and load them into data lakes or data warehouses, functioning in a hybrid business environment composed of on-premise and cloud-based transaction-process systems. Broadly, data ingestion tools use one or more source actions to ingest data, followed by any of multiple available data transformation technologies and functions, to load the data into a target database on a single functional step. The approach here to design data ingestion pipelines respects the different business use cases: some pipelines are no data transformation processes, loading data directly from one source into one target or even direct, transactional processing systems. For such cases, data transformation occurs in batch processes orchestrated to execute the planned business operation overnight off-hours.

5.4.3. Data Storage Solutions

A data pipeline needs a read-write storage, which takes in data products produced by ingestion and prepares data products requested by processing (and possibly also ingestion). Business, systems, and software storage requirements inform our storage choices. In particular, how will the storage be used: temporary staging storage (which is frequently refreshed) or permanent asset storage (which is less frequently accessed, modified, or refreshed)? What methods will be used for writing data (one shot, incremental update, periodic refresh from core tables)? What methods will be used for reading data (full table extract, incrementally processing changes, querying for snapshots, subset extracts)? What types of data (data schema and file size) will be stored? What frequency and volume of data will the storage be handling? Is access latency a concern? What business domain is being modeled (and what storage requirements do those domains have)?

At the storage system level, the assets used can be local vs. distributed systems (in terms of file and database system architecture), on-server vs. external systems (in terms of physical read-write requirements), and disk vs. tape or flash RAM (in terms of data access latency). There are various combinations of answers to the above questions that collectively enumerate our choices. Based on the resolutions chosen, we can select an underlying storage management system to fulfill our data product requirements. Use log-based systems for temporary staging storage, and SELECT, JSON, and other asset templates for permanent asset storage. Use cloud object storage for permanent historical storage, instance storage for staging storage in batch pipelines, and staging storage in pipeline steps with frequent but small amounts of data. Use SQL (for known query patterns locally) and other APIs (for arbitrary queries externally) with SQL storage engines as wrappers for external data files.

5.5. Data Quality and Governance

Importance of Data Quality

Data governance and data quality support organizational goals and enhance the character and presentation of data. Diligent data quality control helps to ensure that quality issues do not have a compounded effect as data moves through the pipeline and across the organization. Understanding the source of data is also critical to ensuring data provenance, especially in the insurance domain, which is heavily regulated and often data-poor due to privacy and confidentiality concerns. An in-depth understanding of data quality issues and geolocation of relevant datasets related to insurance issues, for example, can only be done when both content and dataviz characters are fully examined. This exposed belief supports traditional supervisory data governance by domain experts.

Some business problems can be overcome with data pipeline infrastructure to do efficient long-term exploratory data analysis and also high-speed, interim more focused data analytics. However, problems with data content must be addressed and resolved through more of a manual process until data are improved. Issues with automated DI and BI can greatly contribute to inefficient executive decision-making, which is a surefire way to harm the organizational bottom line.

The value of BI to the enterprise is that the value-add from high-quality and timely data via data pipelines is without question. By ingesting and processing better data more quickly, business intelligence systems are empowered to provide information for improved executive decision-making that promotes competitive advantage.

Data management in the insurance domain is heavily regulated, exposing the need for secure data and clean outside data partners. Data heuristics and data provenance involve making sound decisions about how data come into the system and built-in tracking of

how data move through the pipelines. Keeping a close watch on business, client, and regulatory data governance is essential to crisis management.

Fig 5 . 2 : Data Quality and Governance

5.5.1. Importance of Data Quality

Data availability is increasingly viewed as a barrier to innovation in the insurance industry, as it is considered simpler to invest in data generation rather than data preparation. Both the availability of third parties providing data and services and new technology for data collection allow for acceleration in this phase. If operationalized successfully, artificial intelligence can have profound effects in areas where the public has always considered insurance outdated, and in desperate need of reinvention. It has also been suggested that actuaries have a critical role to play in this transformation, as the first phase of innovation, product creation, is supported by the new technology for data collection. However, insurers need to keep in mind that the new data are often of low quality, which may result in effects opposite to innovation. Operations management researchers found that data quality is a key barrier to the performance of data analytics, both in general and specifically in companies from the services sector. It has been

pointed out that this is particularly the case for new data collections based on novel technologies.

Actuarial services are defined by their specific tasks, which incorporate core insurance activities such as product responsibility, risk matching, collateral and margin management, capital charge, and external comparability. These tasks and others require external comparisons of company results, often driven by data availability, but not testing the actual external assumptions that back the reported audited account. This is in clear contrast with the activities of IT units or consultants, which are defined by their administrative no-decision tasks, providing the infrastructure for any decision. The quality of the decisions made by the actuarial department regarding core insurance activities is based on the quality of the data that is considered, but the responsibility for the quality of the data is usually assigned to an IT department, which does not regard itself liable for the correctness of external data, or of active externalization tasks such as implementing new risk-based capital regimes, working with payer/provider negotiations as a consulting service for managed care organizations, or providing portfolio quantification for hedge funds. Working as consultants by core insurance companies, the company's actuaries are responsible for the actuarial work but do not carry out the data preparation and processing.

5.5.2. Data Governance Framework

We define a data governance framework for teams in insurance analytics and provide implementation guidelines. A data governance framework is an essential element of an enterprise data strategy, to enable effective discovery, understanding, consumption, and implementation of data by potential data consumers or implementers. Lack of it leads to wasted investments, failures, or friction from improper uses of data, or implications. Because insurance companies maintain such a vast array of actuarial data to comply with regulatory and management reporting requirements over long time horizons, insurance analytics team stakeholders often lack the knowledge or insight to make the best decisions when choosing data. The absence of reference information like business key definitions, scope limitations, data lineage, and information on subjects and sources of derived attributes in shared metadata repositories amplifies the possibility of such outcomes. An implementation of a data governance framework for insurance analytics fills these gaps, providing information to potential data consumers to make informed decisions.

A data governance framework increases the likelihood of project success, by establishing reference information for data and metrics commonly used in insurance analytics projects. Reference information captured by a data governance framework expands the capabilities of data catalogs and metadata repositories, which contain technical attributes

only. Analyses performed by analytical teams can modify business metadata of data sources, through approval workflows to achieve trust and responsibility for the quality of the data. Approved metadata entries are then available through automated data catalogs and repositories, integrating technical and business metadata. Approved business metadata lessens the dependency on data engineering teams to answer user requests for business context information, reducing responsiveness risks from an overburdened data engineering team. Contrast the ease with which data can be misused when no data governance framework is in place with the constraints on data consumption with a data governance framework.

5.6. ETL Processes in Insurance Analytics

Insurance companies are data-rich, but information-poor. Modern insurance firms can have terabytes or even petabytes of not only transactional data but also customer records and their interactions with a firm through various channels. Moreover, insurance companies may use other diverse unstructured or structured data sources. Data in insurance is wide and deep but is often fragmented across vertical silos, risk lines, and organizational units or divisions. It is often the case in some firms that actuaries need to start each project from scratch, identify, and query transactional databases and data marts, or even ask IT to extract the required datasets marshaled towards appropriate analytical models used in the projects. Oftentimes, the scale of the required extraction work is too large or the data may be not readily available to enable timely responses to requests or where, or even when the extractions are feasible, the clients may incur exorbitant costs.

With automatic data engineering pipelines in place, the repetitive extraction work is automated with one-click responses daily weekly, or month-end setups as required by the actuaries or data scientists. The ETL pipeline presents an appropriate automated solution to extracting data from multiple data pools, tables of interest, and dimensions of interest, and transforming the data for each of the projects, leading into datasets laden with the requisite variables for the analytical modeling task at hand. With the data pipeline in place, insurance analytics becomes a no-code or low-code development exercise, especially with GUI-enabled analytical modeling environments. Data pipelines are an integral enabler for such a data-driven information-fueled or algorithmic approach to creating competitive differentiation and sustained value in the business.

5.6.1. Extracting Data from Various Sources

Data warehouse and ETL tools collect data for analysis in one single place, cleaning it up so that it is properly formatted and filling in the data gaps from various source

systems. An ETL stack may consist of multiple components and a variety of sources from which data is to be ingested. Data extracted can come from structured relational databases, NoSQL document databases, messages from cross-platform connectivity tools, streaming events, batch files dropped into a file export folder, along with Web scraping, and any other system available. Data export commonly takes place in the form of database scripts generating CSV files, reports printed in PDF format, etc.

NoSQL support for document and graph databases is also necessary along with direct connectivity to event stream messaging systems. ETL tools that require significant setup are usually best when the source system does not change often over time and the data extraction is not mission-critical to the core system. If a company has a reporting system used for analysis, then it may also be beneficial to pull from there but make sure to discuss with your report administrator whether or not the reports will still be needed going forward – pulling from older reports that are being phased out may be counterproductive. Pulling from reports may help you prototype early reports quickly, but when time is sensitive it is usually better to work directly with the source system.

5.6.2. Transforming Data for Analysis

When the data has been extracted, the second major component of ETL, Data Transformation, comes next. Here, the various data mappings and transformations are performed to convert the data into the wanted shape. The operations are based on the underlying model structures of the different resident big data types. The transformation step may involve performing one or more of the following operations - deduplication, records selection, filtering out unwanted records, joining data from numerous resident logical tables, aggregating, validating with various constraints, re-shaping of columns, data type conversions and attribute mapping and other reshaping/schema change operations. Most Insurance datasets are either summarized to be used for analytics or are used at a detailed transactional level. Actuarial Modeling is mostly done using the summarized data set supported by the real observed transactions. Hence, the Data Transformation step creates models at the right summary level so that they can be easily joined with plan financials to explain the movements in reserves.

Considerable care needs to be taken at this step to ensure that the data models are correctly validated concerning the specific business area. If the transformations are not done correctly, it can lead to wrong and invalid results breaking the p-CM of Actuarial Theory and Business practice followed in the process of Making an Insurance. Another aspect of this step is the assumption of the employed approximations. For example, averaging, sampling, or other similar approximate techniques should be done with caution as they are not mathematically factored in the various process areas. These assumptions must be documented well in these models because of the frequent and usual

questioning of these assumptions by Field underwriters, Account Managers, Pricing Actuaries, Renewal Actuaries, and Reserving Actuaries during development, presentation, and file reviews.

5.6.3. Loading Data into Analytical Models

In this section, we concentrate on supervised predictive modeling tasks, which include the majority of use cases of commercial interest forecast in several stages of the data engineering pipeline. There are many unsupervised structure discovery tasks, typically clustering and association rules mining, which are also commercially important but involve use case matters in the insurance analytics sector. The fourth and final stage is the loading of the fully engineered data to be used in a supervised predictive modeling problem for performing the prediction and depending on the problem, the data requirements can be rather simple as with a general logistic regression model or far more complicated as with the generalized boosted regression models. Thus, what is defined as engineered data is all data that is used in the analysis of response and predictor variables and a potential list of the engineered data tables with whom we are quite familiar with a decade-plus of use including a clarification of the feature question for GBM. A massive data-loading process may take one or more numerous engineered data tables and load each to a related analytical modeling package. For machine status checks and any automated reporting to work, it is necessary to control for which engineering data tables have analysis responses and also to keep a careful tab on which engineered data and subsequently answer datasets are built toward working with the related modeling task feature.

5.7. Data Modeling Techniques

Data modeling is the process of taking the raw information that was extracted from the data source and converted into a usable format, after which the next step is to use it to do something interesting. In its various shapes and forms, data modeling is where the final purpose of data engineering is located. The process could be seen as the last segment of the data engineering pipeline, or a last-touche process carried over the outputs of the data engineering, obtaining data products. In general, two categories can be found in the data modeling phase: dimensional modeling and predictive modeling.

Dimensional modeling comprises the actions needed to turn the prepared data into data sources for queries and reports, as well as for dashboards of performance indicators. Business Intelligence solutions typically load their data from databases conforming to the dimensional format, and the reports querying that data present information selected for users and decision-makers. All the data in dimensional databases is structured,

coming from fact tables storing counts of events and measured quantities as numerical data types, and from dimension tables storing categorical data types, directly related to the data facts for those queries. A mandatory part of the dimensional model asks for the star schema format, presenting a fact table related to the rest of the database structure by a series of dimension tables. For performance purposes concerning high data volumes, some dimensional databases can also be designed in a snowflake schema, but that could imply an increase in complexity and lower manageability and usability. Other known data modeling types mandated by big data and non-relational databases offer greater flexibility concerning data formats and structures, allowing changes over time and data in natural language. Predictive modeling allows analysts to build robust probabilistic models that allow them to evaluate, forecast, and sometimes, control, relevant business outcomes. Classic regression modeling techniques can be applied, as also might be other techniques known as machine learning or data science techniques.

5.7.1. Dimensional Modeling

Dimensional modeling is an approach for designing data warehouses and data marts that facilitates online analytical processing. Unlike traditional databases that are optimized for online transaction processing, data marts serve to store analytical data, optimized for reporting queries, which are typically read-centric. Heightened interest in low-latency analytics has led to the explosive growth of data marts, as well as data warehouses of larger scope and complexity, in industry practice. Data marts can serve as independent data sources, or arrive at a derived structure from a larger data warehouse. The particular model selected, whether independent or derived from a larger multi-subject data warehouse, often reflects tradeoffs related to ETL effort, storage cost, data analytics scope, data quality, skewed access patterns, and use-case complexity. Hence, dimensional modeling is relevant throughout the data lifecycle, from ETL through storage structure to the data analyst's consumption of analytics.

Data marts with star schema logical organization for storing analytical data are popular among organizations requiring low-latency OLAP. Compared to a normalized OLTP schema that minimizes storage cost, predicting organized data marts optimizes query performance, ensuring that multiple dimensional paths define a rich analytic structure for users. Star schemas offer simplicity, as compared to blizzard schemas that support use cases requiring a super join to analyze a pair of business metrics against a combination of hierarchical business dimensions. Low-latency, filtration-centered analytics are defined with star schemas, typically supported by data loads with second currents on business dimensions. The intuitive appeal of star schemas translates into common heuristics, such as the answer to all but the simplest analytic queries appearing in the fact table.

5.7.2. Predictive Modeling

Predictive modeling and machine learning techniques have gained popularity in recent years for implementing predictive analytics pipelines, mainly due to the increased availability of data, enhanced data processing capabilities, and standardization of tools and libraries. Unlike conventional actuarial methods, these new-age techniques are often deployed as a black-box solution due to their inherent complexities. ML algorithms create smart models that can generalize well to make predictions using unseen data. Owing to their flexible nature and propensity for higher accuracy, ML pipelines have been extensively used in business areas for applications like customer retention, underwriting, premium pricing, fraud detection, claims classification, and reserving.

Predictive models can typically be divided into two categories: supervised and unsupervised. Supervised models are often used for inference and interpretability, and are typically trained on data where we know the label or the output. These models aim to correctly map features or predictors to an output. The output can be either categorical or continuous. Unsupervised models find latent relationships in the data based on densely populated clusters. Such informative clusters can have different forms including categorical clusters or continuous clusters. These models estimate an informative mapping that captures the densities and provides generalization on unseen data.

5.8. Machine Learning in Actuarial Science

Supervised learning deals with prediction tasks given an input variable X as well as an output variable y. Typical actuarial prediction tasks are the prediction of the claim amount given several explanatory variables, the prediction of the claim frequency or loss ratio on an insurance segment level given a set of explanatory variables, or the prediction of the probability of a given person dying at age x given several age x related individual characteristics. Apart from predictive applications, there are, however, also descriptive supervised learning applications. A typical example is the construction of individual health risk profiles or insurance fraud detection. Even ratio range transformation for predictive actuary tasks could be viewed as a supervised learning task.

There is much overlap between well-known actuarial models and supervised learning or more precisely statistical learning algorithms implementing these models. Here, traditional actuarial models denote regression and generalized linear models and more recent implementations include additively linked generalized multivariate models like generalized additive models and gradient boosting machines, and tree models like recursive partitioning or tree boosting. Also, classical and modern machine learning methods and actuary models share the same ultimate aim, which is a good prediction and especially a good judgment of the accuracy of the prediction. Apart from risk estimation,

risk preference, risk accumulation by a company, portfolio transfer with another company, latent risk profile separation, or risk and fraud detection could also be seen as actuary objectives.

5.8.1. Supervised Learning Techniques

Supervised learning approaches assume the presence of structured data with observations mapped to their target responses. In this section, we will review some supervised techniques in application to actuarial content over the years. The main categories of supervised learning techniques typically applied to the actuarial domain include regression techniques, tree-based methods, support vector methods, and neural networks. They have all undergone various levels of actuarial exploration, and unlike other, less representative actuarial content, it is not possible to address each exhaustive or quote works related to the modeling strategy of such techniques using most traditional actuarial-focused journals alone.

Linear regression has been applied to areas such as poisson modeling, prediction and reserving for bodily injury for a valuation of property damage, forecasting new automotive claims using linear regression, demand forecasting and loss reserving, frequency modeling of natural hazard insurance, and utilizing linear models in individual claim reserving. Generalized linear models have been applied by actuaries mainly for modeling and reserving property insurance. In a study, it has been examined the added value of using additional Fourier terms in refined linear modeling of modeling block heating load predictions. In the context of catastrophe events, a comparison of the prediction accuracy of various linear models in hurricane forecasting has been conducted.

5.8.2. Unsupervised Learning Applications

More than 80% of the actuarial value added in the context of big data comes from its analysis. But the innovation does not come only from predictive models; in fact, there is still scope for innovation with segmentation models, only this time these are no longer formatting of original data theory, but rather big data processing models. Consider for example the "flock" of text, insurance supporting documents, market complaints, and the "happy few" are the core claims of this corpus of insurance documents. The innovation here relies not on estimable coefficients but rather on semantic networks built using algorithms and models. Similarly, boosting topic modeling has created for years now marketing segmentation for clients who buy life and long-term care insurance. What other application it is possible to carry out with the wide variety of unsupervised machine learning modules?

In insurance analytics pre-processing, dimensionality reduction, and topic detection/modeling are the three stages of unsupervised learning. Topic detection/modeling used alone is not the only task achievable; other tasks can be achieved such as keystroke modeling through fighting against the first order hypothesis; generative collaborative filtering; sound and image modeling – the first order type in a horocycle, but the majority of applications remain topic modeling on related sets and heavy-tailed dataset. You cannot only produce modeling algorithms; you can utilize also them for insurance document verification processes for news economy articles and via explicit word-by-word functions. Of course, these different vicious cycles of application with task-specialized algorithms (and usually task-unspecialized datasets) have multiple consequences especially for banking or insurance marketing segmentation because they produce prosperous or depressed economic segments.

5.9. Conclusion

Our book discusses the need for data engineering solutions in property and casualty insurance analytics. We illustrate their usage through modern high-volume insurance data streams, such as connected devices or IoT. The book aims to discuss the methods and techniques required to design, develop, and maintain data pipelines from data sources to analytic teams and tools. Insurance-specific pipeline needs are described with a focus on the use of pipelines for both predictive analytics and operational machine learning. The aim is to allow data engineers, data scientists, and actuaries to collaborate better for the ultimate goal of the insurance business, to estimate risks, and to satisfy policyholders. In Section 2 of the book, we focus on the use of pipe pipelines both within insurance and in neighboring fields, such as actuarial science. In Section 3 we provide a few equations describing the nature of problems using insurance data and the possible machine learning solutions, the empirical risk minimization on which insurance modeling has been developed. Introducing a special case, the generalized linear model allows us to use very few assumptions with either linear or generalized regression pipelines supported by available software, making insurance modeling easily accessible. Section 4 describes the input side of data pipelines, and how data flows from source to storage supporting further analytics. The use of cloud services has revolutionized how data can be procured in insurance pipelines and allows us to build cheaply scaled hubs. Expectations in data engineering around data hub usage, such as transparency and reliability, current tools and technologies, as well as decision support system designs are also described to provide a stereo view of data hub pipelines.

Fig 5 . 3 : Insurance Claims Modeling

5.9.1. Summary and Future Directions in Insurance Data Engineering

Data engineering is the set of activities associated with the ingestion of raw data into a data ecosystem, including data storage, structuring, and accessibility for data analytics. In this monograph dedicated to data engineering in insurance, we summarized extensive discussions we had in the past 10 years, to answer questions like "What data should insurance companies make available for analytic purposes, from which sources, and how", "What analytic tools should be made available for the inspection and exploration of insurance data" and "how to set a data architecture, in terms of storage, design, accessibility, and integration, to facilitate agile interaction of business users with data". It includes a brief overview of insurance analytics, its internal preparation functions, and associated tooling, with special emphasis on actuarial modeling activities, which have long led discussions inside the actuarial profession about the role of data in actuarial work. With that insurance analytics overview, we explore and attempt answers to the questions presented previously, organizing the discussions and results in four main areas,

that correspond to critical success factors for the use of data in the insurance business: Sources, Architecture, Tools, and Workflow.

While data analytics is in itself a discipline with its ethical pitfalls, exposing the dangers of utilizing misinterpreted results, totally disregarding the limitations of the data on which conclusions are drawn, and the inherent biases introduced when data are collected and used to formulate and associate correlations with insurance damages, which are the basis for model creation and prediction of future event probabilities and severity, this monograph aims to present a few thoughts and directions on how to create and curate the data insurance analytics operationalizes. While it is in itself just a subset of a larger picture, a data quality focus, it is a differentiating factor for an insurance company towards effective differentiation strategies, core to the creation of competitive advantage using data.

References

Henckaerts, R. J., Verbelen, R., Antonio, K., & Claeskens, G. (2022). Machine Learning for Mortality Modeling. Insurance: Mathematics and Economics, 106, 154–170. https://doi.org/10.1016/j.insmatheco.2021.10.005

Wüthrich, M. V. (2020). Machine Learning in Individual Claims Reserving. Scandinavian Actuarial Journal, 2020(1), 1–23. https://doi.org/10.1080/03461238.2019.1681390

Esposito, C., Castiglione, A., & Aloisio, G. (2021). A Data Engineering Approach for Big Data Analytics in Insurance. Future Generation Computer Systems, 114, 387–397. https://doi.org/10.1016/j.future.2020.08.024

Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unravelling the Predictive Power of Claims History Using a Copula-Based Approach. Scandinavian Actuarial Journal, 2018(4), 319–337. https://doi.org/10.1080/03461238.2017.1332990

Cernaianu, S., & Corbos, R. A. (2019). Optimizing Insurance Data Pipelines with Cloud-Based Architectures. Journal of Business Research, 98, 223–233. https://doi.org/10.1016/j.jbusres.2019.01.058