

# What should an actuary know about Artificial Intelligence?

---

## 1. Preface

Actuaries are living through the birth of large-scale adoption of Large Language Models (LLMs) like ChatGPT. Next to enormous promises, this technology carries risks and threats. Actuaries working in business, pensions, social security and other areas need to familiarize themselves with novel tools to understand how they can revolutionize practices also in the actuarial domain. As well actuaries need to understand how they can, based on their long experience with complex models, bring their expertise to this novel area, and also warn of the possible dangers with models that might be used when assumptions they are based on are not valid or when the tools are used in areas where their validity has not been tested.

In addition to LLMs there are additional more general phenomena in the areas of Artificial Intelligence, and still more generally, in Data Science:

- Data availability
  - Data explosion or Big Data: Unprecedented amount of data in different areas in a variety of forms together with the data processing capacity needed to process this data.
  - Pervasive digitalization through continuously connected sensors everywhere: Real-time information on what is happening around (visibility on customers, business activities and market trends).
  - Growing ecosystems across industries will lead to an increase in open-source data and data exchange, connecting e.g., insurers with banks, big techs, or other consumer brands
- Technology boost
  - Enabling technologies with exponentially increasing efficiency: Processing power, storage, robotics ready for AI and automation, open-source AI software.
- Market and consumer changes
  - Changing consumer behaviour: Passion to engage with brands/companies anytime, anywhere, in need of immediate service 24/7.
  - New market forces in a 'winner takes all' economy: Tech giants taking the markets. These AI-oriented players have access to top-notch talent, alternative data pools, immense funding from investors, and the ability to disrupt certain parts of the value chain
  - InsurTechs challenging incumbents.
  - Sharing economy revolutionizing the need to own.
  - Regulatory harmonization trying to keep up-to-speed with different developments.

To make these concepts more concrete, let us fast-forward to the year 2030 and let us look at how journeys of different stakeholders could look like:

- **Isabel the actuary:** She works in product development and is designing a new pay-as-you-drive product for self-driving cars. Having access to vehicle, driving, and claims data, she analyses the impact on claims depending on the chosen route, on choosing autopilot, and on weather conditions. Isabel designed this product together with the car manufacturer such that optimal routes are chosen automatically through the navigation system – associated premium savings are directly shown on the driver’s screen. She embedded feedback loops to life and health products as well, such that clients with multiple covers can directly benefit from reduced claims expectations through discounts on their e.g., accident cover. Additionally, Isabel can test the impact of this new product on forecasted internal and external KPIs in real-time.
- **Nolan the sales agent:** Nolan gets a trigger that one of his customers, Livia, got a new self-driving car. Hence, Nolan gives her a call and talks her through potential insurance solutions for her car to figure out what is best suited for her needs. Based on the ongoing call, the sales LLM provides real-time suggestions and quotes to Nolan to provide him best-advice options. After learning, that Livia moved to a new apartment a few weeks ago, the sales LLM automatically updates Livia’s policyholder profile with the latest data, only requiring Nolan to check the update and press ok.
- **Livia the customer:** After owning the car for few months, Livia drives to her parents’ place, not using autopilot. Unfortunately, her windshield gets hit by a small rock. Her insurance app guides her to take a photos. The car’s diagnostics show that she can continue to drive but needs to go to a repair shop straight away – the nearest one is shown on the navigation system and only five minutes away. Once she arrives there, a replacement car is already available such that she is able to arrive at her parent’s place just in time.
- **Peter from claims management:** The company’s GenAI assistant creates a full report within a few minutes including damage assessment, fraud potential, plan coverage, estimated payment, and the repair shop where Livia is taking her car. Peter reviews the report, conducts minor adjustments, and settles the claim shortly thereafter.

While this is a made-up example, many of these technologies already exist today and will become reality in the near future. Hence, rather than a burden, this is a unique opportunity for actuaries to strengthen their role, broaden their areas of influence, and safeguard responsible use of AI in critical areas. AI and adjacent trends will fundamentally transform how companies operate and will allow them to optimize sales, distribution, pricing, claims management, and many more. As many of these applications require actuarial expertise paired with data, business and communication skills, actuaries are optimally positioned already to claim central roles and shape this change. Taking a stronger stance, you could even argue that actuaries have the duty to step up in AI and protect what matters most – safeguarding welfare and protection of customers and societies.

Actuaries are traditionally well-trained to answer descriptive and predictive questions:

- WHAT are renewal rates?
- WHO are riskier clients?

A core issue with AI is that predictions become cheaper (cheaper mushrooming data, cheap storage, cheap computing power) which certainly helps actuaries in all of their work. According to economic theory, when the price of something drops the demand for it will increase – and thus better and cheaper predictions are used more and also in novel areas.

While AI can improve answers to these actuarial core questions, its true value-add lies within prescriptive applications. Answering such “HOW-questions” can create huge benefits for customers and societies:

- How can we target the right customers at the right time?
- How can we prevent lapse?
- How should our pensions and social security systems better react to changes in the societies and to changing demand?
- How can insurers and societies build a more inclusive environment without risk differentiation leading to direct or indirect discrimination, especially with intersectionally vulnerable minorities?

This paper of the Actuarial Association of Europe attempts to give a condensed overview of the most important concepts connected to the area of Artificial Intelligence. We will then build on these concepts to highlight our ideas on how actuaries could make best use of AI.

This fairly short paper does not try to tell everything that is crucial. We hope it will anyway be a good starter and pave the way for actuaries to learn more of the subject through references to recommended further reading.

## 2. Artificial Intelligence and Data Science in a nutshell – main concepts clarified

### 2.1. Data Science (DS)

*According to Wikipedia, ‘Data science is an interdisciplinary field focused on extracting knowledge from typically large data sets and applying the knowledge and insights from that data to solve problems in a wide range of application domains. The field encompasses preparing data for analysis, formulating data science problems, analyzing data, developing data-driven solutions, and presenting findings to inform high-level decisions in a broad range of application domains. As such, it incorporates skills from computer science, statistics, information science, mathematics, data visualization, information visualization, data sonification, data integration, graphic design, complex systems, communication and business.’*

Data Science (DS) can be considered as the operational tool to solve business problems through machine learning or other data-driven models, extracting information and knowledge from structured or unstructured data. Seen from a business perspective, DS enables the translation of a business problem into a research and analysis project and then transform it, again with the help of data into a practical solution. A DS project can be represented through a process of data analysis and interpretation that must be seen as iterative rather than linear, subject to continuous verification.

Data Science is useful in answering five types of fundamental questions:

- How much or how many? (regression)
- Which category? (classification)
- Which group? (grouping)
- Is it strange? (anomaly detection)
- Which option should be taken? (recommendation/prescription)

From a business perspective, these questions become, for example:

- Who are the best customers?
- Why are they buying 'that' product?
- How to predict whether a customer will buy another type of product?
- Why have those customers not been buying for a long time?

Data science does not necessarily require sophisticated algorithms and multi-core cloud computing but (depending on the problem) solid understanding of the business problem & data, good data handling skills and ability to bring it into action in the organisation.

## 2.2. Artificial Intelligence, Machine Learning, Deep learning, and Generative AI

The terms Artificial Intelligence, Machine Learning and Deep learning are terms that are often used synonymously. The simplest and most effective way to explain the difference is to refer to the Chinese box system where one domain is a component of the one before it. Machine Learning (ML) is a domain of Artificial Intelligence (AI) and Deep Learning (DL) in turn is a domain of Machine Learning. ML is simply a way of achieving AI and DL is one of the many approaches related to ML. In other words, one can consider AI as the basic discipline and ML and DL the techniques, or rather, the models that enable its application. More formally, we can have the following definitions:

- AI involves all those operations that are characteristic of the human intellect and performed by computers. These include planning, language understanding, object and sound recognition, learning and problem solving.
- ML is an area of AI that focuses on the ability of machines to receive a set of data and learn on their own, modifying algorithms as they receive more information about what they are processing. Machine learning is thus a way of 'educating' an algorithm so that it can learn from various situations. Education, or even better training, involves the use of huge amounts of data and an efficient algorithm to adapt (and improve) according to the situations that occur.
- DL is one of the approaches to machine learning that originates from the brain morphology and functioning, i.e. the interconnection of the various neurons. Deep learning uses huge models of neural networks with various processing units; it exploits computational advances and training techniques to learn complex patterns through huge amounts of data. Common applications include image and speech recognition.

An additional issue to mention is that generative AI utilizes deep learning algorithms to create new data, such as realistic images and coherent texts, for different applications. Combined models like chatGPT enable more natural interactions. Language models like LLMs, trained on vast text data, generate meaningful content by predicting the next word based on context, playing a vital role in Generative AI's ability to produce coherent and relevant text.

## 2.3. Machine Learning (ML) deep dive – Supervised, unsupervised and reinforced learning

ML has at its base a series of different algorithms that, starting from primitive notions, will know how to make a specific decision rather than another, or perform learned actions over time. Depending on the type of algorithm used to enable machine learning, i.e. on how the machine learns and accumulates data and information, one can subdivide ML into three different learning systems: supervised, unsupervised and reinforcement learning. See (Wüthrich & Merz, 2022) for an extensive overview of modern machine learning methods.

### 2.3.1. Supervised Learning

Supervised learning consists of providing the machine's computer system with a series of specific and codified notions, i.e. models and examples that allow it to build a real database of information and experiences. In this way, when the machine is faced with a problem, all it has to do is to draw on the experiences stored in its system, analyse them, and decide what answer to give on the basis of already codified experiences. This type of learning is, in a way, provided already packaged and the machine only has to be able to choose which is the best response to the stimulus given to it. In short,

the operation of this algorithm consists of a goal/outcome variable (or dependent variable) that must be predicted by a given set of predictors (independent variables). Using this set of variables, we generate a function that maps the inputs to the desired outputs. The training process continues until the model reaches the desired level of accuracy on the training data. Examples of supervised learning: regression, decision tree, random forest, KNN (The k-nearest neighbours algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point), logistic regression, etc. Algorithms that make use of supervised learning are used in many fields, from medicine to speech identification: they have the ability to make inductive hypotheses, i.e. hypotheses that can be obtained by scanning a series of specific problems to obtain a suitable solution to a general problem.

### 2.3.2. Unsupervised Learning

In unsupervised learning the information entered into the machine is not encoded. The machine is able to draw on certain information without having any examples of its use and, therefore, without having knowledge of the expected results depending on the choice made. The machine itself must catalogue all the information in its possession, organise it and learn its meaning, its use and, above all, the result to which it leads. Unsupervised learning offers greater freedom of choice to the machine, which will have to organise the information intelligently and learn which results are best for the different situations that arise. The operation of this algorithm is characterised by the fact that we have no target or outcome variable to predict/estimate. Examples of unsupervised learning: Apriori algorithm (an algorithm for frequent item set mining and association rule learning over relational databases), K-means (k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition  $n$  observations into  $k$  clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster), etc.

### 2.3.3. Reinforcement Learning

Reinforcement learning is the most complex learning system, which requires the machine to be equipped with systems and tools to improve its learning and, above all, to understand the characteristics of its environment. In this case, therefore, the machine is provided with a series of support elements, such as sensors, cameras, GPS, etc., which enable it to detect what is happening in its surroundings and make choices to better adapt to its environment. The machine is trained to make specific decisions, exposing it to an environment in which it continuously trains itself through trial and error. It learns from past experience and tries to acquire the best possible knowledge to make accurate business decisions. An example of reinforcement learning is a Markov decision-making process.

## 2.4. Deep learning deep dive

Those wishing more information about deep learning are advised to check, e.g., <https://www.linkedin.com/pulse/what-generative-ai-llm-luis-escalante>. Some concepts there are:

- Convolutional Neural Networks (CNNs): often used for image recognition and classification tasks by applying filters to extract features that will be passed through a series of layers to produce a classification output.
- Recurrent Neural Networks (RNNs): these are commonly used for sequential data and work by processing one element of the input sequence at a time and using the previous state of the network to inform the processing of the current element. RNNs are useful for language modeling, speech recognition, and sentiment analysis tasks.

- Generative Adversarial Networks (GANs): used for generative tasks, such as image synthesis and text generation, by training two neural networks, one to generate fake data (Generator) and the other to discriminate between real and fake data (Discriminator). The two networks are trained in opposition to each other with the goal of improving the generator's ability to generate realistic data.
- Transformers: designed specifically for natural language processing tasks by using self-attention to selectively focus on different parts of the input sequence, allowing them to process long sequences of text efficiently so it can be used for language modeling, machine translation, and question answering.

## 2.5. Generative AI and Large Language Models (LLMs)

Generative AI, with its ability to create original text, images, audio, and more in seconds, is poised to transform various industries. Its power surpasses existing enterprise technologies, and its capabilities have rapidly advanced in the past few years. Already, generative AI can engage in human-like conversations, generate diverse styles of images from text, and even pass exams. It can create high-quality videos and compose music across instruments and genres.

As there are risks such as intellectual property infringement and biased outputs, responsible AI approaches need to be taken to mitigate them. It needs to be remembered that an LLM does not 'understand' the problems it tackles – instead it basically does a statistical analysis and predicts based on the training material (Internet content) the probable answer to the question posed to it. Organizations need to define rules, integrate generative AI into workflows, and train teams to use these tools responsibly. Early experimentation and ongoing learning will be crucial for long-term success with generative AI.

In the insurance industry, generative AI can revolutionize the full value chain, e.g., claims management, underwriting, and customer service. By leveraging generative AI, insurance companies have the potential to reduce time, save costs, and improve accuracy in various processes. However, to realize the full value of generative AI, organizations need to address roadblocks such as limited understanding of value, data governance challenges, digital platform limitations, skill fragmentation, and change management.

As generative AI becomes more prevalent, it will lead to the emergence of new roles, changes in operating models, increased productivity, and revamped talent acquisition and performance management practices. Personalized training will be essential to equip employees with the skills to effectively use generative AI tools.

## 2.6. Data Science in practice

We can identify seven key steps for the practical application of DS:

- ask a precise question
  - answer needs to be measurable
  - start small with a simple question
  - have a hypothesis in mind
- gather and assemble data
  - generate a single point of truth
  - take GDPR into account
  - normalise
  - make sure you have the resources
- clean and enhance data
  - check data quality (missing values, sums are correct, outliers, duplicates, blanks, errors...)

- harmonise scale if necessary
- perform feature selection
  - remove redundant or duplicate features
  - remove noisy features (e.g., with few data points)
  - keep your hypothesis in mind
- prepare infrastructure and tools
  - choose appropriate infrastructure and tools for the problem ahead
  - try to keep things as simple as possible
  - make sure that every step will be repeatable
- train, test and assess model
  - no hard coding
  - comment your code
  - understand your model
  - if possible, provide ranges
  - always test you trained model
  - no in-sample testing
- infer business actions
  - check that your result gives an answer to your question
  - iterate the process and refine/extend your results
  - make your message clear, simple and precise

2.6.1. Data Selection and Data Collection: retrieve the 'raw data' needed for the problem identified.

This stage of the process requires some attention because it involves both thinking a priori about what data will be needed, and the actual 'retrieval' of data from a plurality of sources (both internal and plurality of sources (both internal but also external datasets). Many elements herein resemble to anything actuaries traditionally do with data analysis. However, while the danger of duplication is recognised, it seems reasonable to restate here also the obvious parts.

The data may be structured data (e.g., from databases and internal applications the company, such as a CRM or an industrial application, e.g., for managing production or warehouse management) or unstructured data (text, images, videos from e-mails, documents, collaboration platforms, but also from external sources such as social networks, open document repositories, web pages, etc.).

2.6.2. Data Cleaning and Data Preparation: processing data for analysis.

The Data Cleaning (or Data Preparation) phase consists of the act of manipulating, pre-process, raw data from a variety of sources and in different formats, to clean them, harmonise them and transform them into data that can be used by analysis tools. A crucial element of this phase is the verification of regulatory compliance.

2.6.3. Data Exploration and Data Transformation

Data Exploration phase means an initial 'exploratory analysis'; in essence, statistical tests are carried out, initial analyses are made and the first Data Visualisation techniques are tested. From here we begin to see the concept of an iterative and non-linear process. In the Data Exploration phase data errors may emerge in the data or an intervention can be needed that 'leads back' to the previous phase of previous phase of data cleaning, preparation, enrichment. Closely related parts of the Data Exploration phase are experimentation and modelling, i.e. the process of identifying and building the analysis model for solving of the specific problem identified in the very first phase of the entire Data Science. These phases involve the 'fine-tuning' and validation (including the choice of algorithms and their possible 'tuning') of the analytical model. The latter is then tested by exploiting the transformed data and based on the generated output (i.e. the insights obtained), its performance and effectiveness in terms of accuracy of information and actual value with respect to the decision-making process.

2.6.4. Data Mining: Advanced Analytics and Machine Learning.

At this point in the process algorithms are used to analyse data, discover hidden patterns or extract interesting knowledge from this data. 'Typical' operations of this phase are parameter identification, processing, modelling, pattern evaluation. One defines here how to extract actual value from large volumes of information, choosing algorithms and 'training' methods to search for patterns in the data (e.g., with machine learning), as well as the form of representation or the set of representations with which the information is to be extracted (classification rules, decision trees, regression classification, decision trees, regression, clustering, etc.). An important part of this phase of the DS process is also to provide business people with all the necessary elements (both quantitative and qualitative) to be able to access information and knowledge that is truly relevant to the problem identified, the possible applicable solution and, therefore, effectiveness with respect to the business decision.

#### 2.6.5. Data Visualisation: communicating and visualising results.

Data visualisation actually comes into play several times during various phases of a typical Data Science process. The 'final' phase of the process concerns the communication of the results deriving from the analyses, understood as the visualisation of such results through analysis systems that must be made available and usable easily by business users. Here Data Visualisation and Data Storytelling, i.e. advanced data analysis systems make it possible to 'read', among hundreds and thousands of pieces of data (of different formats and structures) information, correlations and patterns. The aim is to unearth 'a story' hidden in this data that can only come 'to the surface' through advanced analysis and can become usable for business people, without specific technical skills, precisely thanks to Data Storytelling and information visualisation.

In addition to the stages of the data science process just described, it is essential to avoid the mistake of thinking that once the analysis system has been put into production the process is concluded. Models and algorithms do not perform indefinitely. It is therefore necessary to continue to monitor the performance of the models with respect to the business objectives.

The above steps are perfectly consistent with what is proposed by the methodological framework CRIP-DM ([Cross Industry Standard Process for Data Mining](#)) which represents a industry standard, which has become popular for its flexibility and possibility of customisation.

#### 2.7. AI-DS in business use in insurance – examples

In the following, we provide an incomplete list of applications of AI along the insurance value chain to show the breadth of its applicability already today. See, for example, (Seehafer, et al., 2021) for an overview on selected actuarial case studies. A lot of these examples require actuarial input or interpretation:

- Product development and pricing
  - Non-linear pricing and other data-driven pricing approaches, with potential benefits for insurability
  - Personalized product development based on other than classical sources of data, better reflecting individual needs for the benefit of the customers
  - Coverage of new risks and leveraging new data sources (e.g., cyber, robotics), in particular to provide solutions for new consumer trends
  - AI-supported product design leveraging LLMs based on market and consumer data, reflecting strategic appetite and market conditions.
  - Dynamic quote modification in real-time
  - Wording and price and/or renewal fine-tuning based on client feedback
- Marketing, distribution, and CRM:
  - AI-based campaign design and management, e.g., optimising which customers to best approach with which products, via which channels, with which messages



- Improved client experience through better-informed sales staff and best offers, enabled through LLMs trained on all previous client interactions
- Improved lead generation for sales staff, partners, and direct channels, leveraging multiple data sources, e.g., daily smart lead lists for agents or real-time leads (next best offer) during calls via LLMs
- Renewal pricing optimization considering customer profile, price elasticity, past interactions, and other factors, see (EIOPA, 2023)
- Real-time and fully integrated quoting for brokers and agents
- Churn detection (along different types of churn) and suggesting prevention measures based on previous experience, sentiment analysis, and expert experience
- Improving customer/agent/broker segmentation through multiple data sources to better steer business decisions
- AI-based derivation of customer lifetime value to define room for CRM measures such as discounts
- Underwriting
  - Leveraging of external (big) data sources to improve underwriting decisions
  - Pattern recognition among claims, generating underwriting insights
  - AI-based risk assessment, e.g., pre-damage check in Kasko using image recognition, documents analysis in risk assessment
  - AI-based underwriting fraud detection
  - Digital brokerage technology such as “UnderwriteGPT”
- Claims and benefits management
  - Automated claim assessment, determining degree of damage and instant generating of appraisal forms
  - Automated claims categorization, e.g., determining automated vs. manual handling, granularity of information needed from customer
  - Self-learning fraud model leveraging AI and analysis of structured and unstructured data, as well as prevention
  - Automated claims management and prevention using Internet of Things (IoT) and LLMs to extract relevant data from e.g., images and emails
  - Claims management through AI, considering customer experience, adjacent extra costs (legal), and potential renewal options
  - AI-based single loss reserving considering various data sources, e.g., claims documents, call protocols, submitted material, pictures
  - AI-driven expert network management and control leveraging also integrated scorings
  - Direct re-imburement powered by AI
  - Improving customer satisfaction through simplifying complex claims processes, e.g., assisted by virtual/augmented reality tools, 3D modelling, AI to scan documents and retrieve relevant info to register claims in the systems
- Operations
  - Automation of administration and services, e.g., through LLM-powered chat-bots and apps

- Acceleration of day-to-day tasks through speech-to-text, machine translation, summarizing of text, LLMs for quick text generation, personalization of emails, classification of text and documents etc.
- Conversational interface, allowing customers to review policy, submit claims, and track status
- Finance and actuarial
  - Data-driven risk analyses to improve the use of reinsurance
  - Legacy product simplification within migrations
  - Automated reports and document generation
  - Real-time financial forecasting
  - Automated invoice processing
  - Automated tailoring of investor reports and communication
  - Automated coding of software
  - Automated regulatory changes monitoring and alerting
  - Automated model documentation, validation, development (coding)
  - Real-time monitoring of KPIs and KRIs, and scenario testing
- ALM, and Investment
  - Research engine based on unstructured data
  - Portfolio adjustments/hedging based on news surveillance
  - Automated monitoring of market trends
  - Summarization of market reports
  - Optimization of SAA, TAA, manager selection
  - Better proxy modelling of assets, liabilities, and interactions

### 3. When to apply AI – how to avoid overuse and issue of AI/ML

#### 3.1. Practical examples and guidelines

- Next to regulatory constraints, there needs to be a business case showing value for the insurer (financial, risk minimizing, increasing customer satisfaction, data quality) and a dedicated owner, responsible for implementation, maintenance, and monitoring
- Agile implementation can help to quickly show a proof of concept before rolling out on a bigger scale, increasing buy-in from management and reducing sunk costs

The value of AI lies within the combination of business sense, data handling, and modelling. Actuaries are perfectly positioned and enabled to unlock this value. Hence, the following anecdote from Matt Lerner from PayPal about an intern who solved a problem should serve as a prime example. PayPal faced the challenge of losing around one million merchants annually – previous analyses had not been able to give an answer as to where this is coming from. To narrow down the reasons behind this churn, they first excluded account closures and focused on "going dark" accounts, which indicates disengagement from the product. They also ruled out "one-and-dones" and addressed onboarding issues for new entrants. They further excluded false positives and non-regretted churn, narrowing down the analysis to well-behaved, non-seasonal merchants. By shifting their focus to larger merchants, they were able to refine their approach and detect various smaller issues that led to churn. By creating a predictive model, they now flag merchants at risk along these smaller issues and proactively manage through their customer service.

This example should highlight that understanding the data, reducing noise, and linking it back to business reality are key success factors in business-related problems. Here specifically,

understanding the true reason of churn has been key to create the appropriate action plan for operations with significant value generated.

So, what does this mean for actuaries concretely? Actuaries can and should actively shape and drive certain AI and data science initiatives. Looking at successful projects in that area, there are a few best practices that increase the chance of success and are straight-forward to apply:

1. Bring together on one table the right people – impact improves with diversity of skills and experience. You as an actuary can act as the glue, keeping everything together
2. Start with a concrete business question you try to solve and write down hypotheses on its impact on strategy, operations, and financials. In most applications, the more the question targets prescriptive aspects, the higher the value of the answer will be. Align with management.
3. Once this is clear, collect, modify, and interpret the data you need to answer this question, of course considering data protection requirements. As shown in the example of PayPal, it is of utmost importance the data matches the business question, because even when applying the “best” model, conclusions can be wrong if data is not properly understood.
4. Then, design the model considering factors of responsibility, explainability, simplicity, accuracy, predictability, and potentially more. Leverage data scientists if needed.
5. Consider the risks that are embedded in answering this business question (e.g. insurability, wrong incentivization, customer satisfaction and cannibalization), in the data (e.g. data protection, data security), and in the model (e.g. explainability, discrimination) and be concrete on limitation of applicability and results.
6. Implement and run the model. Try to push for early output in an agile way, even when first results will not be perfect.
7. Interpret, challenge, and validate results, putting particular focus on previously highlighted risks. Iteratively refine results and document.
8. Answer the question you asked at the beginning and communicate results in an understandable way to management
  - Be clear on the impact on strategy, operations, and financials.
  - Explain deviations from initial hypotheses.
  - Highlight areas of concern and limitations.
  - If applicable, provide next steps on how roll out, generalize, or transfer the learnings

#### **4. Overview of European regulation on the topic**

##### **4.1. Risks of AI**

Application of AI poses several risks. In this chapter we want to highlight a few pitfalls that actuaries should have in mind when handling data and AI:

- Explainability: AI has the potential to generate highly persuasive but factually incorrect responses, leading to misinformation and misinterpretation.
- Discrimination: The probabilistic nature of many AI applications can lead to unforeseen behaviours and capabilities during deployment, necessitating careful monitoring and management to ensure desired outcomes. Moreover, if models are trained on biased real-

world data, these models can perpetuate and amplify existing biases if deployed without proper oversight, exacerbating social and cultural inequalities.

- Data security and intellectual property: Cloud-based training of AI models poses security risks as it involves the transmission of proprietary data, increasing the potential for data breaches and unauthorized access. In general, improper usage of data and tools without adequate guidance and supervision can result in unintended consequences and ethical dilemmas.
- Cybercrime: The instant generation of convincing phishing emails and deepfakes facilitated by Generative AI enhances the ease of cybercrime, requiring heightened vigilance in online security.

These risks require responsible ways on how to use and apply AI. While recent and upcoming regulation will embed principles to safeguard sustainable use of AI, actuaries will play a crucial role as well. Key success factors will include the following

- A sound governance to navigate ethical, legal, and technological risks such as clear rules on where, when (not), and how to apply certain tools
- A framework with clear roles and responsibilities to support decision making
- Expertise, professionalism, and training around the topic of AI
- Suitable tools to monitor and manage AI risks

#### 4.2. General horizontal Regulation in the EU

**Direct discrimination** occurs when a person is treated less favourably than another person simply because one of their protected characteristics is not the same. If the person's corresponding risk factor is not used by insurers, such discrimination can be completely avoided. See, for example, the Directive 2004/113/EC ("Gender Directive") (Council of the European Union, 2004).

**Indirect Discrimination.** After avoiding direct discrimination, indirect discrimination occurs when a person is still treated disproportionately than another person by virtue of implicit inference from their protected characteristics, based on an apparently neutral practice such as using proxy variables from the non-protected characteristics of policyholders (i.e. identifiable proxy), or opaque algorithms (i.e. unidentifiable proxy). See, for example, the Directive 2004/113/EC ("Gender Directive").

**Algorithmic discrimination** refers to the biased outcomes or decisions produced by algorithms and is usually considered as a subset of indirect discrimination. The European Insurance and Occupational Pensions Authority (EIOPA, 2019) conducted a thematic review on the use of Big Data Analytics (BDA) based on 222 participated motor or health insurers from 28 European jurisdictions. The thematic review has revealed that 31% of insurance firms already actively used BDA tools and another 24% of firms plan to use them within the next three years. These new data analytics tools are generally used on pricing and underwriting, claims management and sales and distribution, whereas insurers have only taken limited approaches to ensure fair and ethical outcomes in the use of BDA in underwriting and pricing.

#### 4.3. Horizontal data oriented regulation in the EU

The General Data Protection Regulation (2016/679, "GDPR") is a Regulation in EU law on data protection and privacy in the EU and the European Economic Area (EEA). The GDPR is an important component of EU privacy law and of human rights law, in particular Article 8(1) of the Charter of Fundamental Rights of the European Union. It also addresses the transfer of personal data outside the EU and EEA areas. The GDPR's primary aim is to enhance individuals' control and rights over their personal data and to simplify the regulatory environment for international business. The regulation contains provisions and requirements related to the processing of personal data of individuals, formally called "data subjects", who are located in the EEA, and applies to any enterprise—regardless of its location and the data subjects' citizenship or residence—that is processing the personal information of individuals inside the EEA.

The GDPR was adopted on 14 April 2016 and became enforceable beginning 25 May 2018. As the GDPR is a regulation, not a directive, it is directly binding and applicable, and provides flexibility for certain aspects of the regulation to be adjusted by individual member states.

Personal data may not be processed unless there is at least one legal basis to do so. Article 6 states the lawful purposes are:

- a. If the data subject has given consent to the processing of his or her personal data;
- b. To fulfill contractual obligations with a data subject, or for tasks at the request of a data subject who is in the process of entering into a contract;
- c. To comply with a data controller's legal obligations;
- d. To protect the vital interests of a data subject or another individual;
- e. To perform a task in the public interest or in official authority;
- f. For the legitimate interests of a data controller or a third party, unless these interests are overridden by interests of the data subject or her or his rights according to the Charter of Fundamental Rights (especially in the case of children).

If informed consent is used as the lawful basis for processing, consent must have been explicit for data collected and each purpose data is used for (Article 7; defined in Article 4). Consent must be a specific, freely-given, plainly-worded, and unambiguous affirmation given by the data subject; an online form which has consent options structured as an opt-out selected by default is a violation of the GDPR, as the consent is not unambiguously affirmed by the user. In addition, multiple types of processing may not be "bundled" together into a single affirmation prompt, as this is not specific to each use of data, and the individual permissions are not freely given.

The European Commission published on 21 April 2021 its draft regulation on artificial intelligence (AI). The Commission designed a horizontal regulatory framework that encompasses any AI system that touches the single market, whether the provider is based in Europe or not. The Artificial Intelligence Act uses a risk-based approach and sets up a series of escalating legal and technical obligations depending on whether the AI product or service is classed as low, medium or high-risk, while a number of AI uses are banned outright. At the time of writing the draft regulation is dealt with in the trilogue negotiations of the Parliament, the Council and European Commission. The earliest possible adoption will be in 2024.

The original draft regulation includes into the high risk category essential private and public services, including access to financial services such as credit scoring systems. I.e., it does not put insurance or pensions into the high risk bucket. An updated draft of the EU AI Act, released in November 2021, classifies "AI systems intended to be used for insurance purposes" under the high-risk category. Specifically, it refers to "AI systems intended to be used for insurance premium setting, underwritings and claims assessments."

The outcome of the trilogues is still not certain but it seems wise to assume that at least certain parts of insurance will fall into the high risk bucket.

For AI applications in the high risk category there will be the following requirements:

- design in line with requirements: Ensure AI systems perform consistently for their intended purpose and are in compliance with the requirements put forward in the Regulation
- conformity assessment: Ex ante conformity assessment
- Post-market monitoring: Providers to actively and systematically collect, document and analyse relevant data on the reliability, performance and safety of AI systems throughout their lifetime, and to evaluate continuous compliance of AI systems with the regulation
- incident report system: Report serious incidents as well as malfunctioning leading to breaches to fundamental rights (as a basis for investigations conducted by competent authorities)
- new conformity assessment: New conformity assessment in case of substantial modification (modification to the intended purpose or change affecting compliance of the AI system with the Regulation) by providers or any third party, including when changes are outside the 'predefined range' indicated by the provider for continuously learning AI systems.

#### 4.4. Vertical regulation in the EU

There is already a comprehensive legislative framework underpinning the activity of insurance firms, which is also applicable to the use of AI within their organisations. This is, particularly, the case of the Solvency II Directive, the Insurance Distribution Directive (IDD), and the upcoming e-privacy Directive (ePD). For example, Article 41 (1) of Solvency II Directive requires "insurance and reinsurance undertakings to have in place an effective system of governance which provides for sound and prudent management of the business." Existing legislation should indeed form the basis of any AI governance framework, but the different pieces of legislation need to be applied in a systematic manner and require unpacking to assist organisations understand what they mean in the context of AI. Furthermore, an ethical use of data and digital technologies implies a more extensive approach than merely complying with legal provisions and needs to take into consideration the provision of public good to society as part of the corporate social responsibility of firms.

#### 5. AI Explainability (XAI) – from black box to white, model agnostic approach

AI explainability, also known as interpretability or transparency, refers to the extent to which the decisions and predictions made by an AI system can be understood and explained by humans. This is an important consideration for a variety of reasons, including accountability, trust, and fairness.

AI explainability is an important topic also specifically in the context of the machine learning (ML) models. It refers to the ability to understand and interpret how an ML model reaches its decisions. If an ML model is not explainable, it can be difficult for users to trust the decisions it makes and for regulators to ensure that it is operating ethically.

Specifically in the actuarial context in the last years there have been some proposals aimed at establishing a framework for explaining ML models to go beyond the confines of the linear models.<sup>1</sup>

---

<sup>1</sup> Towards Explainability of Machine Learning Models in Insurance Pricing – Kevin Kuo, Daniel Lupton - 03/2020.

There are several different categories of indicators that have been proposed for measuring the explainability of AI systems. These include:

- **Intelligibility:** This refers to the degree to which the internal workings of an AI system are understandable to users. This can be assessed by looking at the complexity of the algorithms and models used by AI, as well as the clarity and transparency of any explanations provided by the system.
- **Auditability:** This refers to the ability of an AI system to provide a clear and traceable record of its decision-making process, allowing users to understand how and why the system reached a particular conclusion. This can be important for ensuring accountability and trust in the system.
- **Transparency:** This refers to the extent to which an AI system reveals its inner workings and decision-making processes to users. This can be achieved through techniques such as feature visualization and sensitivity analysis, which allow users to see how the system is using different inputs to make its predictions.
- **Explainability:** This refers to the ability of an AI system to provide clear and understandable explanations for its decisions and predictions to users.
- **Trustworthiness:** This refers to the degree to which an AI system is perceived as reliable and trustworthy by users. This can be influenced by a variety of factors, including the system's accuracy, reliability, and explainability.

Those categories of indicators are applied/developed, as we will see in the next paragraphs, in five different XAI macro-areas<sup>2</sup>: 1) Feature Interaction and Importance 2) Attention Mechanism 3) Dimensionality Reduction 4) Knowledge Distillation and Rule Extraction 5) Intrinsically Interpretable Models.

AI interpretability and explainability play a pivotal role in the actuarial field, offering valuable insights and enabling effective decision-making. Model understanding provides actuaries with essential tools to:

- debug ML models effectively;
- identifying potential issues in data preprocessing, feature engineering, or model training.

This empowers actuaries to rectify these issues and enhance model performance.

Additionally, model understanding is instrumental in detecting biases that may lead to disparate impacts on certain groups. By uncovering such biases, actuaries can take corrective measures, mitigating discrimination and upholding fairness in decision-making processes.

Furthermore, in actuarial applications where ML models' predictions significantly impact individuals' financial well-being and livelihoods, model understanding becomes a vital tool to provide recourse. By revealing the reasoning behind predictions, individuals can comprehend the factors influencing their outcomes and challenge decisions if necessary. This transparency empowers individuals to seek redressal for potentially unjust outcomes.

In high-stakes actuarial scenarios, trust in ML model predictions is paramount. Model understanding enables actuaries to evaluate the reliability of model outputs and make informed decisions based on the model's reasoning. By verifying the soundness of the model's reasoning with respect to auxiliary criteria, actuaries can gain greater confidence in using the model's predictions for critical decision-making.

---

<sup>2</sup> Explainable Artificial Intelligence (XAI) in Insurance – Systematic Review

## 5.1 Local and global explanations

Local explanations are specific to a single prediction made by an ML model. For example, a local explanation might show which input features had the greatest influence on a model's decision to classify an image as a dog (e.g., LIME, i.e., Local Interpretable Model-agnostic Explanations).

Feature importance is a measure of how important each input feature is to an ML model's predictions. For example, an ML model might assign a higher importance to an image's color than its shape when classifying objects (e.g., Attribution Maps, SHAP)

Global explanations provide an overview of an ML model's behavior across multiple predictions. For example, a global explanation might show which input features are generally important for an ML model's decision-making process (e.g., Permutation Importance).

Counterfactual explanations: These are explanations that show how a model's prediction would change if one or more input features were altered. For example, a counterfactual explanation might show how an ML model's prediction of a customer's creditworthiness would change if their income were increased (e.g., Partial Dependency Plot)

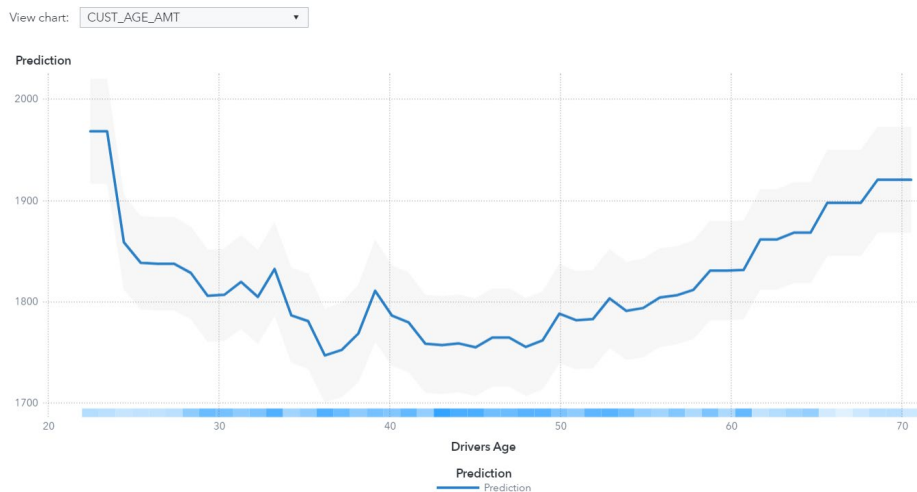
## 5.2 Partial Dependency

The partial dependency plot is a way to visualize the marginal effect of a feature on the predicted outcome of a machine learning model. Every statistical indicator (also the relativity plot for the classic GLM) has some assumptions, for the PD the main one is the independence: the PD assumes that the feature of interest is independent of other relevant features, except when explicitly varied in the plot.

An example of PD, in the context of insurance could be used to understand how the predicted probability/severity of a claim changes with the value of a particular feature, such as the age of the policy holder. It will be used a severity model as example, specifically a Random Forest model applied to predict the average claim amount.

To create a partial dependency plot, it is necessary first to fit a model. This could be a decision tree, random forest, or any other type of model. Once the model is trained, it can be used to make predictions on a grid of values for the feature of interest. For example, if we are looking at the effect of age on the predicted probability of a claim, we could make predictions for a range of ages from 20 to 70. Next, we plot the predicted severity of a claim on the y-axis and the age of the policy holder on the x-axis. This will show us how the predicted probability changes with age.





The plot above shows the relationship between the policyholder age and the predicted target, i.e., claim average severity, averaging out the effects of the other inputs. For interval inputs, the 95% confidence interval for the average target prediction is indicated by the shaded band around the line.

As said previously, one limitation of the partial dependency plot is that it only shows the marginal effect of a single feature on the predicted outcome. It does not take into account the potential interactions between features. To understand the interactions between features, we would need to use more advanced techniques such as partial dependence plots with interaction terms.

From a mathematical point of view the structure of the Partial Dependency Plot (PDP) is a function that maps a single feature to the target variable. This function is typically constructed by fixing the values of all other features in the model, and then averaging the predicted values of the target variable for a range of values of the feature of interest.

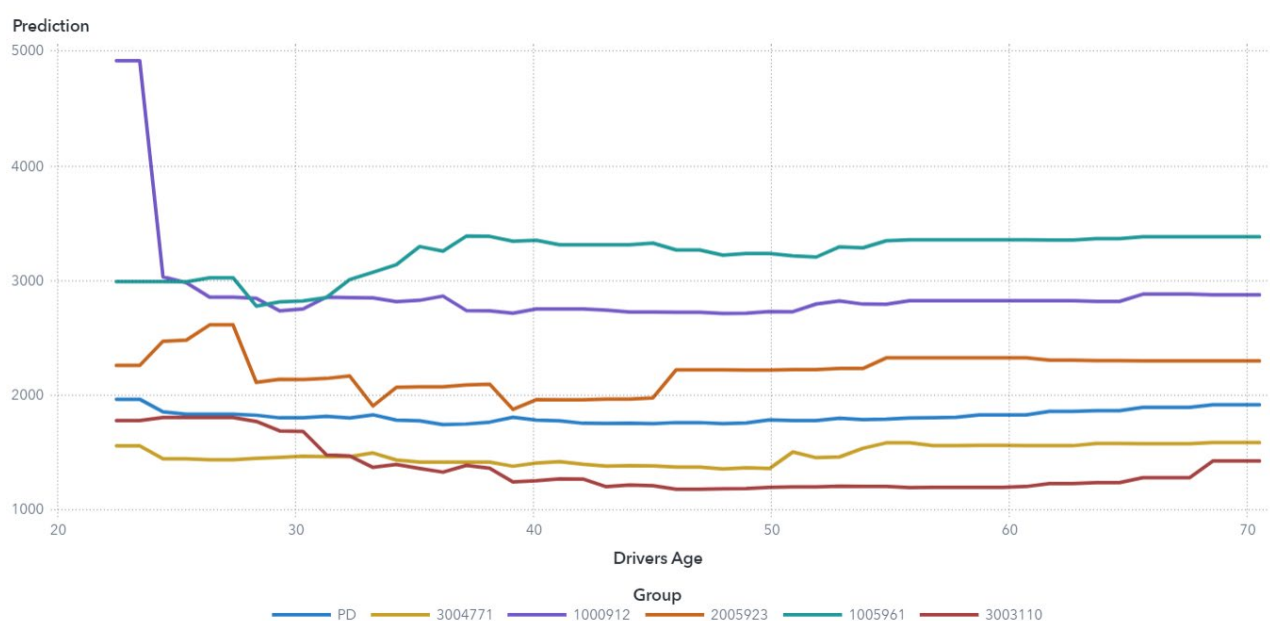
### 5.3 Individual Conditional Expectation (ICE)

The ICE explainability indicator is based on the idea that a model is more interpretable if its predictions can be accurately explained by a small number of input features. An ICE plot visualizes the dependence of the prediction on a feature for each instance separately, resulting in one line per instance, compared to one line overall in partial dependence plots. A PDP is the average of the lines of an ICE plot.

In the context of insurance the ICE can be used to evaluate the interpretability of a machine learning model that predicts the likelihood that a policyholder will file a claim based on their individual characteristics. To use the ICE, we first calculate the ICE curve for each input feature, which shows the relationship between the feature and the model's predictions.

If the ICE curves are relatively simple and can be accurately explained by a small number of features, this indicates that the model is highly interpretable, as its predictions can be easily understood in terms of the input features. On the other hand, if the ICE curves are complex and cannot be accurately explained by a small number of features, this indicates that the model is less interpretable, as its predictions are less transparent and more difficult to understand.

Following the same example done above:



The plot shows the partial dependency (PD) and the relationship between policyholder age and the predicted target for each individual observation.

#### 5.4 Small Perturbation

The Small Perturbation explainability indicator is a method for evaluating the interpretability of a machine learning model. It is based on the idea that a model is more interpretable if small changes to the input data result in small, predictable changes to the model's predictions.

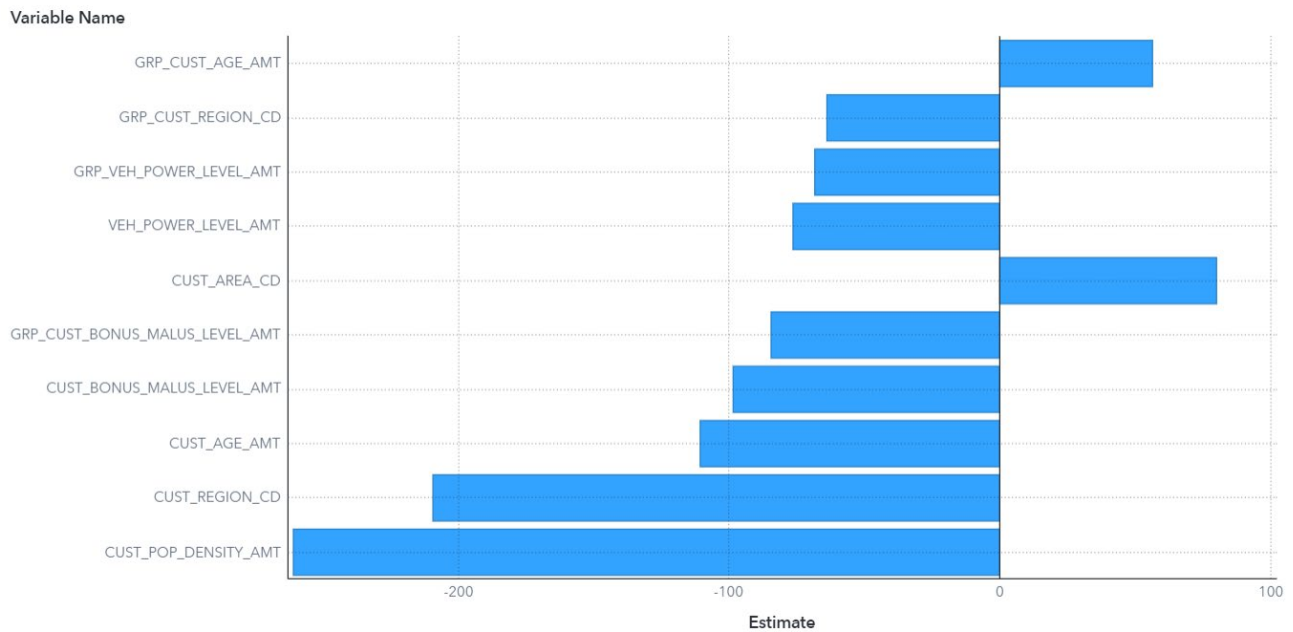
To use the Small Perturbation explainability indicator, we first choose a small, fixed perturbation value. Next, we make small changes to the input data by adding or subtracting the perturbation value from each feature. For each perturbed input, we calculate the model's prediction and compare it to the original prediction.

If the model's predictions are largely unchanged by the small perturbations to the input data, this indicates that the model is highly interpretable, as small changes to the input data result in small, predictable changes to the model's predictions. On the other hand, if the model's predictions change significantly with the small perturbations to the input data, this indicates that the model is less interpretable, as small changes to the input data can result in large, unpredictable changes to the model's predictions.

#### 5.5 Local Interpretable Model-Agnostic Explanations (LIME)

LIME is a technique used to explain the predictions of machine learning models in a way that is interpretable to users. The basic idea behind LIME is to fit a simple, interpretable model to the neighborhood around the prediction. Once the interpretable model has been fit it is used to explain the prediction. Following the same example used above for the Partial Dependence, we suppose to use a Random Forest model to predict the average claim amount.

We can use LIME to explain the model's prediction for a specific data instance, specifically in the graph reported below we see the LIME result for an instance whose predicted claim amount is 1.732€, the LIME results are:



This LIME plot displays the regression coefficient (estimate) for the inputs selected in the local surrogate linear regression model for fitting the predicted value of the target Average Claim Amount for each individual observation. The inputs are ordered by significance in the chart (GRP stands for grouped), with the most significant input for the local regression model appearing at the bottom of the chart.

In the instance considered as example in the graph above the most important variable, contributing to the prediction is the CUST\_POP\_DENSITY\_AMT (the population density where the policy holder lives), the second most important is CUST\_REGION\_CD (the region where the policy holders lives) etc.

In this way, LIME allows us to explain the predictions of complex machine learning models in a simple, interpretable way. This can be useful for understanding the model's behavior and for detecting any potential biases or errors in the model's predictions.

### 5.1. Shapley Additive explanation SHAP

The SHAP value for a given input feature in a machine learning model is derived using a method called the Shapley value from game theory. The Shapley value is a measure of the contribution of each player in a cooperative game and it has been adapted for use in machine learning to measure the contribution of each input feature to a model's predictions.

To calculate the SHAP value for a given input feature, the following steps are performed:

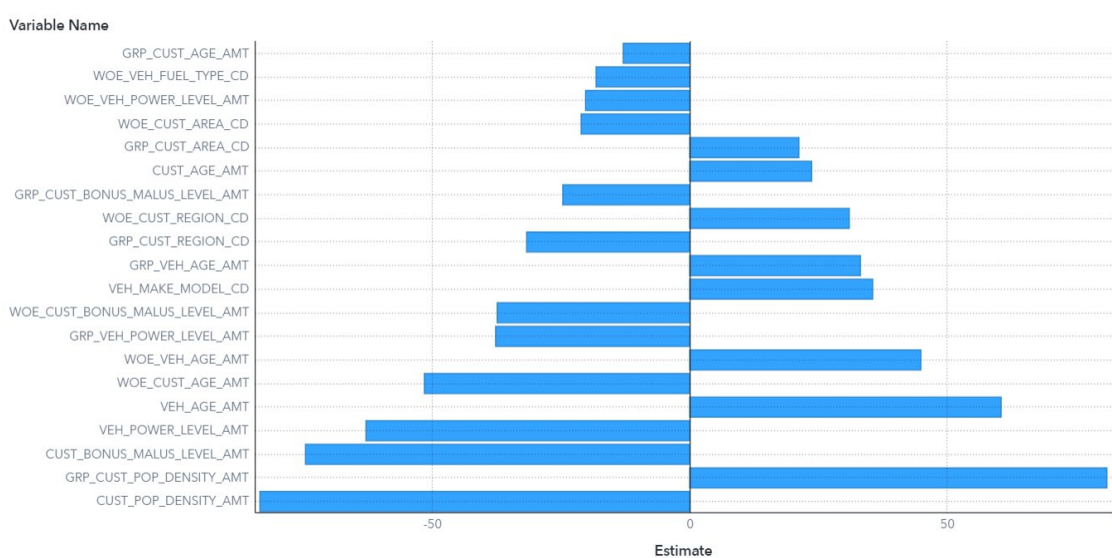
- The model's prediction is calculated for the input data.
- The average prediction of the model is calculated across all possible inputs.
- The difference between the model's prediction and the average prediction is attributed to the contribution of the input feature.

This process is repeated for each input feature to calculate the SHAP value for each feature. The SHAP values for the input features can then be used to evaluate the interpretability of the model, as described below.

In general, the SHAP value for a given input feature measures the degree to which the feature contributes to the model's predictions and can be used to evaluate the interpretability of the model. If the SHAP values for

the input features are relatively small and consistent in direction, this indicates that the model is highly interpretable, as its predictions can be accurately explained by the contributions of individual input features. On the other hand, if the SHAP values are large and inconsistent in direction, this indicates that the model is less interpretable, as its predictions are less transparent and more difficult to understand.

Still following the example considered in the previous paragraphs, the prediction of the average claim intensity, below we see an example of displaying the SHAPValue:



For each individual observation, an input's Shapley value is the contribution of the observed value of the input to the predicted value of the target AVERAGE\_CLAIMS\_AMT. The inputs are displayed in the chart ordered by importance according to the absolute values.

## References for further readings: papers, books

### References

- EIOPA. (2023). *Supervisory statement on differential pricing practices in non-life insurance lines of business*. [https://www.eiopa.europa.eu/system/files/2023-03/EIOPA-BoS-23-076-Supervisory-Statement-on-differential-pricing-practices\\_0.pdf](https://www.eiopa.europa.eu/system/files/2023-03/EIOPA-BoS-23-076-Supervisory-Statement-on-differential-pricing-practices_0.pdf).
- Seehafer, M., Nörtemann, S., Offtermatt, S., Transchel, F., Kiermaier, A., Külheim, R., & Weidner, W. (2021). *Actuarial Data Science*. De Gruyter STEM.
- Wüthrich, M. V., & Merz, M. (2022). *Statistical Foundations of Actuarial Learning and its Applications*. Springer Actuarial.