Deloitte.

AI, Model Interpretability, and the Future of Insurance

Actuarial and Insurance Solutions

January 2024

"Interpretable Machine Learning refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans."

- Christoph Molnar, 2019*

(Interpretable Machine Learning: A Guide for Making Black Box Models Explainable)

**Christoph Molnar* is an Independent Researcher and Author who writes about machine learning*

Introduction

Machine learning and artificial intelligence (ML/AI) are revolutionizing the insurance industry and the work of actuaries. These powerful tools have the potential to transform the way insurers make business decisions, but they also come with risks. As ML/AI models become more sophisticated, there is a threat of inaccurate predictions and challenges to integrate them into existing business processes and products. However, model-agnostic interpretability methods can help mitigate this, providing insights into how the models make their predictions.

Model interpretability is the practice of understanding the rationale behind an AI model's predictions, by identifying the variables and the degree of their impact on a certain prediction. Thus, they have the potential to play an important role in insurance and can be used to investigate and correct bias by adjusting the model or collecting additional data.

ML/AI algorithms are designed to build accurate predictions, and in many cases they successfully do so. However, they also suffer from the possibility of developing a biased model; for example, from (either consciously or unconsciously) choosing training data that underrepresents certain demographics.

Simply having a technically accurate prediction is not enough; understanding the "why" behind an ML/AI model prediction is important for preventing biases and regulating the use of ML/AI in the insurance industry. Demand is growing for actuaries to develop AI expertise, such as model interpretability; this article introduces the concepts and methods and considers two large groups of model interpretability: Global and Local agnostic methods.

To illustrate, consider the scenario of constructing a model to automate the approval or rejection of housing loan applications. Since the model is trained with historical data, it may go unnoticed that there is an inadequate representation of women within the dataset. As a result, the model will be discriminatory, perhaps resulting in fewer women being approved for housing loans and perpetuating discriminatory practices.

WHAT

What does model interpretability mean?

Methods of model interpretability are a useful debugging tool for detecting bias and building general understanding of predictions in machine learning models. **Actuaries are uniquely positioned to play a leading role in interpreting and evaluating the socio-economic impact of AI/ML models, going beyond traditional actuarial requirements.**

WHY

Why do actuaries need to learn model interpretability?

AI/ML models are used by companies to make decisions that influence financial outcomes. Model interpretability is a tool that can be leveraged by actuaries to understand AI/ML models.

Let's take an example to illustrate how model interpretability can be used by Actuaries to detect bias. Let's say we are building a model to predict creditworthiness for loan applicants. We train the model on historical data that includes factors such as income, credit score, employment status, and age. After training the model, we notice that it consistently denies loans to applicants over the age of sixty.

To investigate this, we can use model interpretability techniques to understand which factors the model is using to make its predictions. We might find that the model is heavily weighting age as a factor in its decision-making process. This could indicate that the model is biased against older applicants, as it is assuming that they are less creditworthy based solely on their age.

The insurance industry is of course subject to strict regulations, requiring insurers to ensure ethical, secure, and reliable implementation of AI/ML technologies.

Model interpretability techniques can be used to investigate and correct bias by adjusting the model or collecting additional data.

HOW

How do we begin interpreting the model?

Although there is no single correct model interpretability method, there are certain rules that an actuary can follow to narrow the approach. **We will consider two large groups of model interpretability methods: Global and Local agnostic methods.**

Figure 1 below will assist with deciding which model-agnostic method(s) to use when considering model interpretability.

Further considerations

Assessing the appropriate model interpretation method

The selection of a model-agnostic interpretation method should be deliberate, when assessing machine learning models. The nature of the ML/AI model itself must be factored in - the advantages and drawbacks of the model interpretation method, as well as the type of data it employs. If the original model is inherently non-interpretable, our assessment goals become pivotal. Consider the following questions:

Questions:
1. What statistical algorithm was used? Can the model outputs be immediately interpreted? E.g., linear regression coefficients can be used to interpret the impact a variable has on predictions. A neural network does not immediately have similar coefficients.
2. How does the trained model make predictions?
3. How do parts of the model affect predictions? E.g., variables that were / weren't included, interaction terms, variable transformations prior to model training, etc.
4. Are we aiming to comprehend the general impact of variables on the model? Or do we want to understand why specific predictions were made for individual observations or groups of observations?
5. Do we seek to quantify how individual variables contribute to specific outcomes?

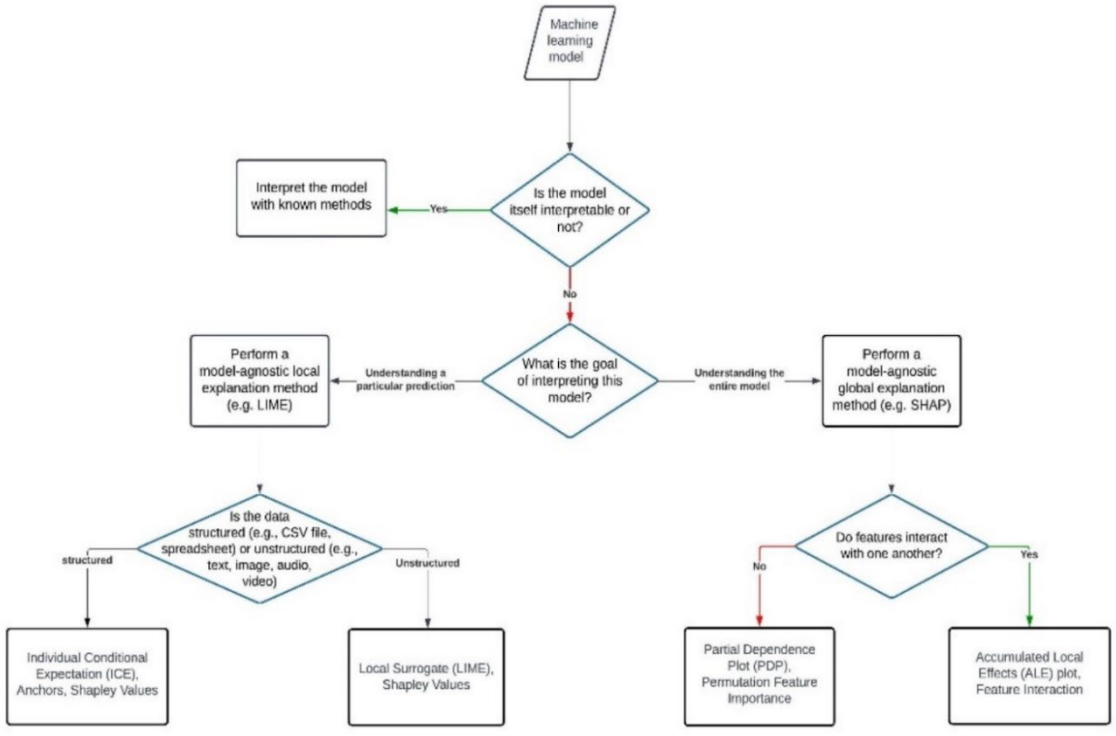

Figure 1: Methods of model interpretability.

Global and local model-agnostic methods

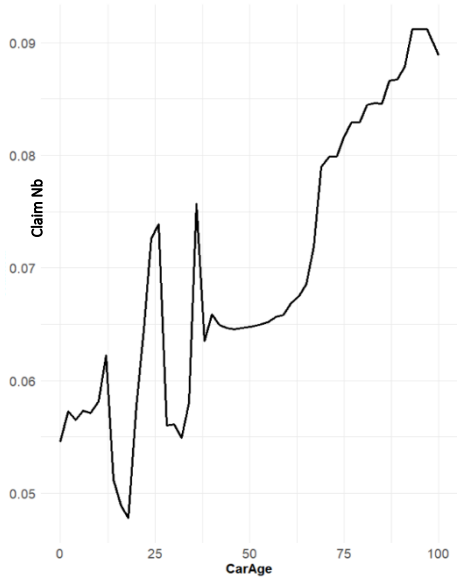

Figure 2: PDP for the claim frequency prediction model, based on car age.

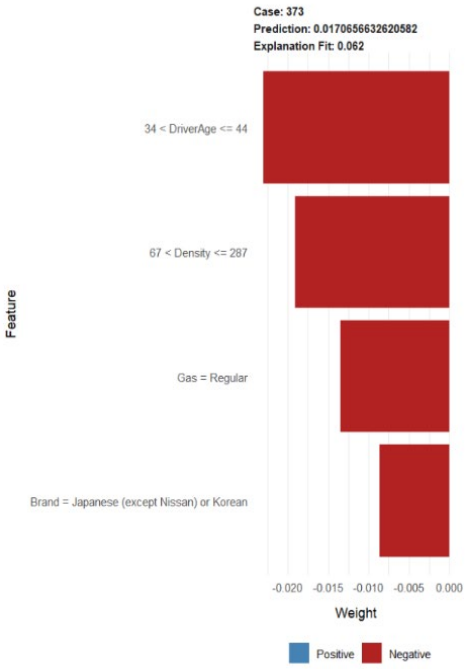

Figure 3: LIME explanation of auto insurance dataset. The x-axis shows the feature effect: the weight times the actual feature value.

Global model explanations help us understand the primary drivers of predictions made by models, aggregated over the whole dataset. For example, a Partial Dependence Plot (PDP) shows the marginal effect of one or two features on the predicted outcomes of the predictive model.

To illustrate, suppose we have a model that forecasts the frequency of auto insurance claims. We are interested in understanding the impacts of variables like car age, driver age, vehicle model, and population density on our predictions. Figure 2 presents an example of a PDP on a hypothetical model predicting auto insurance claim frequency. Based on the graph, we notice that there are drastic fluctuations between the relationship of car age and claim frequency. At least intuitively, we would anticipate the claim frequency to increase steadily with increasing car age. There is an unexpected dip in claim frequency predictions after around vehicle ages twenty and thirty. As a pricing actuary, we would be compelled to understand what are leading to these predictions, before moving forward with pushing this ML/AI pricing model into production.

Contrary to global explanations, a local explanation provides information about a prediction for a single observation. An example of a local method is local interpretable model-agnostic explanations (LIME), which focus on training local surrogate models to explain individual predictions.

Instead of using the training data, LIME assumes that only the black box model is available, which can then be "queried" to input data points and obtain the corresponding predictions. The objective is to comprehend the rationale behind a specific prediction generated by the machine learning model. LIME accomplishes this by observing the impact on predictions when data variations are fed into the machine learning model.

Again, in the context of our claim frequency auto insurance example, suppose we are interested in the degree to which each factor positively or negatively contributed to the predicted claim frequency for a specific observation. In Figure 3, generated by LIME, driver age (33<driver age <= 44) has a negative (i.e., reduces the) contribution to the predicted claim number. Furthermore, LIME categorizes numerical features into bins because categorical features are easier to interpret than numerical features. Numerical values in the same bins have the same positive or negative effect on the outcomes.

Conclusion

ML/AI has the potential to transform the insurance industry by improving accuracy, efficiency, and profitability. However, it is essential for insurers and actuaries to stay up-to-date with the latest developments in ML/AI and use these tools responsibly to ensure that they are making accurate and ethical decisions. By doing so, they can continue to provide value to their clients while minimizing the risks associated with these powerful technologies.

One of the biggest advantages of ML/AI in the insurance industry is the ability to process large amounts of data quickly and accurately. This can help insurers make more informed decisions about risk assessment, pricing, and claims processing. However, as with any technology, there are risks associated with ML/AI. One of the biggest concerns is the potential for unwanted predictions. For example, an ML/AI model might predict that a certain group(s) of people are more likely to make a claim, based on factors such as age or gender - that could lead to discrimination and ethical concerns, especially if the predictions are inaccurate or biased.

Another challenge is integrating ML/AI into existing processes and products. Insurers and actuaries need to ensure that the models are accurate, reliable, and transparent, and that they can be easily integrated into existing systems. Model-agnostic interpretability methods can help address these challenges by providing insights into how the models make their predictions. This can help insurers and actuaries better understand the models and identify any potential biases or errors.

To learn more about how Deloitte's Pricing Centre of Excellence can help your organization with model interpretability, please contact:

Antonio Ferreiro
Partner, National Leader P&C
Audit & Assurance
aferreiroc@deloitte.ca

Shayan Sen
Senior Manager
Audit & Assurance
shasen@deloitte.ca

We'd like to thank **Harrison Jones, Kimon Karavias, Arjun Batra, Lily Fatykova and Sofia Colella** for their help with this publication.

Sources

An, Qi, Carl Lussier, Joshua Snow, Erik Christianson, Alexandre Monette-Pagny, and Alina Rogozhnikova. "Bias and Fairness in Pricing and Underwriting of Property and Casualty (P&C) Risks." Canadian Institute of Actuaries, April 2023. https://www.cia-ica.ca/docs/default-source/2023/223056e.pdf.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. "Fairness and Machine Learning." MIT Press, 2023. https://fairmlbook.org/.

jphall663. "Awesome Machine Learning Interpretability." GitHub, 2023. https://github.com/jphall663/awesome-machine-learning-interpretability.

Molnar, Christoph. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." christophm.github.io, August 21, 2023. https://christophm.github.io/interpretable-ml-book/.

Smith, Logan T, Emma Pirchalski, and Ilana Golbin. "Avoiding Unfair Bias in Insurance Applications of AI Models." SOA Research Institute, 2022. https://www.soa.org/4a36e6/globalassets/assets/files/resources/research-report/2022/avoid-unfair-bias-ai.pdf.

Deloitte.