MILLIMAN WHITE PAPER

Explainable AI in practice

Build trust and encourage adoption

Jan Thiemen Postema,
Raymond van Es

In the past five years, the interest in *explainable AI* (XAI) has taken flight. With this increased interest came an influx in algorithms that attempted to explain the outcomes of increasingly intricate black-box machine learning models, where we can only observe the input and outputs, not the inner workings. Within this new environment, as is often the case, there's not one technique to rule them all.

Applications rooted in artificial intelligence (AI) have stirred up a revolution in almost every industry including the insurance sector, which has traditionally been highly data-driven. Despite the great predictive powers machine learning (ML) algorithms possess, a major hurdle limits their widespread adoption: they lack transparency.

This lack of transparency limits both the public's and your company employees' trust in such algorithms. In this article we discuss the techniques available to mitigate this issue and what should be considered whilst implementing them.

Those techniques aim to provide some transparency by generating explanations, based on the results of the underlying models, hence the name explainable AI. However, before we can decide what XAI technique is most suitable for us, we ought to ask ourselves, what is an explanation? Because the scholarly world hasn't been able to give us a clear answer on this yet, we'll limit ourselves to another question: When is an algorithm interpretable? For our purposes, we'll use the following answer:

An algorithm is interpretable when the explanation is sufficiently clear to the audience to trust the outcome.

In other words, whether an algorithm is interpretable is fully dependent on the audience and what it considers to be a clear explanation of that algorithm's outcomes.

Simply put, for different situations we need different strategies. Therefore, we'll illustrate our stance on the topic through three potential audiences for our explanations. Each has its own goals and target audience, and each leads to a different solution.

Laypersons

WHY IS MY QUOTE HIGHER THAN MY NEIGHBOUR'S?

Everybody has a right to an explanation. That isn't just a social right, it's also enshrined in law. For example, the General Data Protection Regulation (GDPR) states that data subjects have a right to "meaningful information about the logic involved" and to "the significance and the envisaged consequences" of automated decision-making. The GDPR also states that data subjects shall have the right not to be subjected to a decision based solely on automated processing. These provisions introduce complex obligations between data subjects and the models processing their data, indicating a right to an explanation.

This is especially true in situations that are generally not thoroughly understood by the layperson. The insurance sector is a prime example. Traditionally a car insurance quote is based on a few easy-to-understand factors such as age, location, car brand and type, and the number of claim-free years. However, when we start to lean more and more on pricing strategies driven by big data, it becomes difficult to explain what factors a quote is based on.

This is true for traditional pricing models as well; however, the problem may be amplified when an opaque ML technique is used. Therefore, we need to explain the origins of the quote in a way that can be understood by the average person, our audience in this case.

Out of the wealth of available techniques, two are especially suitable for this kind of problem: scoped rules and counterfactual examples. The first, as the name implies, explains an individual prediction by creating rules that "anchor" it in place. In other words, so long as the rules are true, the prediction won't change. For example, if the model predicts a quote of €60 per month, an explanation could be: "age is under 23 and claim-free period is under two years."

The counterfactual examples technique, however, take a similar yet completely opposite approach. Instead of explaining why the customer gets a certain quote, it explains when he or she would have gotten a different outcome. In the example above the explanation could be: "the quote would have been 10% lower if you were two years older."

Both approaches are equally sound and can be complementary to each other. Which one is most suitable depends on the specific situation and the customer's preferences. There are, of course, other options that are suitable in this situation, as long as they can be presented in such a way that they are understood by the audience.

Regulators

IS MY MODEL FIT FOR ITS PURPOSE?

The insurance sector is under constant scrutiny and the companies operating in it should uphold the highest standards in risk management. That is why many of these players have adopted innovative frameworks for managing the risks stemming from the use of AI. Additionally, the insurance industry is heavily regulated and ML models need to comply with the same stringent regulations as traditional models. Recent guidelines by national and international regulators such as the US National Association of Insurance Commissioners (NAIC) and the Dutch national bank provide some tools to guide us through this process. An important aspect is explainability; knowing why certain decisions are made.

In traditional settings documentation for the regulators is created by actuaries. They are professionals in the actuarial sciences, a subfield of finance and applied mathematics. However, at its core it is still a very theoretical field. Data scientists, on the other hand, practice a more applied form of science, more akin to that of chemistry, where everything resolves around experiments.

These two very different approaches also create a certain tension as one focusses on theory whilst the other focusses on observations. This is exemplified by the available XAI techniques. Of the popular options, only one is rooted in a strong theoretical basis: SHapley Additive exPlanations (SHAP) values. The SHAP option would appear to be the only suitable technique when seeking regulatory approval as this has historically been the realm of theory-driven decision making. However, the SHAP method has its drawbacks, among others a difficult interpretation and high computing costs, often requiring approximated solutions instead.

Domain experts

HOW CAN WE INTEGRATE ML INTO AN EXPERT MODEL?

XAI also opens up new possibilities that weren't available before. One of these possibilities is to integrate the knowledge gathered by an ML model into an expert model. This could be useful when there is a lot of available data but, due to external forces, the use of a black-box ML model is infeasible. In this case, instead of implementing the model directly, a domain expert such as an actuary could use the insights gathered by the black-box model to create an expert model.

Using internal domain experts as an audience the primary goal would be to provide an explanation that is clear on both a global scale (the whole model) and the local scale (one prediction). Because there is no single technique that satisfies both these needs, we must use a combination of techniques. For example, using Permutation Feature Importance in this situation would point to the importance of each feature in the model. Also, Accumulated Local Effects (ALE) Plots, which show the effect of a certain feature on the prediction, might greatly help domain experts find their way through today's massive data sets and incorporate new patterns that they normally wouldn't have discovered.

Conclusion

The insurance sector is built on trust, which might be one of the most valuable currencies for insurance companies. However, to gather trust from the public we need to be as transparent as possible about our models and how they reach decisions. This is increasingly difficult in the world of black-box algorithms and deep neural networks.

Luckily, due to the advent of explainable AI, we're now well equipped to provide explanations to the users of our models. However, the type of explanation we use is completely dependent on the audience—there is no one-size-fits-all solution. This and the fact that the field is constantly evolving makes explainability a very nontrivial problem.

Final remarks

There is a saying in the world of data science: "Life isn't a Kaggle competition." That is to say, there is more to an ML model than achieving the highest possible score on some kind of metric. Even though we've mainly considered advanced black-box models, there are plenty of inherently interpretable ML models, such as linear models and decision trees. Even though these models may not provide the same performance as more advanced versions, the trade-off might be worth making and should always be considered.

CONTACT

Raymond van Es
raymond.vanes@milliman.com

Jan Thiemen Posteman
janthiemen.postema@milliman.com