SHAP and LIME for machine learning model interpretation

Why bother?

It seems that nowadays, companies from all sectors are rushing to sprinkle some machine learning on their products and services. The insurance sector is no different.

Putting semantics and categories aside, we know that ML algorithms like GBMs, neural networks and random forests do a lot better in modelling real world scenarios (insurance pricing and reserving included) compared to our more traditional models like linear regression or GLMs.

Why is ML more accurate?

This is based on the simple fact that these aforementioned ML algos are capable of handling more complex patterns in data, where a linear regression model is limited to a straight line (or plane or any D above) and a GLM would be constrained by its assumption of errors and link function.

Why is ML a black box?

As we all know, this accuracy comes at a cost of interpretability. Where both a linear regression and GLM can be expressed and completely described as a closed form solution INSERT EXAMPLE HERE, something like a neural network cannot.

There are many reasons for this, but this is based on the underlying fact that we currently do not have a good way of scaling up the interpretability of composition of functions. It is this, that gives the neural network its power, and its mystery.

Why is GBM a black box?? Same reason? Why can’t a GBM be expressed in closed form?

Implications on this in the insurance industry and how it compares to other industries?

Keeping all this in mind, we can start to appreciate why senior stakeholders in an insurance company would be hesitant on applying the same level of reliance on these ML algos as a company in a different sector, say Spotify.

Imagine building an extremely accurate fraud detection system into your underwriting process and the slide pack saying: “We are very sure that this guy is committing fraud, but we have no idea why.” Minority report much?

What is the way forward in terms of explainable ML?

The ultimate goal would definitely be to have a closed form solution of GLMs with accuracy of NNs. But as of 2020. This would mean global interpretability, which we currently do not have yet. This is not to say that there haven’t been advances in this area. A lot of research has gone into interpreting model results locally through isolating changes and observing the marginal changes in the response. 2 of the popular methods are LIME and SHAP values.

Intro to SHAP and LIME

Lime is short for Local Interpretable Model-Agnostic Explanations

SHAP is short for Shapley Additive Explanations

Worked Example

What dataset? Something simple

Explain output

Limitations, advantages

Compare to things like PDP/ICE/ Variable importance

Use Binder?

Links

<https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30>

<https://github.com/slundberg/shap>