**Using SHAP and LIME for 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.

Give example of linear + interactions

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

Give example on racist chatbot and sexist AI,and then say how it could get more serious in insurance

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 1

Before trying to interpret SHAP outputs for a more complicated machine learning model, let’s start with one of the simplest models, the simple linear regression model.

Here, we will se sci-kit learn framework’s linear regression class, and since model building is not the focus of article, we will skip all the details and optimisations steps.

One form of model validation that most people will be familiar with are the fitted coefficients of the model, which can provide a very crude interpretation of a feature importance score.

The main reason that this is crude, is because there is an underlying assumption that the inputs are either on the same scale or have been wrangled in some form to make them of the same scale. Just as an example, if we had measured age in hours, we’d realize that the coefficients would be huge, though the importance of the number of hours a policyholder has been alive would most definitely not be more important than the number of years.

The main concept behind SHAP is using a game theoretical approach to allocate contributions for a model’s output to its inputs. Wont go into too much detail?

Main idea of SHAP is that: in order to evaluate the contribution from a single feature, we must consider the model results from each of the combinations of features.

Some definitions we need to do:

* Players – features, if a player is in the game, then we consider the feature
* Game – model outcome
* Power set

Dataset used

https://github.com/sharmaroshan/Insurance-Claim-Prediction

Links

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

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