**Using SHAP values to understand model outputs**

Why bother?

* Insurance is a lot more conservative than a typical tech company in terms of modelling
* Insurers will still trade of accuracy for model transparency, which is why GLMs are still the bread and butter of insurance pricing
* As data becomes increasingly available and insurance products get more complex, the use of more complex models to handle these interactions is inevitable
* The best explanation for a model is the model itself, e.g SLR coefficients. However there are downsides to this, and more complex models do not have a straight forward interpretation.

What are SHAP values?

* SHAP is short for Shapley Additive Explanations
* Stems from a game theory approach from Lloyd Shapley
* It is introduced as a class of explanation models, furthermore, also claims to unify previous explanation models
* Maybe use the uber share ride as an example?

Differences between LIME and SHAP?

Lime is short for Local Interpretable Model-Agnostic 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>