mSHAP
Explaining Two-Part Models

Spencer Matthews

Joint work with Brian Hartman

June 2021

Brigham Young University

Motivation

- Two-part models are used by actuaries to set insurance rates, and therefore must be explainable
- Newer "black-box" methods (such as the gradient boosted forest) provide greater accuracy to these pricing models
- Although methods exist to explain individual models, there is not a good methodology to explain the predictions of a two-part model

A Brief Introduction to SHAP Values

https://github.com/slundberg/shap

Definitions

- Three Models: f, g, and h, where h is the product of f and g.
- Input Matrix: A where A_i is the ith row of A and A is $n \times p$ where n is the number of observations and p is the number of predictors.
- $f(A_i) = \hat{x}_i$, $g(A_i) = \hat{y}_i$, and $h(A_i) = \hat{z}_i$ and the contribution of the jth predictor to $\hat{x}_i$ as $s_{x_i j}$.
- μ_f, μ_g, μ_h signify the average model prediction over the data (known as the baseline term or expected model output)
- Based on the property of local accuracy:

$$\hat{x}_i = \mu_f + s_{x_i 1} + s_{x_i 2} + \ldots + s_{x_i p}$$

and

$$\hat{y}_i = \mu_g + s_{y_i 1} + s_{y_i 2} + \ldots + s_{y_i p}$$

Local Accuracy

$$x_i = \mu_f + \sum_{j=1}^{p} s_{x_i j}$$

$\Rightarrow$

The sum of the SHAP values and the expected model output must equal the model prediction

Local Accuracy

$$x_i = \mu_f + \sum_{j=1}^{p} s_{x_i j}$$

$\Longrightarrow$

The sum of the SHAP values and the expected model output must equal the model prediction

A Brief Example

Local Accuracy

$$x_i = \mu_f + \sum_{j=1}^{p} s_{x_i j}$$

$\Rightarrow$

The sum of the SHAP values and the expected model output must equal the model prediction

A Brief Example

$$\hat{x}_i = \mu_f + s_{x_i 1} + s_{x_i 2}$$

$$\hat{y}_i = \mu_g + s_{y_i 1} + s_{y_i 2}$$

Local Accuracy

$$x_i = \mu_f + \sum_{j=1}^{p} s_{x_i j}$$

$\Longrightarrow$

The sum of the SHAP values and the expected model output must equal the model prediction

A Brief Example

$$\hat{x}_i = \mu_f + s_{x_i 1} + s_{x_i 2}$$

$$\hat{y}_i = \mu_g + s_{y_i 1} + s_{y_i 2}$$

$$\hat{x}_i \cdot \hat{y}_i$$

$$\neq$$

$$\mu_f \mu_g + s_{x_i 1} s_{y_i 1} + s_{x_i 2} s_{y_i 2}$$

Expansion of Terms: $\hat{x}_i \times \hat{y}_i$

	s_{x_i1}	+	s_{x_i2}	+	s_{x_i3}	+	$\dots$	+	s_{x_ip}	+	μ_f
s_{y_i1}	$s_{x_i1}s_{y_i1}$		$s_{x_i2}s_{y_i1}$		$s_{x_i3}s_{y_i1}$		$\cdots$		$s_{x_ip}s_{y_i1}$		$\mu_f s_{y_i1}$
+											
s_{y_i2}	$s_{x_i1}s_{y_i2}$		$s_{x_i2}s_{y_i2}$		$s_{x_i3}s_{y_i2}$		$\cdots$		$s_{x_ip}s_{y_i2}$		$\mu_f s_{y_i2}$
+											
s_{y_i3}	$s_{x_i1}s_{y_i3}$		$s_{x_i2}s_{y_i3}$		$s_{x_i3}s_{y_i3}$		$\cdots$		$s_{x_ip}s_{y_i3}$		$\mu_f s_{y_i3}$
+											
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\ddots$		$\vdots$		$\vdots$
+											
s_{y_in}	$s_{x_i1}s_{y_ip}$		$s_{x_i2}s_{y_ip}$		$s_{x_i3}s_{y_ip}$		$\cdots$		$s_{x_ip}s_{y_ip}$		$\mu_f s_{y_ip}$
+											
μ_g	$s_{x_i1}\mu_g$		$s_{x_i2}\mu_g$		$s_{x_i3}\mu_g$		$\cdots$		$s_{x_ip}\mu_g$		$\mu_f\mu_g$

Expansion of Terms: $\hat{x}_i \times \hat{y}_i$

	s_{x_i1}	+	s_{x_i2}	+	s_{x_i3}	+	$\cdots$	+	s_{x_ip}	+	μ_f
s_{y_i1}	$s_{x_i1}s_{y_i1}$		$s_{x_i2}s_{y_i1}$		$s_{x_i3}s_{y_i1}$		$\cdots$		$s_{x_ip}s_{y_i1}$		$\mu_f s_{y_i1}$
+											
s_{y_i2}	$s_{x_i1}s_{y_i2}$		$s_{x_i2}s_{y_i2}$		$s_{x_i3}s_{y_i2}$		$\cdots$		$s_{x_ip}s_{y_i2}$		$\mu_f s_{y_i2}$
+											
s_{y_i3}	$s_{x_i1}s_{y_i3}$		$s_{x_i2}s_{y_i3}$		$s_{x_i3}s_{y_i3}$		$\cdots$		$s_{x_ip}s_{y_i3}$		$\mu_f s_{y_i3}$
+											
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\ddots$		$\vdots$		$\vdots$
+											
s_{y_in}	$s_{x_i1}s_{y_ip}$		$s_{x_i2}s_{y_ip}$		$s_{x_i3}s_{y_ip}$		$\cdots$		$s_{x_ip}s_{y_ip}$		$\mu_f s_{y_ip}$
+											
μ_g	$s_{x_i1}\mu_g$		$s_{x_i2}\mu_g$		$s_{x_i3}\mu_g$		$\cdots$		$s_{x_ip}\mu_g$		$\mu_f\mu_g$

Expansion of Terms: $\hat{x}_i \times \hat{y}_i$

	s_{x_i1}	+ s_{x_i2}	+ s_{x_i3}	+ $\dots$	+ s_{x_ip}	+ μ_f
s_{y_i1}	$s_{x_i1}s_{y_i1}$	$s_{x_i2}s_{y_i1}$	$s_{x_i3}s_{y_i1}$	$\dots$	$s_{x_ip}s_{y_i1}$	$\mu_f s_{y_i1}$
+						
s_{y_i2}	$s_{x_i1}s_{y_i2}$	$s_{x_i2}s_{y_i2}$	$s_{x_i3}s_{y_i2}$	$\dots$	$s_{x_ip}s_{y_i2}$	$\mu_f s_{y_i2}$
+						
s_{y_i3}	$s_{x_i1}s_{y_i3}$	$s_{x_i2}s_{y_i3}$	$s_{x_i3}s_{y_i3}$	$\dots$	$s_{x_ip}s_{y_i3}$	$\mu_f s_{y_i3}$
+						
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
+						
s_{y_in}	$s_{x_i1}s_{y_ip}$	$s_{x_i2}s_{y_ip}$	$s_{x_i3}s_{y_ip}$	$\dots$	$s_{x_ip}s_{y_ip}$	$\mu_f s_{y_ip}$
+						
μ_g	$s_{x_i1}\mu_g$	$s_{x_i2}\mu_g$	$s_{x_i3}\mu_g$	$\dots$	$s_{x_ip}\mu_g$	$\mu_f \mu_g$

Expansion of Terms: $\hat{x}_i \times \hat{y}_i$

	s_{x_i1}	$+$	s_{x_i2}	$+$	s_{x_i3}	$+$	$\cdots$	$+$	s_{x_ip}	$+$	μ_f
s_{y_i1}	$s_{x_i1}s_{y_i1}$		$s_{x_i2}s_{y_i1}$		$s_{x_i3}s_{y_i1}$		$\cdots$		$s_{x_ip}s_{y_i1}$		$\mu_f s_{y_i1}$
$+$											
s_{y_i2}	$s_{x_i1}s_{y_i2}$		$s_{x_i2}s_{y_i2}$		$s_{x_i3}s_{y_i2}$		$\cdots$		$s_{x_ip}s_{y_i2}$		$\mu_f s_{y_i2}$
$+$											
s_{y_i3}	$s_{x_i1}s_{y_i3}$		$s_{x_i2}s_{y_i3}$		$s_{x_i3}s_{y_i3}$		$\cdots$		$s_{x_ip}s_{y_i3}$		$\mu_f s_{y_i3}$
$+$											
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\ddots$		$\vdots$		$\vdots$
$+$											
s_{y_in}	$s_{x_i1}s_{y_ip}$		$s_{x_i2}s_{y_ip}$		$s_{x_i3}s_{y_ip}$		$\cdots$		$s_{x_ip}s_{y_ip}$		$\mu_f s_{y_ip}$
$+$											
μ_g	$s_{x_i1}\mu_g$		$s_{x_i2}\mu_g$		$s_{x_i3}\mu_g$		$\cdots$		$s_{x_ip}\mu_g$		$\mu_f\mu_g$

Expansion of Terms: $\hat{x}_i \times \hat{y}_i$

	s_{x_i1} $+$	s_{x_i2} $+$	s_{x_i3} $+$	$\dots$ $+$	s_{x_ip} $+$	μ_f
s_{y_i1}	$s_{x_i1}s_{y_i1}$	$s_{x_i2}s_{y_i1}$	$s_{x_i3}s_{y_i1}$	$\dots$	$s_{x_ip}s_{y_i1}$	$\mu_f s_{y_i1}$
$+$						
s_{y_i2}	$s_{x_i1}s_{y_i2}$	$s_{x_i2}s_{y_i2}$	$s_{x_i3}s_{y_i2}$	$\dots$	$s_{x_ip}s_{y_i2}$	$\mu_f s_{y_i2}$
$+$						
s_{y_i3}	$s_{x_i1}s_{y_i3}$	$s_{x_i2}s_{y_i3}$	$s_{x_i3}s_{y_i3}$	$\dots$	$s_{x_ip}s_{y_i3}$	$\mu_f s_{y_i3}$
$+$						
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$+$						
s_{y_in}	$s_{x_i1}s_{y_ip}$	$s_{x_i2}s_{y_ip}$	$s_{x_i3}s_{y_ip}$	$\dots$	$s_{x_ip}s_{y_ip}$	$\mu_f s_{y_ip}$
$+$						
μ_g	$s_{x_i1}\mu_g$	$s_{x_i2}\mu_g$	$s_{x_i3}\mu_g$	$\dots$	$s_{x_ip}\mu_g$	$\mu_f\mu_g$

Expansion of Terms: $\hat{x}_i \times \hat{y}_i$

	s_{x_i1}	+	s_{x_i2}	+	s_{x_i3}	+	$\cdots$	+	s_{x_ip}	+	μ_f
s_{y_i1}	$s_{x_i1}s_{y_i1}$		$s_{x_i2}s_{y_i1}$		$s_{x_i3}s_{y_i1}$		$\cdots$		$s_{x_ip}s_{y_i1}$		$\mu_f s_{y_i1}$
+											
s_{y_i2}	$s_{x_i1}s_{y_i2}$		$s_{x_i2}s_{y_i2}$		$s_{x_i3}s_{y_i2}$		$\cdots$		$s_{x_ip}s_{y_i2}$		$\mu_f s_{y_i2}$
+											
s_{y_i3}	$s_{x_i1}s_{y_i3}$		$s_{x_i2}s_{y_i3}$		$s_{x_i3}s_{y_i3}$		$\cdots$		$s_{x_ip}s_{y_i3}$		$\mu_f s_{y_i3}$
+											
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\ddots$		$\vdots$		$\vdots$
+											
s_{y_in}	$s_{x_i1}s_{y_ip}$		$s_{x_i2}s_{y_ip}$		$s_{x_i3}s_{y_ip}$		$\cdots$		$s_{x_ip}s_{y_ip}$		$\mu_f s_{y_ip}$
+											
μ_g	$s_{x_i1}\mu_g$		$s_{x_i2}\mu_g$		$s_{x_i3}\mu_g$		$\cdots$		$s_{x_ip}\mu_g$		$\mu_f\mu_g$

Expansion of Terms: $\hat{x}_i \times \hat{y}_i$

	s_{x_i1}	+	s_{x_i2}	+	s_{x_i3}	+	$\cdots$	+	s_{x_ip}	+	μ_f
s_{y_i1}	$s_{x_i1}s_{y_i1}$		$s_{x_i2}s_{y_i1}$		$s_{x_i3}s_{y_i1}$		$\cdots$		$s_{x_ip}s_{y_i1}$		$\mu_f s_{y_i1}$
+											
s_{y_i2}	$s_{x_i1}s_{y_i2}$		$s_{x_i2}s_{y_i2}$		$s_{x_i3}s_{y_i2}$		$\cdots$		$s_{x_ip}s_{y_i2}$		$\mu_f s_{y_i2}$
+											
s_{y_i3}	$s_{x_i1}s_{y_i3}$		$s_{x_i2}s_{y_i3}$		$s_{x_i3}s_{y_i3}$		$\cdots$		$s_{x_ip}s_{y_i3}$		$\mu_f s_{y_i3}$
+											
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\ddots$		$\vdots$		$\vdots$
+											
s_{y_in}	$s_{x_i1}s_{y_ip}$		$s_{x_i2}s_{y_ip}$		$s_{x_i3}s_{y_ip}$		$\cdots$		$s_{x_ip}s_{y_ip}$		$\mu_f s_{y_ip}$
+											
μ_g	$s_{x_i1}\mu_g$		$s_{x_i2}\mu_g$		$s_{x_i3}\mu_g$		$\cdots$		$s_{x_ip}\mu_g$		$\mu_f\mu_g$

Expansion of Terms: $\hat{x}_i \times \hat{y}_i$

	s_{x_i1}	+	s_{x_i2}	+	s_{x_i3}	+	$\dots$	+	s_{x_ip}	+	μ_f
s_{y_i1}	$s_{x_i1}s_{y_i1}$		$s_{x_i2}s_{y_i1}$		$s_{x_i3}s_{y_i1}$		$\dots$		$s_{x_ip}s_{y_i1}$		$\mu_f s_{y_i1}$
+											
s_{y_i2}	$s_{x_i1}s_{y_i2}$		$s_{x_i2}s_{y_i2}$		$s_{x_i3}s_{y_i2}$		$\dots$		$s_{x_ip}s_{y_i2}$		$\mu_f s_{y_i2}$
+											
s_{y_i3}	$s_{x_i1}s_{y_i3}$		$s_{x_i2}s_{y_i3}$		$s_{x_i3}s_{y_i3}$		$\dots$		$s_{x_ip}s_{y_i3}$		$\mu_f s_{y_i3}$
+											
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\ddots$		$\vdots$		$\vdots$
+											
s_{y_in}	$s_{x_i1}s_{y_ip}$		$s_{x_i2}s_{y_ip}$		$s_{x_i3}s_{y_ip}$		$\dots$		$s_{x_ip}s_{y_ip}$		$\mu_f s_{y_ip}$
+											
μ_g	$s_{x_i1}\mu_g$		$s_{x_i2}\mu_g$		$s_{x_i3}\mu_g$		$\dots$		$s_{x_ip}\mu_g$		$\mu_f\mu_g$

Proposed Approach

Modified contributions from all variables

The mean prediction of the two-part model

$$\hat{z}_i = \left(\sum_{j=1}^{p} s'_{z_i j} \right) + \alpha + \mu_h$$

The output of the two-part model $(\hat{x}_i \cdot \hat{y}_i)$

The difference between $\mu_f \mu_g$ and μ_h

where

$$s'_{z_i j} = \mu_f s_{y_i j} + s_{x_i j} \mu_g + \frac{1}{2} \sum_{a=1}^{p} (s_{x_i j} s_{y_i a} + s_{y_i j} s_{x_i a})$$

Distributing α

Uniformly Distributed:

$$s_{z_ij} = s'_{z_ij} + \frac{\alpha}{p}.$$

Raw Weights:

$$s_{z_ij} = s'_{z_ij} + \frac{s'_{z_ij}}{\hat{z}_i - \mu_f \mu_g}(\alpha).$$

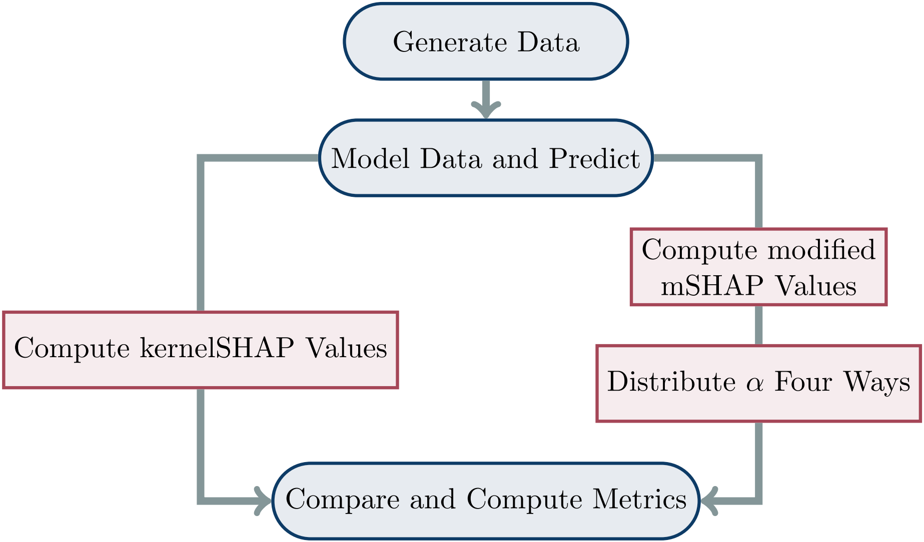
Absolute Weights:

$$s_{z_ij} = s'_{z_ij} + \frac{|s'_{z_ij}|}{\sum_{k=1}^{p} |s'_{z_ik}|}(\alpha).$$

Squared Weights:

$$s_{z_ij} = s'_{z_ij} + \frac{(s'_{z_ij})^2}{\sum_{k=1}^{p} (s'_{z_ik})^2}(\alpha).$$

Simulation Study



Simulation Results

Method	Score	Pct Same Sign	Pct Same Rank
Weighted by Absolute Value	**2.27**	**84.8%**	**62.5%**
Weighted by Squared Value	2.21	81.8%	60.8%
Uniformly Distributed	2.20	83.7%	59.4%
Weighted by Raw Value	1.99	71.4%	56.2%

Final mSHAP Equation

Thus, the final equation for the mSHAP value of the jth predictor on the ith observation can be written as

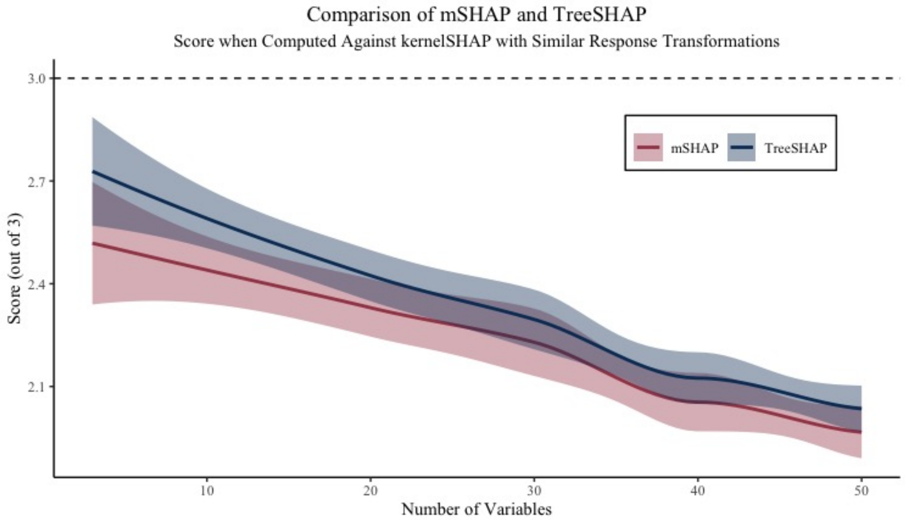
$$s_{z_ij} = \mu_f s_{y_ij} + s_{x_ij}\mu_g + \frac{1}{2}\left[\sum_{a=1}^{p}(s_{x_ij}s_{y_ia} + s_{y_ij}s_{x_ia})\right] + \frac{|s'_{z_ij}|}{\sum_{k=1}^{p}|s'_{z_ik}|}(\alpha).$$

And the overall prediction is

$$\hat{z}_i = \mu_h + \sum_{j=1}^{p} s_{z_ij}$$

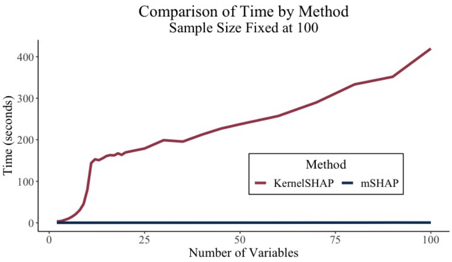
This is implemented in the R package {mshap}, which is available on CRAN and on github at www.github.com/srmatth/mshap

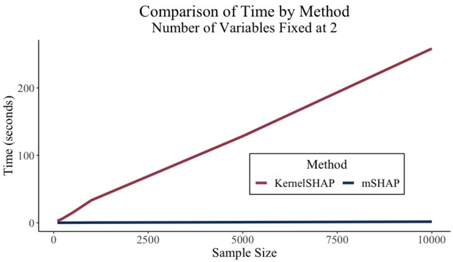
Comparison to kernelSHAP



Comparison to kernelSHAP

A dramatic increase in speed and computational efficiency:





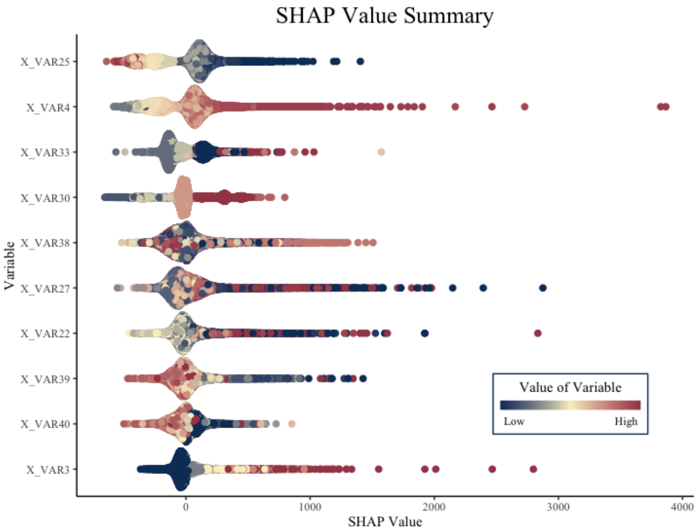
Practical Example:

- 5,000,000 Rows with 45 Covariates
- 131 Days vs. 3 Hours

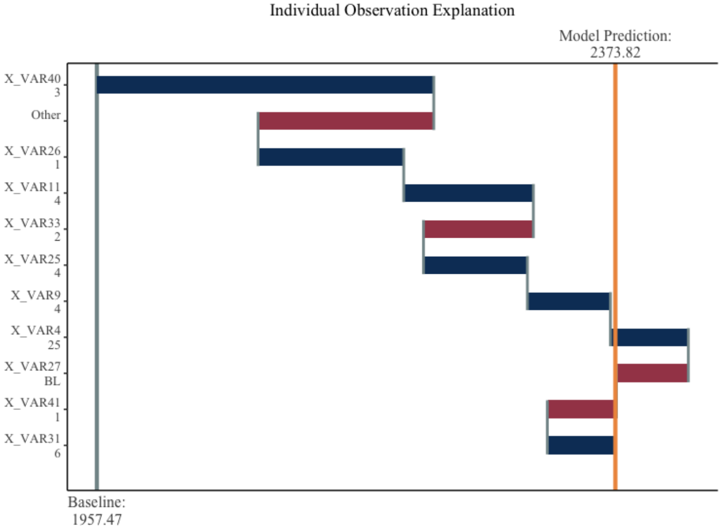
Application

- Obtained a property damage insurance data set which we then cleaned and split it into train, validation, and test sets (R)
- Trained a two-part model where both parts were random forests and the ultimate response of the model was the expected cost of a policy (Python)
- Computed the SHAP values for each individual model part using TreeSHAP (Python)
- Computed and visualized the contributions to the expected cost of a policy using mSHAP (R)

Application



Application



Conclusion

- kernelSHAP is unable to feasibly explain model predictions for two-part models
- mSHAP provides a framework for obtaining model explanations for two-part models, using the SHAP values of the individual model parts
- mSHAP will allow two-part models made up of tree-based models to be used in regulated industries such as insurance

Acknowledgments

This work was supported by:

- The Casualty Actuarial Society (CAS) Individual Grant
- The Statistics Department computing cluster at Brigham Young University

The paper is available on arxiv.org and srmatth.github.io
The code is available at www.github.com/srmatth/mshap

All plots were created with the {mshap} R package, which is available on CRAN and at www.github.com/srmatth/mshap

mSHAP
Explaining Two-Part Models

Spencer Matthews

Joint work with Brian Hartman

June 2021