



# Delta Boosting Machine and its application in Actuarial Modeling

**Simon CK Lee, Sheldon XS Lin**  
**KU Leuven, University of Toronto**

*This presentation has been prepared for the Actuaries Institute 2015 ASTIN and AFIR/ERM Colloquium.  
The Institute Council wishes it to be understood that opinions put forward herein are not necessarily those of the Institute and  
the Council is not responsible for those opinions.*

# Data Modeling

---

- **Data generating process in ratemaking models**

$$x \rightarrow \boxed{\text{nature}} \rightarrow y$$

- x: driver, vehicle and policy characteristics.
- y: claim frequency, claim severity, loss cost, etc.

- **The data modeling culture**

$$x \rightarrow \boxed{\text{Poisson, Gamma, Tweedie}} \rightarrow y$$

- **The algorithmic modeling culture**

$$x \rightarrow \boxed{\text{unknown}} \rightarrow y$$

- **Objectives of statistical modeling**

- Accurate Prediction
- Extract useful information

# Boosting methods

## GBM

---

- **In particular, Gradient Boosting Trees provide ...**
  - Accuracy comparable to Neural Networks, SVMs and Random Forests
  - Interpretable results
  - “Little” data pre-processing
  - Detects and identifies important interactions
  - Built-in feature selection
  - Results invariant under order preserving transformations of variables
    - No need to ever consider functional form revision (log, sqrt, power)
  - Applicable to a variety of response distributions
  - Not too much parameter tuning
- **Boosting idea**
  - Based on “strength of weak learnability” principles
  - Combination of weak learners → increased accuracy

# Boosting $\supseteq$ Additive Model $\supseteq$ Linear Model

---

**Linear Model** :  $E(y|x) = f(x) = \sum_{j=1}^p \beta_j X_j$

**Additive Model** :  $E(y|x) = f(x) = \sum_{j=1}^p f_j(x_j)$

**Boosting** :  $E(y|x) = f(x) = \sum_{t=1}^T \beta_t h(x; a_t)$

where the functions  $h(x; a_t)$  represent the weak learner, characterized by a set of parameters  $a = \{a_1, a_2, \dots\}$ .

Parameter estimation in Boosting amounts to solving

$$\min_{\{\beta_t, a_t\}_1^T} \sum_{i=1}^M L(y_i, \sum_{t=1}^T \beta_t h(x_i; a_t))$$

# Gradient boosting in detail

## Algorithm 1 Gradient Boosting

1. Initialize  $f_0(x)$  to be a constant,  $f_0(x) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^M L(y_i, \beta)$

2. for  $t = 1$  to  $T$  do

3. Compute the negative gradient as the working response

$$r_i = - \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x) = f_{t-1}(x)}, i = \{1, \dots, M\}$$

4. Fit a regression tree to  $r_i$  by least-squares using the input  $x_i$  and get the estimate  $a_t$  of  $\beta h(x; a)$

5. Get the estimate  $\beta_t$  by minimizing  $L(y_i, f_{t-1}(x_i) + \beta h(x_i; a_t))$

6. Update  $f_t(x_i) = f_{t-1}(x) + \beta_t h(x; a_t)$

7. end for

8. Output  $\hat{f}(x) = f_T(x)$

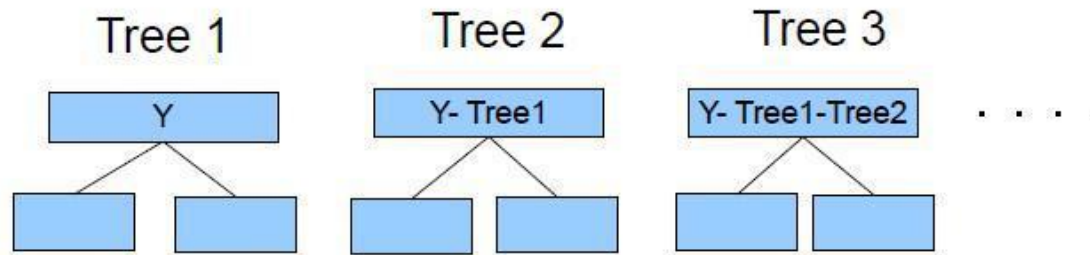
# Gradient boosting in detail

- For squared-error loss, the gradient of L is just the usual residuals

$$L = (y_i - f(x_i))^2$$

$$\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} = 2(y_i - f(x_i)) = r_i$$

- In this case, the gradient boosting algorithm simply becomes



$$\hat{f}(x) = Tree_1(x) + Tree_2(x) + \dots + Tree_T(x)$$

# Strength and Weakness of GBM

Advantages	Disadvantages
Very intuitive	Not very fast
Predictive	Deficient in dataset with many zeros when using exponential form
Output with interpretation	Some distributions are not easily available – e.g. Tweedie distribution
Robust	

# Improvement over GBM

## Delta Boosting

---

- Delta Boosting Machine (DBM) is invented to
  - Inherit the advantages of GBM ...
  - Remove the major weakness
- To describe DBM:
  - It is a modified version of GBM
  - It is faster as it requires fewer calculations at each iteration
  - It is more predictive as it derives a more accurate adjustment at each iteration
  - The algorithm is more robust with data having many zeros
  - Tweedie distribution is incorporated

# Delta Boosting in detail

---

- Recall the mechanism of GBM:
  - It calculates the gradient for each observation
  - Partitions the dataset that max out the difference in the group average of gradient
  - Obtains the group Loss function minimizer
  - Applies shrinkage factor
- DBM instead
  - It calculates the adjustment required to minimized the loss for each observation
  - Partitions the dataset that minimizes the total losses
  - ~~Obtains the group Loss function minimizer~~
  - Applies shrinkage factor

# Delta boosting in detail

## Algorithm 2 Delta Boosting for Tweedie Distribution

1. the Loss function to be negative of log likelihood of Tweedie distribution with exponential form:  $L(y, f(x)) = \sum \frac{y_i e^{(1-p)f(x_i)}}{1-p} - \frac{e^{(2-p)f(x_i)}}{2-p}$
2. Calculate the Group loss minimizer,  $h_i = \ln\left(\frac{\sum y_i e^{(1-p)f(x_i)}}{\sum e^{(2-p)f(x_i)}}\right)$
3. Linear Approximation through Taylor's expansion,  $h = \frac{\sum y_i e^{(1-p)f(x_i)}}{n} - \frac{\sum e^{(2-p)f(x_i)}}{n}$
4. Pseudo loss minimizer  $h_i = y_i e^{(1-p)f(x_i)} - \sum e^{(2-p)f(x_i)}$
5. Initialize  $f_0(x)$  to be a constant,  $f_0(x) = \ln(\sum y_i)$
6. for  $t = 1$  to  $T$  do
7. Compute the pseudo loss function minimizer,  $h_i$
8. Fit a regression tree to fit  $h_i$  by least-squares using the input  $x_i$  and get the estimate  $a_t$
9. Update  $f_t(x) = f_{t-1}(x) + h_i$
10. end for
11. Output  $\hat{f}(x) = f_T(x)$

# Delta boosting in detail

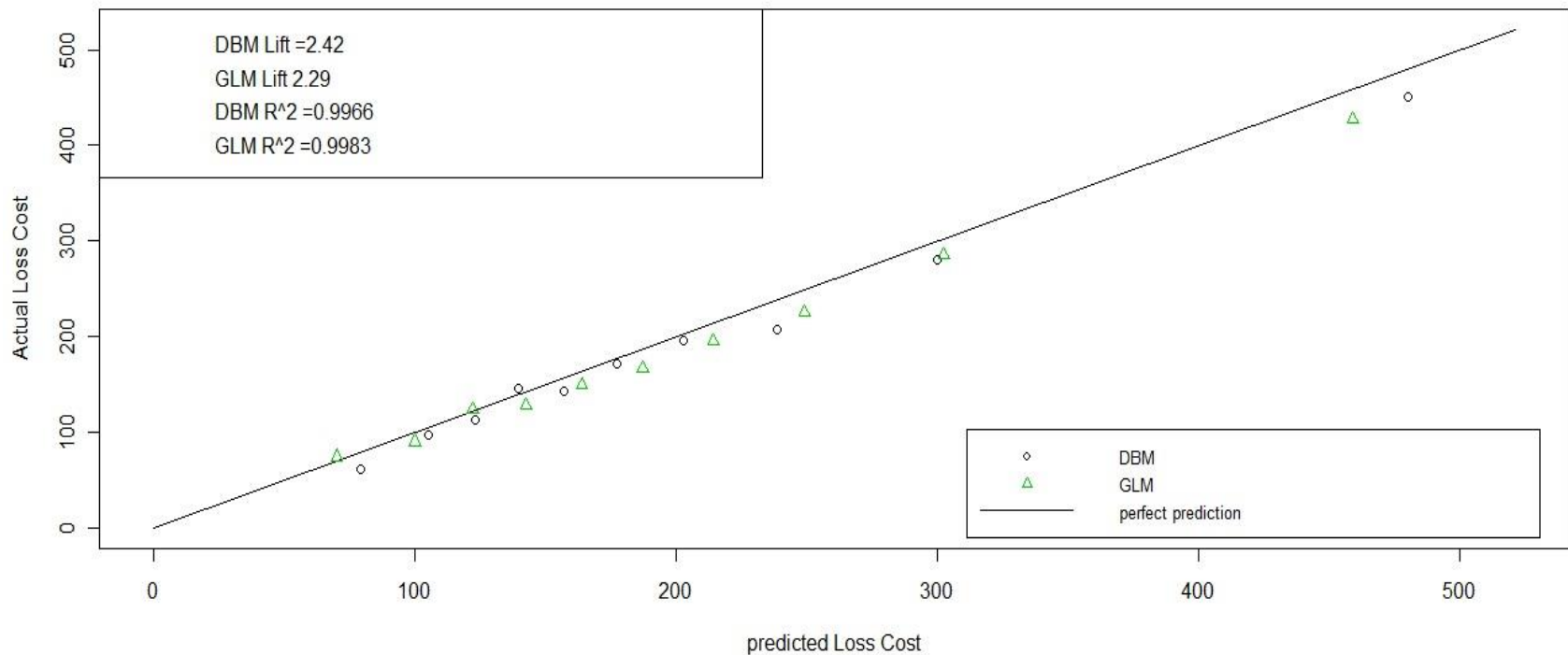
## The predictive power: Retention modeling

- The performance of various models are tested using same data and input variables.
- The model predicts the probability of churn (or renew). For predictive models, we have 40/30/30 for training / validation/ testing.

Model	Lift (top decile churn/ average churn)	ROC Area
Decision Tree	2.6692	0.6981
GLM – Logistic	3.0332	0.7275
Support Vector Machines	3.0520	0.7312
Neural Net	3.0828	0.7293
GBM – Poisson	3.0879	0.7304
GBM – Logistic	3.1016	0.7330
<b>DBM – Poisson</b>	<b>3.1306</b>	<b>0.7330</b>

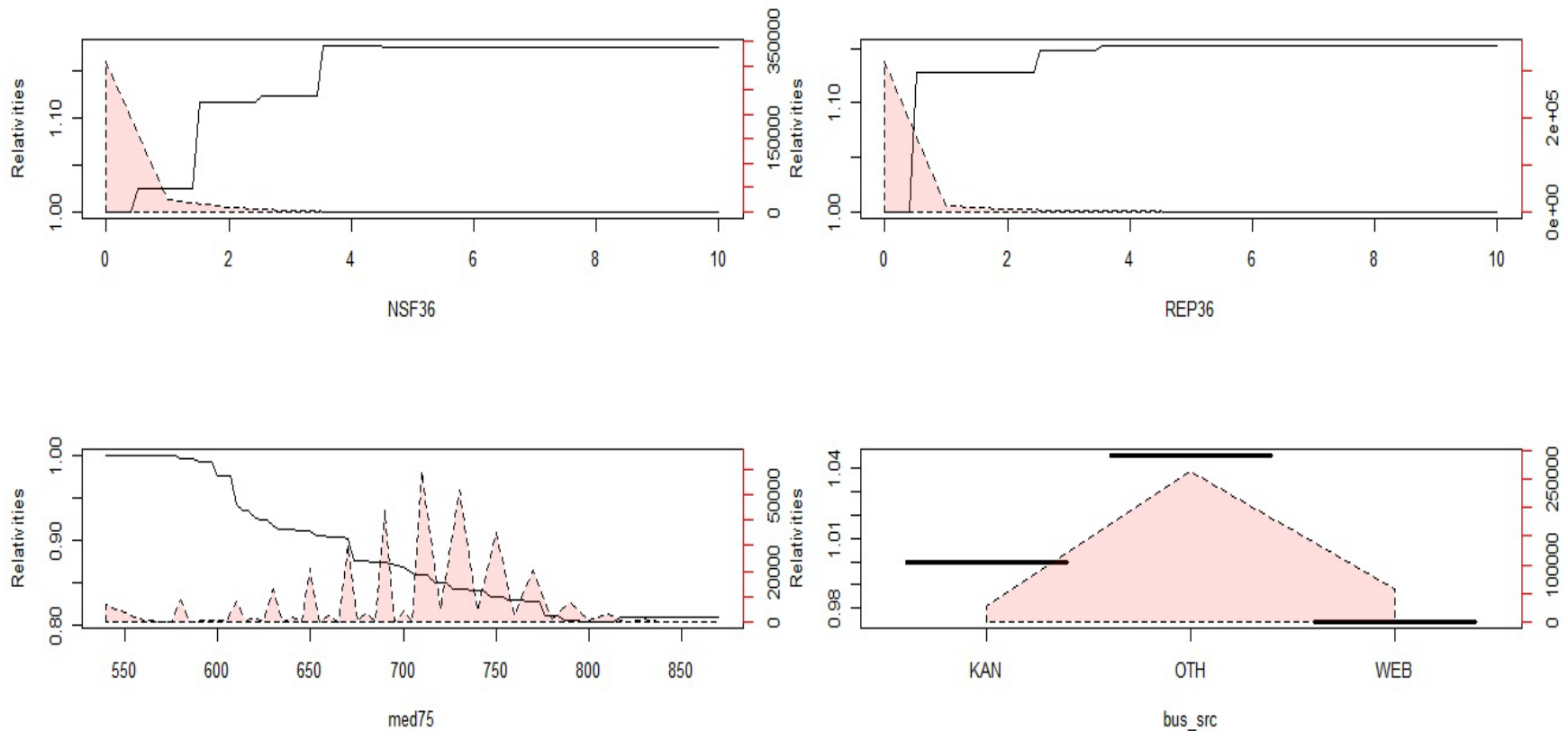
# Delta Boosting vs Gradient Boosting

Performance on Testing Data



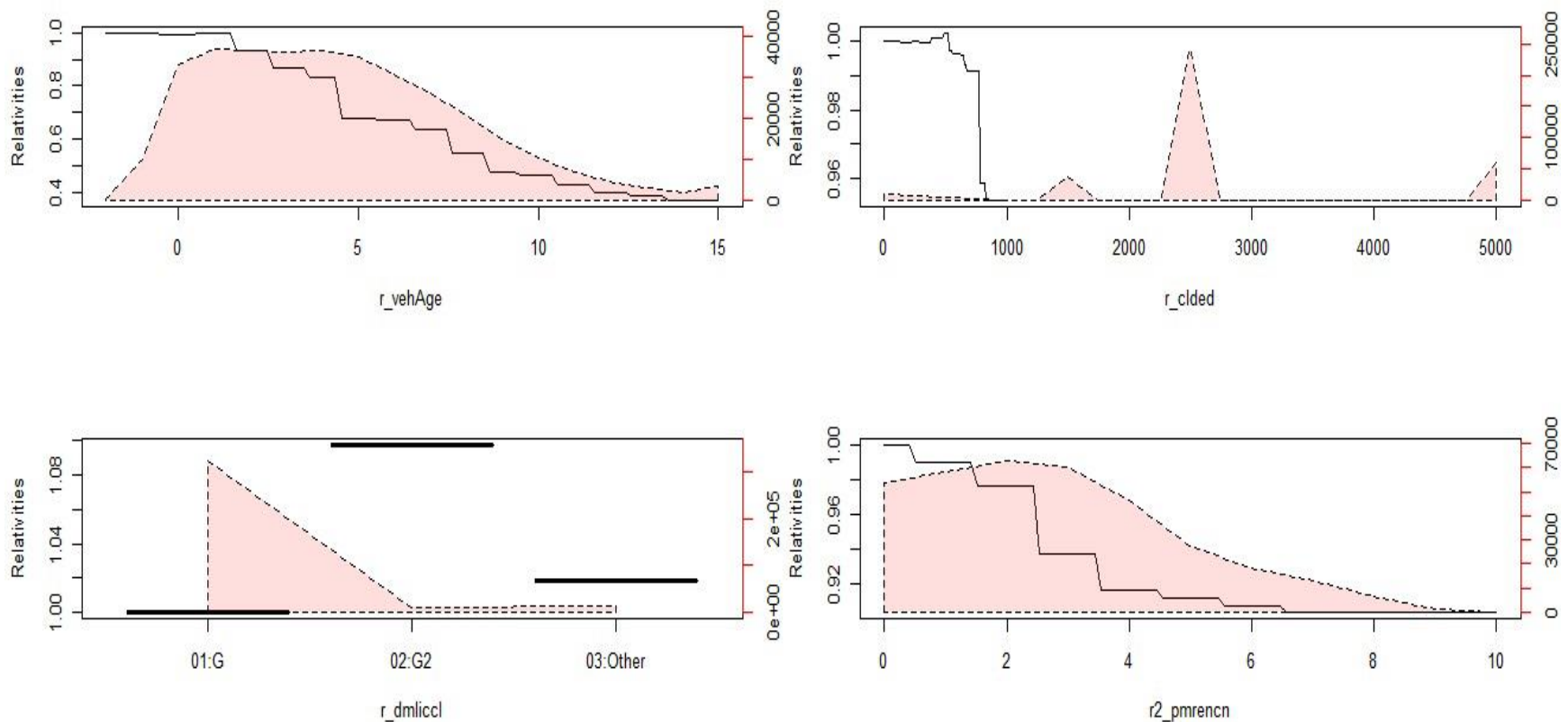
# Delta Boosting Relativities at a Glance

## Relativities for variables



# Delta Boosting Relativities at a Glance

## Relativities for variables



# Delta Boosting in detail

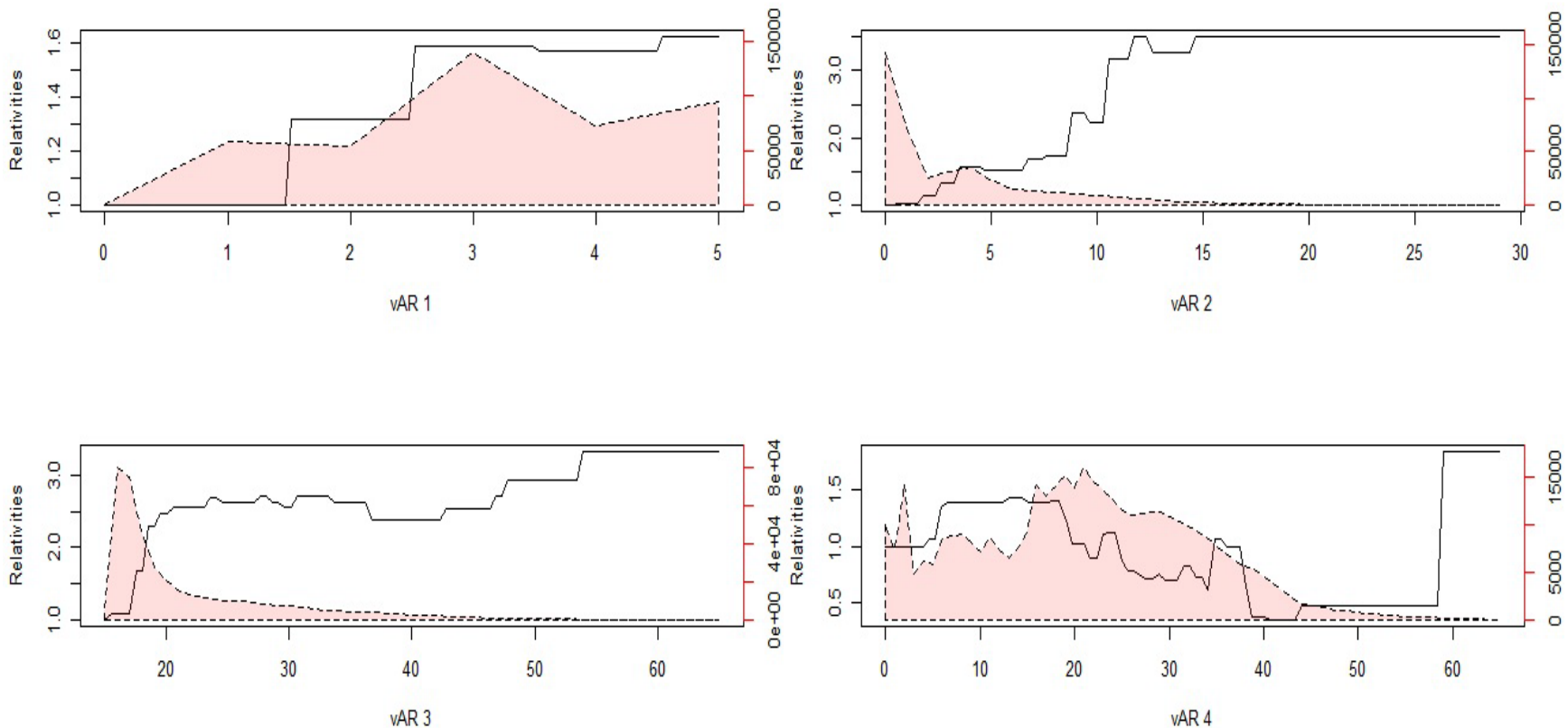
## Additional features

---

- With the above form, DBM is already more predictive than any other predictive models in all 6 of the datasets that we have tried. However, there are some more additional features that help make the model predictive
- Monotonic constraint
  - In many occasions, some of the patterns are desirable. E.g., loss cost decreasing with years licensed
  - This additional feature tells the machine not to split the data in case of reversal.
  - The improvement is promising

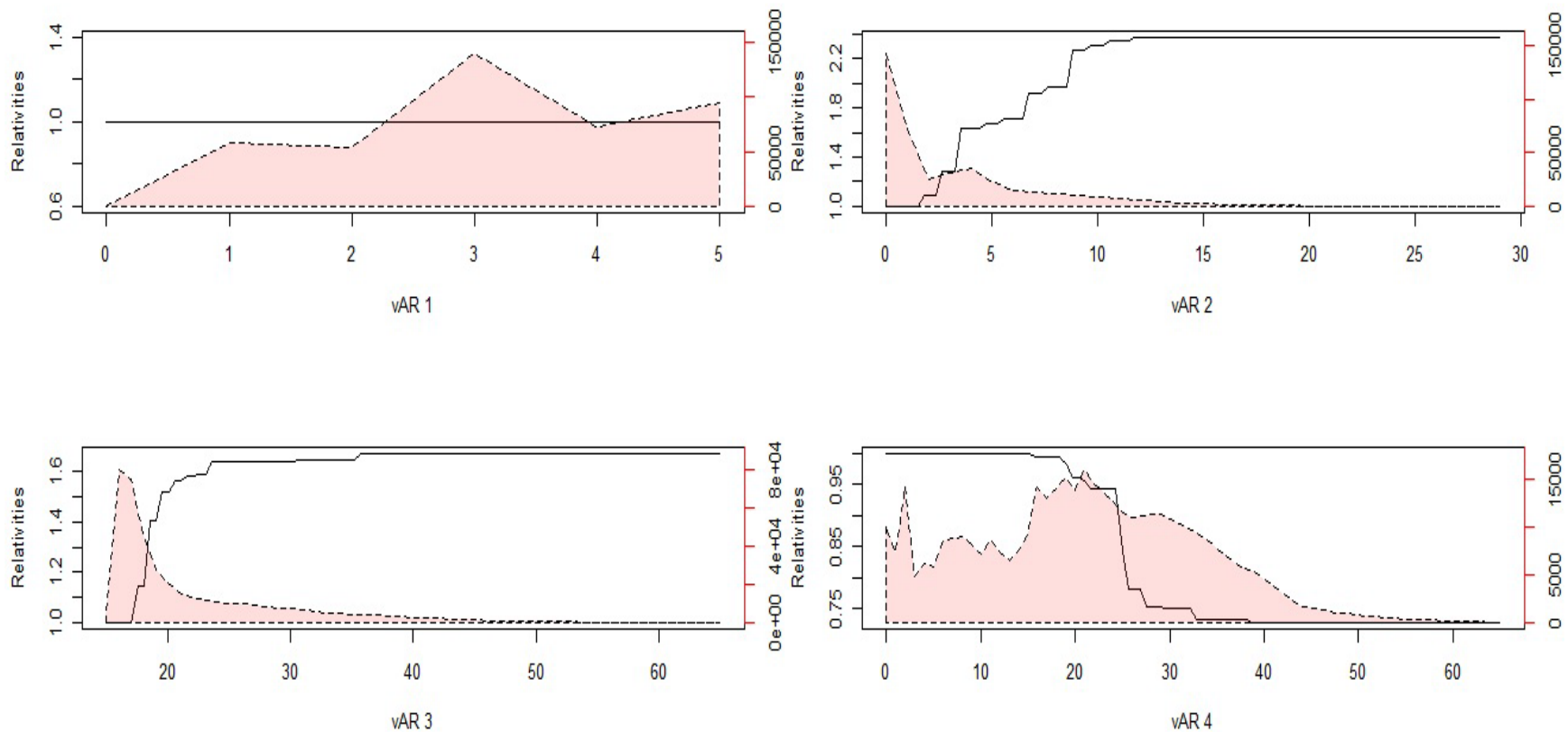
# Monotonic Constraint

## AB: Relativities for variables



# Monotonic Constraint

## AB: Relativities for variables



# Delta Boosting in detail

## Additional features

---

- Interaction constraint
  - The well promoted advantage of data mining techniques is to model any interaction to any degree
  - However, it can be a double-edged sword. It is most often that the interactions are generated from noise.
  - We are working towards the flexibility to allow users to select meaning interaction
  - An example is the model only fit 4 groups of interaction,
    - Group 1 – vehicle related
    - Group 2 – driver's related
    - Group 3 – location related
    - Group 4 – user's specified

# Your questions...

---

Questions ?