GLM Expression of GBM predictive models

Unveiling the black box of a machine learning methodology

© Hiroki Shibata

This presentation has been prepared for the 33rd International Congress of Actuaries.
The Institute of Actuaries of Japan (IAJ) and the International Actuarial Association (IAA) assume no responsibility for the opinions or content expressed in these presentation slides. Furthermore, the opinions and content contained herein do not necessarily reflect the views of the IAJ or the IAA.

Section 00:

BREAK THE ICE

Raise your hand if:

- You are familiar with

Mathematics.

Raise your hand if:

- You are familiar with

Generalized Linear Models.

Raise your hand if:

- You are familiar with

Machine Learning.

Raise your hand if:

- You are familiar with

Deep Learning and/or Generative AI (as theory).

Not familiar with ML because:

- I do not know what it does.
- Prediction is black-box.
- Not accepted by regulators etc. etc.

Purpose of this session:

- Remove fear against ML (GBM) by showing similarity between GBM/GLM under some assumptions.
- Evoke interest in machine learning as an introduction for AI-enabled actuaries.

Predictive model notation

$$(\mathrm{E}[y] =)\ \mu = f(x_1, \ldots, x_P)$$

- y: Target variable
- x_p: Explanatory variables $(p = 1, \ldots, P)$
- P: Number of Explanatory variables

Section 01:

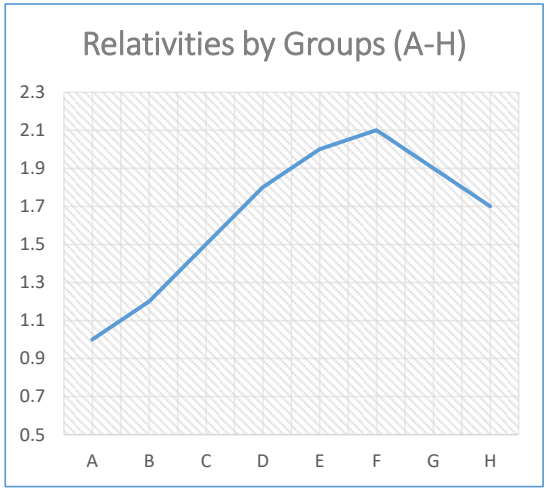
ASSUMPTIONS

Assumption (1)

- All the explanatory variables are:
  - Categorical variables or
  - Ordered categorical variables.
- Numerical variables are binned into finite number of categories.

Assumption (1)

- Ordered: Adjacent levels has similar relativities (continuity).
- Some categorical levels can be ordered by nature.



Assumption (1)

- Ex1: Mileage (binned)

Level Index	Levels
1	<1,000km
2	<3,000km
3	<7,000km
4	<10,000km
5	≥10,000km

- Ex2: Feedback (ordered)

Level Index	Levels
1	Totally Agree
2	Agree
3	Neutral
4	Disagree
5	Totally disagree

Assumption (2)

- All the explanatory variables have their base levels (reference levels).
- Base levels will be used for the normalization of relativities in GLM prediction.

Assumption (2)

- Ex1: Mileage

Index	Levels	Relativity
1	<1,000km	0.75
2	<3,000km	0.80
3	<7,000km	0.90
4	<10,000km	1.00
5	≥10,000km	1.50

- Ex2: Feedback

Index	Levels	Relativity
1	Totally Agree	1.10
2	Agree	1.05
3	Neutral	1.00
4	Disagree	0.85
5	Totally disagree	0.70

Assumption (3)

- At first, we treat predictive models without interactions for better understandability.

N.B. Refer to section 06 for models with interactions.

Section 02:

RECAP GLM

Prediction with GLM

$$(\mathrm{E}[y] =)\ \mu = g^{-1}(\eta(x_1, \ldots, x_P))$$

- g: link function (identical/log/logit…)
- η: linear predictor (LP)

Prediction with GLM

- Expression with simple factors only:

$$\eta = \eta_1(x_1) + \cdots + \eta_P(x_P)$$

- $\eta_p(x_p) = \beta_p(1)I_p(1) + \cdots + \beta_p(l_p)I_p(l_p)$
- $\beta_p(l)$: coefficient for var x_p taking l-th level
- $I_p(l)$: takes 1 when $x_p = x_{p,l}$ otherwise 0
- l_p: number of levels with x_p

Prediction with GLM

- Normalization with base levels:

$$\eta = \beta_B + \eta_1(x_1) + \cdots + \eta_P(x_P)$$

- β_B: predicted LP at base levels
- $\beta_p(b_p) = 0$ for all x_p
- b_p: index for the base level of x_p

Example: TPPD Freq in Auto

- Log link

$$\beta_B = -3$$

Age	l	$\beta_1(l)$
<20	1	1.0
<30	2	0.5
<40	3	0.3
<50	4	0.1
≥50	5	0

Mileage	l	$\beta_2(l)$
<3,000km	1	-0.7
<5,000km	2	-0.4
<10,000km	3	0
<15,000km	4	0.5
≥15,000km	5	0.8

Rating Area	l	$\beta_3(l)$
A	1	0.3
B	2	0
C	3	0.2

Example: TPPD Freq in Auto

- Actual value also available with Log link

$$e^{\beta_B} = 0.050$$

Age	l	$e^{\beta_1(l)}$
<20	1	2.718
<30	2	1.649
<40	3	1.350
<50	4	1.105
≥50	5	1.000

Mileage	l	$e^{\beta_2(l)}$
<3,000km	1	0.497
<5,000km	2	0.670
<10,000km	3	1.000
<15,000km	4	1.649
≥15,000km	5	2.226

Rating Area	l	$e^{\beta_3(l)}$
A	1	1.350
B	2	1.000
C	3	1.221

Def: GLM Expression

- GLM Expression:

$$\eta = \sum_{p=1}^{P} \sum_{l=1}^{l_p} \beta_p(l) \cdot I_p(l)$$

- Normalized GLM Expression:

$$\eta = \beta_B + \sum_{p=1}^{P} \sum_{l=1}^{l_p} \beta_p(l) \cdot I_p(l) \ \left(\beta_p(b_p) = 0\right)$$

Proposition 1: Normalization

- If a model has a GLM Expression:

$$\eta = \sum_{p=1}^{P} \sum_{l=1}^{l_p} \beta_p(l) \cdot I_p(l)$$

- Then it has unique Normalized GLM Expression:

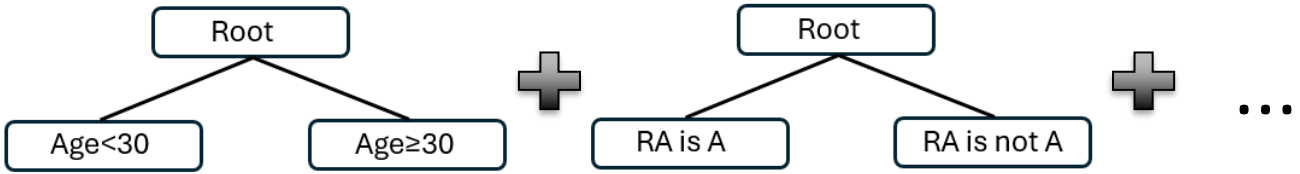
$$\eta = \beta_B + \sum_{p=1}^{P} \sum_{l=1}^{l_p} \beta_p(l) \cdot I_p(l) \ \left(\beta_p(b_p) = 0\right)$$

Section 03:

INTRO GBM

GBM (GBDT)

- Gradient Boosting Machine or
- Gradient Boosting Decision Tree



A Decision Tree

- Separate data into (two) groups A/B, using one of explanatory variables.
- Try to optimize the predicted value for each group (scalar/fixed).
  ex: minimize square error $\sum(\hat{y} - y)^2$ for groups A and B, respectively

A Decision Tree

- Total square error (after optimization)

$$\sum_A (\hat{y}_A - y)^2 + \sum_B (\hat{y}_B - y)^2$$

  depends on the separation.
- A decision tree tries to find the BEST separation with minimum total error.

Possible separation?

- Ex: Age and Rating Area

Age	l
<20	1
<30	2
<40	3
<50	4
≥50	5

(Ordered)

Rating Area	l
A	1
B	2
C	3

(Categorical)

Possible separation?

Variable	Group A	Group B	Meaning
Age	$l < 2$	$l \geq 2$	Age<20 or not
Age	$l < 3$	$l \geq 3$	Age<30 or not
Age	$l < 4$	$l \geq 4$	Age<40 or not
Age	$l < 5$	$l \geq 5$	Age<50 or not
Rating Area	$l = 1$	$l \neq 1$	RA is A or not
Rating Area	$l = 2$	$l \neq 2$	RA is B or not
Rating Area	$l = 3$	$l \neq 3$	RA is C or not

GBM Process

- Create & Update predictive model

$$\mu = F_t(x_1, \ldots, x_P)$$

by decision trees $t = 0, 1, \ldots, T$

GBM Process

1) Initial estimate is start value

$$F_0(x_1, \ldots, x_P) = \beta_0$$

to minimize the total square error for all the data.

$$\beta_0 = \text{argmin}_{\hat{y}} \sum (\hat{y} - y)^2$$

GBM Process

2) For 1st decision tree, select possible separation A/B and find best value update for each group to minimize total square error within group A/B:

$$\sum_A \{(F_0 + u_A) - y\}^2 \qquad \sum_B \{(F_0 + u_B) - y\}^2$$

GBM Process

3) The total square error

$$\sum_A \{(F_0 + u_A) - y\}^2 + \sum_B \{(F_0 + u_B) - y\}^2$$

depends on the separation: find the BEST separation to minimize it to determine F_1.

GBM Process

4) Iterate the process: find the best separation to minimize the total square error

$$\sum_A \{(F_t + u_A) - y\}^2 + \sum_B \{(F_t + u_B) - y\}^2$$

to obtain F_{t+1}.

GBM Example

- Target: TPPD Sev in Auto
- No of trees: 2
- Explanatory variables:

Age	l
<20	1
<30	2
<40	3
<50	4
≥50	5

(Ordered)

Rating Area	l
A	1
B	2
C	3

(Categorical)

GBM Example

- Start value: $F_0 = 7000$.
- Separation with 1st tree: age < 30 (gr A) or not (gr B)

$$u_A = 3000, \qquad u_B = -1000$$

- Separation with 2nd tree: RA = C (gr A) or not (gr B)

$$u_A = -500, \qquad u_B = 600$$

- Predicted value for age=35 & rating area=B is:

$$7000 - 1000 + 600 = 6600.$$

GBM 'Generalization'

- Introduction of link function and LP:

$$\mu = g^{-1}(\eta_t(x_1, \dots, x_P)).$$

- Generalization of loss function: negative log likelihood of Poisson/Gamma etc.
- GBM: update η_t by trees gradually to keep reducing total loss function and move closer to the best predictive model.

Section 04:

GLM EXPRESSION OF GBM

How to express GBM

- Number of trees can be very large: 10K~100K or more, making GBM recognized as black-box.
- What if we can convert trees into table structure like GLM models?

How to express GBM

- Example: age and rating area.
- Parametrization (of LP) for GLM Expression:

Age	l	$\beta_1(l)$
<20	1	$\beta_1(1)$
<30	2	$\beta_1(2)$
<40	3	$\beta_1(3)$
<50	4	$\beta_1(4)$
≥50	5	$\beta_1(5)$

Rating Area	l	$\beta_2(l)$
A	1	$\beta_2(1)$
B	2	$\beta_2(2)$
C	3	$\beta_2(3)$

How to express GBM

- Initiate parameters with zero values.

Tree	Description	$\beta_1(1)$	$\beta_1(2)$	$\beta_1(3)$	$\beta_1(4)$	$\beta_1(5)$	$\beta_2(1)$	$\beta_2(2)$	$\beta_2(3)$
-	Initiation	0	0	0	0	0	0	0	0

How to express GBM

- As 0th tree, apply 3 as start value:

Tree	Description	$\beta_1(1)$	$\beta_1(2)$	$\beta_1(3)$	$\beta_1(4)$	$\beta_1(5)$	$\beta_2(1)$	$\beta_2(2)$	$\beta_2(3)$
-	Initiation	0	0	0	0	0	0	0	0
0	Start value = 3	3	3	3	3	3	0	0	0

N.B. select one variable and apply the same value for all the levels

How to express GBM

- For 1st tree, select separation with age<30 or not

Tree	Description	$\beta_1(1)$	$\beta_1(2)$	$\beta_1(3)$	$\beta_1(4)$	$\beta_1(5)$	$\beta_2(1)$	$\beta_2(2)$	$\beta_2(3)$
-	Initiation	0	0	0	0	0	0	0	0
0	Start value = 3	3	3	3	3	3	0	0	0
1	If age<30 then +2	5	5	3	3	3	0	0	0
1	If age>=30 then -1	5	5	2	2	2	0	0	0

How to express GBM

- For 2nd tree, select separation with RA=B or not

Tree	Description	$\beta_1(1)$	$\beta_1(2)$	$\beta_1(3)$	$\beta_1(4)$	$\beta_1(5)$	$\beta_2(1)$	$\beta_2(2)$	$\beta_2(3)$
-	Initiation	0	0	0	0	0	0	0	0
0	Start value = 3	3	3	3	3	3	0	0	0
1	If age<30 then +2	5	5	3	3	3	0	0	0
1	If age>=30 then -1	5	5	2	2	2	0	0	0
2	If RA=B then -2	5	5	2	2	2	0	-2	0
2	If RA is not B then +1	5	5	2	2	2	1	-2	1

How to express GBM

- For 3rd tree, select separation with age<20 or not

Tree	Description	$\beta_1(1)$	$\beta_1(2)$	$\beta_1(3)$	$\beta_1(4)$	$\beta_1(5)$	$\beta_2(1)$	$\beta_2(2)$	$\beta_2(3)$
-	Initiation	0	0	0	0	0	0	0	0
0	Start value = 3	3	3	3	3	3	0	0	0
1	If age<30 then +2	5	5	3	3	3	0	0	0
1	If age>=30 then -1	5	5	2	2	2	0	0	0
2	If RA=B then -2	5	5	2	2	2	0	-2	0
2	If RA is not B then +1	5	5	2	2	2	1	-2	1
3	If age<20 then +3	8	5	2	2	2	1	-2	1
3	If age>=20 then -0.5	8	4.5	1.5	1.5	1.5	1	-2	1

How to express GBM

- After 3rd tree, the GBM prediction is:

Age	l	$\beta_1(l)$
<20	1	8
<30	2	4.5
<40	3	1.5
<50	4	1.5
≥50	5	1.5

Rating Area	l	$\beta_2(l)$
A	1	1
B	2	-2
C	3	1

This is GLM Expression!

Theorem 1: GLM Expression of GBM

- Any GBM prediction can be converted into GLM Expression.
- GBM model can also be converted into Normalized GLM Expression according to Proposition 1.

Section 05:

ALGORITHM TO OBTAIN NORMALIZED GLM EXPRESSION

Normalized GLM Expression of GBM

- Adding up parameters from 100K~ trees is annoying.
- Easy algorithm to obtain parameters for the unique normalized GLM Expression is available.
- Assumption: a GBM model to predict LP is available:

$$\eta_T(x_1, \ldots, x_P)$$

Normalized GLM Expression of GBM

- Relativity at base: apply base levels.

$$\beta_B = \eta_T(b_1, \dots, b_P)$$

- Relativity for x_p: take difference.

$$\beta_p(l) = \eta_T(b_1, \dots, b_{p-1}, l, b_{p+1}, \dots, b_P) - \beta_B = \eta_p(l) - \beta_B$$

Section 06:

EXPANSION FOR INTERACTIONS

Def: Tensors of order D

- Relativity: $\beta_{\boldsymbol{p}}^{D}(\boldsymbol{l}) = \beta_{p_1,\dots,p_D}(l_1,\dots,l_D)$
- Indicator function: $I_{\boldsymbol{p}}^{D}(\boldsymbol{l}) = I_{p_1,\dots,p_D}(l_1,\dots,l_D)$

- D: Number of explanatory variables participating
- $\boldsymbol{p} = (p_1,\dots,p_D)$: Participating explanatory variables
- $\boldsymbol{l} = (l_1,\dots,l_D)$: Indexes for $\boldsymbol{p}$.

Property of normalized β

- When normalized at base levels, relativity $\beta_{\boldsymbol{p}}^{D}(\boldsymbol{l})$ can have non-zero values only when all the levels $\boldsymbol{l}$ are different from their base levels.
(relativity is zero if at least one participating explanatory variable takes its base level)

$\beta_{1,2}(l_1, l_2)$	Rating Area	A	B	C
Age	$\boldsymbol{l}$	1	2	3
<20	1	1.1	0	0.6
<30	2	0.5	0	0.3
<40	3	0.3	0	0.2
<50	4	0.2	0	0.1
≥50	5	0	0	0

Def: GLM Expression of order D

- GLM Expression of order D:

$$\eta = \sum_{d=1}^{D} \sum_{\boldsymbol{p},\boldsymbol{l}} \beta_{\boldsymbol{p}}^{d}(\boldsymbol{l}) \cdot I_{\boldsymbol{p}}^{d}(\boldsymbol{l})$$

- Normalized GLM Expression of order D:

$$\eta = \beta_B + \sum_{d=1}^{D} \sum_{\boldsymbol{p},\boldsymbol{l}} \beta_{\boldsymbol{p}}^{d}(\boldsymbol{l}) \cdot I_{\boldsymbol{p}}^{d}(\boldsymbol{l})$$

($\beta_{\boldsymbol{p}}^{d}(\boldsymbol{l}) = 0$ whenever $\boldsymbol{l}$ includes base level(s).)

Proposition 1: Normalization

- If a model has a GLM Expression of order D:

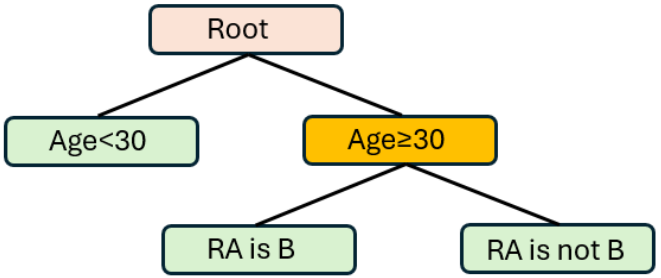
$$\eta = \sum_{d=1}^{D} \sum_{\boldsymbol{p},\boldsymbol{l}} \beta_{\boldsymbol{p}}^{d}(\boldsymbol{l}) \cdot I_{\boldsymbol{p}}^{d}(\boldsymbol{l})$$

- Then it has unique Normalized GLM Expression of order D (refer to the paper for proof):

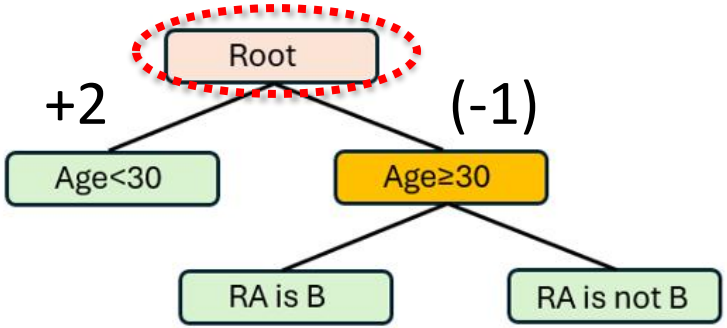
$$\eta = \beta_B + \sum_{d=1}^{D} \sum_{\boldsymbol{p},\boldsymbol{l}} \hat{\beta}_{\boldsymbol{p}}^{d}(\boldsymbol{l}) \cdot I_{\boldsymbol{p}}^{d}(\boldsymbol{l})$$

Decision tree with depth D

- A decision tree with maximum depth D can continue splitting up to D times.
- In the 2nd split of example tree, both Age and Rating Area are participating



Decision tree parametrization

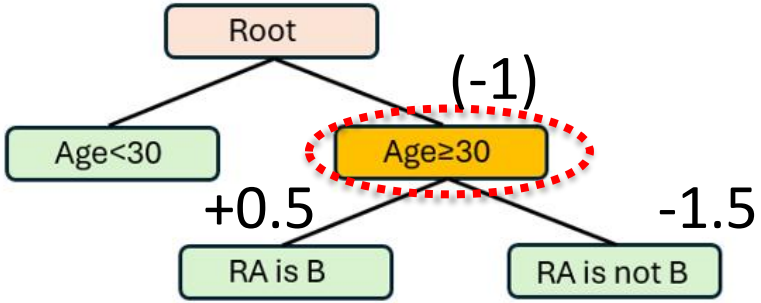


- 1st split:

Age	l	$\beta_1(l)$
<20	1	2
<30	2	2
<40	3	0
<50	4	0
≥50	5	0

N.B. parameter update for 'Age≥30' will be applied at leaf nodes

Decision tree parametrization



- 2nd split:

$\beta_{1,2}(l_1, l_2)$	Rating Area	A	B	C
Age	l	1	2	3
<20	1	0	0	0
<30	2	0	0	0
<40	3	-1.5	0.5	-1.5
<50	4	-1.5	0.5	-1.5
≥50	5	-1.5	0.5	-1.5

Theorem 1: GLM Expression of GBM

- Any GBM prediction of depth D can be converted into GLM Expression of order D, so further converted into Normalized GLM Expression of order D.
  (refer to the thesis for proof)
  N.B. the order in GLM Expression is mostly the same as the maximum depth of original GBM, and in limited cases less than the GBM depth.

Algorithm for GBM to Normalized GLM

- LP Coefficient of base: apply base levels.

$$\beta_B = \eta_T(b_1, \dots, b_P)$$

- LP Relativity at base for $\boldsymbol{p} = (p_1, \dots, p_d)$:

$$\beta_{\boldsymbol{p}}^d(\boldsymbol{l}) = \eta_{\boldsymbol{p}}^d(\boldsymbol{l}) - \sum_{\boldsymbol{s} \subsetneq \boldsymbol{p}} \beta_{\boldsymbol{s}}^d(\boldsymbol{l}')$$

- Ex: $\beta_{1,2}^2(l_1, l_2) = \eta_{1,2}^2(l_1, l_2) - \beta_1^1(l_1) - \beta_2^1(l_2) - \beta_B$
  ($\beta_{\boldsymbol{s}}^d(\boldsymbol{l}') = \beta_B$ when $\boldsymbol{s} = \phi$.)

Algorithm for GBM to Normalized GLM

- The algorithm can be used as the most powerful interpretation tool for GBM models.
- The cost of calculation is extremely low compared to PDP/ALE/ICE etc.
- LP Relativity at base for interaction terms of order d does not depend on higher orders.

Section 07:

WRAP UP

Summary & Challenge

- Showed that GBM predictive models can be expressed in normalized GLM format.
- Introduced an easy and efficient algorithm to derive relativities normalized at base levels.
- Challenge: large number of applicable terms when interactions are allowed (understandable but too complex).

Possible solutions for the complexity

- Reduce the number of candidate explanatory variables by factor penalization (GBM technique).
- Give prioritization to interaction terms:
  - Restrict interaction terms in original GBM (GBM technique)
  - Extract beta parameters only from prioritized interaction terms

Questions?

Thanks!