XGBoost in Non-Life Insurance Pricing

Berlin, 17.01.2025

Hans Alcahya Tjong
570795

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

Inhaltsverzeichnis

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

Motivation

Rating factors

Insurance premium depends on *rating factors*.

The rating factors can be distinguished in three characteristics:

1. Characteristics of the policyholder:
   Gender, age, marital status
2. Characteristics of the insured objects:
   Car manufacturer, engine size, model year
3. Characteristics of the geographical region:
   City, postal code

Key ratios

Duration n_o

Policy years

Claim number N

Claim frequency SH

Number of claims divided by the duration

Key ratios

Total claim amount S

Claim severity SD

Total claim amount divided by the number of claims (average cost per claim)

Pure premium SB

Total claim amount divided by the duration (average cost per policy year)

Key ratios

Exposure $[\omega]$	Target value $[X]$	Normalized target value $\left[Y = \frac{X}{\omega}\right]$
n_o	N	$SH = \frac{N}{n_o}$
N	S	$SD = \frac{S}{N}$
n_o	S	$SB = \frac{S}{n_o}$

Tabelle: Connection between key ratio.

Multiplicative model

Let M denote number of rating factors and i_k denote the class of the rating factor k.
The multiplicative model is defined by:

$$\mu_{i_1,i_2,\dots,i_M} = \gamma_0 \prod_{k=1}^{M} \gamma_{k,i_k}$$

where $\gamma_0 \in \mathbb{R}$ is basis premium and $\gamma_{k,i_k} \in \mathbb{R}$ are price relativities.

Example

Age	Location	i_1	i_2	μ_{i_1,i_2}
21-40	Urban	1	1	$\gamma_0\gamma_{1,1}\gamma_{2,1}$
21-40	Small town	1	2	$\gamma_0\gamma_{1,1}\gamma_{2,2}$
41-60	Urban	2	1	$\gamma_0\gamma_{1,2}\gamma_{2,1}$
41-60	Small town	2	2	$\gamma_0\gamma_{1,2}\gamma_{2,2}$
61-80	Urban	3	1	$\gamma_0\gamma_{1,3}\gamma_{2,1}$
61-80	Small town	3	2	$\gamma_0\gamma_{1,3}\gamma_{2,2}$

Tabelle: Example of a multiplicative model.

Methods

Methods

1. Generalized Linear Models (GLM)
2. Extreme Gradient Boosting (XGBoost)

Components of GLM

1. Distribution
2. Systematic component
3. Link-Function

Components of GLM

1. Distribution

Let the response variable Y_i have a distribution from the exponential family:

$$f_{Y_i}(y_i;\theta_i,\phi) = exp\left(\frac{y_i\theta_i - a(\theta_i)}{\phi/\nu_i}\right) c(y_i,\phi,\nu_i)$$

Components of GLM

1. Distribution

where:

$$
\begin{aligned}
\theta_i &= \text{canonical parameter (real value)} \\
\phi &= \text{Dispersion parameters (positive)} \\
\nu_i &= \text{Weight (positive)} \\
a(\cdot) &= \text{Cumulant function (bijective and 2 time continuously differentia} \\
c(\cdot) &= \text{Positive normalizing function}
\end{aligned}
$$

Components of GLM

2. Systematic component

The explanatory variables result in a linear predictor:

$$\eta_i = X_i^T \beta = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j$$

Components of GLM

3. Link-Function

The bijective and two time continuously differentiable link function $g(\cdot)$ links the random component and the systematic component:

$$g(\mu_i) = \eta_i \Leftrightarrow \mu_i = g^{-1}(\eta_i)$$

where $\mu_i = \mathbb{E}(Y_i|X)$

XGBoost: Example

i	Age	Location	$\hat{\mu}_i$	S_i	$Error$	$Error^2$
1	25	Urban	1000	1300	+300	90000
2	40	Small town	1000	500	-500	250000
3	32	Urban	1000	1000	0	0

XGBoost: Example

- Tree 1: If Age < 35, then $+100$. Otherwise -200
- Tree 2: If Location $=$ 'Urban', then $+150$. Otherwise -300

XGBoost: Example

i	Age	Location	$\hat{\mu}_i$	S_i	*Error*	*Error*^{2}
1	25	Urban	1250	1300	+50	2500
2	40	Small town	500	500	0	0
3	32	Urban	1150	1000	-150	22500

XGBoost

- Objective function

$$obj(\Theta) = L(\Theta) + \Omega(\Theta)$$

where:

Θ	Parameters
$L(\Theta)$	loss function
$\Omega(\Theta)$	regularization term

XGBoost

- Loss function $L(\Theta)$ measures how predictive the model is with respect to the training data, e.g. MSE:

$$L(\Theta) = \sum_i (y_i - \hat{y}_i)^2$$

- Regularization term $\Omega(\Theta)$ measures the complexity of the model

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

XGBoost

- The model can be mathematically written as

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

where:

$\hat{y}_i^{(t)}$	prediction value at step t
$f_t(x_i)$	function of t-th tree

XGBoost: Example

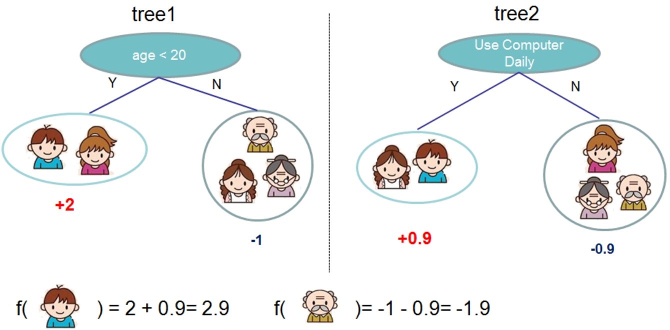

Abbildung:
https://xgboost.readthedocs.io/en/stable/tutorials/model.html

XGBoost

- Objective function of the t-th tree to be optimized using MSE as loss function

$$
\begin{aligned}
obj^{(t)} &= \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)))^2 + \sum_{k=1}^{t} \omega(f_k) \\
&= \sum_{i=1}^{n} \left[2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2 \right] + \omega(f_t) + \epsilon \\
&\stackrel{T}{=} \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \omega(f_t) + \epsilon
\end{aligned}
$$

XGBoost

where:

- $\omega(f_k)$ the complexity of the tree f_k
- $g_i = \delta_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$
- $h_i = \delta^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$

XGBoost

- The complexity of the tree f is defined by

$$\omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$$

where:

γ	minimum loss reduction parameter
T	number of leaves
λ	Ridge regularization parameter
ω_j	Score of leaf

XGBoost

- After re-formulating

$$obj^{(t)} \approx \sum_{j=1}^{T} \left[G_j \omega_j + \frac{1}{2}(H_j + \lambda)\omega_j^2 \right] + \gamma T$$

where:

$$G_j = \sum_{i \in I_j} g_i$$

$$H_j = \sum_{i \in I_j} h_i$$

I_j set of indices of data points assigned to the j-th leaf

XGBoost

- The best ω_j and objective function obj^* are

$$\omega_j^* = -\frac{G_j}{H_j + \lambda}$$

$$obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T$$

XGBoost: Example

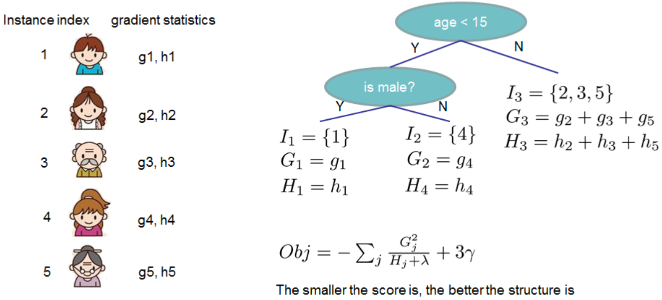

Abbildung:
https://xgboost.readthedocs.io/en/stable/tutorials/model.html

Data

Data Profile

- Available in `R` package `dataOhlsson`: `insuranceData`.
- Derived from period 1994-1998
- Partially comprehensive insurance for motorcycles.
- 64548 observations, each corresponding to one insurance policy.

Features

agarald	Age of the policyholder (0-99)
kon	Gender (K/M)
zon	Zone (1-7)
mcklass	Vehicle type (1-7)
fordald	Vehicle age (0-99)
bonuskl	Bonus class (1-7)
duration	Policy year
antskad	Claim number
skadkost	Total claim amount

Results

Claim Frequency

Model	Features
GLM0	(Intercept model)
GLM1	Alter + Geschlecht + Zone + Fahrzeugtyp + Fahrzeugalter + Bonusklasse
GLM2	Alter + Geschlecht + Zone + Fahrzeugtyp + Fahrzeugalter + Bonusklasse + Bonusklasse * Zone
XGB	Alter + Geschlecht + Zone + Fahrzeugtyp + Fahrzeugalter + Bonusklasse + Schadenaufwendungen

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

Claim Frequency

Model	Claim Frequency (Test)		Poisson Deviance		RMSE
	Actual	Predicted	Train	Test	
GLM0	1.00%	1.09%	8.75%	7.50%	0.1066
GLM1	1.00%	1.10%	8.49%	6.83%	0.1055
GLM2	1.00%	1.10%	8.38%	6.76%	0.1055
XGB	1.00%	1.01%	0.00%	0.12%	0.0253

Claim Severity

Model	Features
GLM0	(Intercept model)
GLM1	Alter + Geschlecht + Zone + Fahrzeugtyp + Fahrzeugalter + Bonusklasse
XGB	Alter + Geschlecht + Zone + Fahrzeugtyp + Fahrzeugalter + Bonusklasse + Versicherungsjahre

Claim Severity

Model	Claim Severity (Test)		RMSE
	Actual	Predicted	
GLM0	4762611.00	5116337.08	38224.70
GLM1	4762611.00	4861973.03	38418.00
XGB	4762611.00	4984814.42	42570.22

Summary

- For predicting claim frequency, XGBoost outperformed GLM
- XGBoost's RMSE in predicting claim severity higher than GLM's RMSE
- XGBoost captures interactions among rating variables, GLM needs manual creation of interaction terms to find a better model
- Parameter tuning in XGBoost could be required

Anlagen

Literaturverzeichnis

- Esbjörn Ohlsson und Björn Johansson. Non-Life Insurance Pricing with Generalized Linear Models. Springer Berlin, Heidelberg, 18. März 2010. ISBN: 978-3-642-10791-7.
- J. A. Nelder und R. W. M. Wedderburn. „Generalized Linear Models“. In: Journal of the Royal Statistical Society. Series A (General) 135.3 (1972), S. 370–384. ISSN: 00359238. URL: http://www.jstor.org/stable/2344614.
- Michel Denuit, Donatien Hainaut und Julien Trufin. Effective Statistical Learning Methods for Actuaries I. Springer Cham, 3. Sep. 2019. ISBN: 978-3-030-25820-7.

htw. Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

Anlagen

Literaturverzeichnis

- Trevor Hastie und Robert Tibshirani. „Generalized Additive Models“. In: Statistical Science 1.3 (1986), S. 297–310. ISSN: 08834237. URL: http://www.jstor.org/stable/2245459.
- Mario Wuthrich und Christoph Buser. „Data Analytics for Non-Life Insurance Pricing“. In: SSRN Electronic Journal (Jan. 2017). DOI: 10.2139/ssrn.2870308
- M.V. Wüthrich und M. Merz. Statistical Foundations of Actuarial Learning and its Applications. Springer Actuarial. Springer International Publishing, 2022. ISBN: 978-3-031-12411-2