AKUR8

AKUR8

A Health Insurance Pricing Use Case

2025-06-30

CONFIDENTIAL

Jan Küthe

Actuarial Data Scientist,
Akur8
Actuary DAV

jan.kuethe@akur8.com

Lars Meder

Sales CEE & DACH
Akur8

lars.meder@akur8.com

Company overview

Founded
2018

Global offices
Paris, NYC, London, Milan, Cologne, Tokyo, Atlanta, Montreal

Employees
200+

Nationalities
25+

Activity
Non-Life Insurance Pricing and Reserving

Customers
300+ in 40+ countries

Akur8: Next-Gen Pricing and Reserving Platform

PRICING

The Complete Pricing Platform
for the Modern Actuary

- Data
- Risk
- Demand
- Rate
- Optim
- Deploy

RESERVING

The Complete Reserving Platform
for the Modern Actuary

- Arius
- Arius Enterprise
- Triangles on Demand

Agenda

Health Insurance Pricing Use Case

1. Pricing - Why and How?
2. Recap - Quiz questions
3. Use Case: How shall we apply the previous topics in Health?
4. Q&A

In Pricing companies can model Risk and Demand of clients

and use these to adjust the Commercial Premium charged

Risk-based pricing

Commercial premiums are built up from the best view of risk, with minimal commercial adjustments on top.

Static demand

Demand (conversion and retention rates) is actively tracked and monitored. Spikes in demand are met with a corresponding commercial adjustment.

Dynamic demand

We not only model demand, but we also measure or model price elasticity of demand and take this into account when setting the commercial premium.

Price optimization*

The commercial premium is built using an optimization routine that takes our best understanding of risk and demand into account.

*In markets where regulation allows.

Why accurate models are important?

Lets start with an example from Motor Insurance!

Adverse selection example

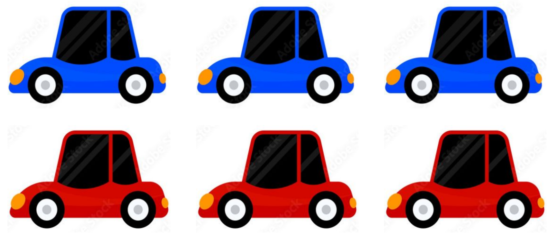
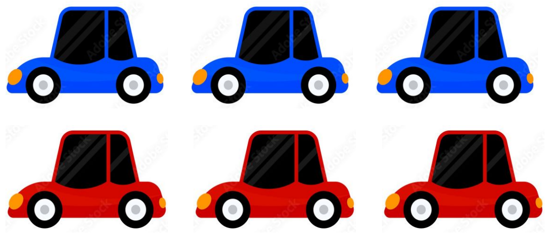
Hypothetical market - two competitors with identical strategy

Claim costs are €80 on average

Company	Alpha Insurance	Beta Insurance
Premium for blue car (€)	100.00	100.00
Premium for red car (€)	100.00	100.00

Adverse selection example

Results for identical strategy are the same

Company	Alpha Insurance	Beta Insurance
Customers		
Av. Claim Cost (€)	80.00	80.00
Av. Premium (€)	100.00	100.00
Loss Ratio (%)	**80.00**	**80.00**

Adverse selection example

Company B updates strategy to better reflect risk

Claim costs are €60 for blue cars

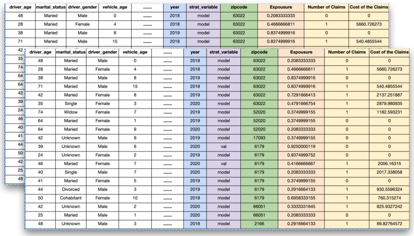
Claim costs are €100 for red cars

Company	Alpha Insurance	Beta Insurance
Premium for blue car (€)	100.00	90.00
Premium for red car (€)	100.00	110.00

Insurance Pricing is Backwards Looking

The closer the Future resembles the Past, the more confidence we can have in our pricing

With historic data…



driver_age	marital_status	driver_gender	vehicle_age	....	year	strat_variable	zipcode	Expousure	Number of Claims	Cost of the Claims
48	Maried	Male	0	....	2018	model	63022	0.2083333333	0	0
28	Maried	Female	4	....	2018	model	63022	0.4666666811	1	5660.726273
38	Maried	Male	8	....	2019	model	63022	0.8374999916	0	0
71	Maried	Male	15	....	2018	model	63022	0.8374999916	1	540.4855544
42										
35										
74										
64										
64										
42										
39										
24										
46										
40										
41										
44										
50										
42										
25										
48										

driver_age	marital_status	driver_gender	vehicle_age	....	year	strat_variable	zipcode	Expousure	Number of Claims	Cost of the Claims
48	Maried	Male	0	....	2018	model	63022	0.2083333333	0	0
28	Maried	Female	4	....	2018	model	63022	0.4666666811	1	5660.726273
38	Maried	Male	8	....	2019	model	63022	0.8374999916	0	0
71	Maried	Male	15	....	2018	model	63022	0.8374999916	1	540.4855544
42	Maried	Female	8	....	2019	model	63022	0.7291666413	1	2137.251887
35	Single	Female	3	....	2020	model	63022	0.4791666754	1	2879.980935
74	Widow	Female	7	....	2019	model	52020	0.3749999155	1	1182.593231
64	Maried	Female	1	....	2019	model	52020	0.3749999155	0	0
64	Maried	Female	9	....	2020	model	52020	0.2083333333	0	0
42	Unknown	Male	9	....	2019	model	17093	0.3749999155	0	0
39	Unknown	Male	6	....	2020	val	6179	0.9250000119	0	0
24	Unknown	Female	2	....	2019	model	6179	0.9374999752	0	0
46	Maried	Female	7	....	2020	val	6179	0.4166666667	1	2006.16315
40	Single	Male	7	....	2020	model	6179	0.2083333333	1	2017.338058
41	Maried	Female	5	....	2019	model	6179	0.3749999155	0	0
44	Divorced	Male	3	....	2019	model	6179	0.2916664133	1	930.5596324
50	Cohabitant	Female	10	....	2018	model	6179	0.6958333155	1	760.315274
42	Unknown	Male	2	....	2020	model	66051	0.3333331645	1	825.9327242
25	Maried	Male	1	....	2020	model	66051	0.2083333333	0	0
48	Unknown	Male	3	....	2018	model	2166	0.2916664133	1	89.82764572

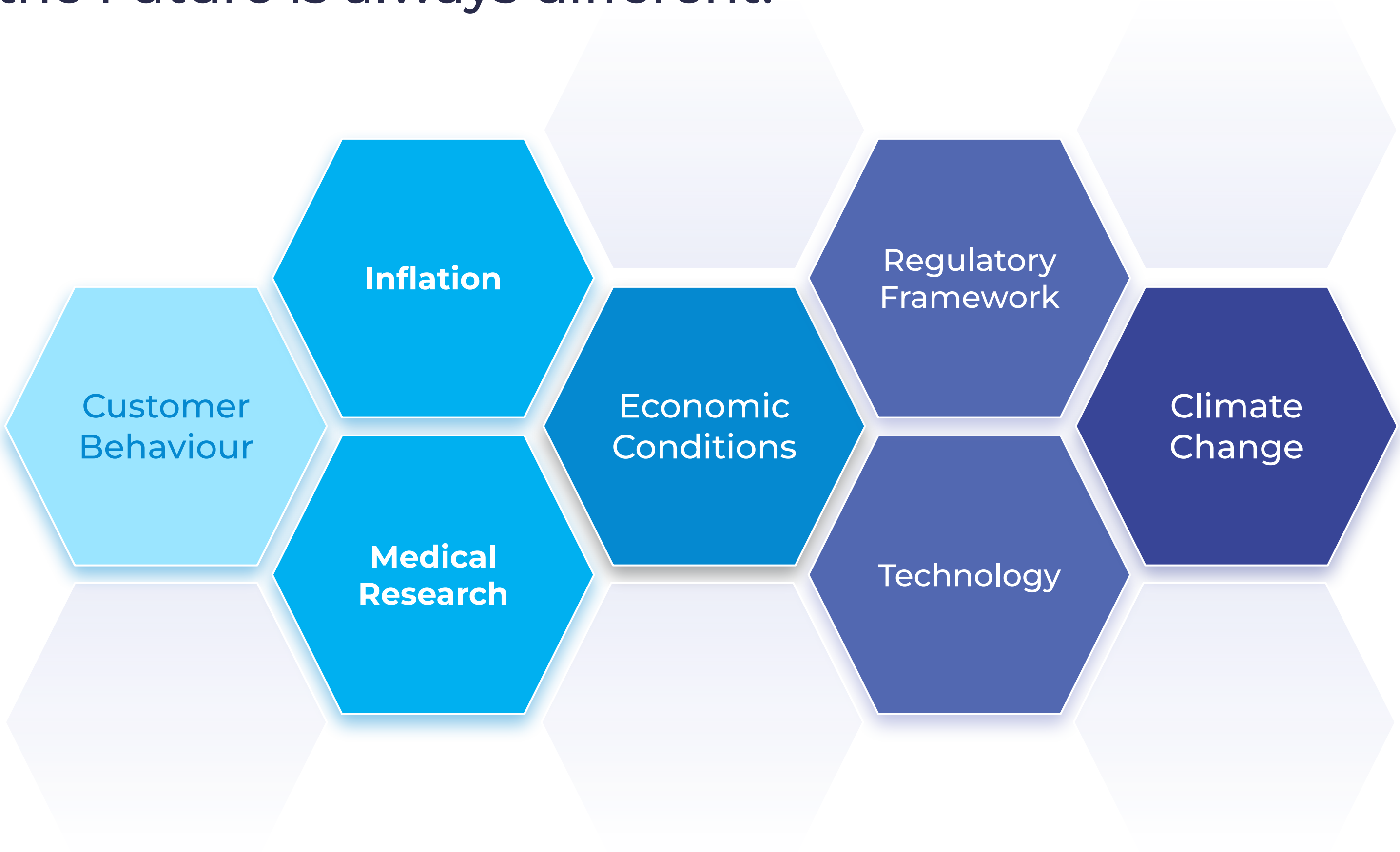
… we build pricing models and assumptions…

… to price for future risks

Past

Future

...but the Future is always different!



Actuarial modeling with GAMs

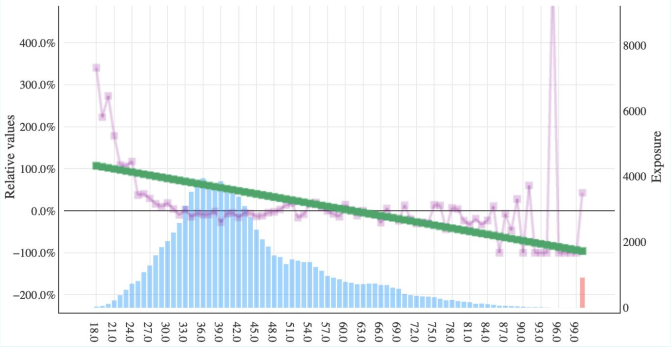
Actuarial modeling with GAMs

Actuarial modeling: from GLMs to GAMs

What GLMs Offer

Generalized Linear Models (GLMs) are, by definition, **linear**.

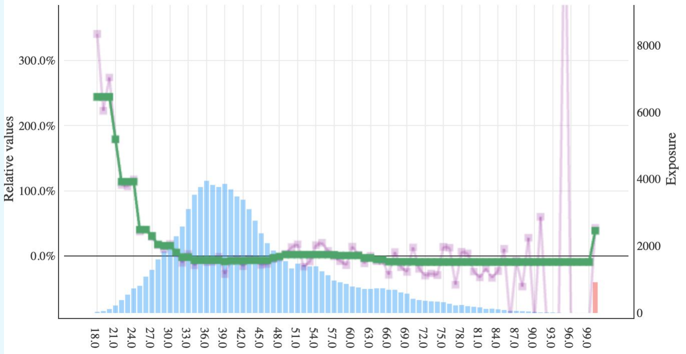
They are **easy to fit** as only one parameter has to be found for every variable



What we want.

We want to capture the **non-linear relations** between the explanatory and predicted variables.

They are **hard to fit** because, for every variable, a large number of parameters has to be found.



Fitting GAMs leveraging credibility

Prior and credibility

A credibility framework is defined by the prior assumptions the modeler has on their model. These **assumption represent a prior probability** distribution for the model coefficients. For instance, **"simpler" models are usually assumed to be "more likely"**.

Classic prior assumptions would be:

- "The coefficients follow a Gaussian distribution, centered on 0"
- "The coefficients follow a Laplace distribution, centered on 0"

The **Maximum of Likelihood approach can incorporate prior information** about the data.
We define the likelihood as the probability of the observed data in a given model $p(y|\hat{y}(X))$

$$\beta^* = Argmax_\beta \ p(y|\hat{y}(X)) \times p_{prior}(\beta)$$

Changing the prior distribution of the coefficients allows the modeler to control for the strength of the prior. Taking the log of this formula provides an easy-to-optimize log-likelihood function, where ***LL*** is the log of the likelihood function.

$$\beta^* = Argmax_\beta \ LL(x, y, \beta) + \log\left(p_{prior}(\beta)\right)$$

Fitting GAMs leveraging credibility

Prior ⇔ Penalized regressions: example of the linear regression

Prior assumptions are at the center of machine-learning methods used to control high-dimensional or correlated data, such as LASSO or Ridge regression. Controlling the distribution (through the λ parameter) allows the modeler to control the overfitting of the models.

Gaussian Hypothesis ⇔ **Prior:** Coefficients follow a Normal distribution N(0, 1/λ): ⇔ **Coefficients Distribution:** $p(\beta) \sim e^{-\lambda\,\beta^2}$ ⇔ **Log-Likelihood (w. prior)** $LL(x, y, \beta) - \lambda\ \beta^2$ ⇔ **Ridge Regression**

Laplace Hypothesis ⇔ **Prior:** Coefficients follow a Laplace distribution L(0, 1/λ): ⇔ **Coefficients Distribution:** $p(\beta) \sim e^{-\lambda\,|\beta|}$ ⇔ **Log-Likelihood (w. prior)** $LL(x, y, \beta) - \lambda\,|\beta|$ ⇔ **LASSO Regression**

Refresher on smoothness selection

Refresher on smoothness selection

What is signal and what is noise?

When trying to do any prediction based on observed data, it is important to make a distinction between:

- Signal: meaningful and systematic patterns or relationships within the data. It is the target information we are interested in predicting or understanding.
- Noise: random variation or error that is unrelated to the underlying signal. It can arise from sources such as measurement errors or sampling variability. Noise can introduce misleading patterns or correlations that are not representative of the true signal.

Striking the right balance between capturing signal and noise involves avoiding overfitting, i.e., not taking noise as if it were signal. By doing so, we can improve the model's ability to generalize and make accurate predictions on unseen data.

Refresher on smoothness selection

Smoothing out the noise

Choose a model that is smooth enough - i.e. no suspicious moves in the coefficients:

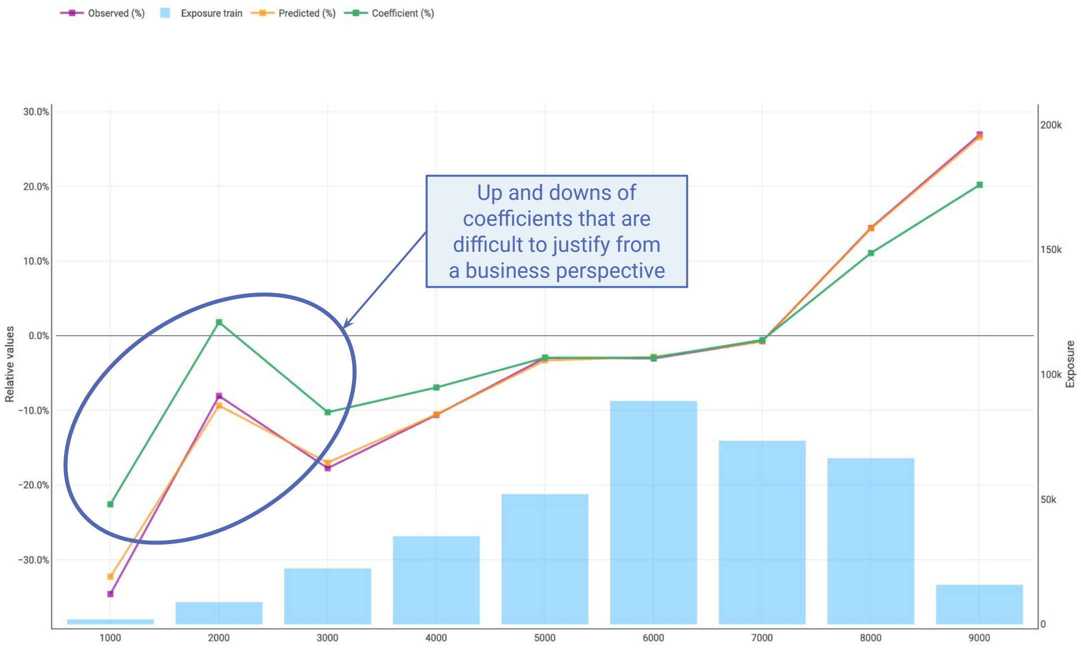
- If a model contains one variable with one dubious coefficient, it captures noise -> move to a smoother model instead,
- A model capturing noise in one coefficient was created with a weak statistical threshold: it may capture noise on all the other coefficient. Its low threshold does not provide good guaranties.

Once you have selected a model that is smooth enough, you need to check that the selected variables are non-redundant. If some variables do not provide value to the model, move to a model with fewer variables. Once you are satisfied you can tag the model.

The following slides will show examples of models for the risk associated with a vehicle mileage with different levels of smoothness.

Refresher on smoothness selection

Example of an overfitted model



This graph is an example of a noisy, overfitted model.

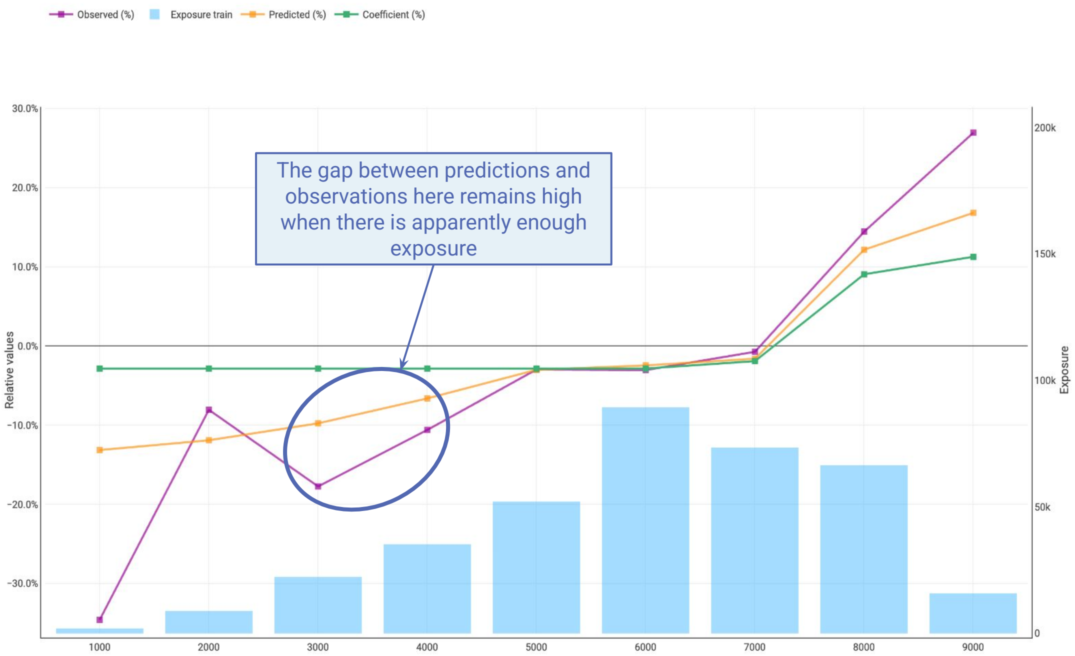
We can see that the coefficients, shown in green, go up and down following every movement of the observed values.

Some of these trends are not very intuitive: why does risk go up from 1000 to 2000 limit, but then goes down again for limits of 3000?

We need to look for a smoother model.

Refresher on smoothness selection

Example of an underfitted model



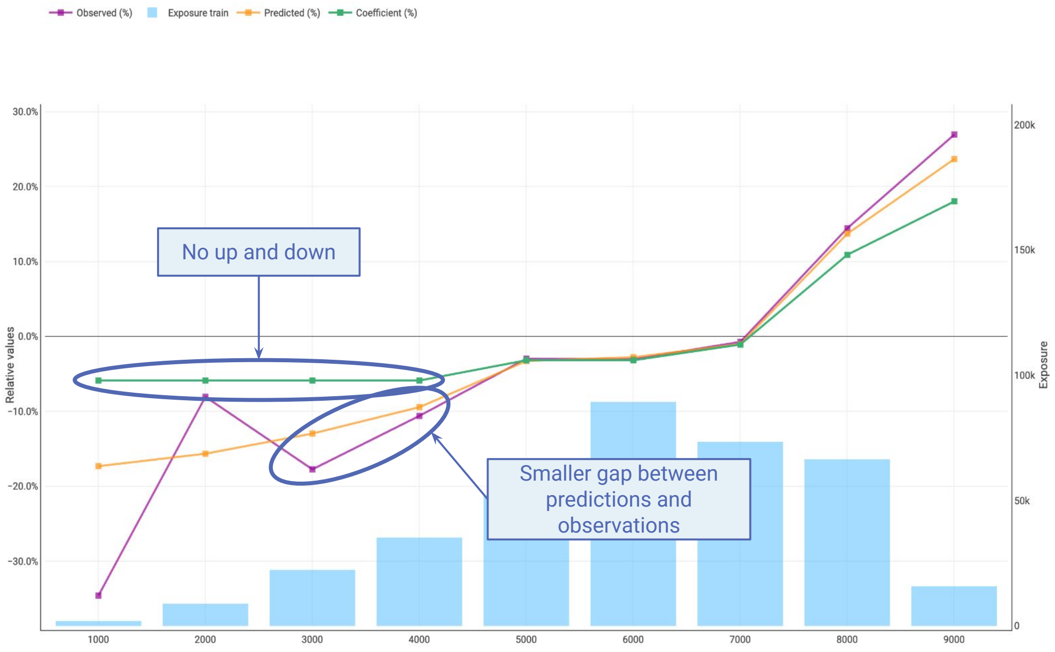
If the model is smoothed too much, we arrive to an underfitted model such as the one presented here.

In this case, the coefficients are flat for all values at or below 6000, so the credible upward trend is not captured. This results in predictions that are far from the observations.

Smoothing too much might prevent the model from capturing desirable signal

Refresher on smoothness selection

A balanced model



This graph shows a well balanced model that is not underfitted neither overfitted.

Coefficients are stable and show a general upward trend, which matches our expectations and the overall pattern of the purple line. The levels corresponding to limits 1000 and 2000 do not provide with signal with respect to limit 3000 and 4000, but this is actually positive given their low exposure.

As mentioned as the beginning of this presentation, the right threshold remains at the discretion of the user. If the smoothness level automatically suggested are not suitable, it is always possible to enrich a grid search with more smoothness steps.

Stay time consistent

Stay time consistent

Refresher on how to access time consistency graph

We first recall how to access the time consistency graphs:

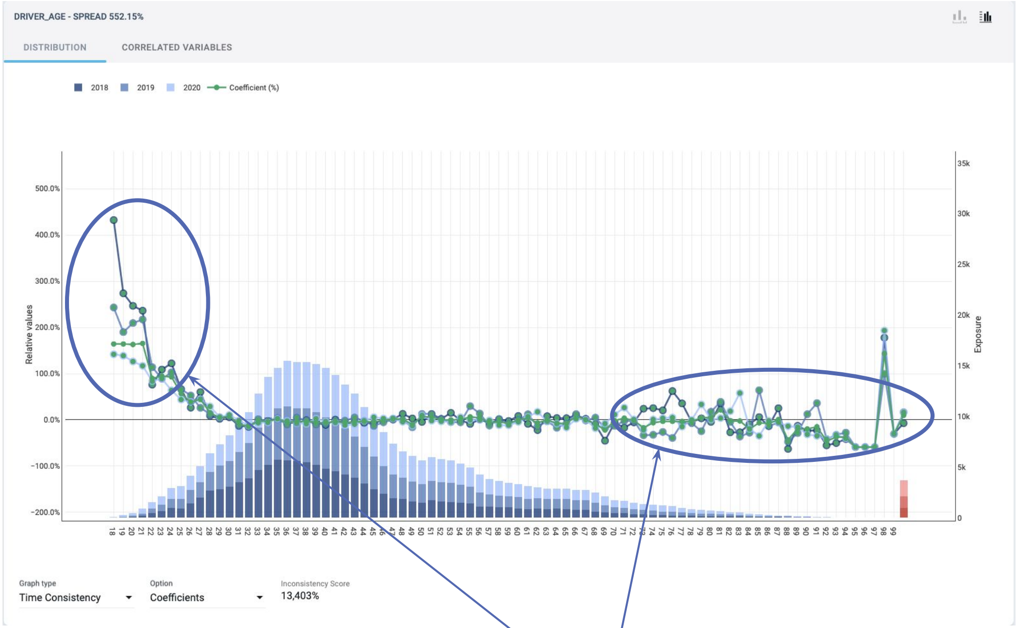
- Click on a model
- Then click on the variable tab
- When the graph of the variable is displayed, select time consistency in the coefficients option

Stay time consistent

Example of an overfitted model

This model is overfitted. It displays a lot of inconsistencies for every variables. These inconsistencies are:

- in a high proportion
- of a high magnitude
- display very different patterns in terms of trends and grouping



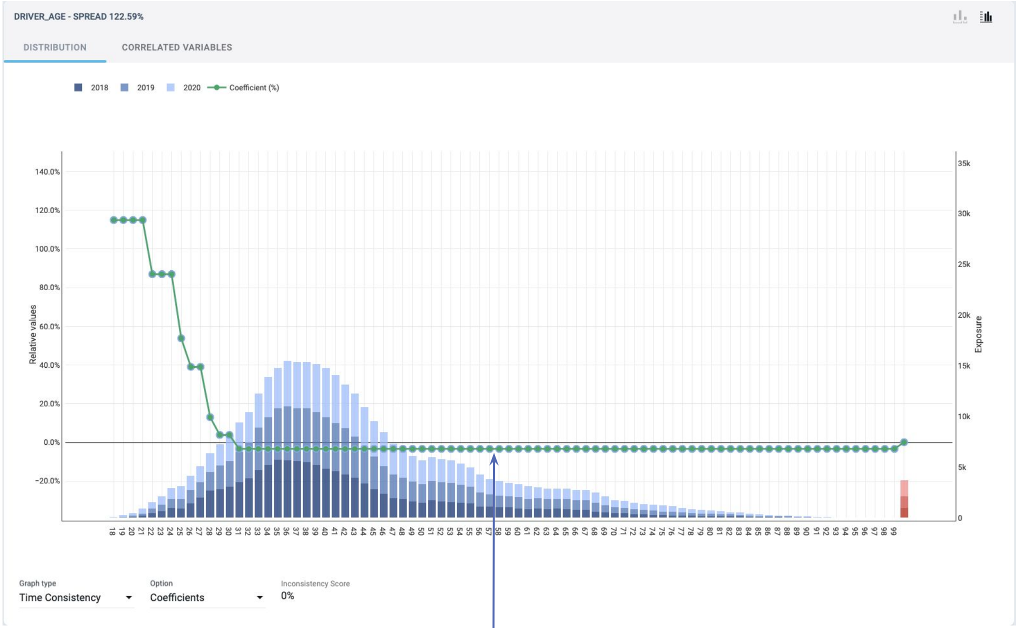
Almost all coefficients are unstable which is weird in particular for segments with a lot of exposure. Besides, the differences between the years are huge and display very different variations

AKUR8

Stay time consistent

Example of an underfitted model

This is an underfitting model. This can be easily spotted because there no inconsistencies at all across all variables which is highly unlikely in reality.



All coefficient curves are superposed whatever the year

Stay time consistent

Example of a balanced model

A balanced model might display some inconsistencies here and there but these inconsistencies are:

- in a small proportion
- not of a high magnitude
- displaying similar patterns in terms or trends and grouping

We have small inconsistencies here but we can see that the groups remain similar across the different year and that the decreasing trends are similar across the years

Gini index and Lorenz curve

The Gini index and the Lorenz curve

Definitions

The **Lorenz curve** is built this way:

- Sort all model predictions from highest to lowest
- Plot the cumulative observed value

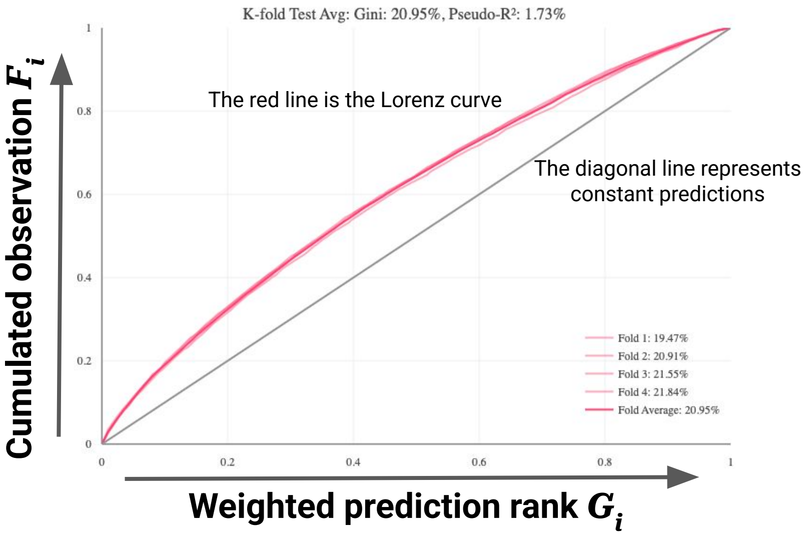
More formally, we have on the x axis:

$$G_i = \frac{\sum_{j=1}^{n} w_j I(\hat{y}_j > \hat{y}_i)}{\sum_{j=1}^{n} w_j}$$

And on the y axis:

$$F_i = \frac{\sum_{j=1}^{n} w_j y_j I(\hat{y}_j > \hat{y}_i)}{\sum_{j=1}^{n} w_j y_j}$$

Thus a constant prediction ($y_i = \bar{y}$) leads to the diagonal line



The **Gini index** is twice the area between the Lorenz curve and the diagonal. It is:

- a classic measure of inequality in economics, adapted to evaluate models in insurance
- useful for interpreting skewed distributions and unbalanced datasets
- a measure of improvement over constant prediction

The Gini index and the Lorenz curve

Gini intuition: perfect model

For a saturated logistic model (all predictions are 1 or 0 and equal to the observations):

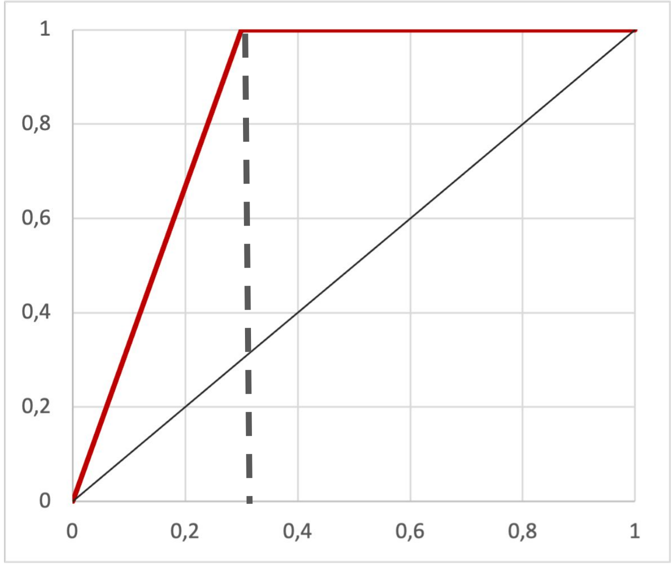
$$F_i = \frac{\sum_{j=1}^{n} y_j I(\hat{y}_j > \hat{y}_i)}{\sum_{j=1}^{n} y_j} = \begin{cases} i/TP \ , \ i < TP \\ 1 \ , \ i > TP \end{cases}$$

In other words, the Lorenz curve looks like in the image:

- All positives are on the left
- Straight line up to the fraction of positives

The exact Gini can be computed easily:

- The total area under the red curve is half the fraction of positives (from the triangle on the left) plus the fraction of negatives (the rectangle on the right): A = (TN+TP/2)/n
- And Gini = 2A - ½ = TN/n = TNR



Fraction of positives

In general, the Gini of a perfect model depends *only* on how balanced or unbalanced is the target

Quiz

Quiz

Question 1

What is the Gini index of a perfect model?

- 1
- Below but close to 1
- Depends on the target
- Depends on the loss type

Quiz

Question 1

What is the Gini index of a perfect model?

Depends on the target

The Gini index of a perfect model depends on how unbalanced the target is. It can be close to 1 for frequency models with very few claims, but much lower in some severity models.

Quiz

Question 2

A model with more variables is always better?

True

False

Quiz

Question 2

A model with more variables is always better?

Okham's razor incite us to favor parsimony that is if two models have identical performance with a different number of variables, we should take the model with fewer variables. If adding variables can increase performance, adding too many variable leads to overfitting.

False

Quiz

Question 3

What is the time consistency check?

It is a check in which we backtest the predictions of the candidate model with the historical data. It is needed to have a well performing model.

It is a check in which we benchmark the candidate model with a GBM containing only the time variable. It is needed to retain full transparency of the model.

It is a check needed to validate the candidate model. We need to be sure that the estimated coefficients are consistent over time.

It is a check in which we compute the time needed to fit the candidate model to be sure that if we refit it in the future, it will not take too long.

Quiz

Question 3

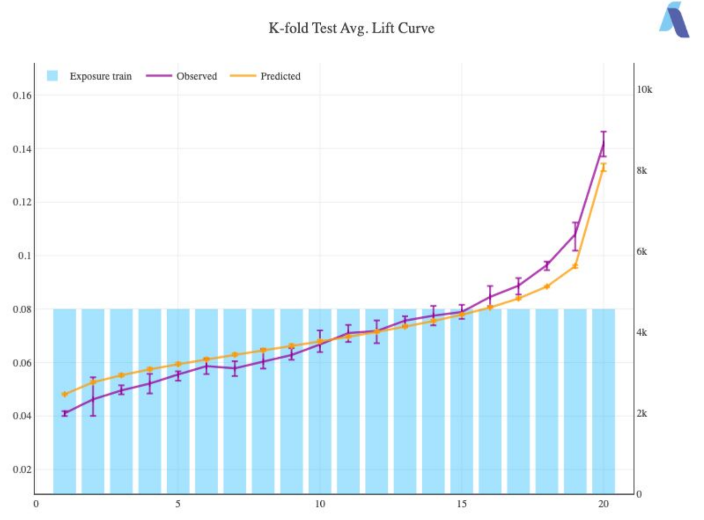
What is the time consistency check?

It is a check needed to validate the candidate model. We need to be sure that the estimated coefficients are consistent over time.

Before putting in production a model, we need to check that all variables entering the model are consistent over time. In Akur8, this is achieved by fitting an interaction between the time variable and the model variables.

Quiz

Question 4



Why is this model likely underfitting?

A: Because it has 20 bins, which are too many

B: Because it underestimates high risk and overestimates low risk

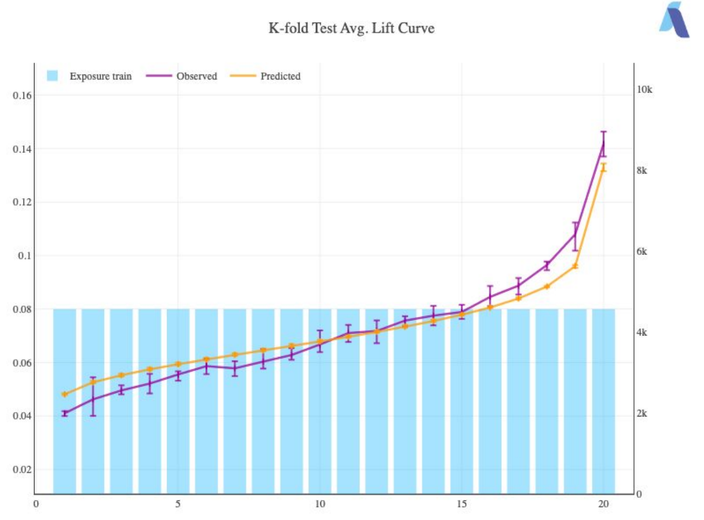
C: Because the predicted curve crosses the observed one

D: Because the exposure is equally distributed along the horizontal axis

Quiz

Question 4

Why is this model underfitting?



This model is underfitting and we can see this in the right and left ends of the graphs. For low risk segments (left-hand side), the predicted line is above the observed one. For high risk segments (right-hand side), we can see the opposite behaviour

B: Because it underestimates high risk and overestimates low risk

Live Use Case

Thank You!

Q&A

28 rue de Londres, 75009 Paris
FRANCE

AKUR8