AKUR8 | Munich RE

Penalization techniques for Optimized GLM modelling

April 25th, 2023

Ruiqi Zeng, Thibault Imbert, Monica Carvajal-Pinto

Agenda

01 Introduction

02 Pricing Models

03 Leveraging Credibility to automate GAMs

04 Q&A Session

Introduction

Thibault IMBERT

Senior Insurance Solutions Manager

Timbert@munichre.com

Biography

Thibault Imbert is **Senior Insurance Solutions Manager** at Munich Re.

After 8 years in primary insurance, holding various positions related to actuarial functions, he joined Munich Re as part of the Global Consulting Unit.

Thibault has extensive experience in modelling, best practices set-up and rates development.

Thibault hold a Master of Science in Applied Mathematics from Ecole des Mines de Saint-Etienne.

AKUR8

Monica Carvajal-Pinto

Actuarial Data Scientist

monica.carvajal-pinto@akur8.com

Biography

Monica is an **Actuarial Data Scientist** at Akur8: she supports non-life insurers worldwide to accelerate and optimize their pricing process thanks to Transparent AI.

Monica had experience in Consulting, Research and Teaching in Risk Theory and Actuarial Models. Monica has a Master degree in Actuarial Science and a PhD in Probability.

AKUR8

Pricing Models

GLMs are a simple extension of the simple Linear Models

Each variable has a **linear effect**, and a **link-function** is applied to the sum of the effects:

$$\hat{y}(X) = g^{-1}(\sum_d \beta_d \times X_d)$$

The link function g used in pure-premium models is often a logarithm, leading to a multiplicative formula (as g^{-1} is an exponential function).

For example:

$$\log(\widehat{ClaimsCost}) = Base + \beta_{SumInsured} \times SumInsured + \beta_{KmDriven} \times KmDriven + \beta_{NbPastClaims} \times NbPastClaims$$

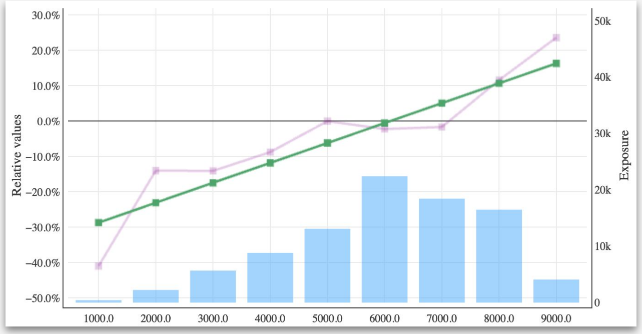
Which leads to:

$$\begin{aligned} \widehat{ClaimsCost} &= \exp(Base + \beta_{SumInsured} \times SumInsured + \beta_{KmDriven} \times KmDriven + \ \ldots) \\ &= \exp(\mathrm{Base}) \times \exp(\beta_{SumInsured} \times SumInsured) \times \exp(\beta_{KmDriven} \times KmDriven) \times \cdots \end{aligned}$$

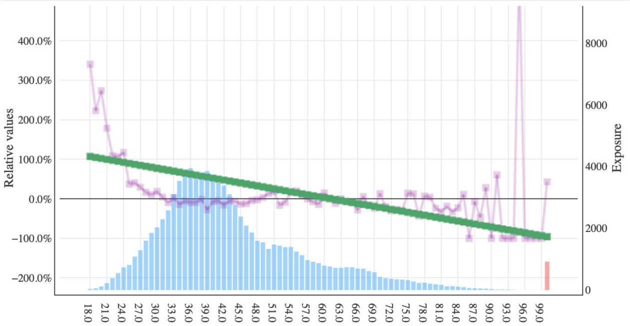
GLMs have limitations in capturing all relations between explanatory and explained variables

This kind of model naturally captures the linear (or exponential, with a log-link) relations between the explanatory and explained variables.

If this relation does not exist, however, **the effects won't be properly captured.**



The Annual Mileage is linearly related to the predicted variable (Claims Frequency, in purple).



The Driver Age is non-linearly related to the predicted variable (Claims Frequency, in purple).

The GLM effects (in green) capture the relation correctly on the left, but incorrectly on the right.

GLMs offer several advantages

They are **easy to understand,** as the effect of each variable is represented by a single value, and can be easily visualized by graphs as the ones presented above.

They are **easy to implement** as the models predictions can be computed with simple sums and products.

They are very **easy to compute**, as only a few coefficients need to be fitted, and well-known algorithms exist to provide efficient solutions. This strength was especially true in the 90's and early 2000's when computing power was scarce.

As they have been widely used during the last 40 years, **actuaries know how to build (and regulators are very used to review) this class of models**.

Generalized Additive Models (GAMs) are an extension of the simple GLMs

Each variable has an effect (**non necessarily linear**), and a link-function is applied to the sum of the effects:

$$\hat{y}(X) = g^{-1}(\sum_d \beta_d(X_d))$$

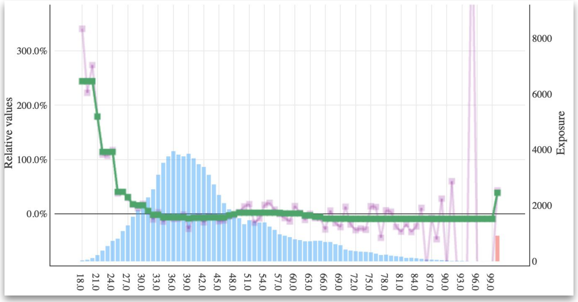
For example:

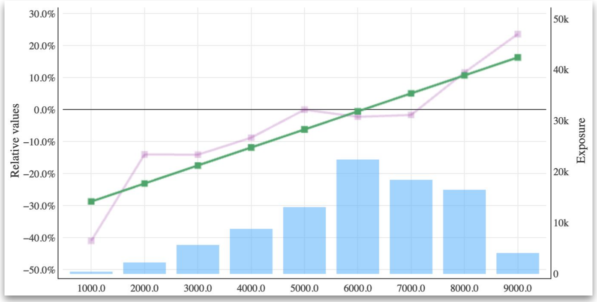
$$\log(\widehat{ClaimsCost}) = Base + \beta_{SumInsured}(SumInsured) + \beta_{KmDriven}(KmDriven) + \beta_{NbPastClaims}(NbPastClaims)$$

It is interesting to note that **GLMs are special cases of GAMs**, where all the functions $\beta_d(\ldots)$ are linear : $\beta_d(X_d) = \beta_d \times X_d$

GAMs generalize the GLMs by allowing non-linear effects through the functions

This allows much richer effects to be captured:





Both effects of the Annual Mileage (left) and Driver Age (right) are well captured by the GAM effects, despite the non-linearity of the relation between the Driver Age and the predicted Claims Frequency (in purple).

The GAMs share the main strengths of the GLMs:

- They are easy to understand, as the effect of each variable is represented by a single value, and can be easily visualized by graphs as the ones presented above.
- They are easy to implement, as the model can be described and put to production as simple rating tables.

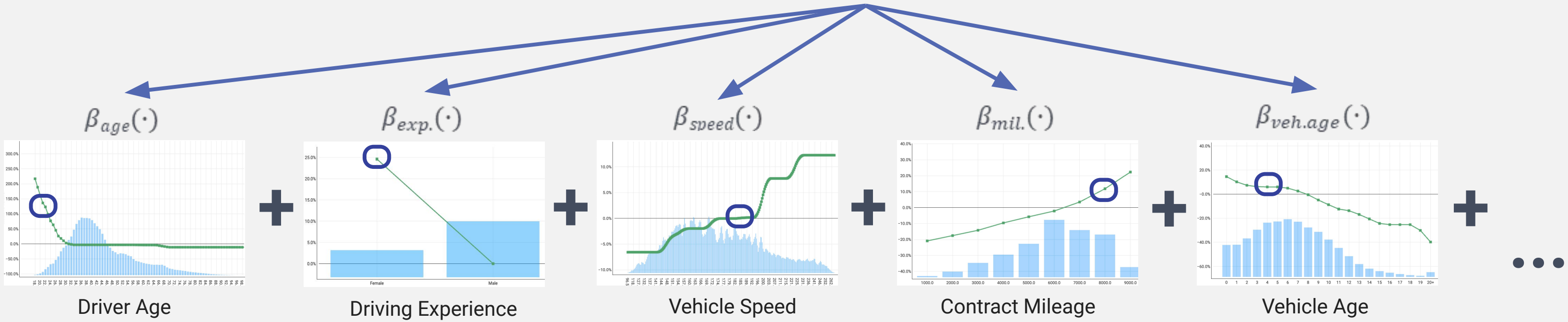
The Transparency of GAMs

Additive models are a great balance between predictive power and adverse selection.

For this reason, they are **currently widely used in the actuarial community**.

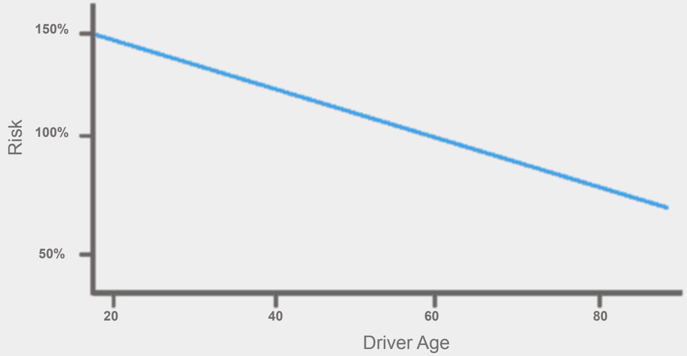
Additive models can be visualized as rating tables, but most remarkably, the **human visualization is convenient** for model review and modification as it displays one function per variable.

$$\hat{y}(X) = g^{-1}(\Sigma_d \beta_d(X_d))$$



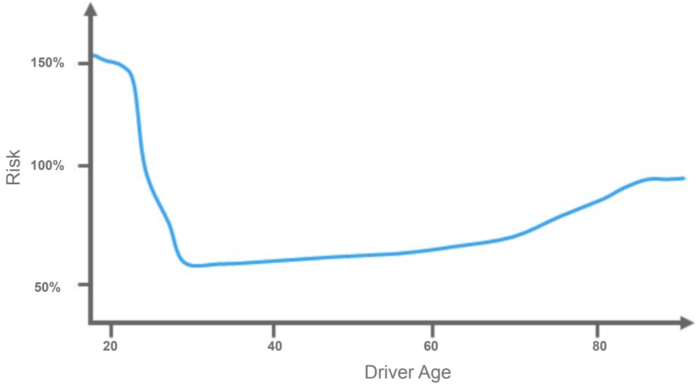
While GAMs offer more flexibility than GLMs, they are harder to fit

Linear Model



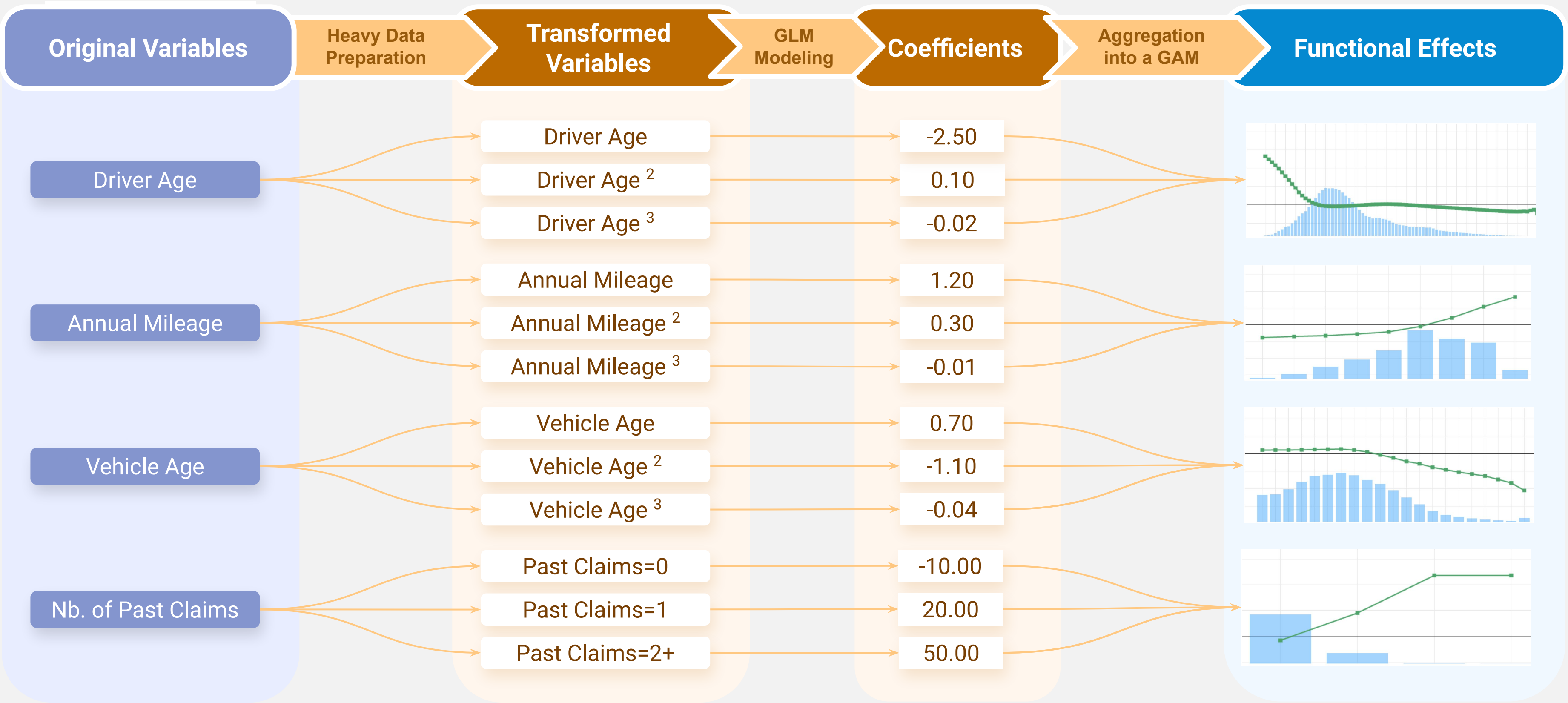
- Simple and well-known technique
- First regression created & learned
- Captures the linear relations in the data
- Simultaneously select the variables and fit the trends

Additive Model



- Much more powerful models
- Captures non-linear effects
- Incorrectly called “GLMs“
- Requires both variables selection and fitting

GAMs can be obtained via variable transformations in GLMs



GLMs and GAMs are equivalent

Linear Models

Variable Transformations

Non-Linear Models

$$\hat{y}(X) = g^{-1}(\textstyle\sum_{i,j} \beta_{i,j} \times I_{X_{i\_j}}))$$

Driver Age=16
Driver Age=17
Driver Age=18
Driver Age=19
Driver Age=20
Driver Age=21
Driver Age=22
Driver Age=23

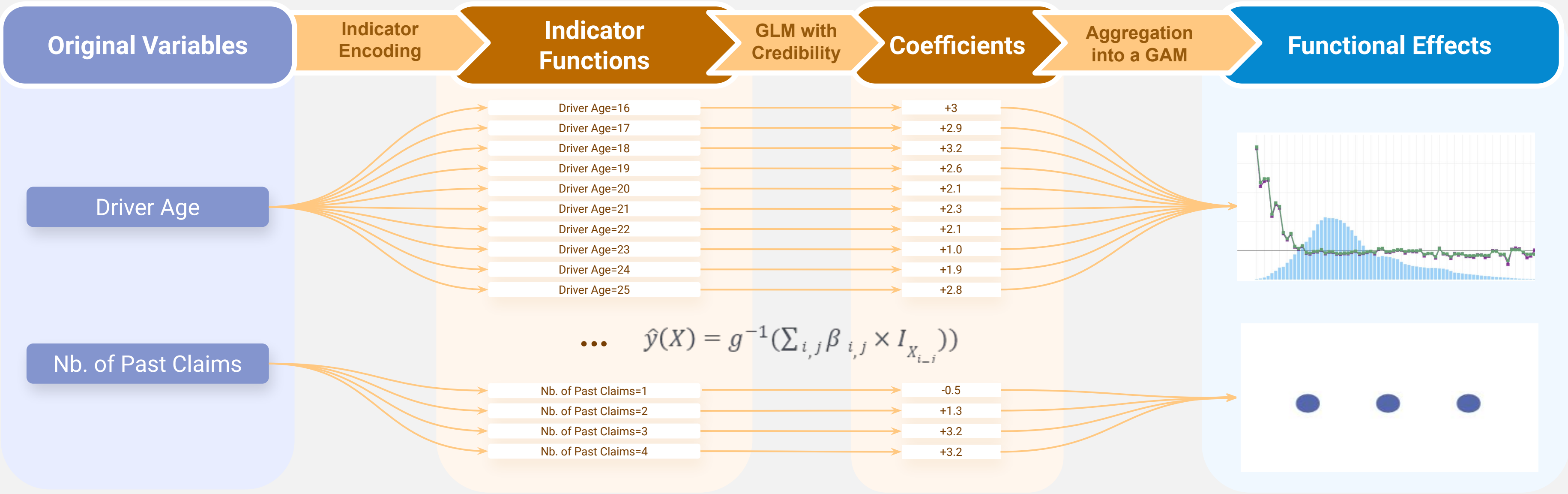
$$\hat{y}(X) = g^{-1}(\textstyle\sum_{j} \beta_{i}(Xi)))$$

$$I_{X_{age\_j}} = \begin{cases} 1 \; if \; Xage = j \\ 0 \; if \; Xage \neq j \end{cases}$$

Coefficients are built for different values of the explanatory variables.

A large number of indicator functions have to be created

- Equal one if a variable equals a given value, zero otherwise.
- Remove time-consuming data preparation.
- The resulting functional effects are **potentially very noisy** if this approach is used naively.



Most regression models are built around the principle of maximum likelihood

The model is built to **maximize the probability of the observations**.

So, the coefficients β^* of the model are the ones maximizing y, the probability of observations. Fitting a model is equivalent to solving:

$$\beta^* = \arg max_\beta \; p(y|\hat{y})$$

(where the probability $p(y|\hat{y})$ is often referred to as the **likelihood** of the model).

This is equivalent to **minimizing the errors** between predictions and observations on the train dataset (where the "errors" are defined as minus the log-likelihood):

$$\beta^* = \arg min_\beta \, Errors(y, \hat{y})$$

However, creating a non-linear model requires **controlling the overfitting** in the fitting process. This can be done either by:

- **Controlling the transformations** created
- Leveraging **credibility** in the fitting process

Leveraging Credibility to automate GAMs

AKUR8

What is Credibility?

> “Credibility, simply put, is the weighting together of different estimates to come up with a combined estimate.”
>
> *Foundations of Casualty Actuarial Science*

Buhlmann credibility is the best-known approach. It is equivalent to a simple **Bayesian** framework, where a prior “knowledge” based on a model is updated based on observations.

Usually the output of a credibility approach is that the predictions are a **weighted average** between the observations and the initial assumption.

The weight will depend on:

- ➔ the **quantity of data** (the larger the data, the higher the weight)
- ➔ the **strength of the prior** assumptions (a very reliable assumption with small variance will have a large weight).

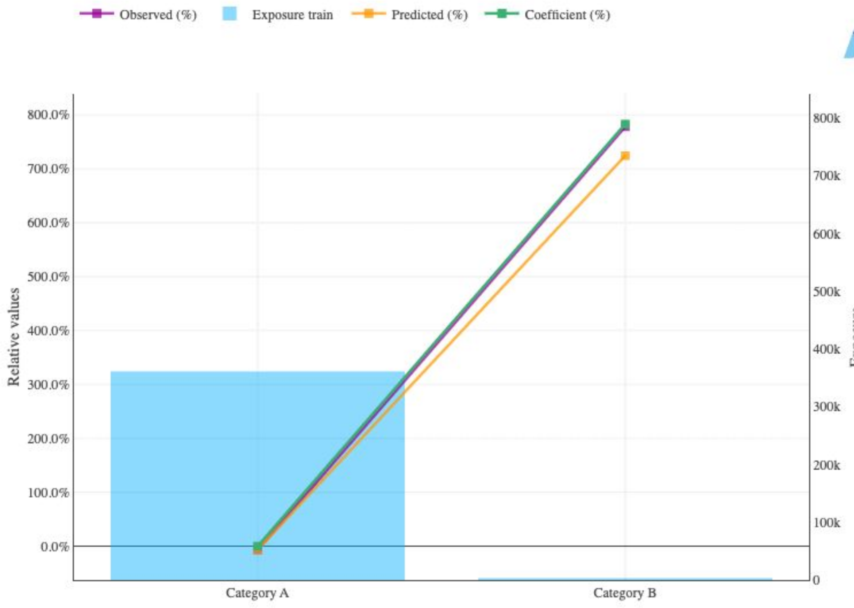
Pulls Coefficients to the Mean like a Magnet

Data Magnet vs. Average Magnet

Data*(Z) + Average*(1-Z)

This is credibility

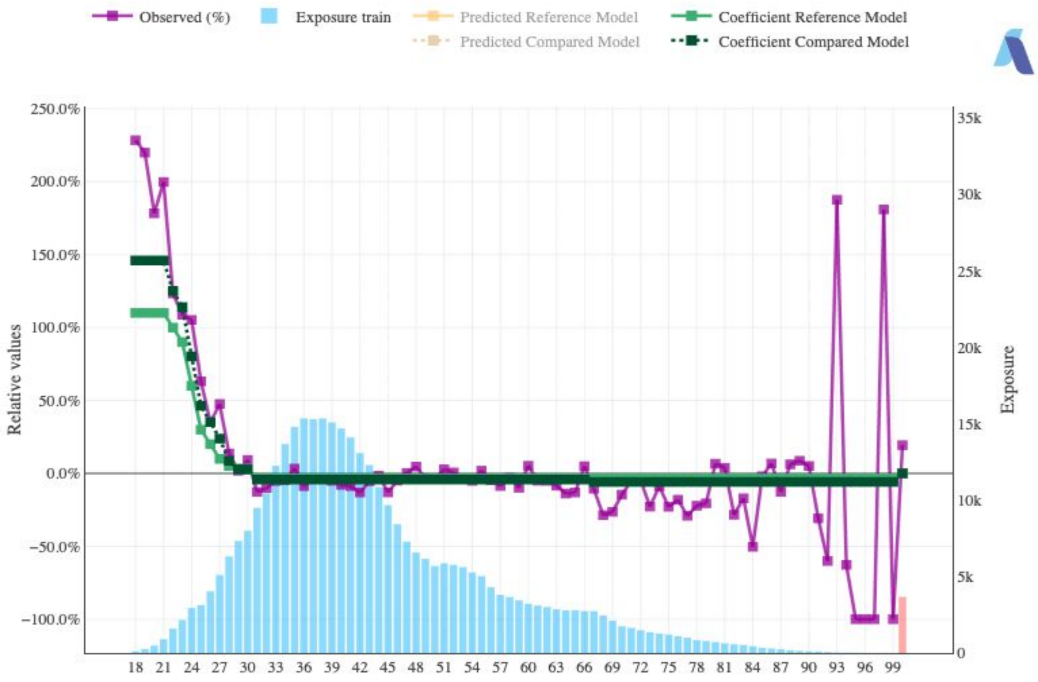
Note: Usually a well known concept in fleet insurance



Driver Age Example

Different levels of credibility

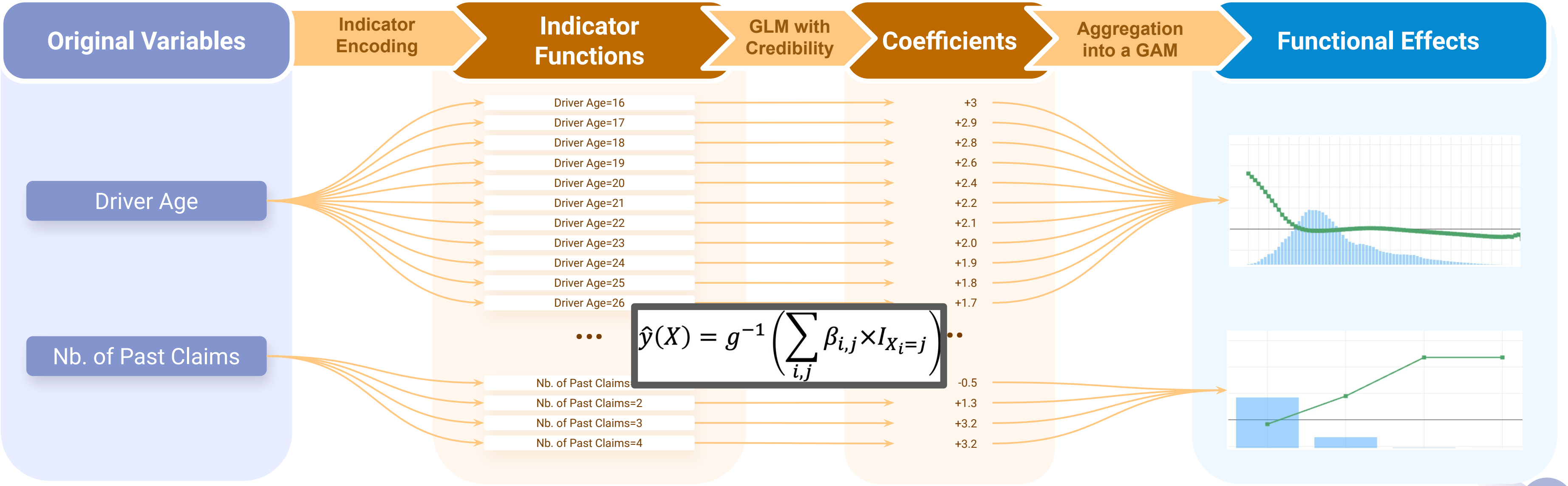
When the modeled coefficient is zero, the complement has full credibility.



Credibility via weighting of levels

In order to remove the heavy and time-consuming data preparation step, a **large number of indicator functions** are created - these functions equal one if a variable equals a given value, zero otherwise.

Then a model **fitted leveraging credibility** ensures the coherence between the different coefficients created.

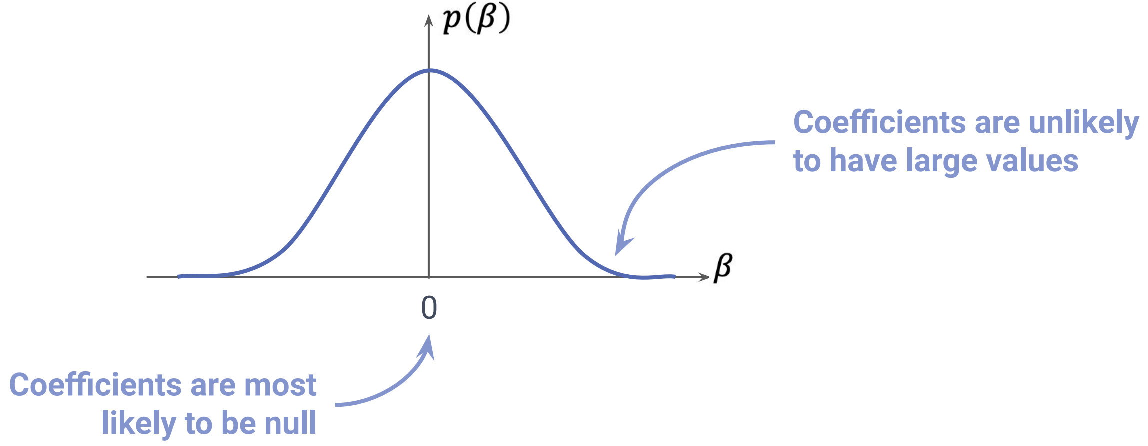


From credibility to priors

A credibility framework is defined by the prior assumptions the modeller has on his model. These **assumptions represent a prior probability** distribution for the model coefficients.

For instance, **“simpler” models are usually assumed to be “more likely”**.

Classic prior assumptions can be: “The coefficients follow a Gaussian distribution, centered on 0”



Credibility-Bayes and penalised regression

All regression models are built around the same main principle:

$$\beta^* = ArgMax\ p(y|\hat{y}_\beta) = ArgMax\ LogLikelihood(x, y, \beta)$$

However, maximizing a likelihood on hundreds of parameters would lead to overfitting,

Integrate priors on the coefficients into the model creation

- The priors will be directly included into the likelihood optimization
- They will reduce the complexity of the models created

Credibility-Bayes and penalised regression

This prior is formalized as a distribution of probability for the coefficients: $p_{prior}(\beta)$

The **Maximum of Likelihood approach directly integrates the prior**:

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

Taking the log, we get the Maximum of Likelihood problem:

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

Or equivalently the Minimization of Error (aka **Penalized Regression**):

$$\beta^* = Argmin_\beta \;\; Errors\ (y, \hat{y}) - Penalty(\beta)$$

Prior ⇔ Penalized Regressions: an illustration

Some examples in the Linear Regression case

Prior assumptions are at the center of penalized regression methods used to control high-dimensional or correlated data, such as Lasso or Ridge Regression. Controlling the distribution (through the λ parameter) leads to controlling the overfitting of the models.

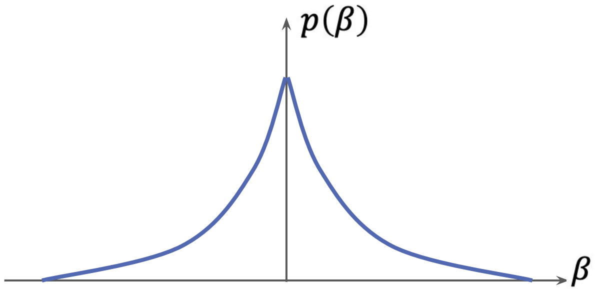
Gaussian Hypothesis ⇔ **Prior:** Coefficients follow a Normal distribution N(0, 1/2λ): ⇔ **Coefficients Distribution:** $p(\beta) \sim e^{-\lambda\,\beta^2}$ ⇔ **Log-Likelihood (incl. prior)** $LL(x, y, \beta) - \lambda\ \beta^2$ ⇔ **Ridge Regression**

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

Sparse priors

Lasso is particularly popular as it is a good tool for variable selection: models created with the Lasso framework are sparse - non-relevant coefficients set to zero.

The Laplace distribution that underlies the Lasso has a peak at zero:



When used on binary explanatory variables, it is also equivalent to **hypothesis testing**:

Null Hypothesis: $\beta = 0$:
"The coefficient is not significantly different from zero."

- If the null hypothesis is **not rejected**, the coefficient value is zero.
- If the null hypothesis is **rejected**, the coefficient has a non-zero value.

Creating new Priors and Penalties

Priors can be easily adapted to the desired structure of the models.

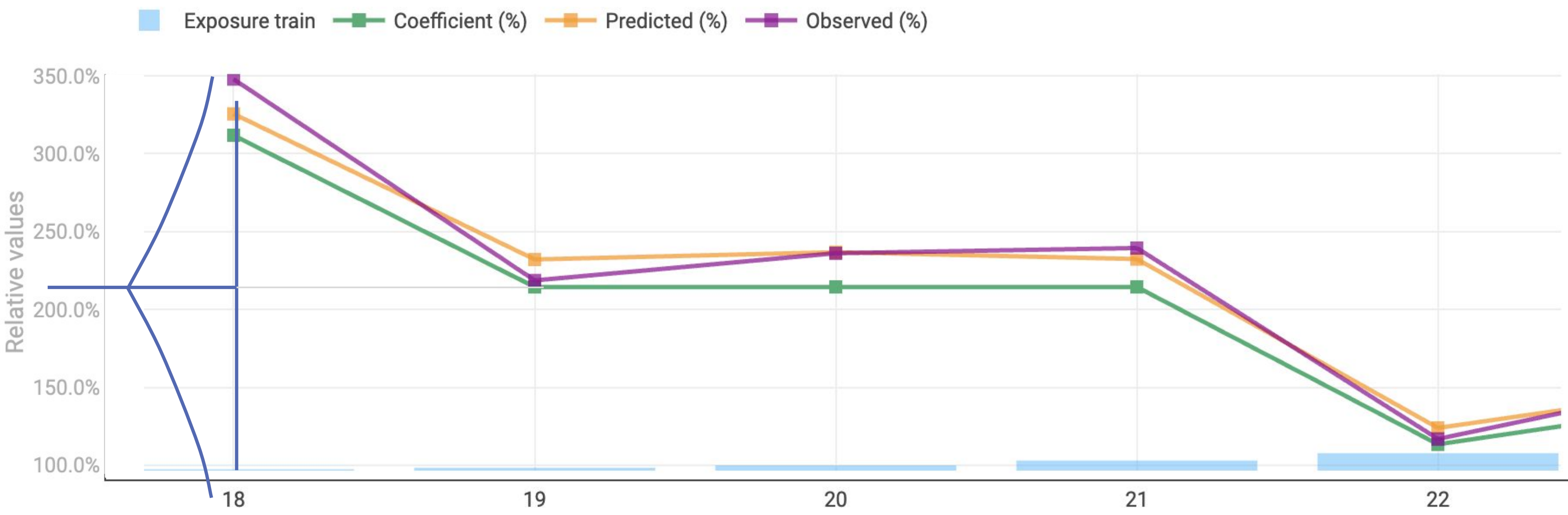
For instance we could want that two consecutive levels are:

- **be more likely to be close than far apart**;
- equal (**have the same coefficients**) if they are not significantly different…

This concept **generalizes the Lasso penalty to continuous function**, providing the high level of flexibility and stability necessary to create GAM models.

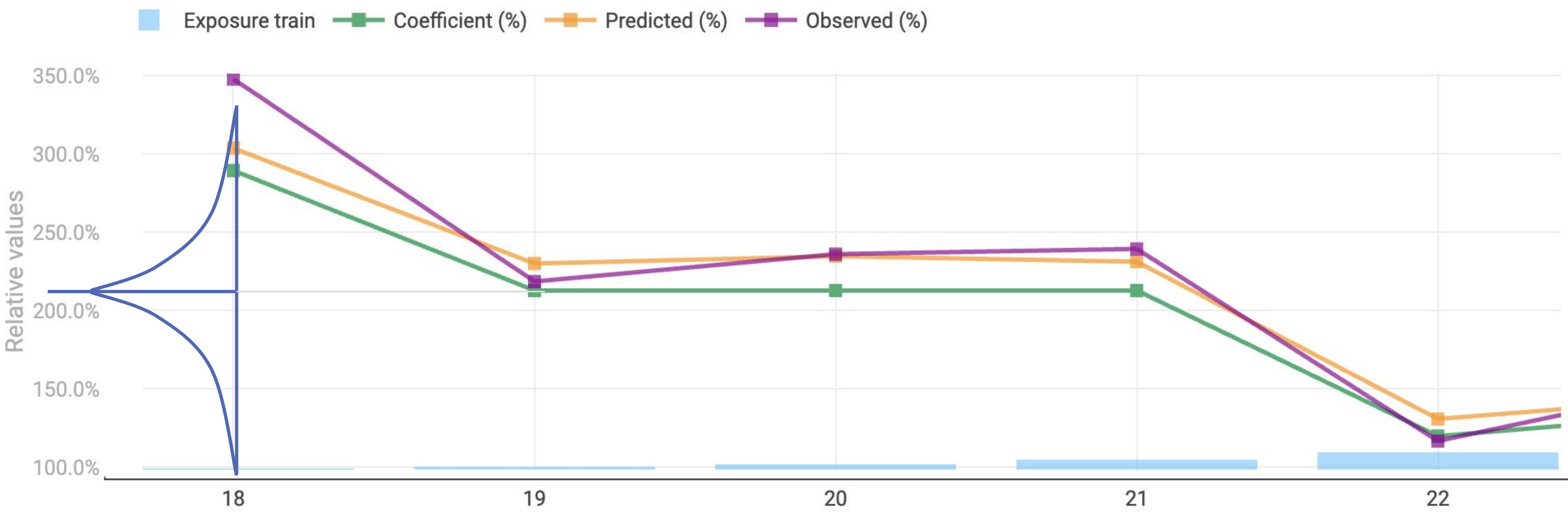
Weak Prior ⇔ Strong reliance on the observation

The prior has a very limited impact on the final model



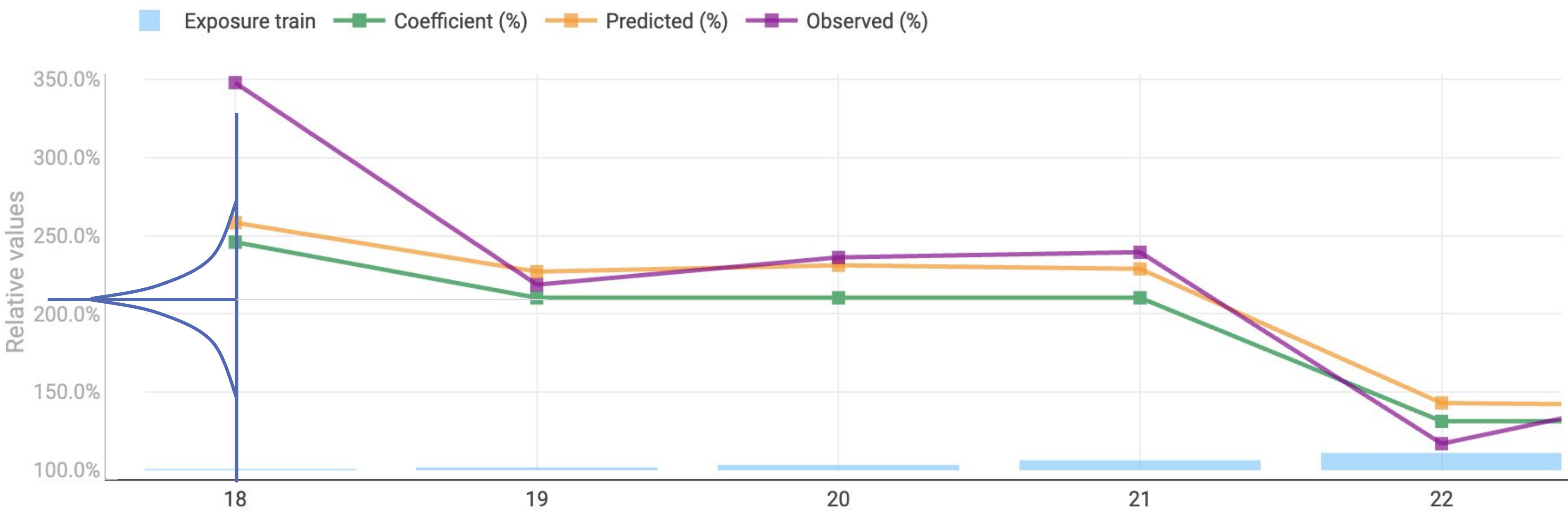
Stronger Prior ⇔ Weaker reliance on the observation

The final model is an average between the most likely coefficients according to the prior and the observations



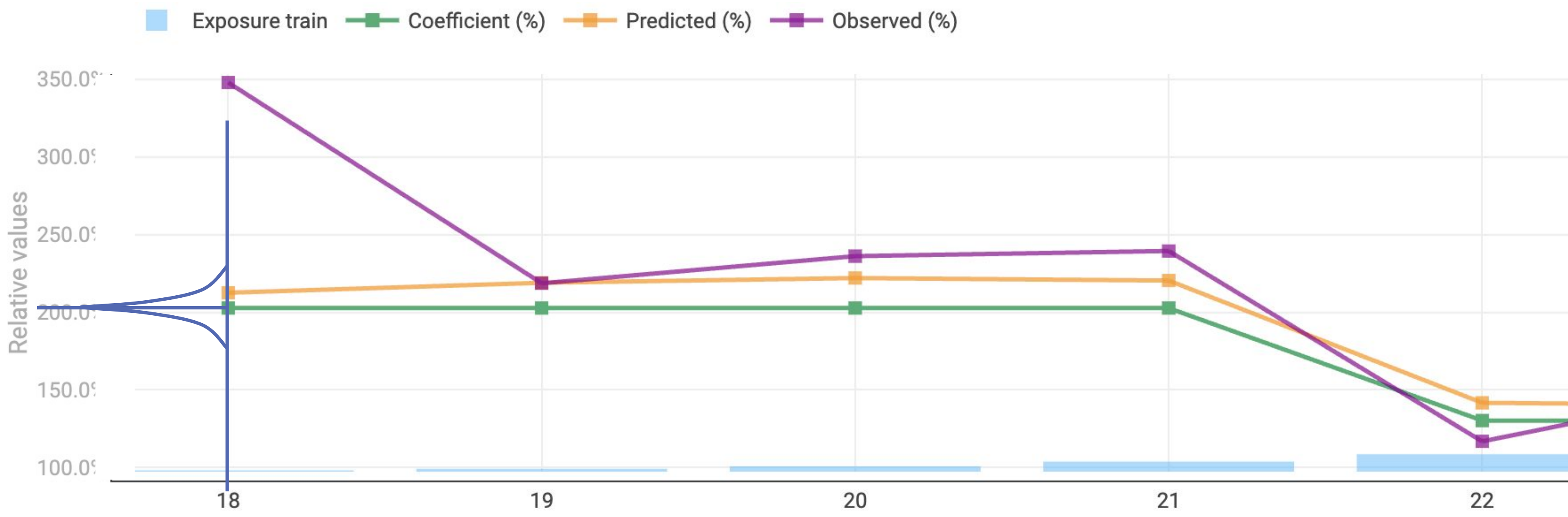
Strong Prior ⇔ Very weak reliance on the observation

The weight of the observation in the model is weaker than the priors



Very Strong Prior ⇔ Full reliance on the prior

The observations can't disprove such a strong prior - more data would be needed



AutoML brings the pricing practitioner at the centre of the decision making process

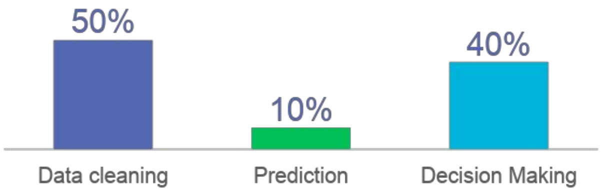
AutoML

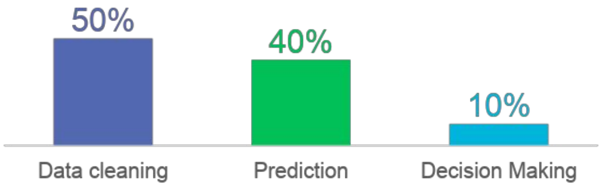
- More choices of models
- Less time is required on modelling work
- Focus on business interpretation

Traditional pricing

- Limited numbers of models
- Large proportion of time spending on modelling
- Limited time on business interpretation

Effort allocation





AutoML does **not** mean that less actuaries or modelers are required, but frees them from iterative modelling work and empowers them to allocate more time on understanding business requirement and modelling result interpretation.

Thank you for attending!

If you have questions feel free to reach out!

monica.carvajal-pinto@akur8.com
Timbert@munichre.com

AKUR8