AKUR8

Transparent Models
with Machine-Learning

2022

Guillaume Béraud-Sudreau
Chief Actuary & Co-Founder of Akur8

Biography

Guillaume is the **Chief Actuary** and **Co-Founder** of Akur8.

He has both a data science and an actuarial background.

Guillaume started researching the potential of AI for insurance pricing as **Head of Pricing R&D** at AXA Global Direct, before being incubated at Kamet Ventures and founding Akur8.

Guillaume is a **Fellow** of the French Institute of Actuaries and holds Master's degrees in **Actuarial Science**, **Cognitive Science** and **Engineering** from Institut des Actuaires, Ecole normale supérieure, and Télécom Paris.

AKUR8

Actuarial Modeling

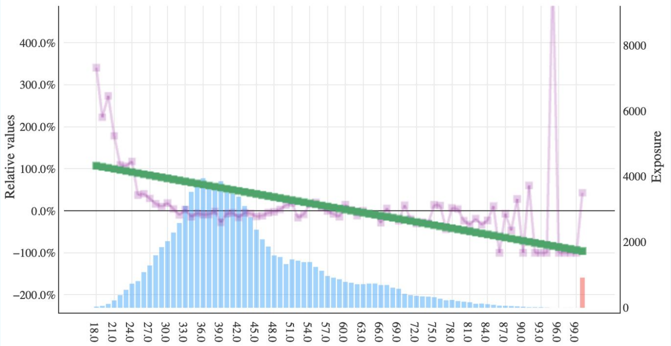
AKUR8

Actuarial Modeling: Capturing Non-Linearities

What GLMs Offer...

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

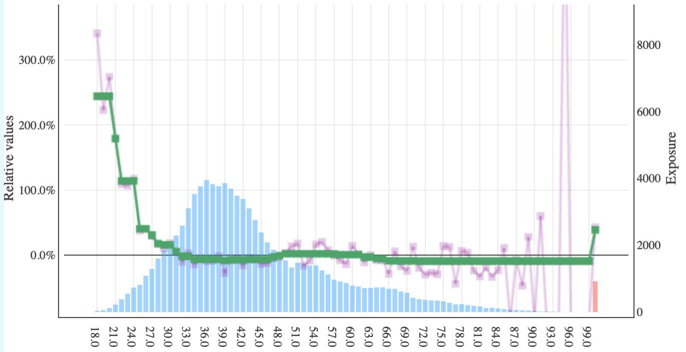
They are easy to fit (as only one parameter has to be found for every variable).



...What we want

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

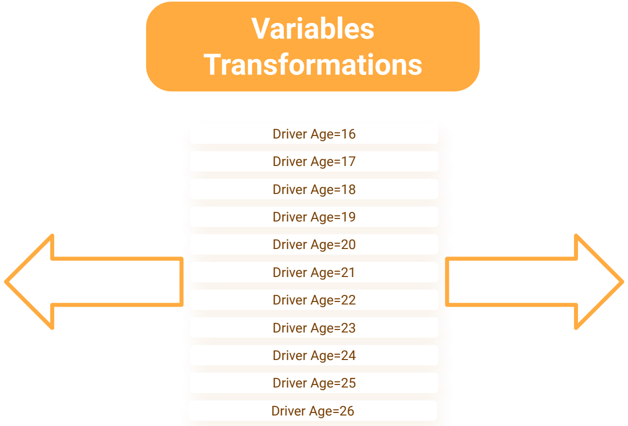
They are hard to fit because, for every variable, a large number of parameters has to be found.



GLMs and Additive Models equivalence

Linear Models

$$\hat{y}(X) = g^{-1}\left(\sum_{i,j} \beta_{i,j} \times I_{X_i=j}\right)$$



Non-Linear Models

$$\hat{y}(X) = g^{-1}\left(\sum_{j} \beta_i(X_i)\right)$$

$$\beta_i(X_i) = \sum_j \beta_{i,j} \times I_{X_i=j}$$

GLMs and Additive Models are equivalent: coefficients are built for different values of the explanatory variables.

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

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

Creating a GLM to capture non-linear relationships

All regression models are built around the same main principle:

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

However, **maximizing the likelihood** on hundreds of parameters would lead to overfitting, which needs to be controlled.

Two main approaches are used by the actuarial community:

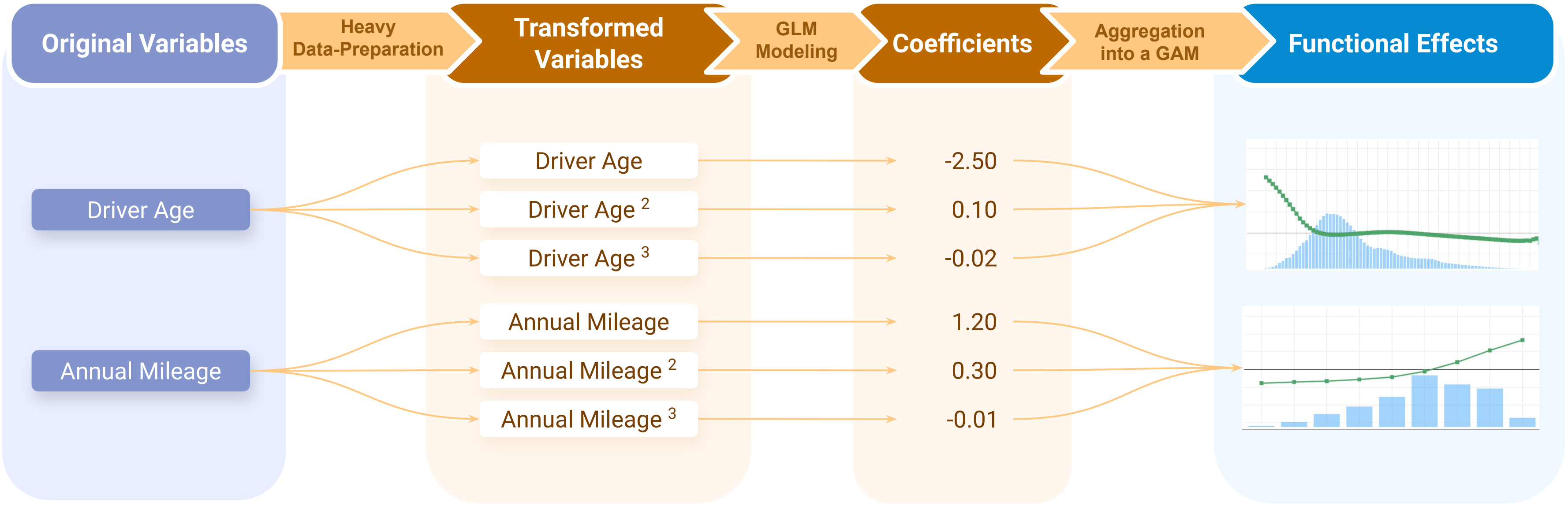
Manage the number of parameters by carefully **selecting which transformations are used**:

- Polynomials
- Groupings
- …

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.

Modeling with variable transformations



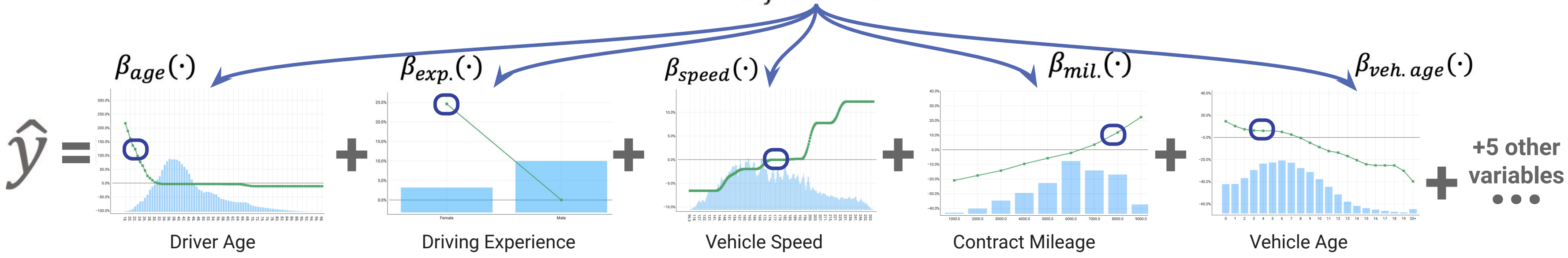
Creating a GLM, visualizing an Additive Model

The models created are GLMs: $$\hat{y}(X) = g^{-1}\left(\sum_{i,j} \beta_{i,j} \times I_{X_i=j}\right)$$

These models can be visualized either as tables (containing all the $\beta_{i,j}$, which can be useful for instance to put the model into production), or as GAMs (Generalized Additive Models).

The **GAM visualization is convenient** for model review and modification as it displays one function per variable.

$$\hat{y}(X) = g^{-1}\left(\sum_{j} \beta_i(X_i)\right)$$



Leveraging Credibility

AKUR8

Creating a GLM to capture non-linear relationships

All regression models are built around the same main principle:

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

However, **maximizing the likelihood** on hundreds of parameters would lead to overfitting, which needs to be controlled.

Two main approaches are used by the actuarial community:

Manage the number of parameters by carefully **selecting which transformations are used**:

- Polynomials
- Groupings
- …

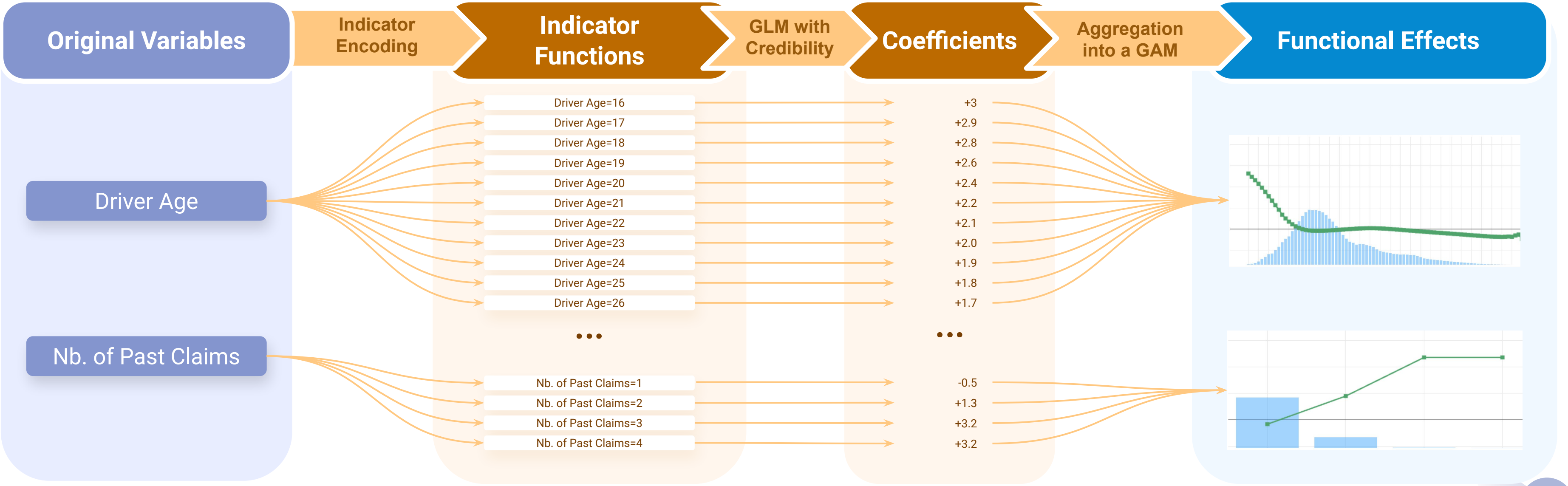
Integrate priors on the coefficients into the model creation:

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

Automatic Modeling with Credibility

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.



Automatic Modeling with Credibility

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** to ensure the coherence between the different coefficients created.

Quick Reminder... 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 (after equations involving conditional probabilities), the output of a credibility approach is that the model 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).

Prior and Credibility

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

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

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

Taking the log of this formula provides an “easy-to-optimize” log-likelihood function:

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

Prior ⇔ Penalized Regressions

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) allows for controlling the overfitting of the models.

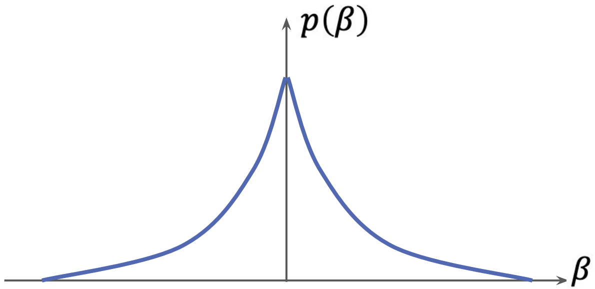
Gaussian Hypothesis ⇔ **Prior:** Coefficients follow a Normal distribution N(0, 1/2λ): ⇔ **Coefficients Distribution:** $p(\beta) \sim e^{-\lambda\,\beta^2}$ ⇔ **Log-Likelihood (incl. prior)** $LL(x, y, \beta) - \lambda\ \beta^2$ ⇔ **Ridge Regression**

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**

Lasso and Hypothesis testing

Lasso is especially popular as it is a good tool for variable selections: models created with the Lasso framework are sparse - all the non-relevant coefficients equal zero.

The Laplace distribution that underlies the Lasso has a maximum 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.

Back to the original problem…

We want to use a GLM leveraging credibility to fit many of coefficients and create a model: $\hat{y}(X) = g\left(\sum_j \beta_{i,j} \times I_{x_j=i}\right)$

Lasso can be used in actuarial modelling…

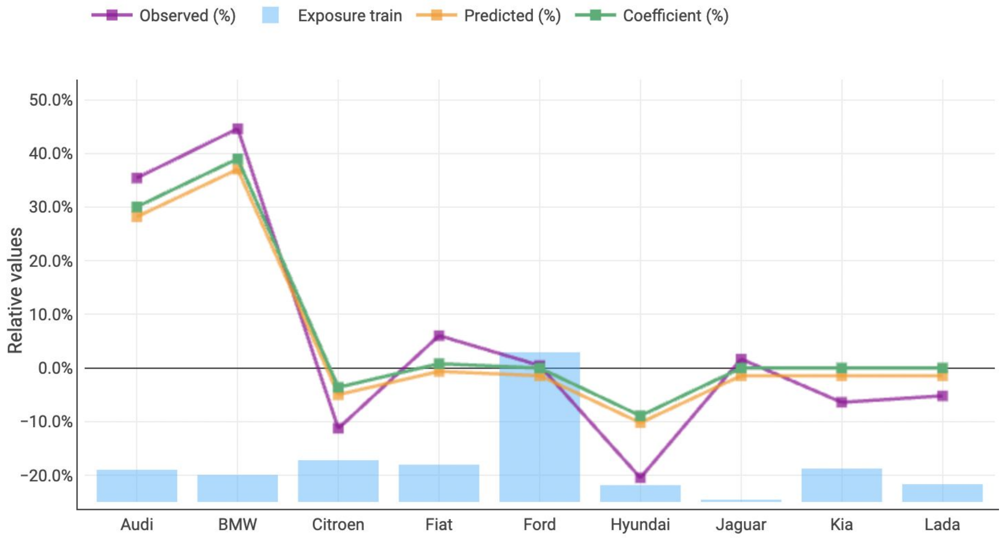
Lasso can be used to capture the signal on **categorical variables**.

Coefficients are created for each level of the data:

$$\hat{y}(X) = g\left(\sum_j \beta_{i,j} \times I_{x_j=i}\right)$$

The result is coherent with a **credibility approach**: predictions are between their “pure GLM” values and the grand-mean of the observations.

Non-significant levels are grouped, with null coefficients.



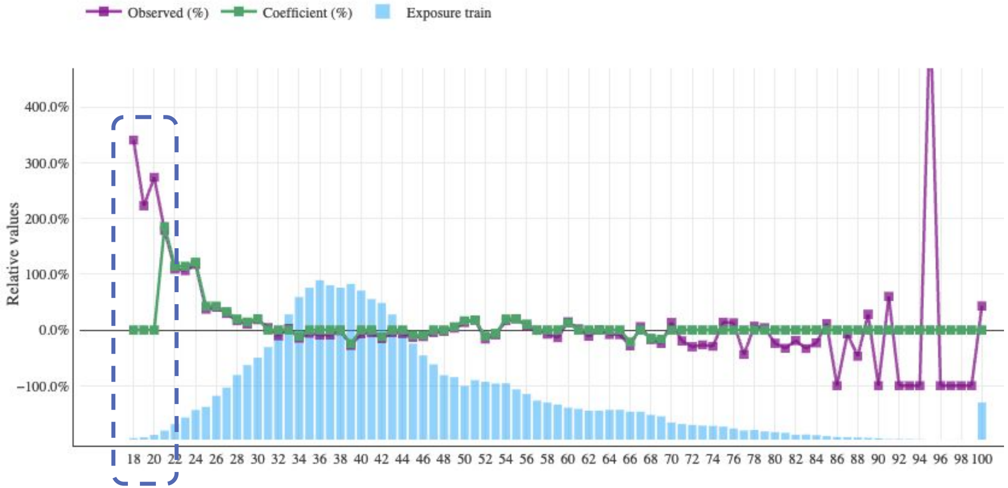
…but Lasso does not capture non-linear effects!

While it is very powerful and well documented, the **Lasso can't be directly applied** to indicator- representation on the data to create a non-linear model:

$$\hat{y}(X) = g\left(\sum_j \beta_{i,j} \times I_{x_j=i}\right)$$

All non-significant coefficients would be grouped at zero, which makes no sense.

A key piece of information: **the order of the levels would be lost in the process**.



No information in the data =
The most likely coefficients
are at zero.

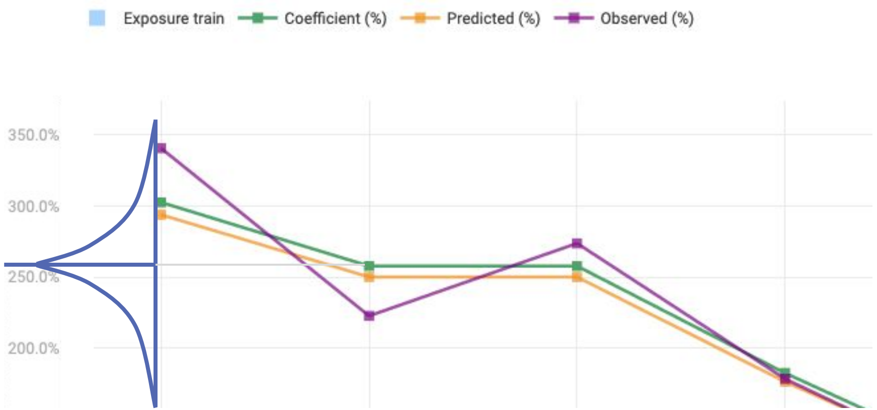
Credibility on Ordered Variables

Creating new Priors and Penalties

New priors have to be considered to take into account the structure of the models created.

In particular, for ordinal variables, two consecutive coefficients should:

- **be more likely to be close than far apart** if they are significantly different.
- or **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.

Creating new Priors and Penalties

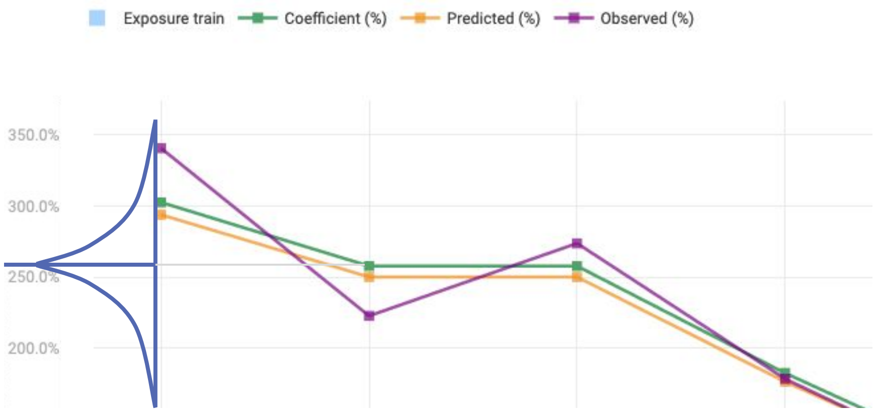
This means that the **derivative of the coefficient function** $\beta'(X)$ **follows a Laplace distribution**:

As the values of the coefficients are discrete, the derivative can be written as:

$$p(\beta)\ \alpha\ e^{-\lambda\,|\beta_i-\beta_{i+1}|}$$

This distribution of probability is used as a prior when maximizing the likelihood to fit a model:

$$\beta^* = Argmax_\beta\ LL(x,y,\beta) - \lambda\,|\beta_i - \beta_{i+1}|$$



Controlling the Prior distribution

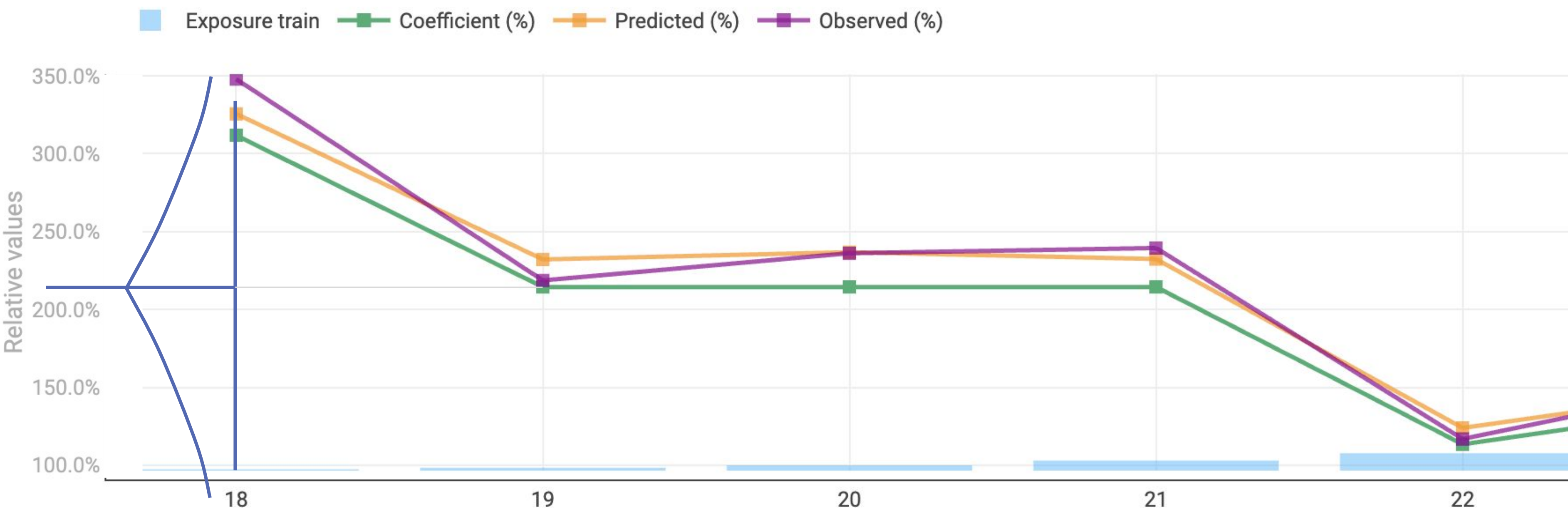
The prior follow a distribution $p(\beta) \propto e^{-\lambda\,|\beta_i - \beta_{i+1}|}$ of variance $2/\lambda^2$

The coefficients should **maximize**: $LL(x, y, \beta) - \lambda\,|\beta_i - \beta_{i+1}|$

Large λ	Prior distribution has a **small variance**.	**Strong a-priori knowledge** on the model.	Large weight is given to the ***smoothness term***.	A **smooth model** is created.
Small λ	Prior distribution has a **large variance**.	**Weak a-priori knowledge** on the model.	Large weight is given to the ***observations term***.	A **noisy model** is created.

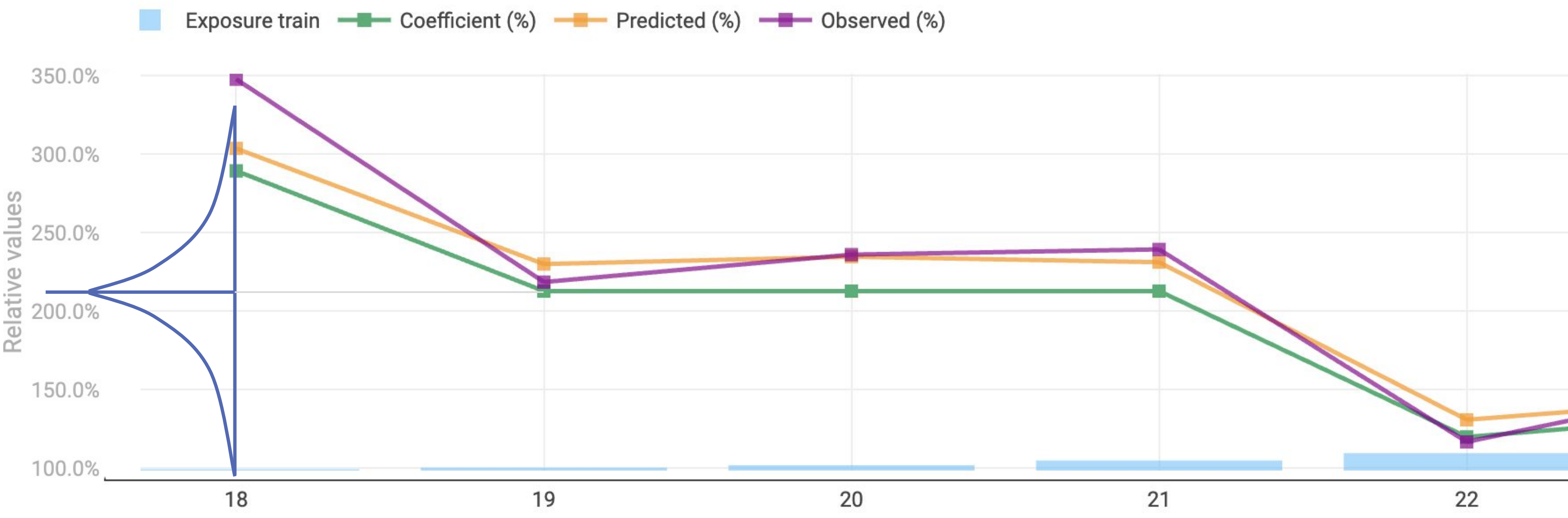
Weak Prior ⇔ Strong reliance on the observation

The prior has a very limited impact on the final model



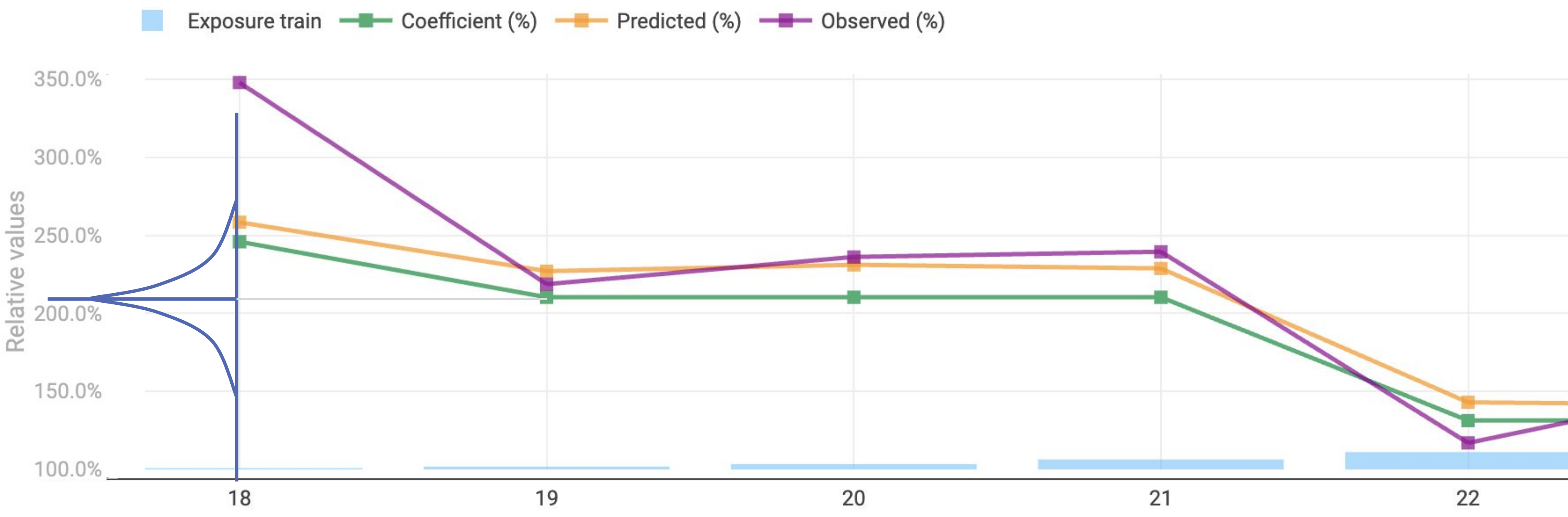
Stronger Prior ⇔ Weaker reliance on the observation

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



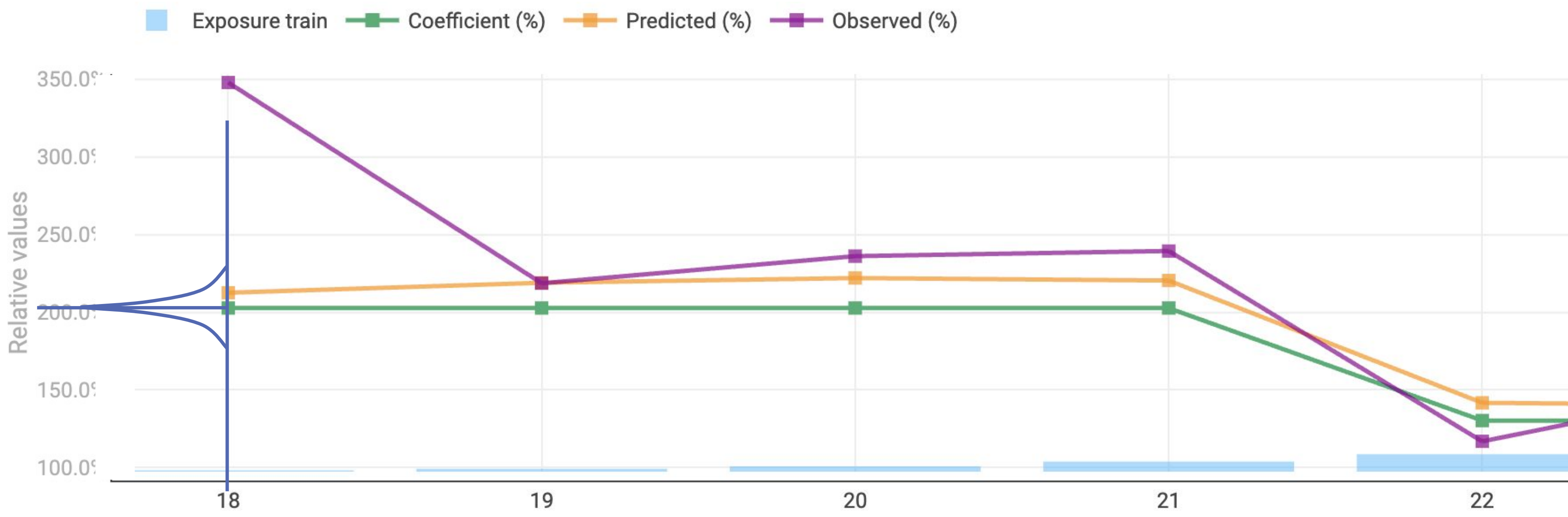
Strong Prior ⇔ Very weak reliance on the observation

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



This is equivalent to failing a significant test against the null hypothesis: "the first two coefficients are equal".

A stronger effect - or more exposure - would be necessary to disprove it, and split the coefficients.

Like for a Lasso, this is equivalent to a test!

The behavior is similar to a hypothesis-testing approach:

A priori, we suppose the null-hypothesis: $\beta_{i+1} - \beta_i = 0$

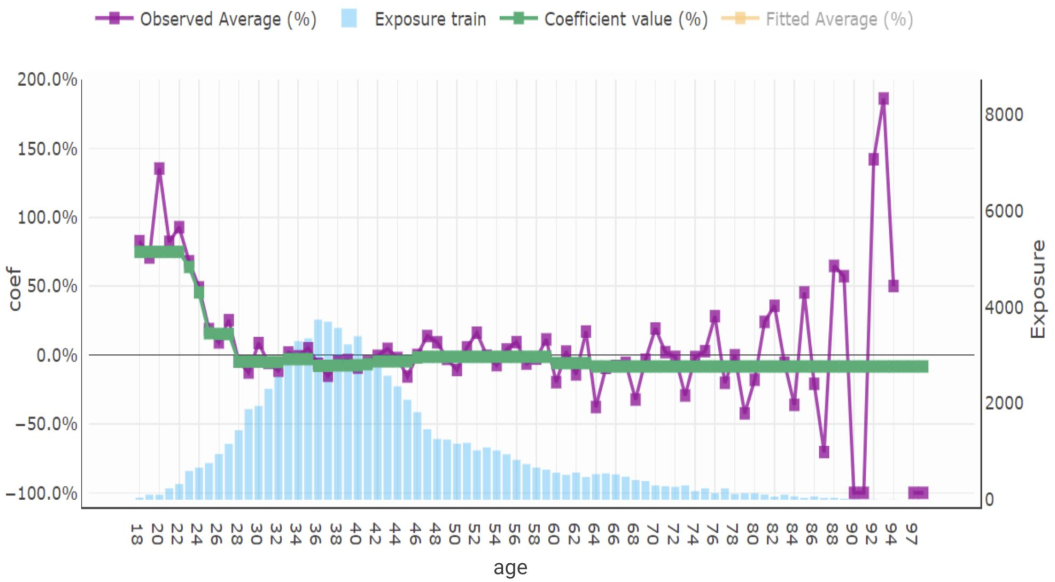
This null hypothesis is tested with the data, and potentially rejected.

This null hypothesis is equivalent to: $\beta_{i,j} = \beta_{i+1,j}$

- If it is not rejected by the data, then the coefficients function is locally constant.
- If it is rejected by the data, then the coefficients function is not constant.

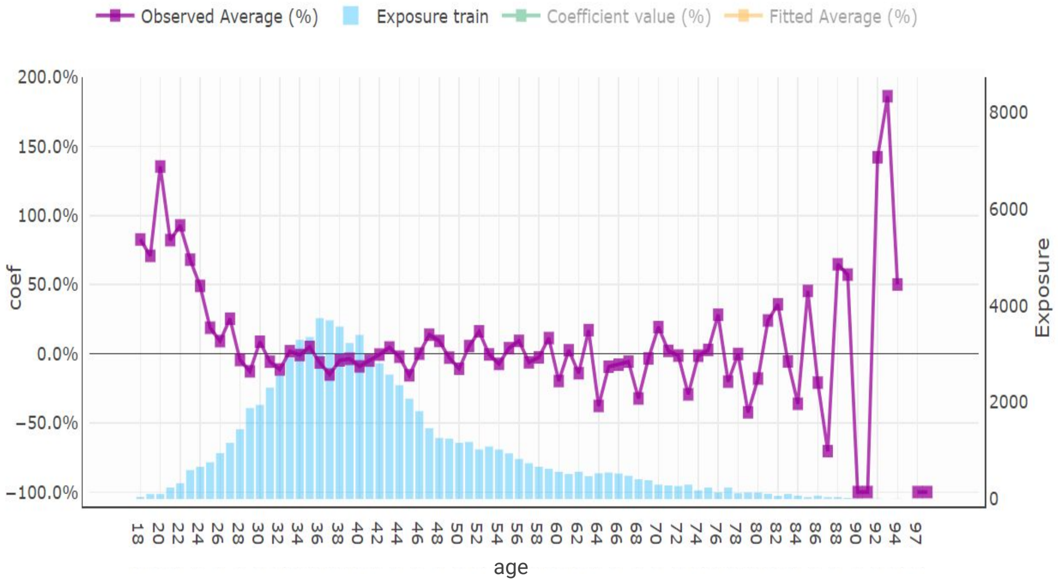
Leveraging the prior on a full model scale

A more **balanced prior**
(with a medium variance)
leads to more **sensitive models**.



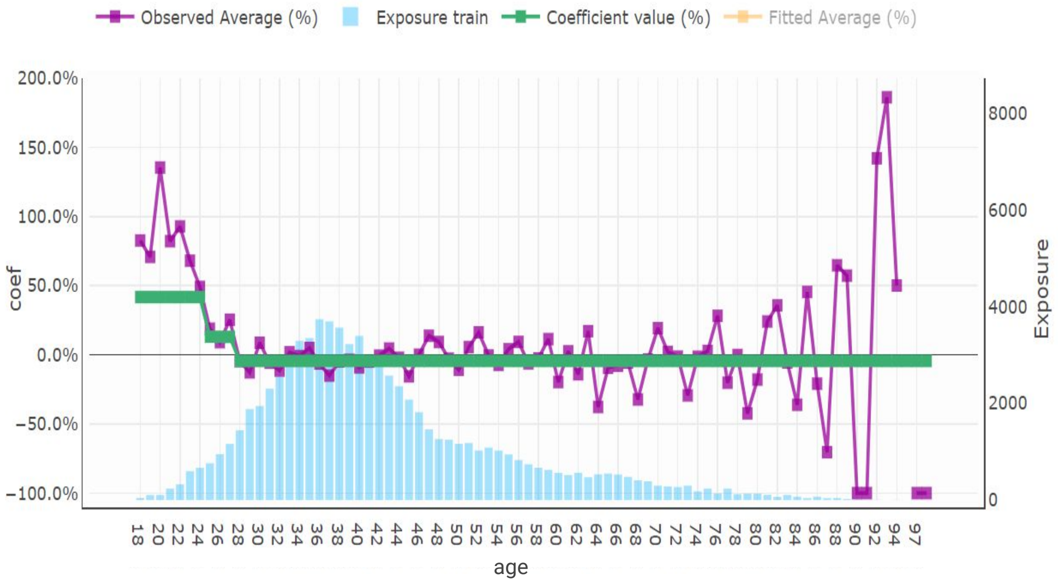
Leveraging the prior on a full model scale

Data used to create the models are **naturally noisy.**



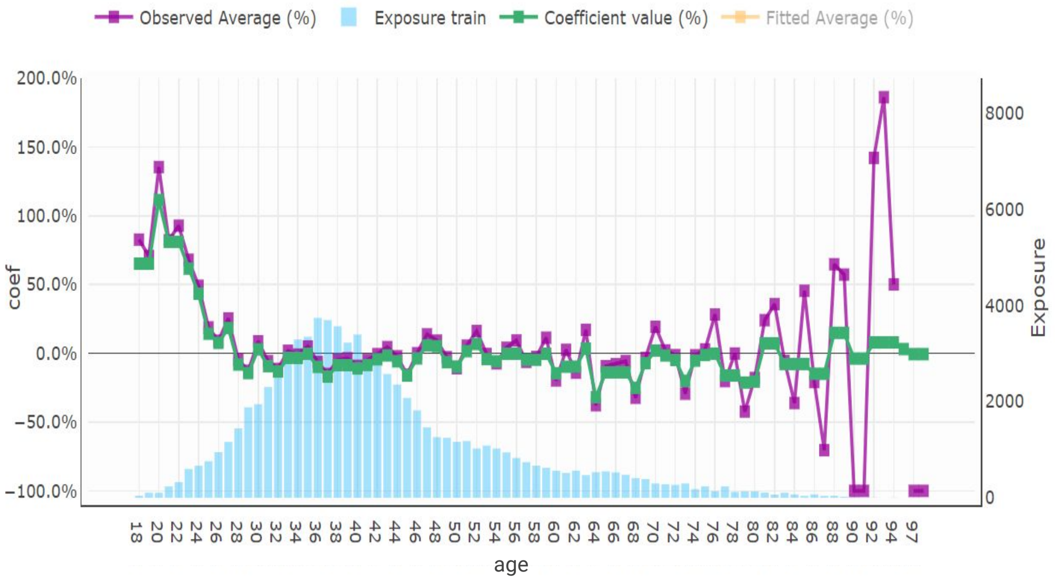
Leveraging the prior on a full model scale

A very **strong prior** (with a small variance) leads to **robust models**.



Leveraging the prior on a full model scale

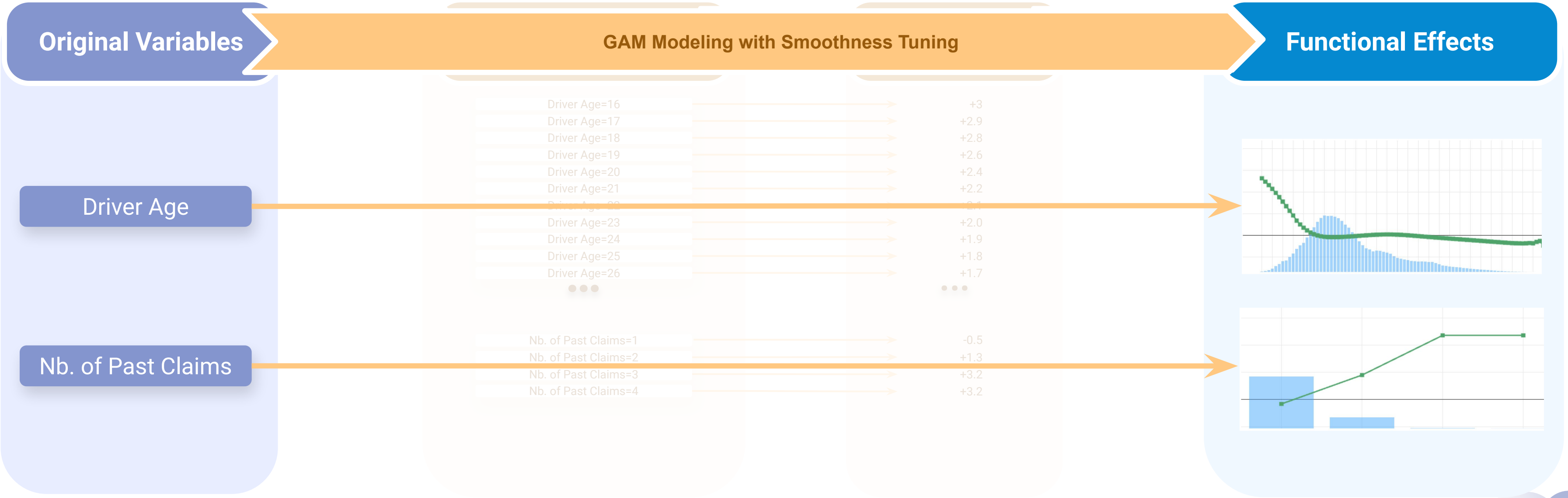
A very **weak prior** (with a large variance) leads to **noisy models**.



Machine-Learning = GLM and Credibility

From a user's point of view, the creation of the models is **fully automated** and provides a unified machine-learning algorithm. As with all **machine-learning** techniques, the one presented today relies on a **solid statistical basis**.

A similar framework **can be leveraged to achieve variable selection**.



Thank You!

Guillaume Béraud-Sudreau
Chief Actuary & Co-Founder of Akur8
guillaume.beraud@akur8-tech.com

Our new white paper “Credibility and Penalized Regression” is available now at www.Akur8.com under “Resources”

https://akur8.com/resources

AKUR8