AKUR8

qx-Club

20. Februar 2024

CONFIDENTIAL

Jan Küthe

Aktuar (DAV) / Actuarial Data Scientist

Biography

Jan ist ein Aktuar (DAV) aus Siegen und arbeitet bei Akur8 als Actuarial Data Scientist, um Versicherern zu helfen, die Potenziale der Pricing-Methoden des 21. Jahrhunderts nutzbar zu machen.

Zuvor hat er drei Jahre bei einer weltweit tätigen Aktuariellen Beratung gearbeitet. Er hat einen Master of Mathematics an der Universität Bonn abgeschlossen und wohnt in Köln.

Daneben ist er ein begeisterter Leser der Werke von Anna Seghers und Dietmar Dath.

Aktuarielle Modellierung

Ein neuer Credibility-Ansatz für transparente Risikomodelle

Zunächst:
Eine gängige Praxis

um dünn besetzte Ausprägungen zu modellieren

SME Liability example

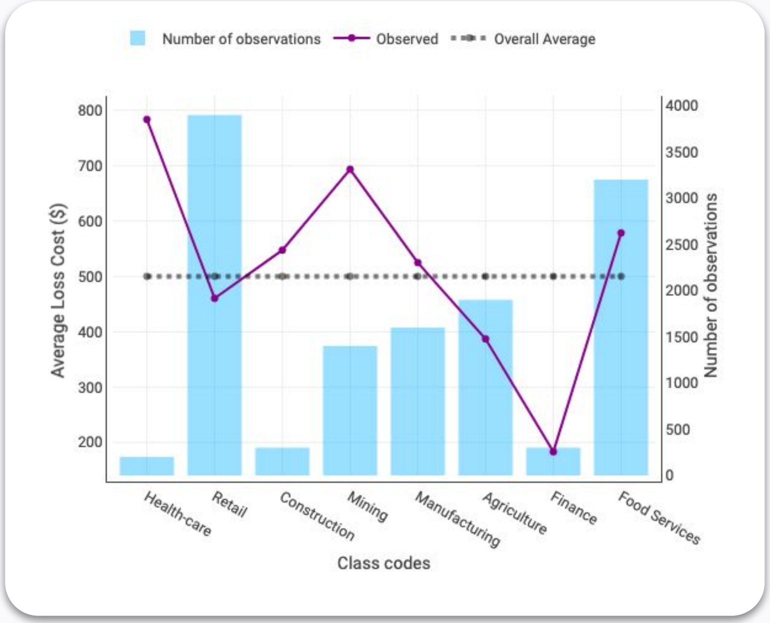
Burning Cost by class code example

Losses and exposures for companies are collected, and we want to compute an estimation of the average loss cost per class code.

The data **can be represented visually**:

- The **blue bars** represent the number of observations for a given class;
- The **purple lines** represent the **Observed Experience** as the average loss cost for each class;
- The **black line** represents the **overall average** (or grand average) of $500 in this example.

Observed loss by class code



GLMs: Univariate estimate

A natural estimate is the average loss cost by class code.

Such estimate may be inappropriate for class Health-Care which has low exposure.

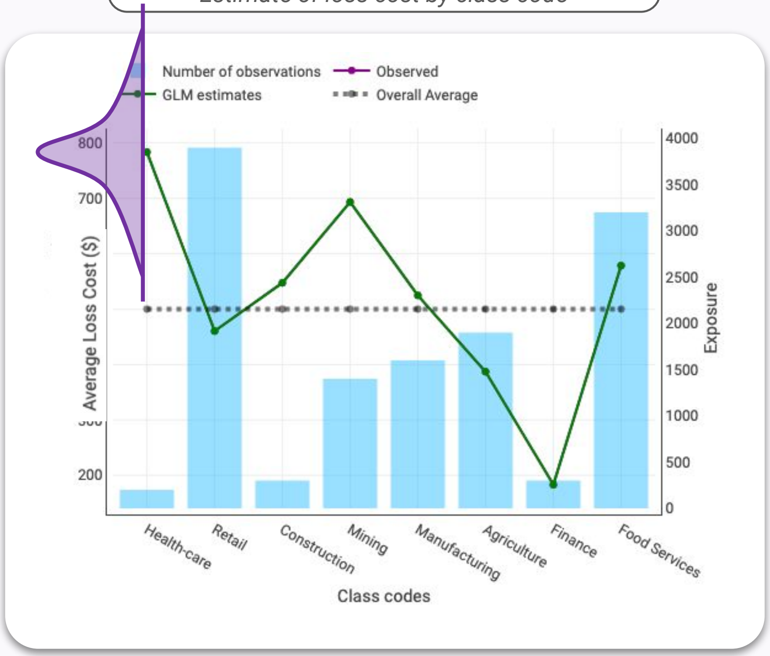
The same argument applies for Finance and Construction.

This approach is followed in the GLM framework, that fully trusts the data:

$$\beta^* = Argmax\ Likelihood(Obs., \beta)$$

In many cases (for instance Poisson-LogLink or Gaussian-IdentityLink) the maximum of likelihood matches the average.

Estimate of loss cost by class code



GLMs: Univariate estimate

A natural estimate is the average loss cost by class code.

Such estimate may be inappropriate for class Health-Care which has low exposure.

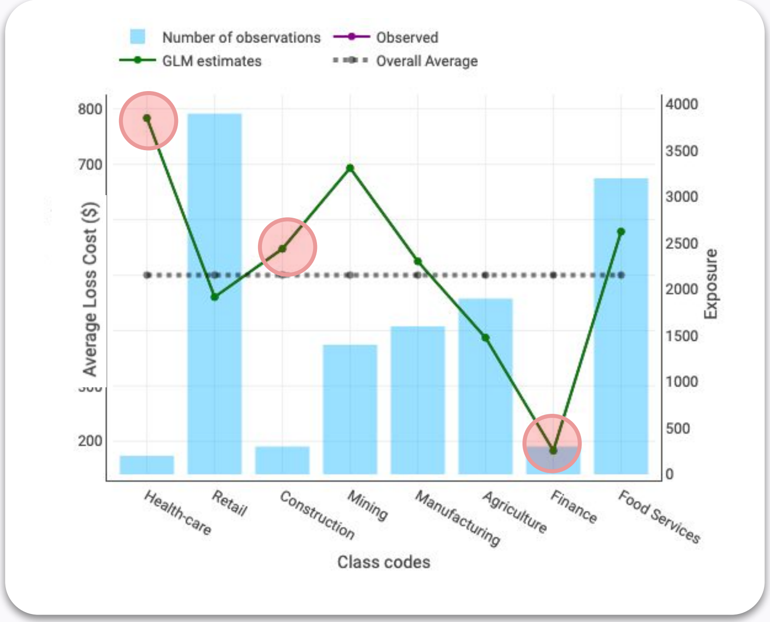
The same argument applies for Finance and Construction.

This approach is followed in the GLM framework, that fully trusts the data:

$$\beta^* = Argmax\ Likelihood(Obs., \beta)$$

In many cases (for instance Poisson-LogLink or Gaussian-IdentityLink) the maximum of likelihood matches the average.

Estimate of loss cost by class code



Eine gängige Praxis

ist das Entfernen dieser Ausprägungen aus dem Modell

Removing low-significance levels

A classic approach is to use the **statistical significance** of the different levels.

Levels that have low exposure (or small effects) are grouped together, or put at the average value.

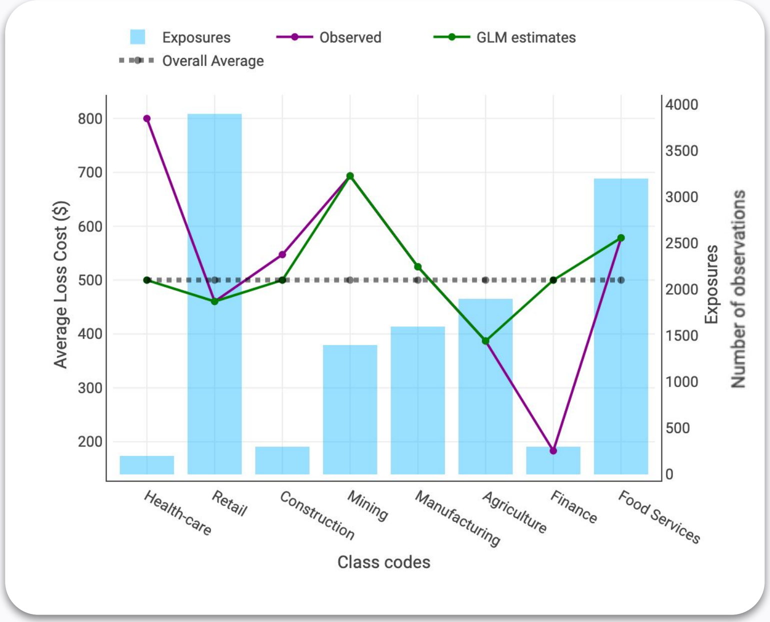
The goal of this approach is to avoid trusting very noisy models with a few observations.

The result obtained will depend on the **significance threshold** above which levels will be kept into the final model or grouped:

- If a level is **more significant** than the threshold, it is **kept;**
- If a level is **less significant** than the threshold, it is **removed.**

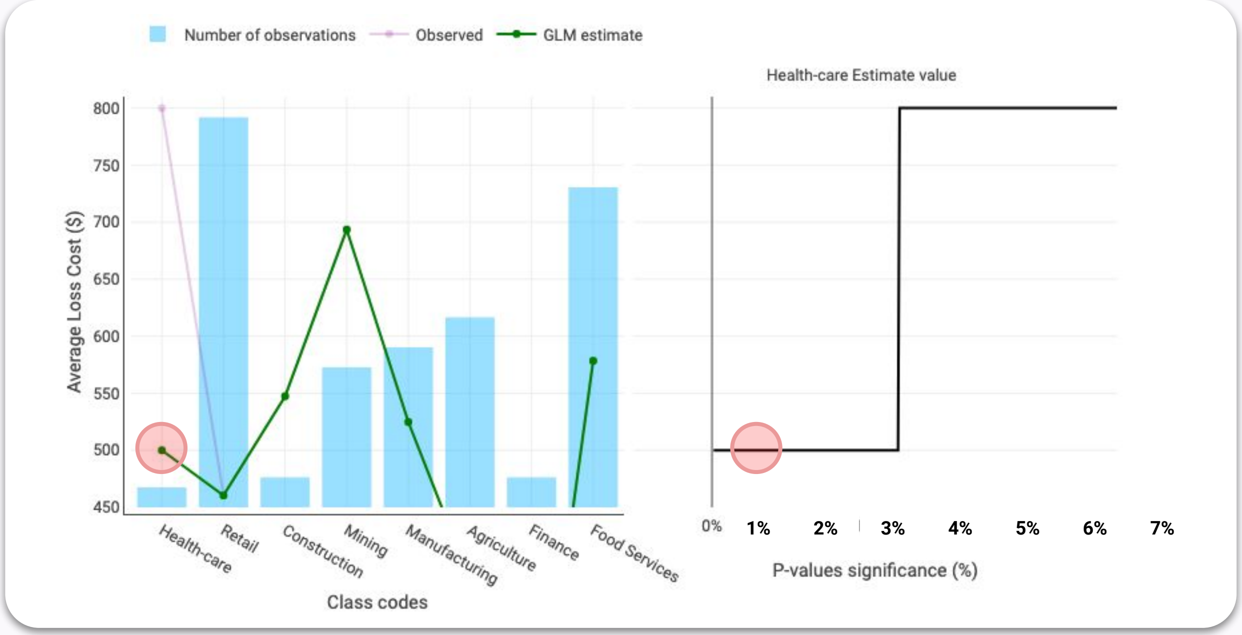
Modelers often use a “5% significance level” but any other value can be selected.

Estimate of loss cost by class code



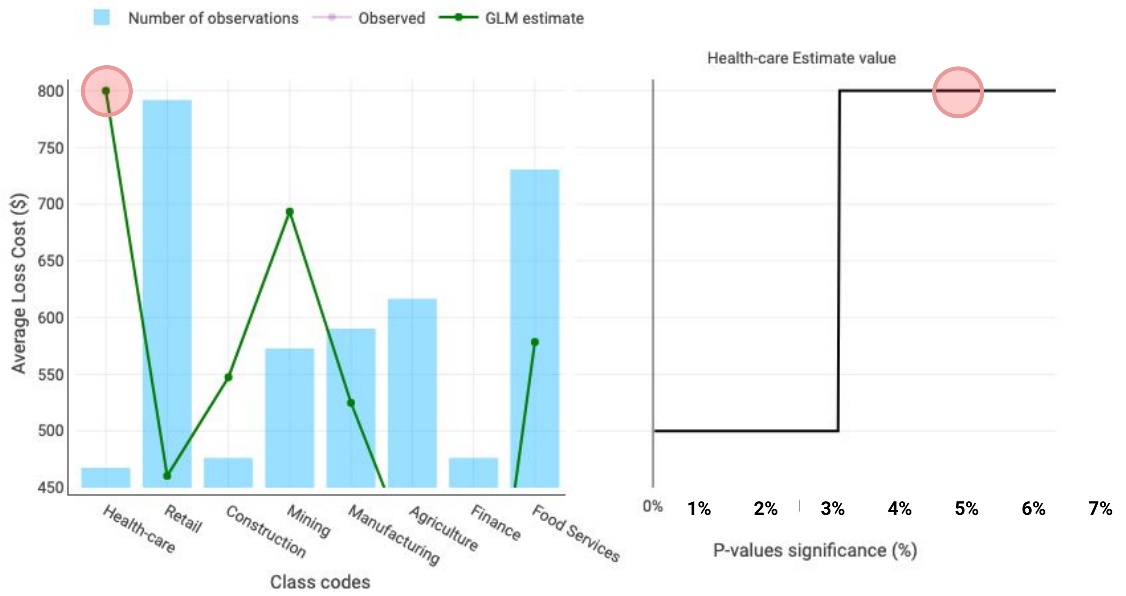
Fitted model depends on the threshold

Strong (low) significance thresholds are hard to validate and lead to a **robust** model.



Fitted model depends on the threshold

Weak (high) significance threshold are easy to validate and lead to a **volatile** model.

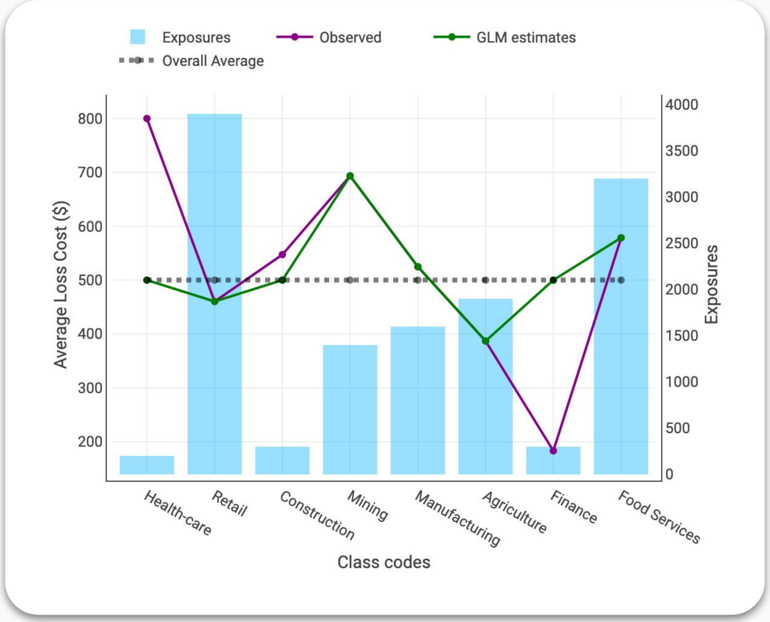


Strengths & limits of levels selection

This approach has well know strengths and limits:

- ✅ It is a binary method, leading to clear decisions;
- ✅ It is very frequently used and widely accepted;
- ✅ It relies on very classic statistics.
- ❌ It is a binary method: it does not use efficiently the limited observations we have on “health-care”;
- ❌ Tests justification rely on hypothesis often not met in practice.

Estimate of loss cost by class code



Geht das nicht besser?

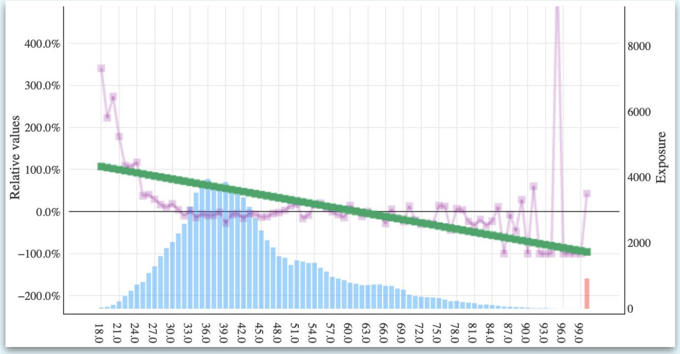
Und wie ist das mit stetigen Variablen?

Actuarial Modeling: Capturing Non-Linearities

What GLMs Offer…

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

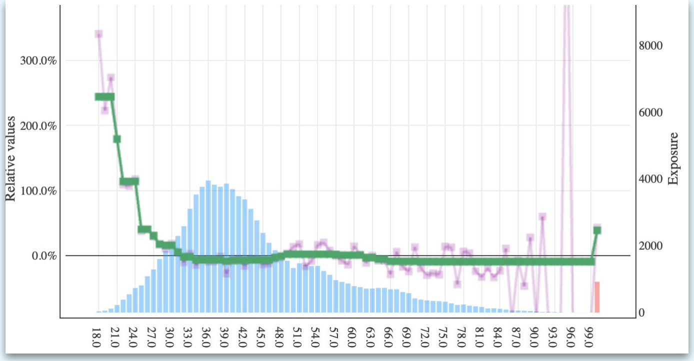
They are easy to fit (as only one parameter has to be found for every variable).



…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.



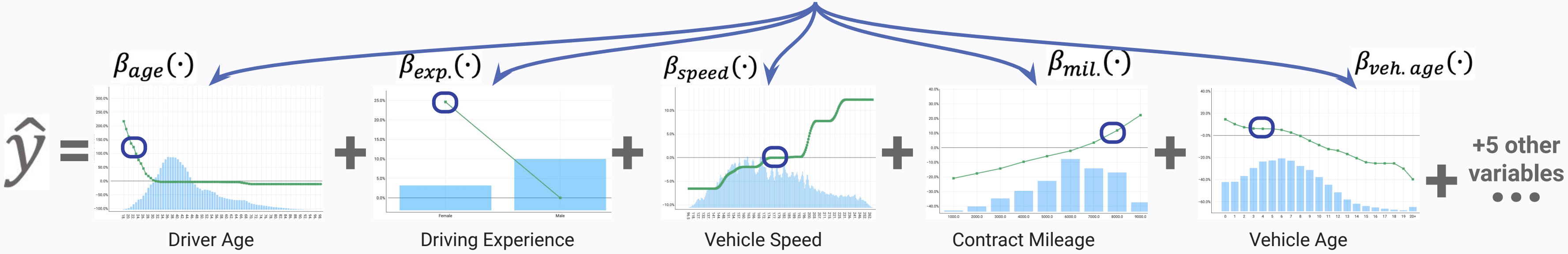
Additives Models: Easy to visualize

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

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.

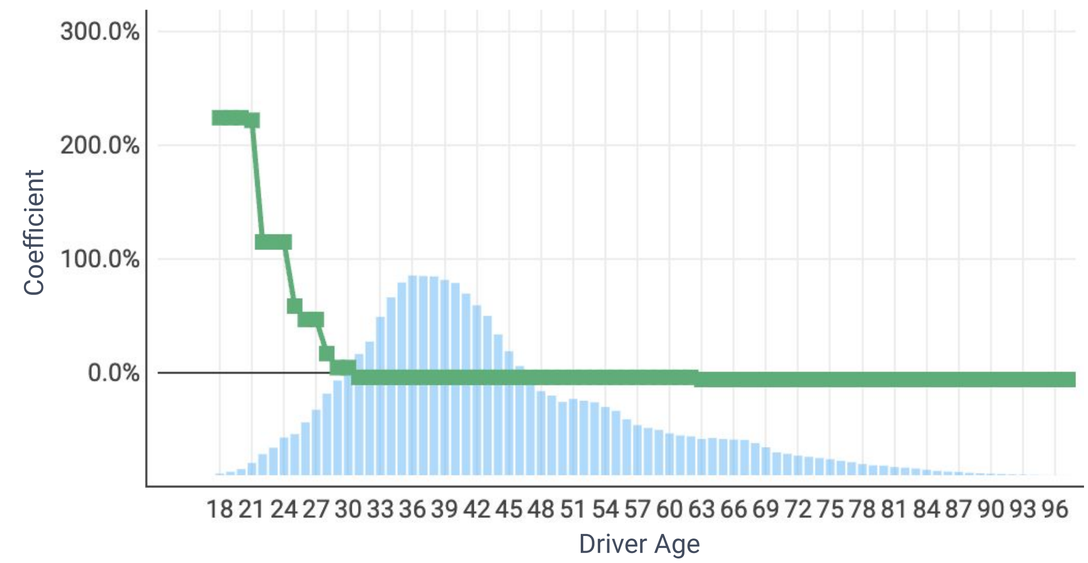
$$\hat{y}(X) = g^{-1}\left(\sum_{d} \beta_d(X_d)\right)$$



Representing Additive Models with Tables

Driver Age	Coefficient
18	224%
19	224%
20	224%
21	222%
22	115%
23	115%
24	115%
25	59%
26	47%
27	47%
28	17%
29	4%
30	4%
31	-4%
32	-4%
33	-4%
34	-4%
35	-4%
36	-4%

The function $\beta_{age}(\cdot)$ is represented by all its values, for all values of the driver age.



GLM and GAM equivalence

Linear Models

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

Variables Transformations

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

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

Non-Linear Models

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

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

The Maximum of Likelihood

All regression models are built around the same main principle:

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

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

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

(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 (with the “errors” are actually defined as minus the log-likelihood):

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

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

Viele manuelle Transformationen

Aber welche ist angemessen?

Creating a GLM to capture non-linear relationships

All regression models are built around the same main principle: to **maximize the probability of the observations**.

$$\beta^* = Argmax_\beta \ p(y|\hat{y})$$

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**:

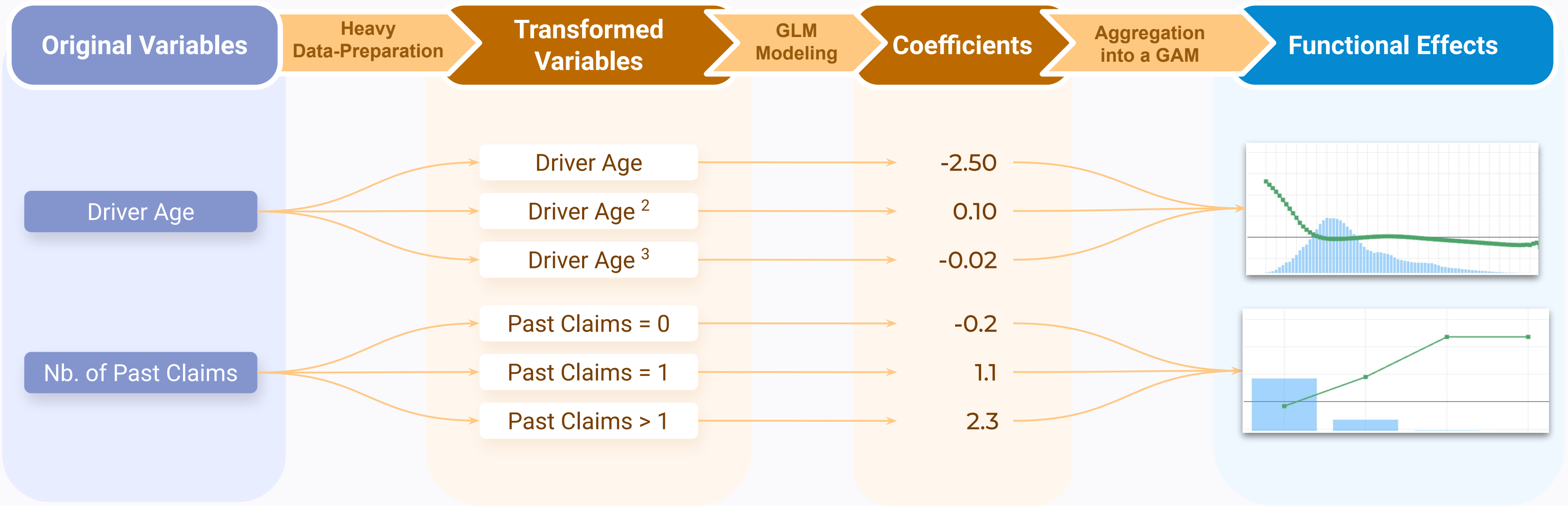
- Polynomials
- Groupings
- …

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

Building the model ⇔ Selecting the transformation

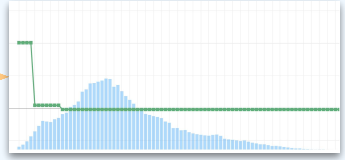


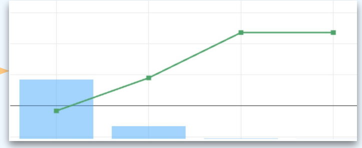
Modeling with variable transformations

Different transformations lead to a different model

Original Variables → Heavy Data-Preparation → **Transformed Variables** → GLM Modeling → **Coefficients** → Aggregation into a GAM → **Functional Effects**

Original Variables	Transformed Variables	Coefficients
Driver Age	Driver Age ≤ 22	1.80
	22<Driver Age<28	0.25
	Driver Age ≥ 28	-0.10
Nb. of Past Claims	Past Claims = 0	-0.20
	Past Claims = 1	1.10
	Past Claims ≥ 1	2.35





Nun zur Credibility!

Und ihrer Verwandtschaft

Creating a GLM to capture non-linear relationships

All regression models are built around the same main principle:

$$\beta^* = Argmax_{\beta}\ Likelihood(y, x, \beta)$$

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
- …

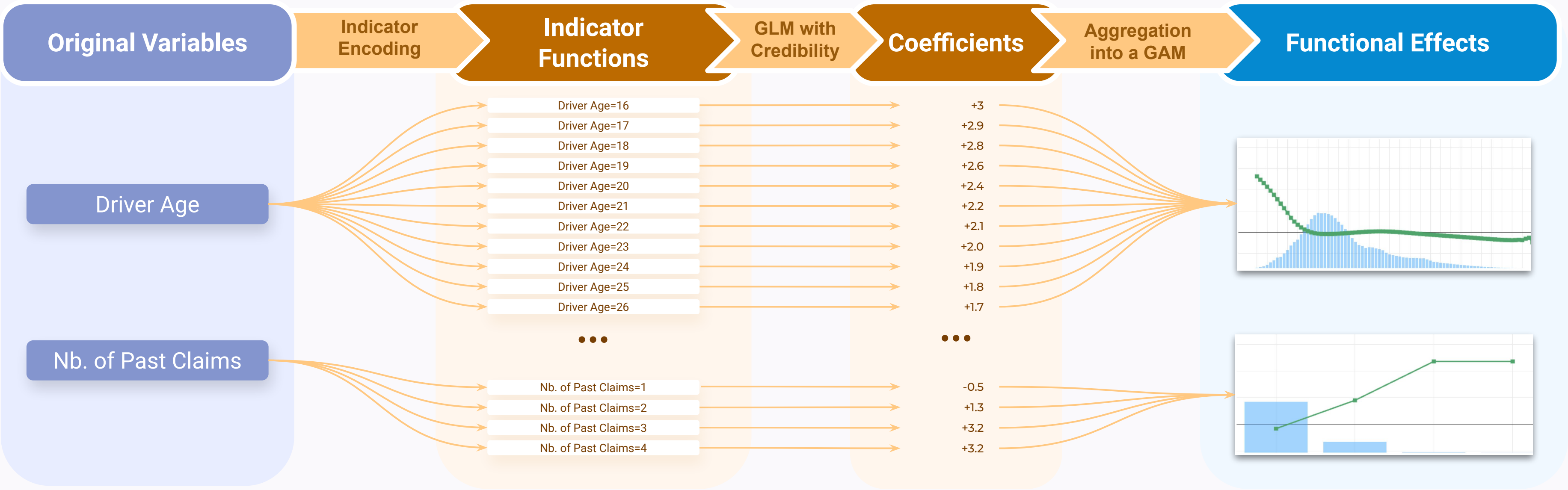
Integrate priors on the coefficients into the model creation:

- 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

+3

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:

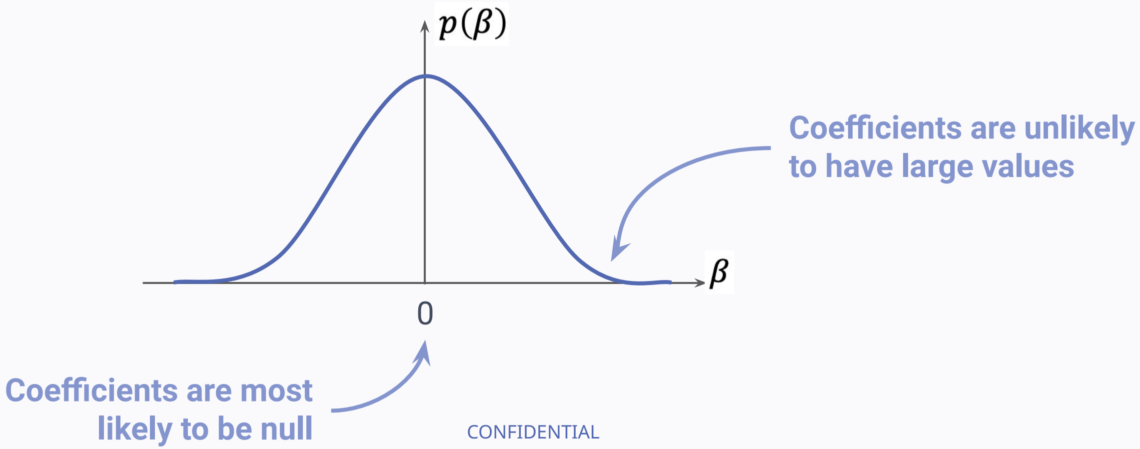
- ➔ the **quantity of data** (the larger the data, the higher the weight)
- ➔ the **strength of the prior** assumptions (a very reliable assumption with small variance will have a large weight).

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”



Prior and Penalized 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 problem (or **Penalized Regression**):

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

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**

Controlling the Prior distribution

With Lasso

In a Lasso, the prior follow a distribution $P(\beta) \sim e^{-\lambda\,|\beta|}$ of variance $^2/_{\lambda^2}$

The coefficients should **maximize**: $LL(x, y, \beta) - \lambda\,|\beta|$

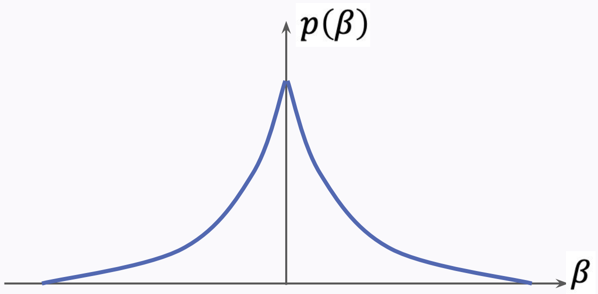
Large λ	Prior distribution has a **small variance**.	**Strong a-priori knowledge** on the model.	Large weight is given to the ***complexity term*** $\lambda\,|\beta|$.	A **simple model** is created.
Small λ	Prior distribution has a **large variance**.	**Weak a-priori knowledge** on the model.	Large weight is given to the ***observations term*** $LL(x, y, \beta)$	A **complex model** is created.

This approach has the **same roots as credibility** (it is based on a Bayesian framework) and the **same consequences** (it “shrinks” the coefficients towards our prior most likely estimate).

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

The Penalized GLM Formula

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

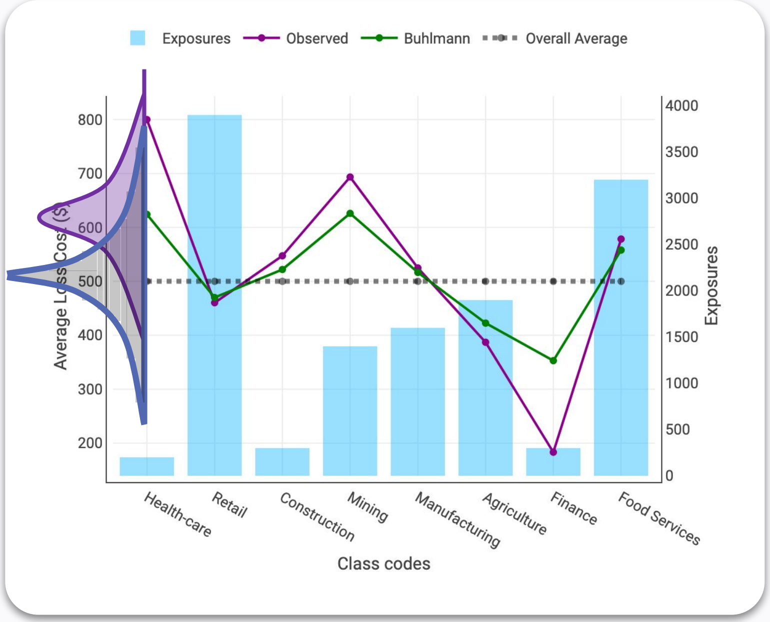
Coefficients are created for each level of the data:

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

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.

Mixing the two distributions



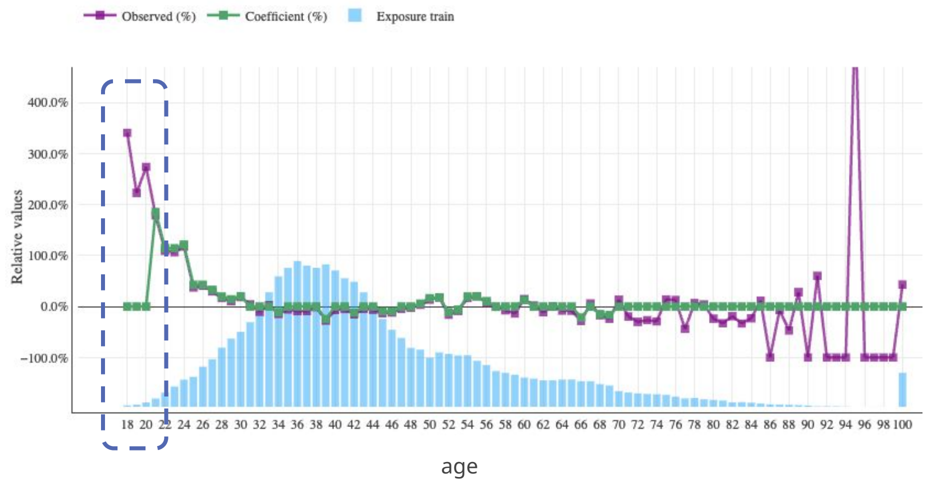
...but Lasso does not capture continuous effects!

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:

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

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 für stetige Variablen

Das *Derivative-Lasso*

Creating new Priors and Penalties

Also see Fused Lasso approaches

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:

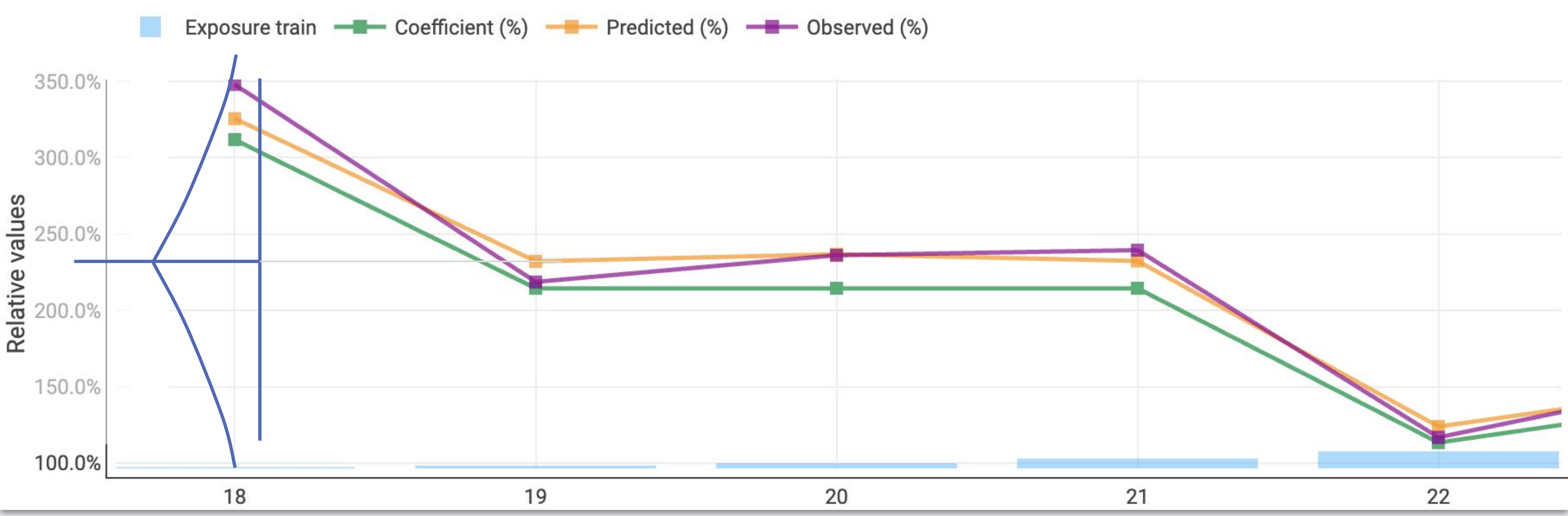
$$p(\beta) \; \alpha \; e^{-\lambda \, |\beta_i - \beta_{i+1}|}$$

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

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

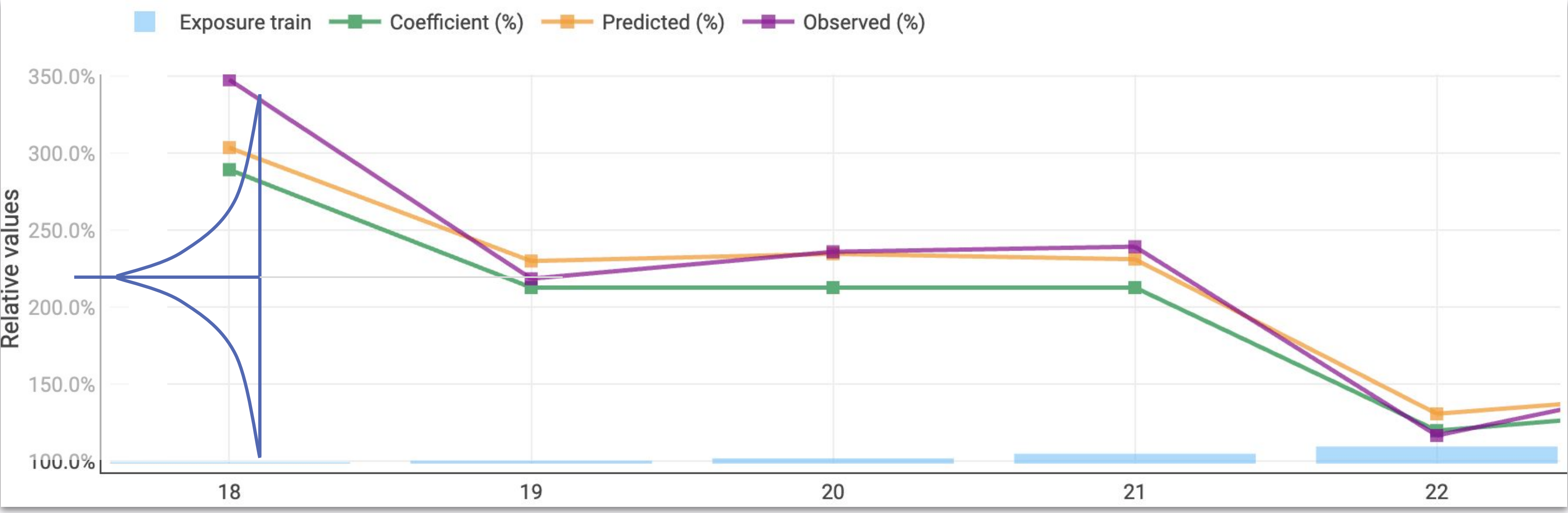
Weak Prior ⇔ Strong reliance on the observation

The prior has a very limited impact on the final model



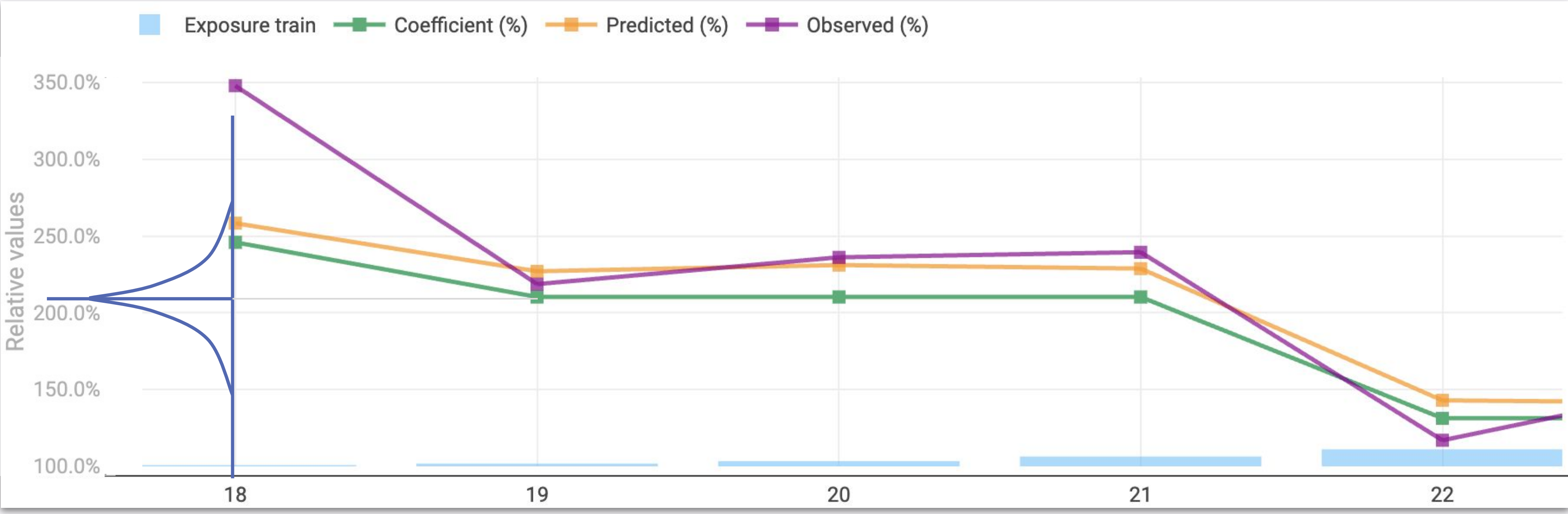
Stronger Prior ⇔ Weaker reliance on the observation

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



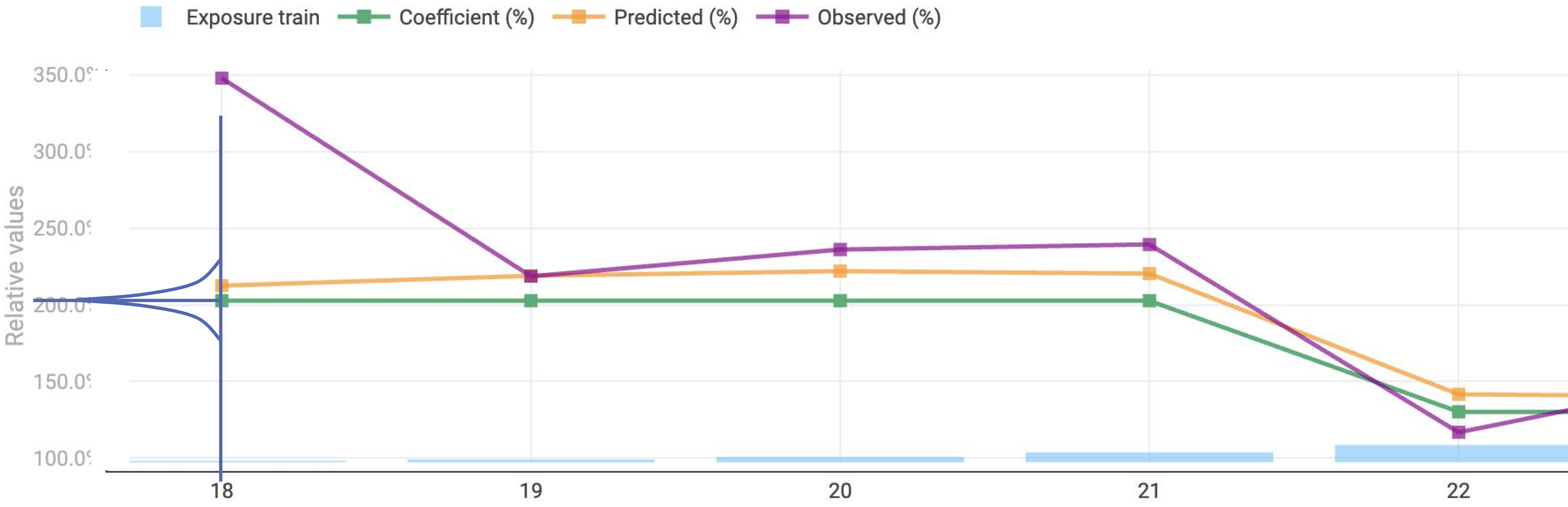
Strong Prior ⇔ Very weak reliance on the observation

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.

Controlling the Prior distribution

To create smooth models

The prior follow a distribution $p(\beta) \, \alpha \, e^{-\lambda \, |\beta_i - \beta_{i+1}|}$ 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.

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: $\beta_{i+1} - \beta_i = 0$

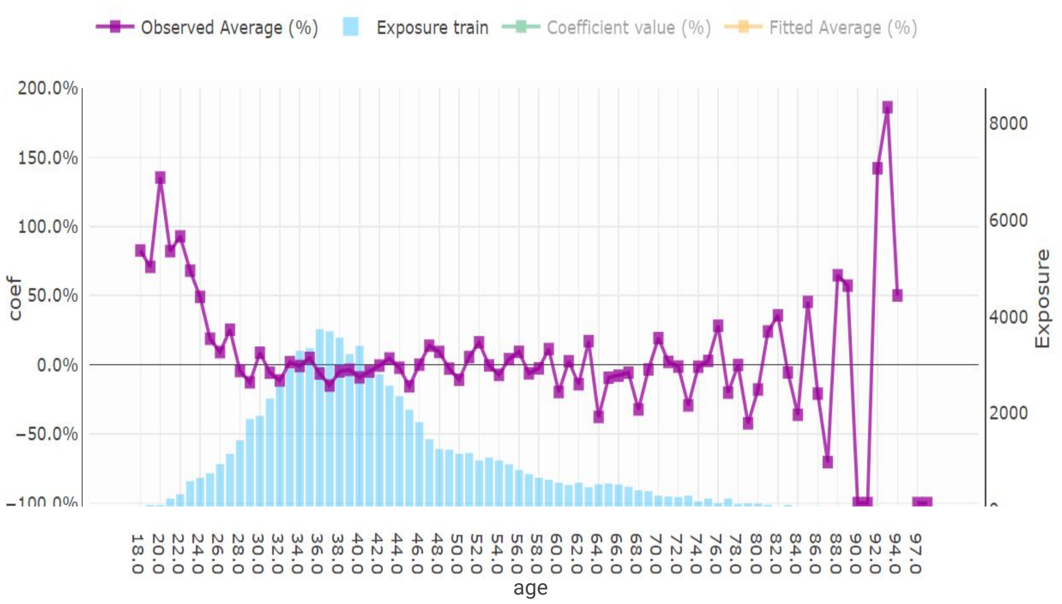
This null hypothesis is tested with the data, and potentially rejected.

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.

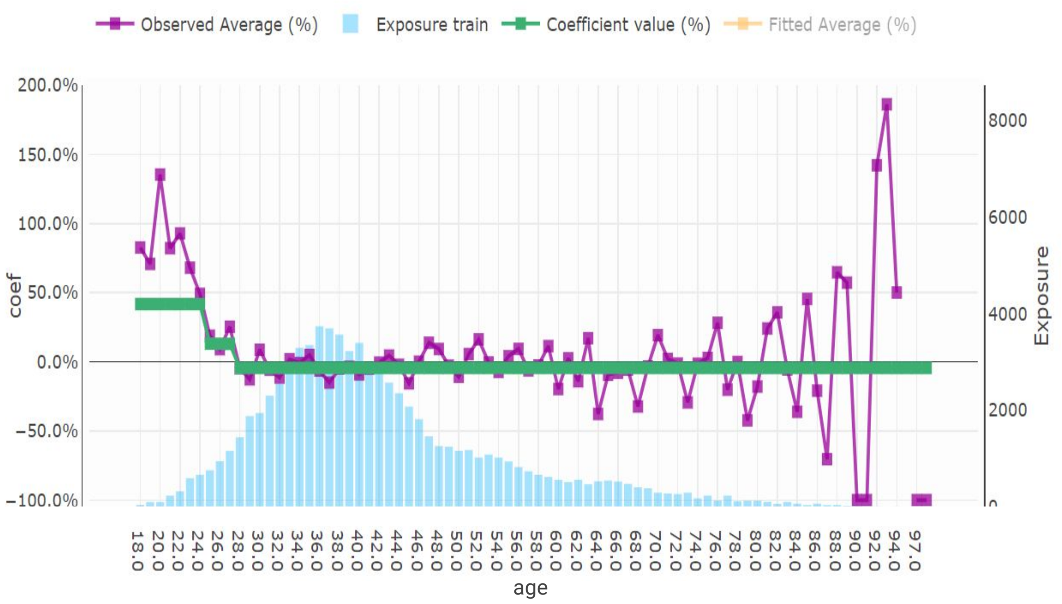
Leveraging the prior on a full model scale

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



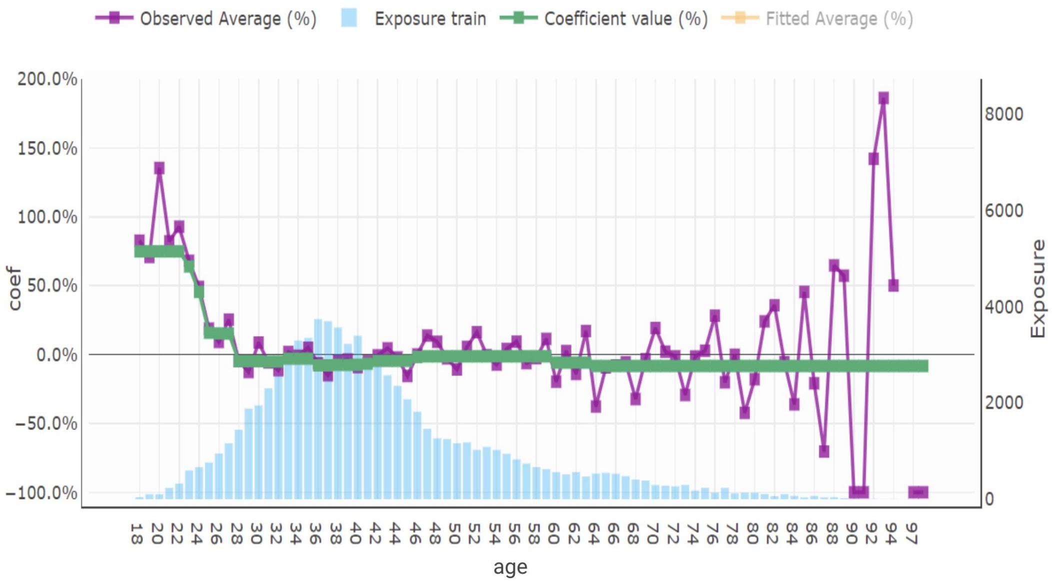
Leveraging the prior on a full model scale

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



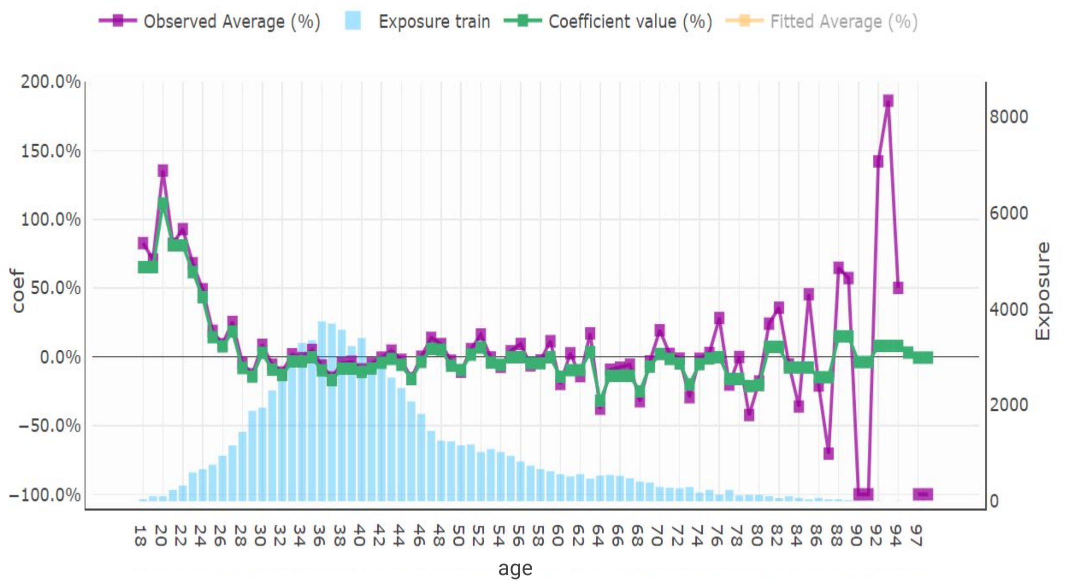
Leveraging the prior on a full model scale

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

Erweiterungen dieses Ansatzes

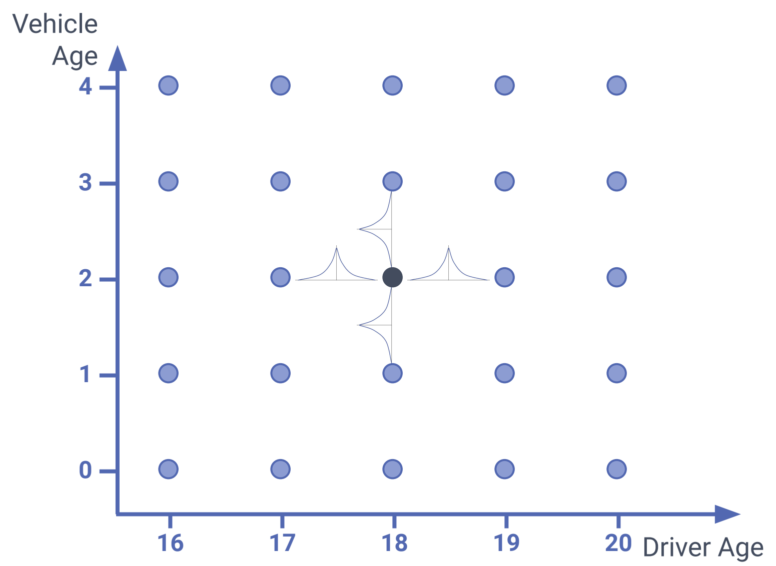
Applying to Interactions

The same principle can be applied in **two dimensions, to fit interactions**. The prior there is slightly different to take into account the 2-D nature of the problem.

For instance, on an interaction between two ordered variables, we could suppose as prior that the differences between all the “connected” levels are supposed to follow a Laplace distribution.

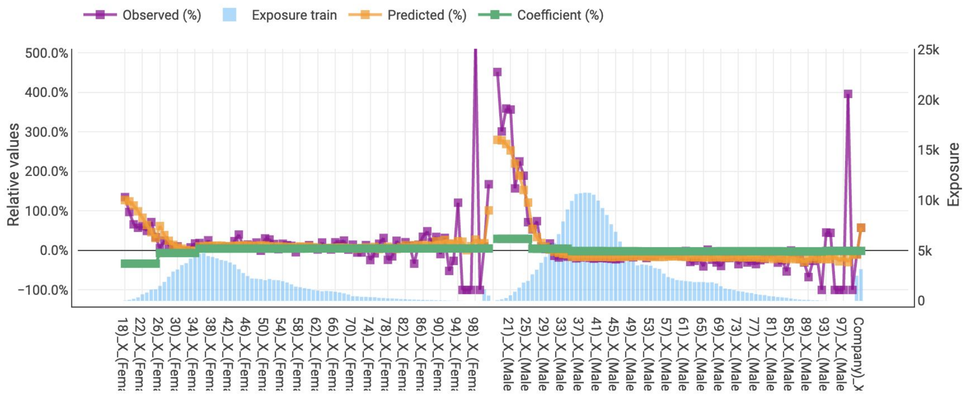
The prior term would become:

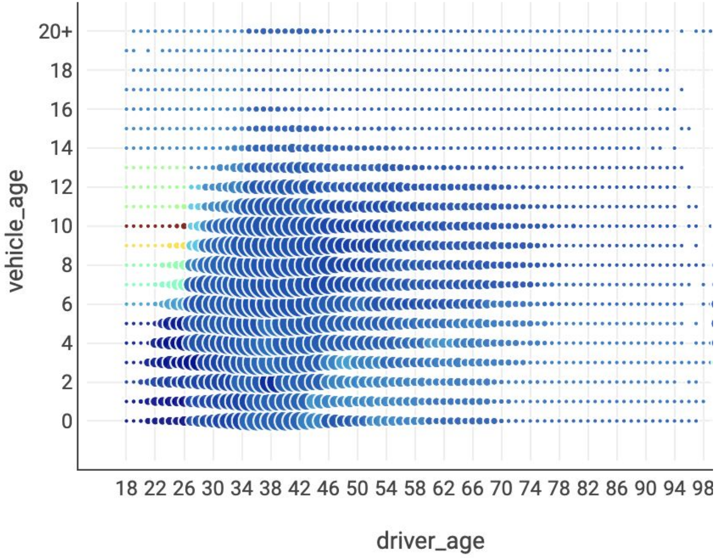
$$
\begin{aligned}
Penalty(\beta) = \cdots &+ \lambda\left|\beta_{18,2} - \beta_{19,2}\right| \\
&+\lambda\left|\beta_{18,2} - \beta_{17,2}\right| \\
&+\lambda\left|\beta_{18,2} - \beta_{18,1}\right| \\
&+\lambda\left|\beta_{18,2} - \beta_{18,3}\right| + \cdots
\end{aligned}
$$



Applying to Interactions

The interactions generated by applying this kind of priors would naturally extend the properties of pure additive models to interactions, allowing to identify the relevant ones and fit them automatically.





Applying to Geography

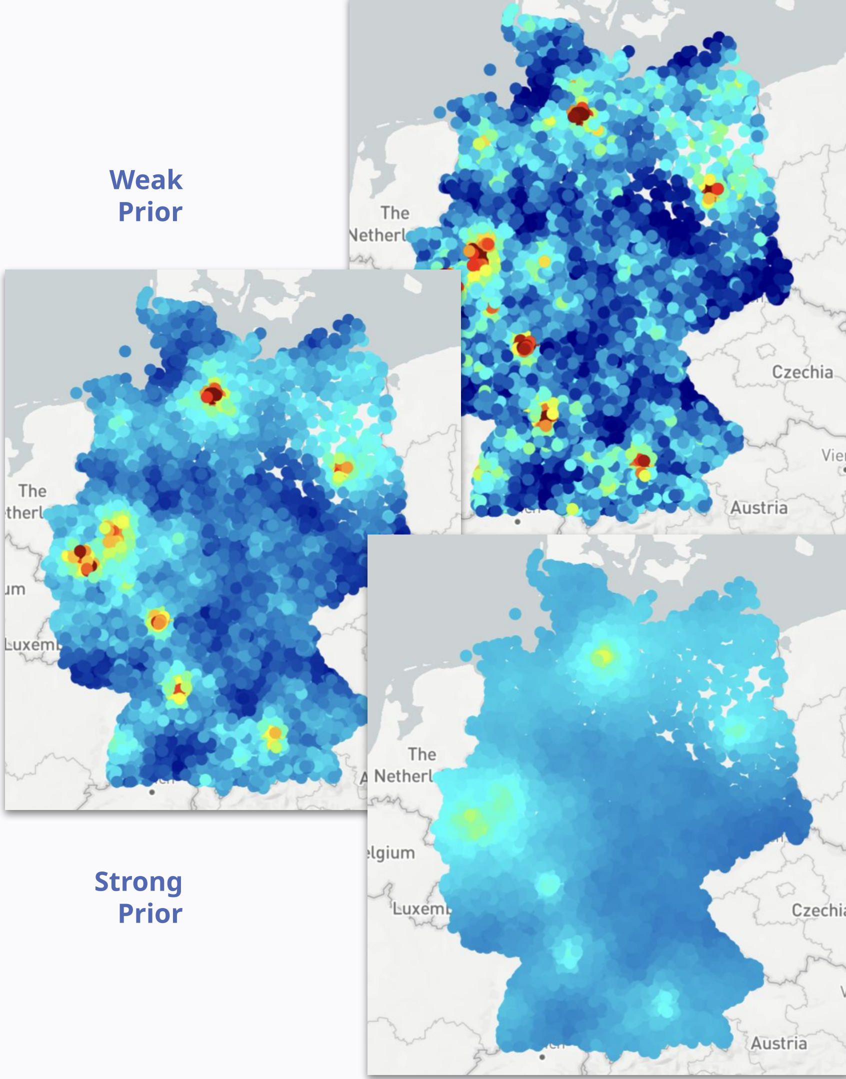
Geographic modeling can also be achieved with a similar method : the prior is that **nearby locations are expected to have similar risk levels**.

This has strong similarities to a **Gaussian Process** modeling.

Weak Prior

**Intermedia
te Prior**

Strong Prior

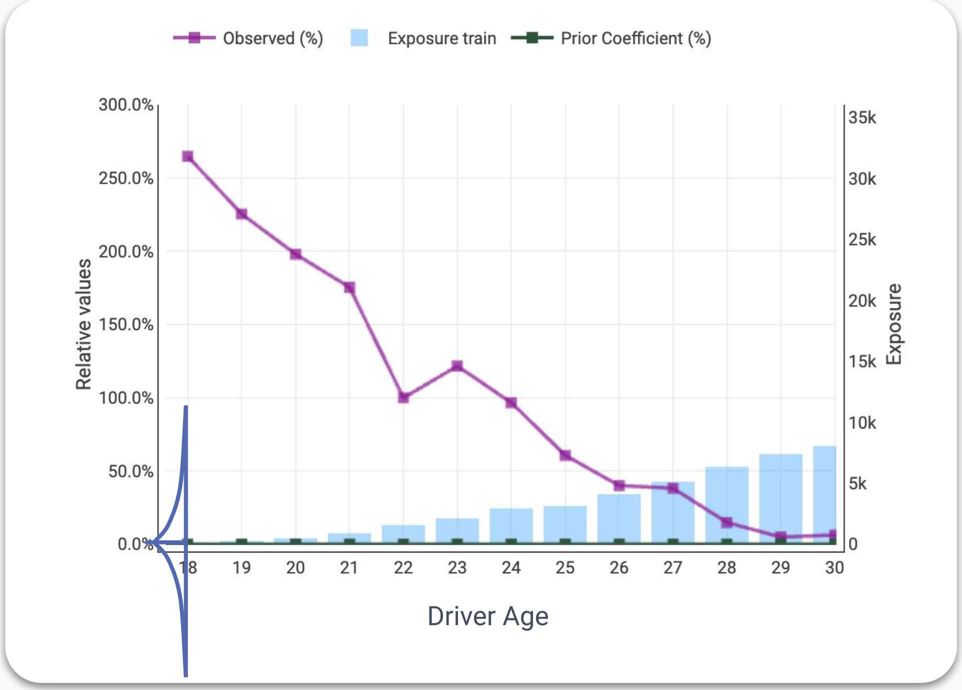


Leveraging a different prior estimate

In all the examples above, the priors used were centered on zero:

- For categorical variables, we assumed the coefficients were **most likely to be null**.
- For ordered variables, we assumed the coefficients were **most likely to be constant**.

In both case, the most likely prior model is null.

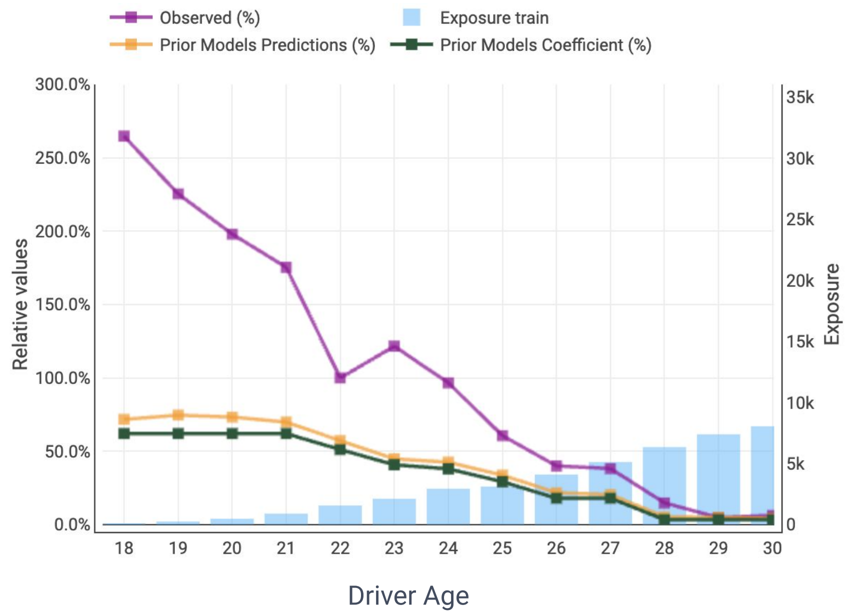


Leveraging a different prior estimate

We could use a different set of assumptions. For instance we can assume that the coefficients are **most likely to follow a reference model**.

The assumed distribution would then be following that model:

- For categorical variables, the updated coefficients can be assumed to be **centered on the reference coefficients**.
- For ordinal variables, the derivative of the coefficients can be assumed to be **centered on the derivative of the reference coefficients**.

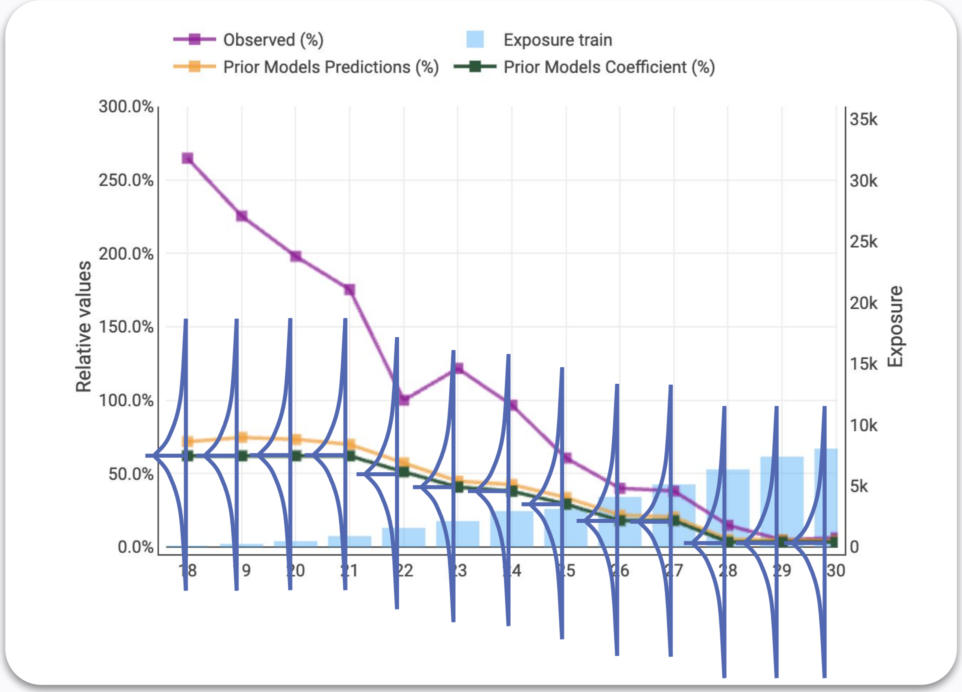


Leveraging a different prior estimate

We could use a different set of assumptions. For instance we can assume that the coefficients are **most likely to follow a reference model**.

The assumed distribution would then be following that model:

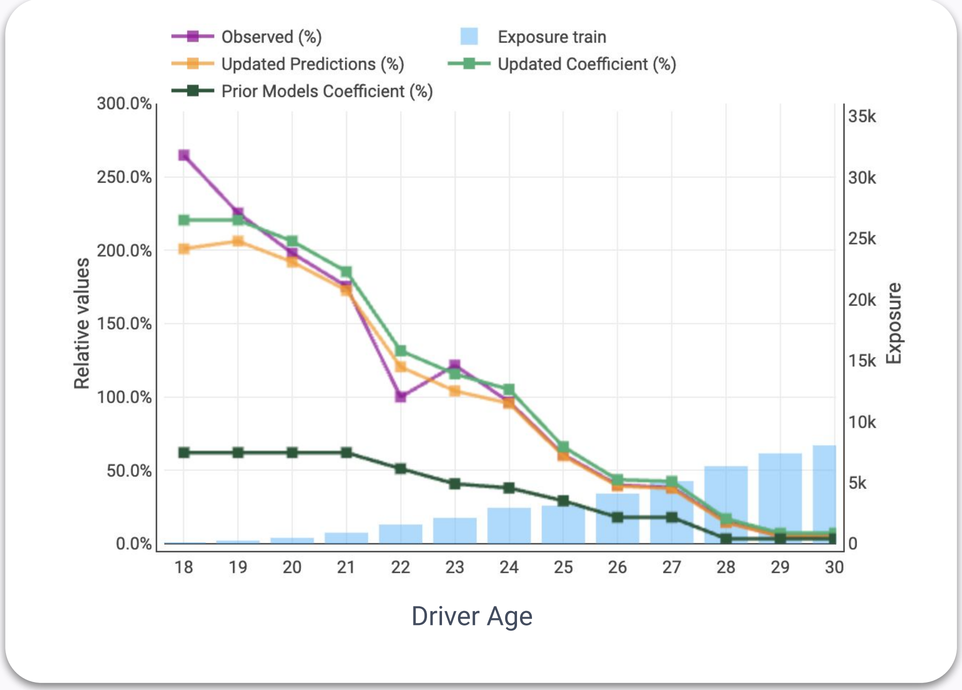
- For categorical variables, the updated coefficients can be assumed to be **centered on the reference coefficients**.
- For ordinal variables, the derivative of the coefficients can be assumed to be **centered on the derivative of the reference coefficients**.



Leveraging a different prior estimate

We could use a different set of assumptions. For instance we can assume that the coefficients are **most likely to follow a reference model**.

The new model would then be a **balance between the reference model and the new observed data** (which is aligned with a credibility-based intuition).

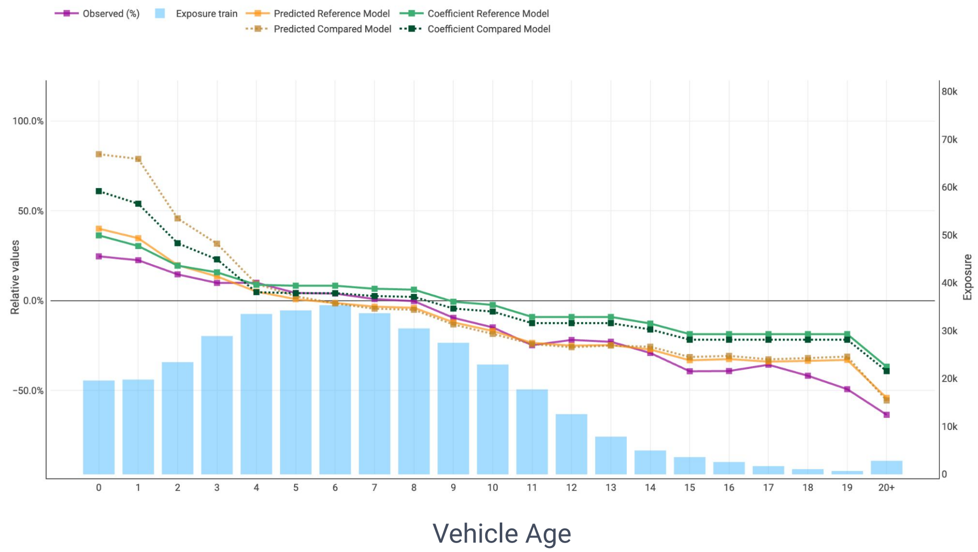


Leveraging a different prior estimate

This approach only updates coefficients if they are proven to be significantly off by the new data.

This ensures that the variables or profiles that are currently predicted well by the model will stay unchanged, while the ones that need an update are properly modified.

It increases the confidence in the model created (by considering the reference model as a relevant source of truth) and avoids unnecessary dislocation.



Data-driven Modelling

A three-step process

It is possible to directly leverage a model right out of the fit process.

This would be similar to a classic data-science approach.

However, handling transparent models opens the possibility of interacting with them, integrating expert knowledge in the modeling.

So the process is (on purpose) mixing elements of:

- Machine-Learning: **automated fit**, purely **data-driven** model creation, acting on **global parameters** to control overfitting.
- Direct interaction with the models: control of all the **effects** captured in the fitting model, analysis and potentially edition of the **effects** to ensure a good extrapolation of the model.

GLOBAL PARAMETERS

- Smoothness level
- Parsimony level

DATA

- Explanatory Variables
- Target Variable

FIT

MODEL PARAMETERS

- Effect functions values
- Fitted from a purely data-driven process

ADAPT

FINAL MODEL PARAMETERS

- Effect functions values
- Adapted based on expertise, to ensure safe extrapolation on low-data segments

Dankeschön!

Ihre Fragen?

jan.kuethe@akur8.com