Swiss Re

Survival of the Fittest

SAA après-midi series, Daniel Meier, 03 September 2025

New tutorial just released

Survival of the Fittest: Classical and Machine Learning Methods for Time-to-Event Modeling

Daniel Meier*   Adam Sturge†

Prepared for:
Fachgruppe "Data Science"
Swiss Association of Actuaries SAV
Version of August 31, 2025

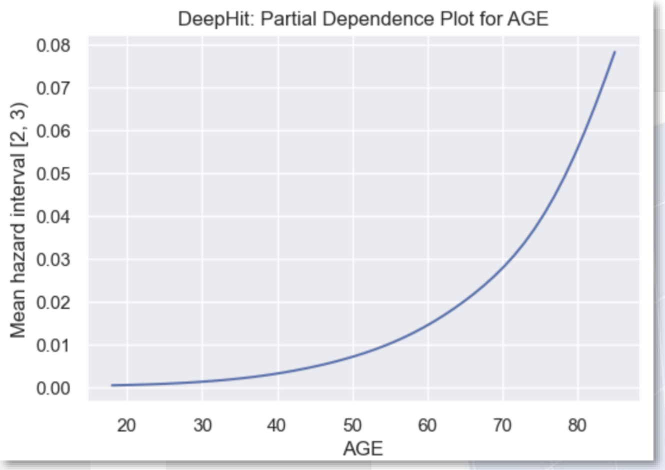
Abstract

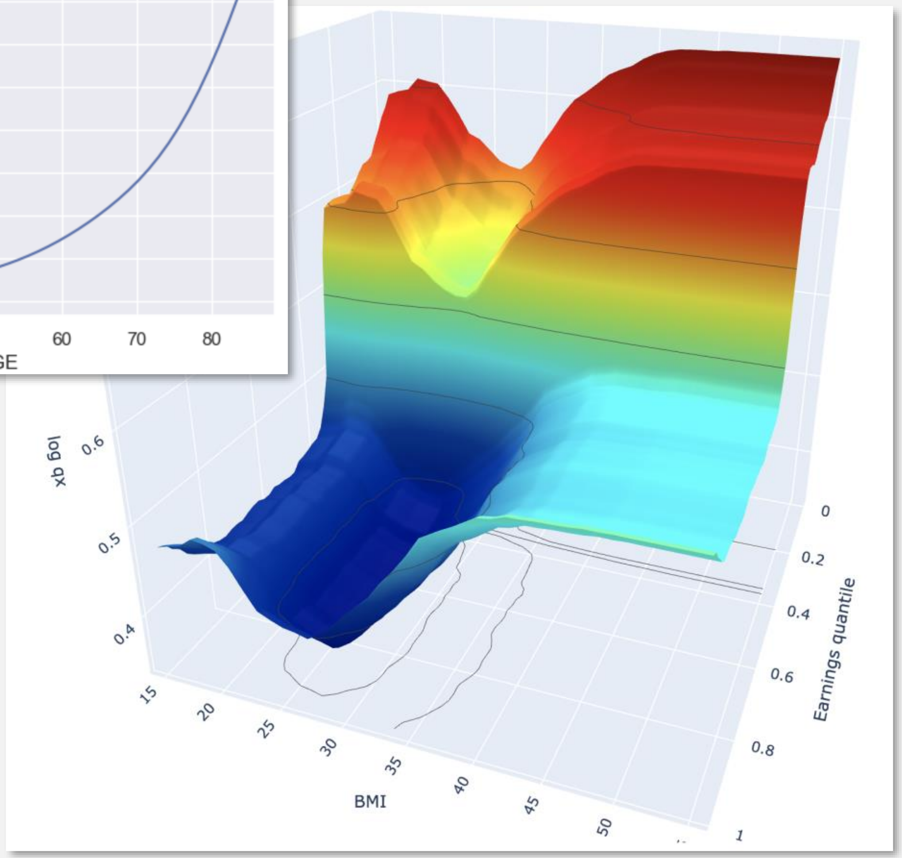
This tutorial provides an overview of classical and machine learning methods for survival modeling. We start with introducing the basic concepts of survival modeling using the Cox proportional hazards model and the accelerated failure time model, highlighting their

Case study 16 on actuarialdatascience.org

Where is survival modelling applied?

- Life & Health Underwriting
- Scenario testing, e.g., weight loss drugs
- US cancer registry SEER: Underwriting
- CIA pensioner mortality tables
- Unemployment times
- Public health
- Any other use case where **time-to-event** is important, e.g., credit default, lapse, engineering, etc.



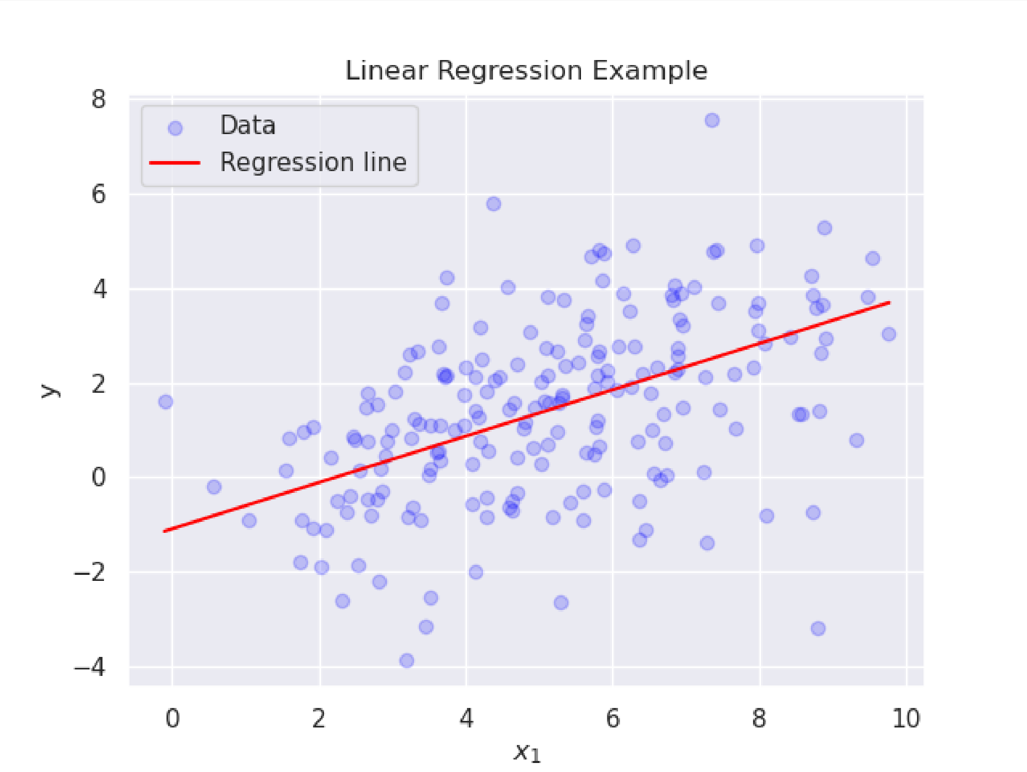


Linear regression

$$y(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M$$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.738
Model:                            OLS   Adj. R-squared:                  0.734
Method:                 Least Squares   F-statistic:                     184.4
Date:                Fri, 15 Aug 2025   Prob (F-statistic):           8.17e-57
Time:                        14:50:57   Log-Likelihood:                -277.39
No. Observations:                 200   AIC:                             562.8
Df Residuals:                     196   BIC:                             576.0
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.1888      0.293      7.458      0.000       1.610       2.768
x1             0.4899      0.034     14.387      0.000       0.423       0.557
x2            -0.3280      0.025    -13.149      0.000      -0.377      -0.279
x3             1.1735      0.068     17.172      0.000       1.039       1.308
==============================================================================
Omnibus:                        0.265   Durbin-Watson:                   2.082
Prob(Omnibus):                  0.876   Jarque-Bera (JB):                0.402
Skew:                           0.064   Prob(JB):                        0.818
Kurtosis:                       2.822   Cond. No.                         48.1
==============================================================================
```

statsmodels summary



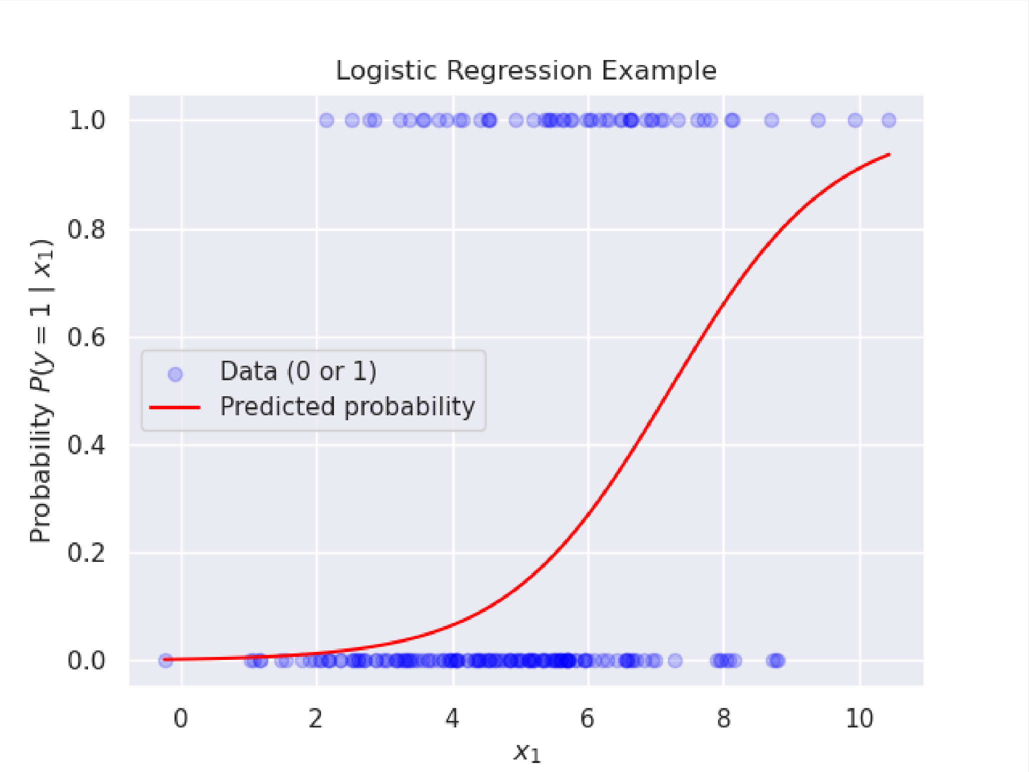
Logistic regression

$$p(\boldsymbol{x}) = \text{logistic}(\beta_0+\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M)$$

$$\text{logistic}(x) = (1+\exp(-x))^{-1}$$

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  200
Model:                          Logit   Df Residuals:                      196
Method:                           MLE   Df Model:                            3
Date:                Fri, 15 Aug 2025   Pseudo R-squ.:                  0.4304
Time:                        14:48:34   Log-Likelihood:                -66.445
converged:                       True   LL-Null:                       -116.65
Covariance Type:            nonrobust   LLR p-value:                 1.266e-21
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.9304      0.915      1.017      0.309      -0.863       2.723
x1             0.8313      0.155      5.354      0.000       0.527       1.136
x2            -0.6647      0.112     -5.961      0.000      -0.883      -0.446
x3             1.1915      0.270      4.413      0.000       0.662       1.721
==============================================================================
```

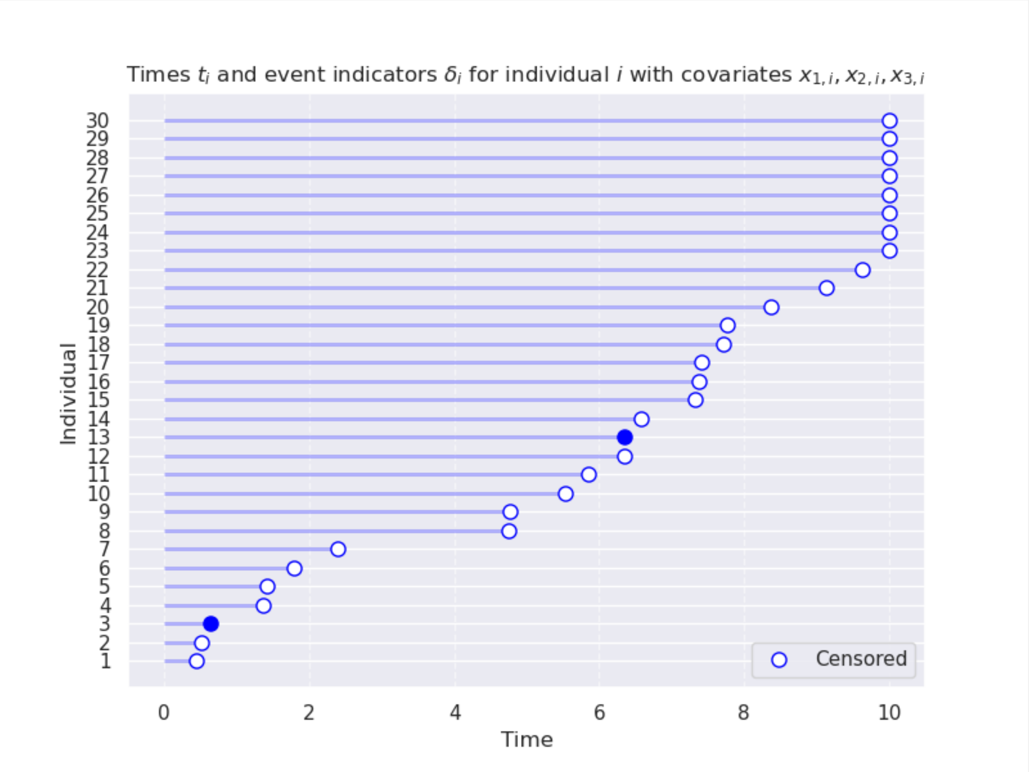
statsmodels summary



Cox regression
(the most common survival model)

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M)$$

- **Data** consists of individuals i with
  - features $x_{1,i}, x_{2,i}, \ldots$
  - time t_i
  - event indicator δ_i, where
    - $\delta_i = 0$ denotes (right-)censoring
    - $\delta_i = 1$ denotes, e.g., mortality
- What is the distribution (CDF F, PDF f) of **survival time** T?

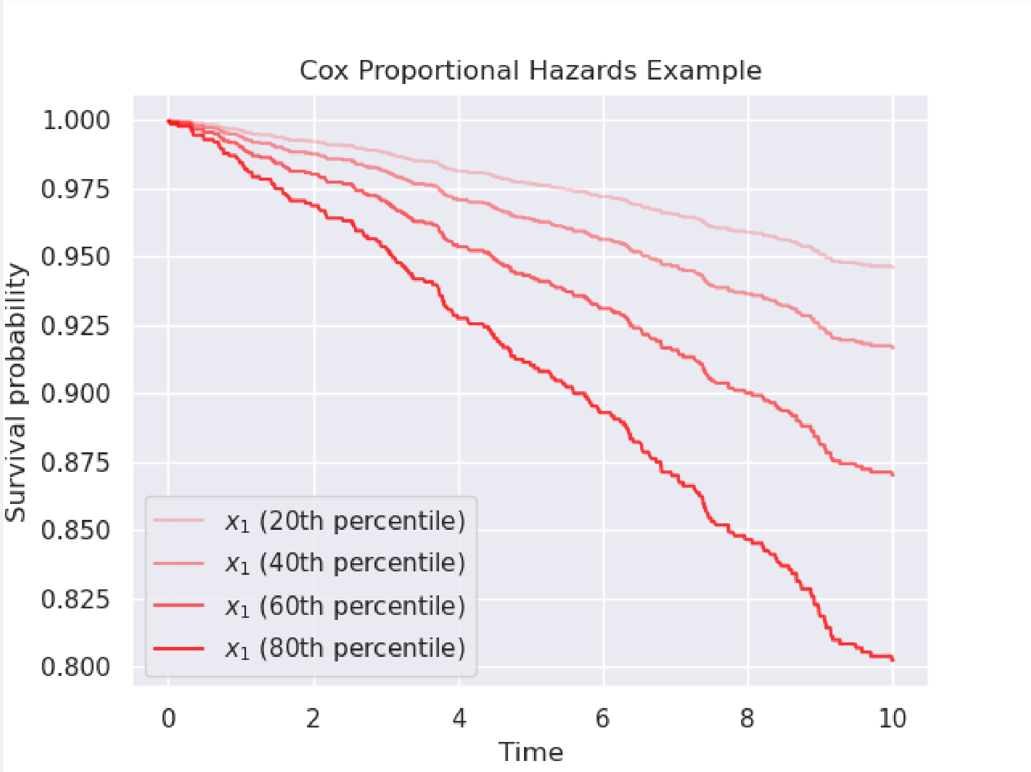


Cox regression
(the most common survival model)

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M)$$

- **Hazard rates** $h(t|\boldsymbol{x})$, correspond to force of mortality $\mu_{\boldsymbol{x}}(t)$ in continuous time and $q_{\boldsymbol{x},t}$ or $m_{\boldsymbol{x},t}$ in discrete time
- **Proportional hazards**: $h(t|\boldsymbol{x}_i)/h\big(t|\boldsymbol{x}_j\big)$ const.
- **Survival probability function** $S(t|\boldsymbol{x})$, corresponds to ${}_tp_{\boldsymbol{x}}$

- $S(t|\boldsymbol{x}) = 1 - F(t|\boldsymbol{x})$
- $h(t|\boldsymbol{x}) = -\frac{\partial}{\partial t}\log S(t|\boldsymbol{x}) = \frac{f(t|\boldsymbol{x})}{S(t|\boldsymbol{x})}$

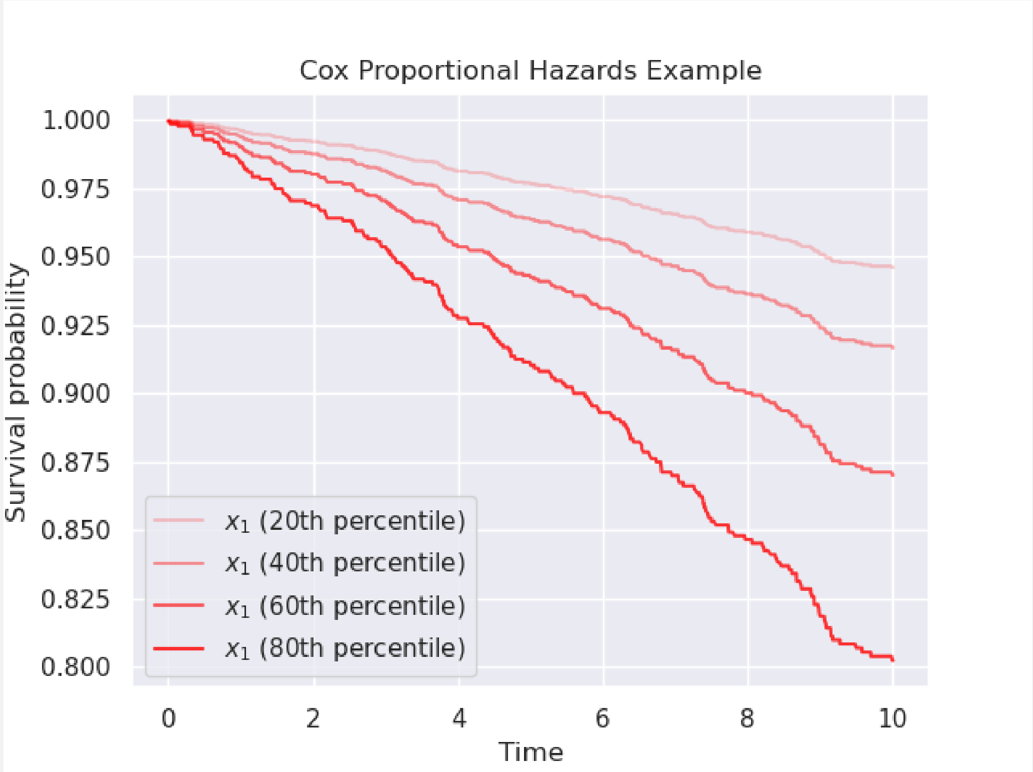


Cox regression
(the most common survival model)

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M)$$

- **Baseline hazard rates** $h_0(t)$ via
  - Kaplan-Meier: $S(t) = \prod_{t_i \le t}\left(1 - \frac{d_i}{n_i}\right)$
  - Nelson-Aalen: $H(t) = \sum_{t_i \le t}\frac{d_i}{n_i}$
- **Coefficients** $\beta_1, \beta_2, \ldots$ via partial likelihood function maximization (Breslow method)

$$\mathcal{L} = \prod_{i:\delta_i=1}\prod_{j:t_j=t_i}\frac{\exp(\beta_1 x_{1,j} + \cdots)}{\sum_{k:t_k \ge t_j}\exp(\beta_1 x_{1,k} + \cdots)}$$



A bit of public health history…

Lester Breslow (1915-2012), the father of Norman Breslow after whom the method was named

Dr. Lester Breslow, a former dean of the UCLA Jonathan and Karin Fielding School of Public Health, professor emeritus of health services, and one of the leading figures in public health for seven decades, died Monday. He was 97.

Breslow was a visionary public health figure with a well-established track record for being ahead of his time. As early as the 1940s, he linked tobacco use to disease in three studies that were later cited in the U.S. Surgeon General's landmark 1964 report.

He is widely known for his early advocacy and research into health promotion and disease prevention. Breslow's pioneering Alameda County studies beginning in the early 1960s were among the first to show that simple health practices — such as getting regular exercise and sleep, not drinking excessively, not smoking, and maintaining a healthy weight — add both years and quality to life.

While these conclusions are taken for granted today, the idea of such a strong connection between lifestyle and health was seen as "bizarre" at the time, Breslow noted decades later. He would smile when recalling the response of the National Institutes of Health panel of scientists that reviewed the initial study proposal: "Unanimous rejection." When the study was completed, even Breslow was shocked at the magnitude of the results, which helped usher in current thinking about health and fitness.

Source: https://ph.ucla.edu/news-events/news/memoriam-dr-lester-breslow-public-health-visionary

Cox regression
(the most common survival model)

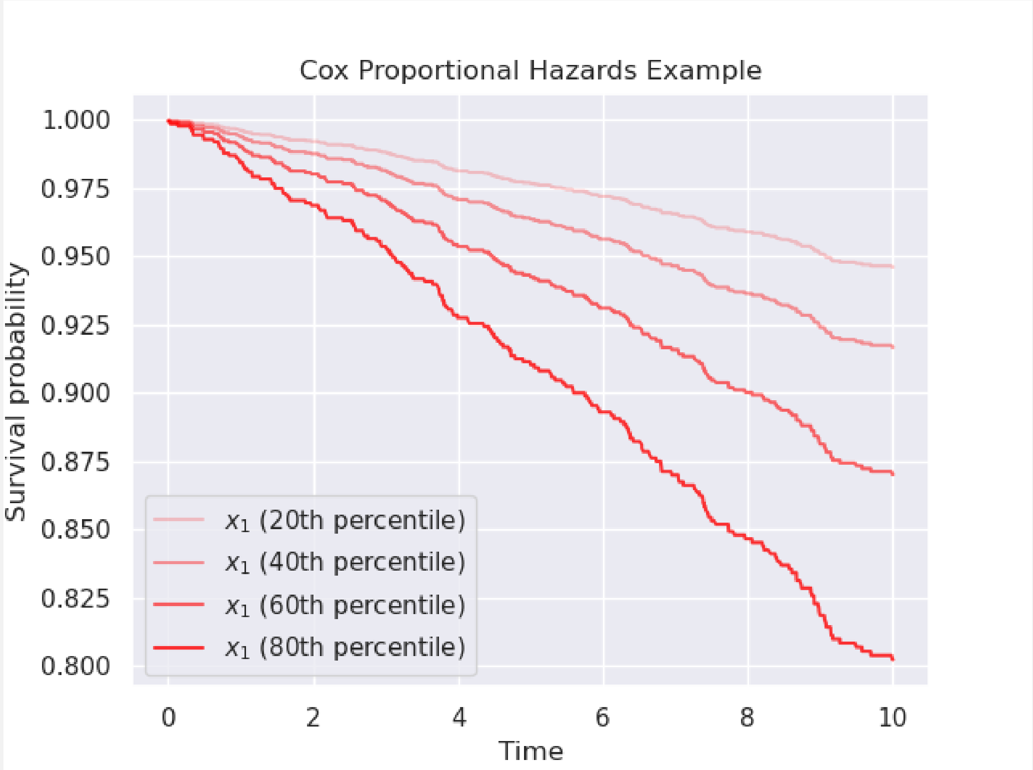
$$h(t|\boldsymbol{x}) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M)$$

model	lifelines.CoxPHFitter
duration col	'time'
event col	'event'
baseline estimation	breslow
number of observations	2000
number of events observed	178
partial log-likelihood	-1248.33
time fit was run	2025-08-18 08:31:08 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
x1	0.09	1.09	0.01	0.07	0.11	1.07	1.12	0.00	7.76	<0.005	46.71
x2	0.20	1.22	0.15	-0.09	0.50	0.91	1.64	0.00	1.34	0.18	2.46
x3	0.04	1.04	0.02	0.00	0.07	1.00	1.08	0.00	2.06	0.04	4.67

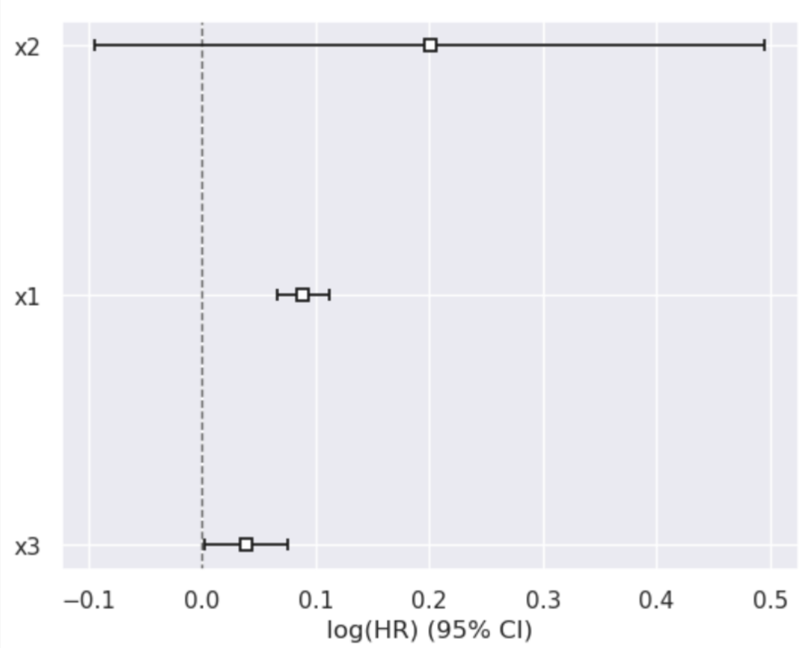
Concordance	0.67
Partial AIC	2502.65
log-likelihood ratio test	72.86 on 3 df
-log2(p) of ll-ratio test	49.77

lifelines summary

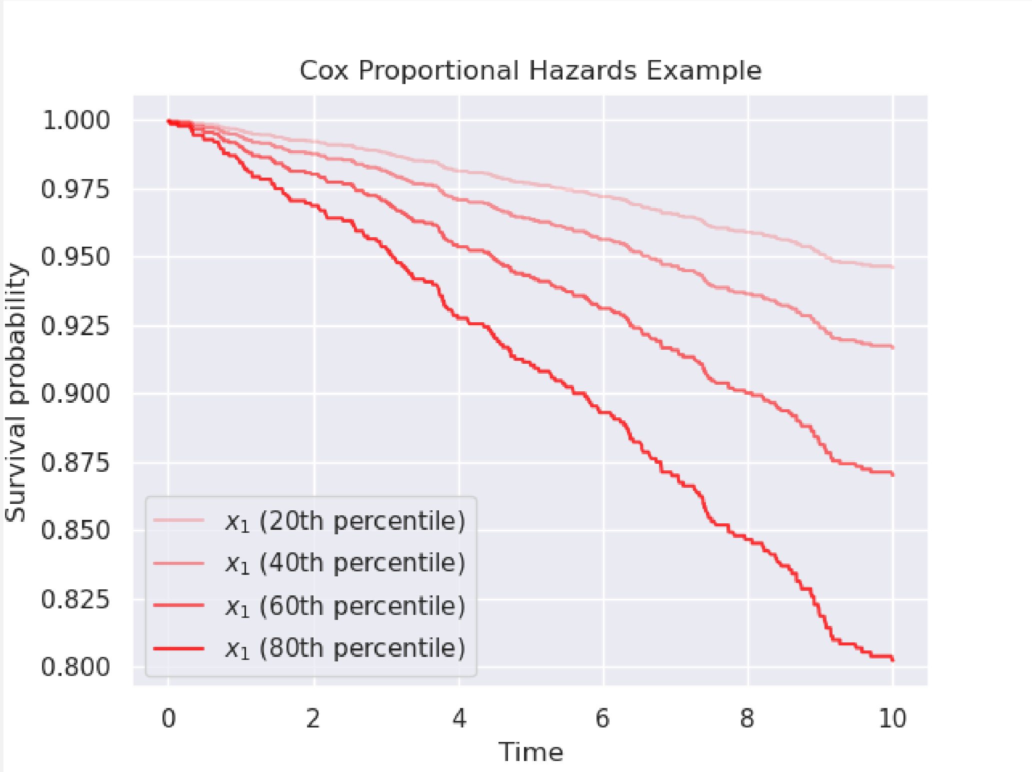


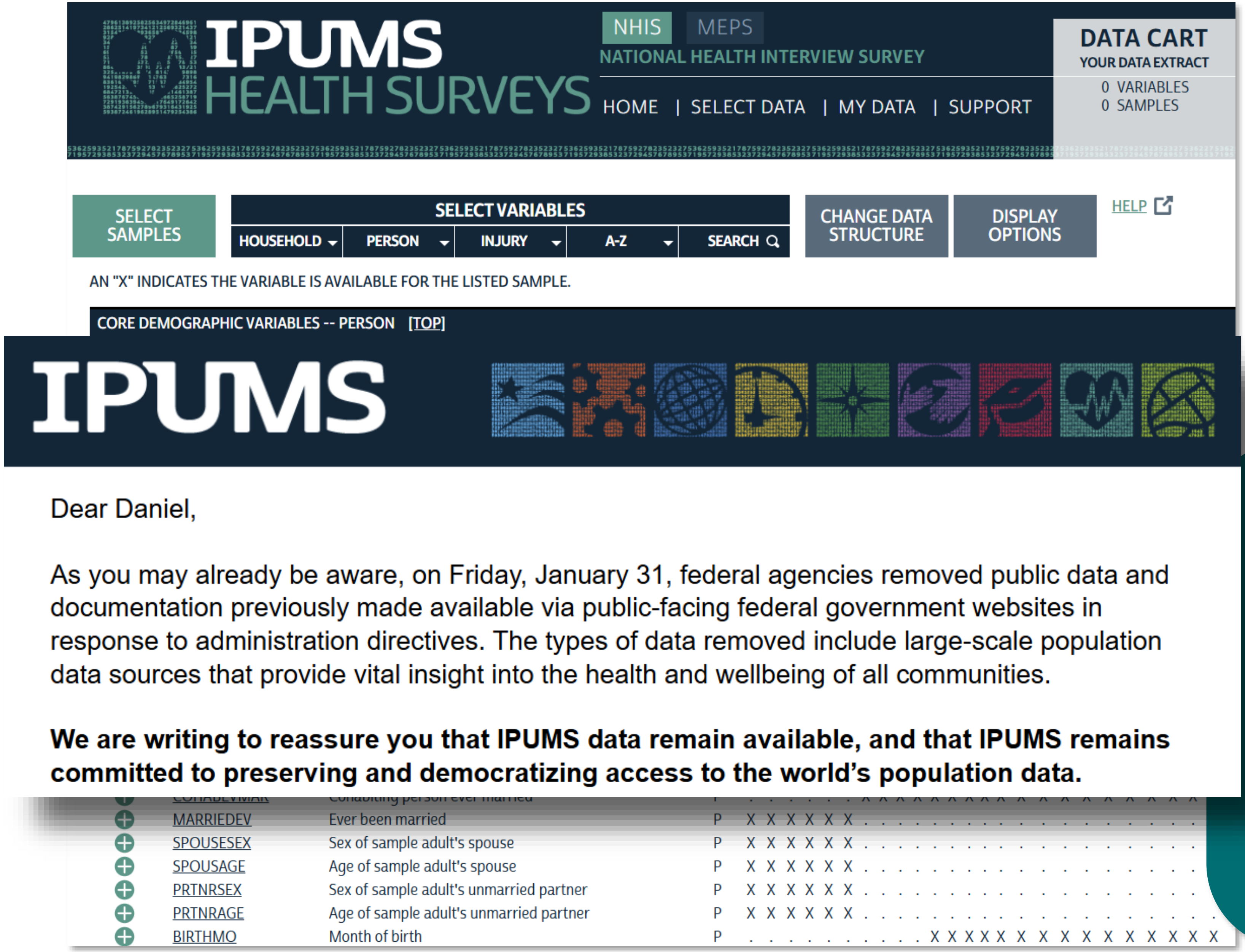
Cox regression
(the most common survival model)

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M)$$



lifelines forest plot



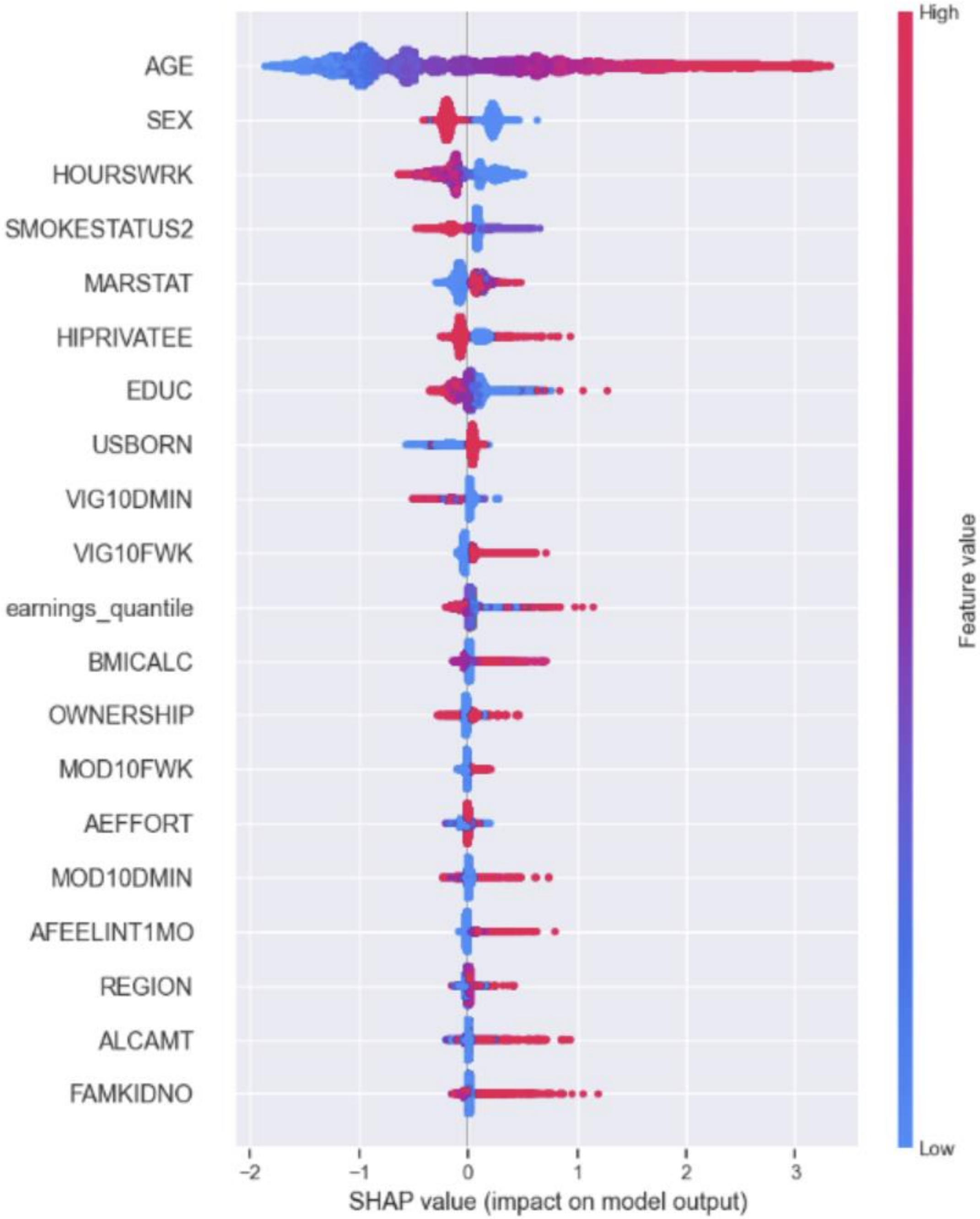


The dataset

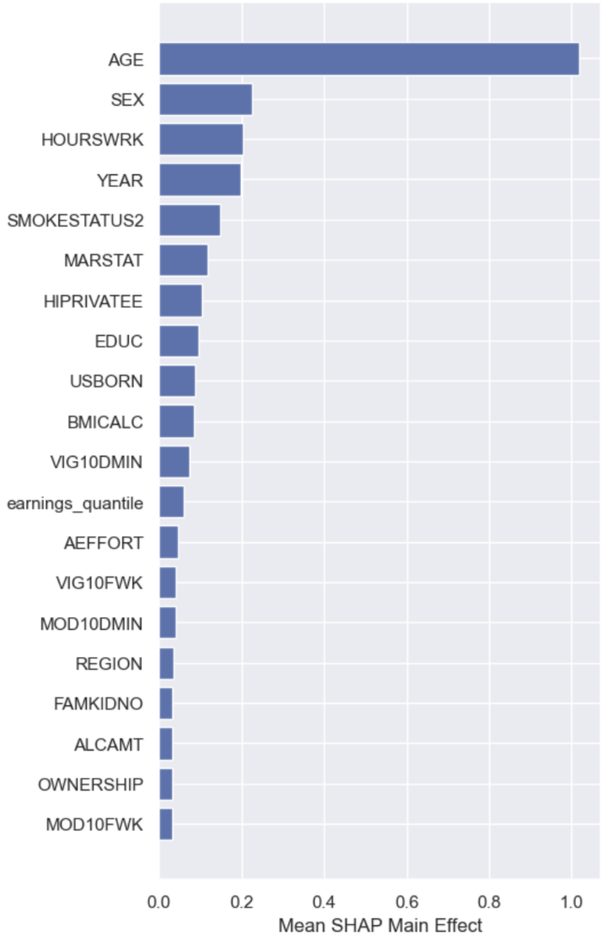
- Data since 1963
- 100k individuals per year
- 500+ features: demographic, socio-economics, health behaviors, mental health, mortality, cause of death, etc.
- Part of our directory of publicly available datasets at actuarialdatascience.org

Source: nhis.ipums.org

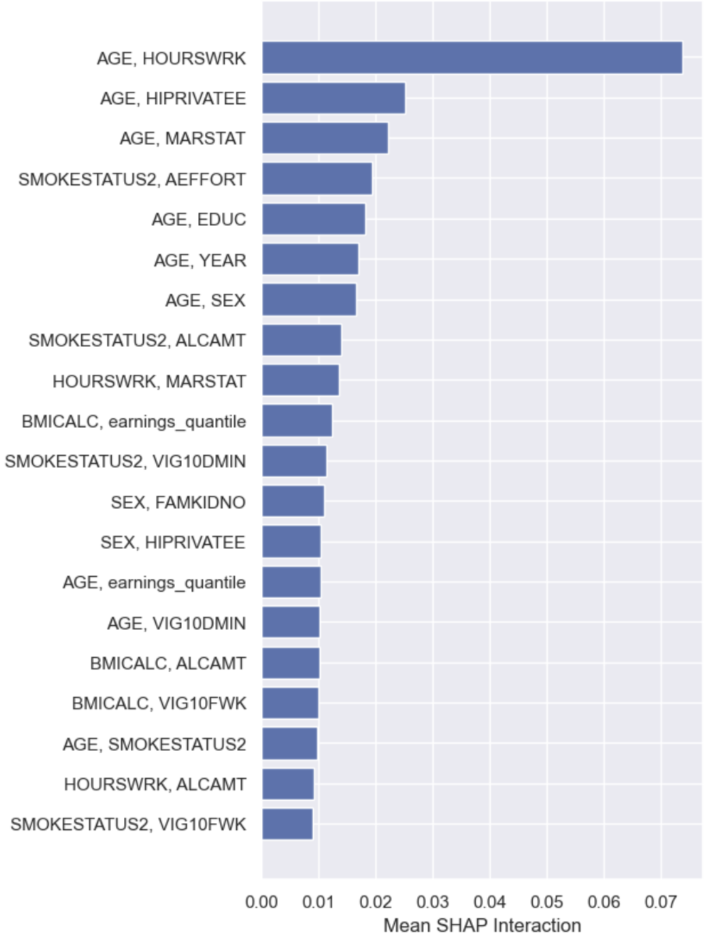
The dataset



High
AGE
SEX
HOURSWRK
SMOKESTATUS2
MARSTAT
HIPRIVATEE
EDUC
USBORN
VIG10DMIN
VIG10FWK
earnings_quantile
BMICALC
OWNERSHIP
MOD10FWK
AEFFORT
MOD10DMIN
AFEELINT1MO
REGION
ALCAMT
FAMKIDNO
−2
−1
0
1
2
3
SHAP value (impact on model output)
Feature value
Low



AGE
SEX
HOURSWRK
YEAR
SMOKESTATUS2
MARSTAT
HIPRIVATEE
EDUC
USBORN
BMICALC
VIG10DMIN
earnings_quantile
AEFFORT
VIG10FWK
MOD10DMIN
REGION
FAMKIDNO
ALCAMT
OWNERSHIP
MOD10FWK
0.0
0.2
0.4
0.6
0.8
1.0
Mean SHAP Main Effect



AGE, HOURSWRK
AGE, HIPRIVATEE
AGE, MARSTAT
SMOKESTATUS2, AEFFORT
AGE, EDUC
AGE, YEAR
AGE, SEX
SMOKESTATUS2, ALCAMT
HOURSWRK, MARSTAT
BMICALC, earnings_quantile
SMOKESTATUS2, VIG10DMIN
SEX, FAMKIDNO
SEX, HIPRIVATEE
AGE, earnings_quantile
AGE, VIG10DMIN
BMICALC, ALCAMT
BMICALC, VIG10FWK
AGE, SMOKESTATUS2
HOURSWRK, ALCAMT
SMOKESTATUS2, VIG10FWK
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Mean SHAP Interaction

Other survival models

Model	Parameterization	Comments
Cox proportional hazards	$h(t\|\boldsymbol{x}) = h_0(t)\exp(\beta_1 x_1 + \cdots)$	Proportional hazards, de facto standard
Accelerated Failure Time (AFT)	$T = \exp(\varepsilon)\exp(\beta_1 x_1 + \cdots)$	Scaling survival time
Survival trees	$S(t\|\boldsymbol{x}) = S_l(t)$, where l is $\boldsymbol{x}$'s leaf	Log-rank test to split tree, Kaplan -Meier
Random survival forest	$S(t\|\boldsymbol{x}) = \frac{1}{B}\sum_b S_l^{(b)}(t)$	Tree ensemble of survival trees
Gradient boosted survival	$h(t\|\boldsymbol{x}) = h_0(t)\exp(f^{(m)}(\boldsymbol{x}))$	Iterative tree refinement $f^{(0)}, f^{(1)}, \ldots, f^{(m)}$
DeepSurv	$h(t\|\boldsymbol{x}) = h_0(t)\exp(z_\theta(\boldsymbol{x}))$	Neural network $z_\theta(\boldsymbol{x})$, likelihood, early stopping
DeepHit	Discrete version of PDF $f(t\|\boldsymbol{x})$	Neural network with softmax as last layer, allows to model **competing risks**
Transformer based survival	Discrete version of PDF $f(t\|\boldsymbol{x})$	Transformer based neural network that can consider full longitudinal data, i.e., history of covariates, e.g., BMI timeseries

Survival model performance metrics

- **C-index**: let P be the set of *comparable* individuals (i, j), i.e., $\delta_i = 1$ and $t_i < t_j$,

$$\text{C−index} = \frac{1}{\#P} \sum_{(i,j)\in P} \mathbb{I}_{h(t_i|x_i)>h(t_j|x_j)}$$

- **Integrated Brier score** (IBS):

$\text{IBS} = \int_0^\tau \frac{1}{n} \sum_{i=1}^n w_i(t) \, (\mathbb{I}_{t_i>t} - S(t|\boldsymbol{x}_i))^2$, where $w_i(t)$ are inverse probability censoring weights

- **Log-loss in time interval** (LL): let $y_i(t_1, t_2)$ be the indicator whether individual i had an event in $[t_1, t_2)$,

$$\text{LL} = -\frac{1}{n} \sum_{i=1}^n y_i(t_1, t_2) \log\big(S(t_1|\boldsymbol{x}_i) - S(t_2|\boldsymbol{x}_i)\big) + (1 - y_i(t_1, t_2)) \log\big(1 - S(t_1|\boldsymbol{x}_i) + S(t_2|\boldsymbol{x}_i)\big)$$

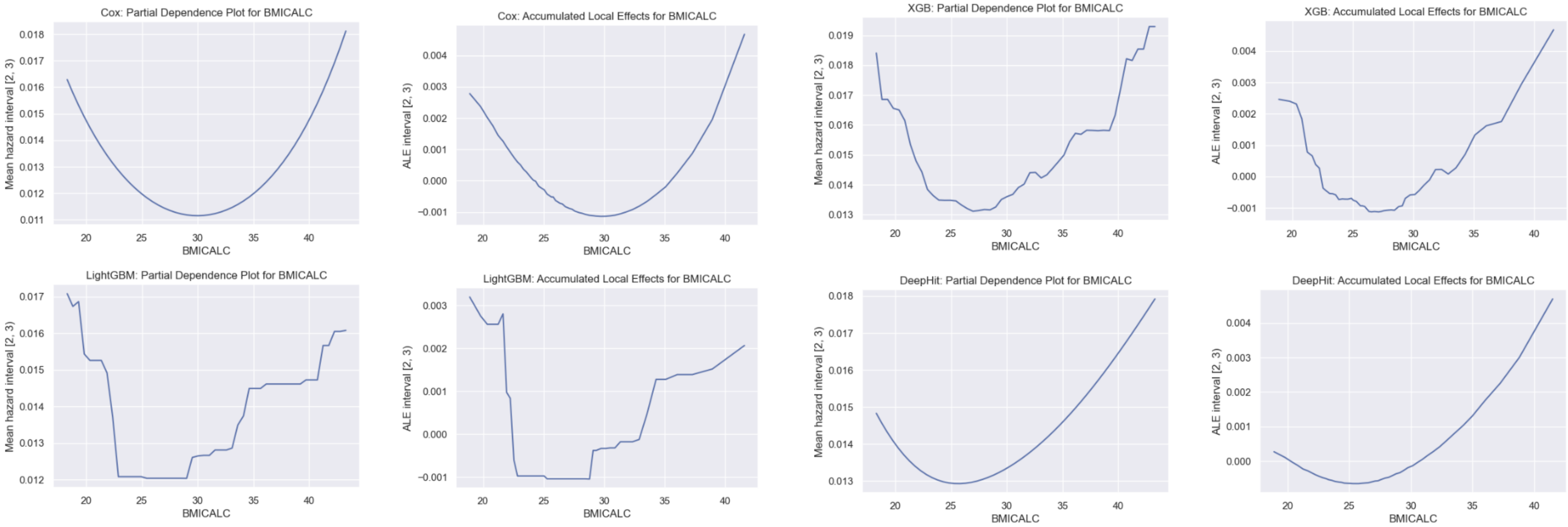
- **Mean squared error** (MSE) of log predictions: let $\mu_i(t_1, t_2)$ denote the ground truth probability

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \left(\log\big(S(t_1|\boldsymbol{x}_i) - S(t_2|\boldsymbol{x}_i)\big) - \log \mu_i(t_1, t_2)\right)^2$$

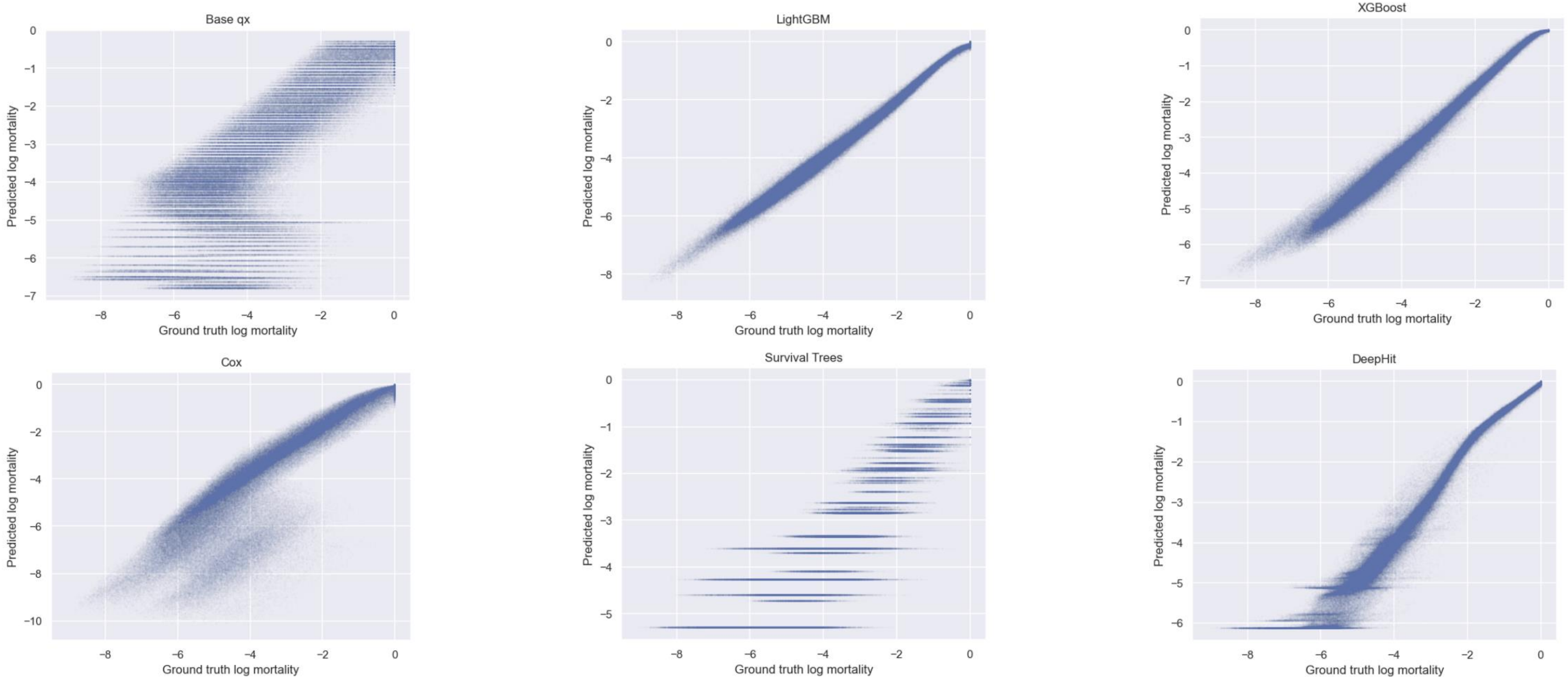
Survival model performance metrics

	C-index ↑	IBS ↓	LL $[2, 3)$ ↓	Time [min] ↓
Base q_x	0.8402	0.0439	0.0519	<1
LightGBM	0.8599	0.0415	**0.0502**	<1
Cox Proportional Hazards	0.8612	0.0417	0.0504	<1
Accelerated Failure Time	0.8612	0.0425	0.0517	<1
Survival Trees	0.8570	0.0415	0.0512	4
Random Survival Forests	0.8682	0.0412	0.0507	491
Gradient Boosted Survival Trees	0.8701	0.0412	0.0507	444
XGBoost Cox	0.8724	0.0410	0.0512	<1
DeepSurv	0.8711	**0.0393**	0.0511	<1
DeepHit	**0.8781**	0.0407	0.0515	4
Deep Survival Machines	0.8705	0.0423	0.0509	5
Transformer Survival Model	0.8689	0.0396	0.0504	337

Partial dependence plots and accumulated local effects



Ground truth vs. predictions on a larger synthetic dataset



Tips and tricks and pitfalls

1. Start with a fast and strong model, e.g., LightGBM (interval event prediction) or XGBoost (survival)
2. For (Life & UW) actuarial purposes, MSE on log predictions is probably the best performance metric – if the ground truth is known
3. If the ground truth is not known, try to predict it with the models from 1., potentially simulating a new dataset – as a learning experience to choose a deep learning model if you have sufficient data
4. Don't underestimate the many pitfalls of survival modelling:
    - off-by-one errors or other discretization issues on the time dimension
    - selection effects for early times
    - missing values (not at random)
    - time-dependencies, e.g., current vs. past BMI
    - miscalibrated models, e.g., overestimating risk of low risk individuals
    - slow running times

Any questions?

Swiss Re

Thank you!

Contact us

Daniel Meier

L&H R&D Manager
daniel_meier@swissre.com

Follow us

Legal notice