Influence of Machine Learning Techniques in Actuarial Sciences

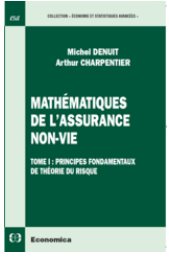
Arthur Charpentier

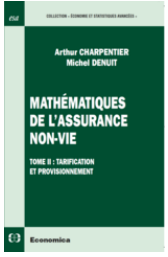
Intel India, August 2021

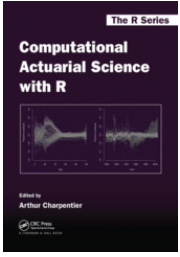
Arthur Charpentier

Université du Québec à Montréal

@freakonometrics
freakonometrics
freakonometrics.hypotheses.org

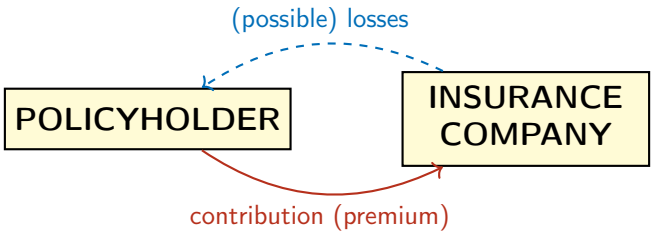






Insurance & Actuarial Science

"*Insurance is the contribution of the many to the misfortune of the few*"



What would be a "*fair contribution*" ? see O'Neill (1997)

- pure actuarial fairness contributions for individual policyholders should perfectly reflect their predicted risk levels $\rightarrow$ predictive modeling
- choice-sensitive fairness contributions should take into account only risks that result from choices - luck-egalitarianism (Cohen (1989) or Arneson (2011))

Agenda

General Insurance & Predictive Modeling (1)

- fraud detection & network data

see e.g. Óskarsdóttir et al. (2019)

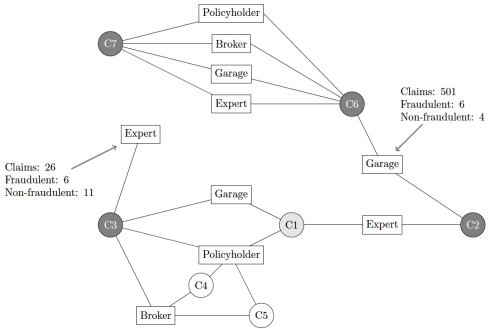
- parametric insurance & satellite pictures

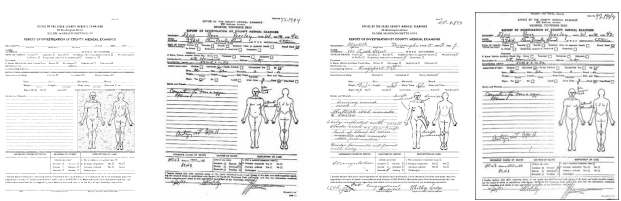
see e.g. de Leeuw et al. (2014)

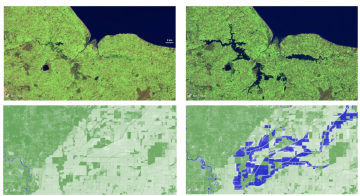
- claims reserving

see e.g. Wüthrich (2018)

- automatic reading (medical reports)



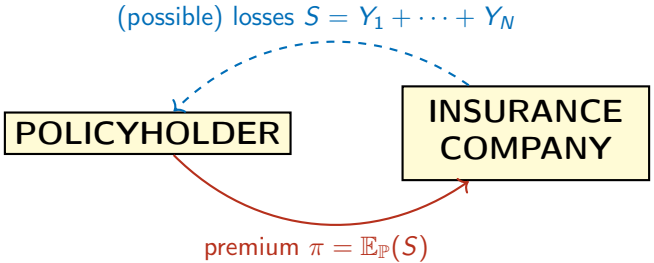




See also Denuit et al. (2019a, 2020, 2019b)

General Insurance & Predictive Modeling (2)

- premium computation, $\pi = \mathbb{E}_{\mathbb{P}}(S)$



or, given some features $\boldsymbol{X} = \{X_1, \cdots, X_p\}$, premium $\pi(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}_{\boldsymbol{x}}}(S) = \mathbb{E}[S|\boldsymbol{X}]$

Under standard assumptions,

$$\pi(\boldsymbol{x}) = \mathbb{E}[S|\boldsymbol{X}] = \underbrace{\mathbb{E}[N|\boldsymbol{X}]}_{\text{frequency}} \cdot \underbrace{\mathbb{E}[Y|\boldsymbol{X}]}_{\text{average cost}} = \underbrace{\mathbb{E}[\mathbf{1}_{S>0}|\boldsymbol{X}]}_{\text{occurence}} \cdot \underbrace{\mathbb{E}[S|\boldsymbol{X}, S>0]}_{\text{individual cost}}$$

From Econometric Techniques to Machine Learning (1)

Merging claims & underwriters databases (per policy), e.g. $(n_i, e_i, \boldsymbol{x}_i)$, $i = 1, \cdots, n_p$

	id	exposure	zone	pwr	agecar	agedrv	model	gas	dsty	region	nbr	occ
1	3227	0.87	C	7	0	56	12	D	93	13	0	0
2	4115	0.72	D	5	0	45	22	R	54	13	0	0
3	5121	0.05	C	6	3	37	17	D	11	13	0	0
4	5142	0.90	C	10	10	42	7	D	93	13	0	0
5	6255	0.12	C	7	0	59	11	R	73	13	0	0
6	8486	0.83	C	5	0	75	7	R	42	13	2	1

Standard model, Poisson regression (possibly zero-inflated, possibly over-dispesersed, etc), related to the Poisson process

$$N_i \sim \mathcal{P}(e_i \cdot \exp[\theta_i]), \quad \text{where } \theta_i = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \cdots + \alpha_p x_{p,i} = \boldsymbol{x}_i^\top \boldsymbol{\alpha}$$

or possibly GAMs (for nonlinearities, or spatial component)

$$\theta_i = \alpha_0 + s_1(x_{1,i}) + \alpha_2 x_{2,i} + \cdots + \alpha_p x_{p,i}, \; s_1(x) = \sum_j \gamma_j \varphi_j(x)$$

From Econometric Techniques to Machine Learning (2)

or $(y_i \boldsymbol{x}_i)$, $i = 1, \cdots, n_c$

```
     id     no cover     cost  zone pwr agecar agedrv model gas dsty
1  1870  17219   1TP  1692.29     C   5      0     52    12   R   93
2  1963  16336   1TP   422.05     E   9      0     78    12   R   27
3  4263  17089   2MT   549.21     C  10      7     27    17   D   19
4  5181  17801   1TP   191.15     D   5      2     26     3   D   91
5  6375  17485   1TP  2031.77     B   7      4     46     6   R   48
```

Standard model, Gamma regression (with a log link function)

$$Y_i \sim \mathcal{G}(\exp[\vartheta_i], b), \quad \text{where } \vartheta_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_p x_{p,i} = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

possible excluding large claims (and pooling on the entiere population)

$$\mathbb{E}[Y|\boldsymbol{X}] = \underbrace{\mathbb{P}[Y \leq s|\boldsymbol{X}]}_{p_s} \cdot \underbrace{\mathbb{E}[Y|\boldsymbol{X}, Y \leq s]}_{\mu_s(\boldsymbol{x})} + \underbrace{\mathbb{P}[Y > s|\boldsymbol{X}]}_{1-p_s} \cdot \underbrace{\mathbb{E}[Y|\boldsymbol{X}, Y > s]}_{\overline{y}}$$

or some Tweedie regression on S (but less flexible), Tweedie (1984).

From Econometric Techniques to Machine Learning (3)

Probabilistic interpretations : estimation using maximum likelihood - Generalized (Mixed) Linear Models (see McCullagh and Nelder (1989) or Jørgensen (1997))

- Individual model: $S = \mathbf{1}_{S>0} \cdot \widetilde{S}$
- Collective model: $S = \sum_{i=1}^{N} Y_i$, $N \sim \mathcal{P}(\lambda)$, $Y_i \sim \mathcal{G}(\alpha, \beta)$ in the exponential family,

with variance function $V(\mu) = \mu^k$, where $k \in [1, 2]$

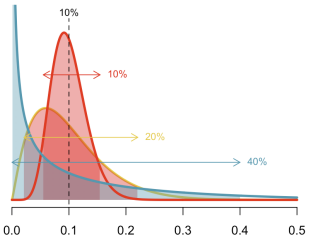
$$\log \mathcal{L} = \sum_{i=1}^{n} \underbrace{y_i \frac{\mu_i^{1-k}}{1-k} - \frac{\mu_i^{2-k}}{2-k}}_{-\ell(y_i, \mu_i)}, \text{ where } \mu_i = \exp[\boldsymbol{x}_i^\top \boldsymbol{\beta}]$$

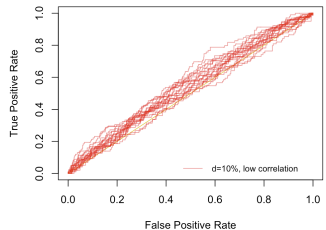
Machine learning : interpret $-\log \mathcal{L}$ as a loss
One can also include some penalty if the number of possible covariates is too large...

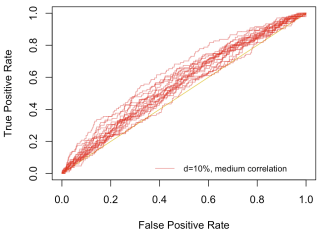
Goodness of Fit & Uncertainty (1)

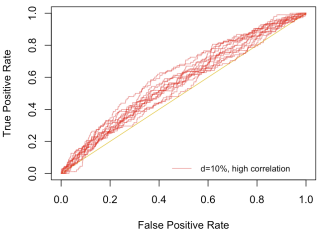
Consider a simple model: number of claims $N \in \{0, 1\}$ and fixed cost, say $\$1,000$.
Simple classification problem.
Assume that $N \sim \mathcal{B}(\theta)$ and we have some covariate x.
Two important features:

- the dispersion of the heterogeneity, i.e. the variance of θ
  e.g. assume that θ has a Beta distribution on $[0, 1]$
- the dependence between heterogeneity θ and covariate x







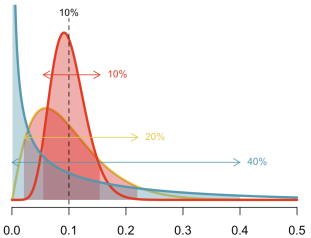


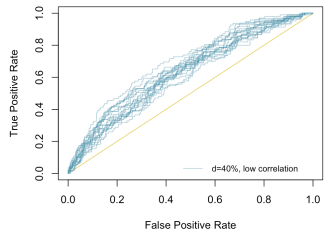
$\leftarrow$ low correlation high correlation $\rightarrow$

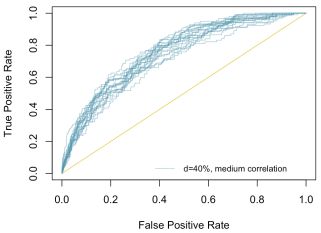
Goodness of Fit & Uncertainty (1)

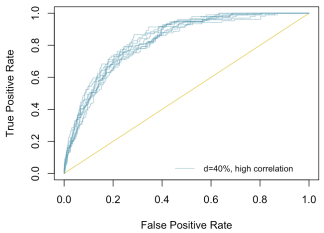
Consider a simple model: number of claims $N \in \{0, 1\}$ and fixed cost, say $\$1,000$.
Simple classification problem.
Assume that $N \sim \mathcal{B}(\theta)$ and we have some covariate x.
Two important features:

- the dispersion of the heterogeneity, i.e. the variance of θ
  e.g. assume that θ has a Beta distribution on $[0, 1]$
- the dependence between heterogeneity θ and covariate x







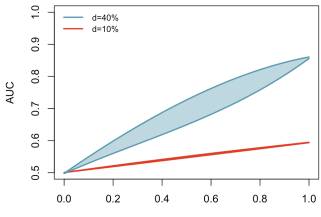


← low correlation high correlation →

Goodness of Fit & Uncertainty (2)

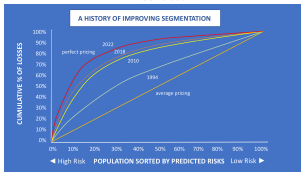
- the dispersion of the heterogeneity, $d = 10\%$ or 40%
- the dependence between heterogeneity θ and covariate x (correlation of the underlying copula function)

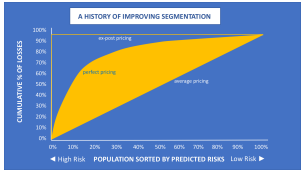
very difficult to reach high AUC (area under the ROC curve)



More complicated for insurance premiums...
Frees et al. (2014) defined a ROC-type curve, inspired by Lorenz curve: given observed losses s_i and premiums $\widehat{\pi}(\boldsymbol{x}_i)$, policyholders ordered by premiums, $\widehat{\pi}(\boldsymbol{x}_1) \geq \widehat{\pi}(\boldsymbol{x}_2) \geq \cdots \geq \widehat{\pi}(\boldsymbol{x}_n)$, plot

$$\{F_i, L_i\} \text{ with } \underbrace{F_i = \frac{i}{n}}_{\text{proportion of insured}} \text{ and } \underbrace{L_i = \frac{\sum_{j=1}^{i} s_j}{\sum_{j=1}^{n} s_j}}_{\text{proportion of losses}}$$



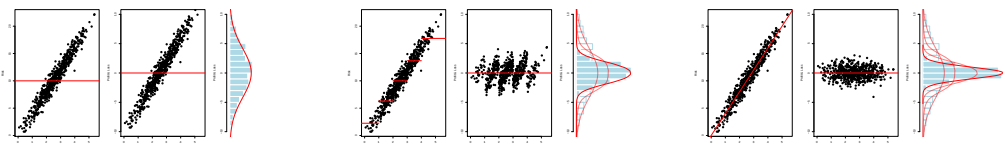


Goodness of Fit & Uncertainty (3)

Parallel with classical OLS models, $S \sim \mathcal{N}(\theta, \sigma^2)$, with covariates $\boldsymbol{X} = \{X_1, \cdots, X_p\}$

$$\mathrm{Var}\Big[S\Big] = \mathbb{E}\Big[\mathrm{Var}[S|\boldsymbol{X}]\Big] + \mathrm{Var}\Big[\underbrace{\mathbb{E}[S|\boldsymbol{X}]}_{\text{premium}}\Big]$$

$$\mathbb{E}\Big[\mathrm{Var}[S|\boldsymbol{X}]\Big] = \underbrace{\mathbb{E}\Big[\mathrm{Var}[S|\Theta]\Big]}_{\text{perfect pricing} = \sigma^2} + \underbrace{\mathbb{E}\left\{\mathrm{Var}\Big[\mathbb{E}[S|\Theta]\Big|\boldsymbol{X}\Big]\right\}}_{\text{misfit}}$$



(1) (θ_i, s_i) and $\mathbb{E}(S_i|\boldsymbol{X_i})$ (2) $(\theta_i, s_i - \mathbb{E}(S_i|\boldsymbol{X_i}))$ (3) distribution of $S - \mathbb{E}(S|\boldsymbol{X})$

Model Interpretation & Explainability

Most machine linearing algorithms are black boxes, Pasquale (2015)
"*providing transparency and explanations of algorithmic outputs might serve a number of goals, including scrutability, trust, effectiveness, persuasiveness, efficiency, and satisfaction* " Diakopoulos (2016)

- Ceteris paribus can be translated into "*all other things being equal*" or "*holding other factors constant*"
- Mutatis mutandis approximately translates as "*allowing other things to change accordingly*" or "*the necessary changes having been made*"

Guidotti et al. (2018) for a survey on methods for explaining back boxes

- explaining the model or explaining the outcome
- providing a *transparent* design (locally or globally)

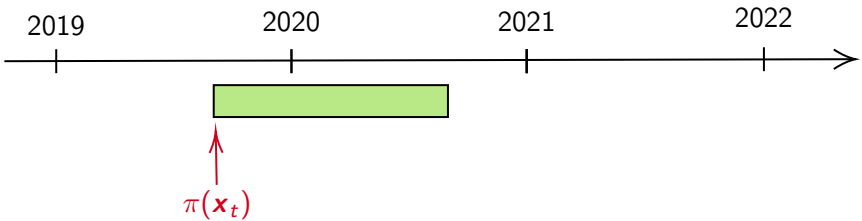
Insurance Premium Fairness

Various concepts and definitions,

- pure actuarial fairness: insurance costs for individuals should directly reflect their level of risk
- choice-sensitive equity: insurance costs for individuals should reflect only the risks that result from individual choices (see luck-egalitarian)
- equity as social justice: insurance of goods that are basic requirements of social justice should be provided irrespective of the risks and choices of individuals

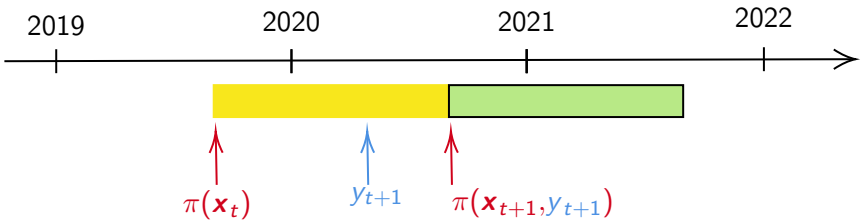
Economic and philosophical issues.... but also statistical ones, see Feldman et al. (2015), Bonchi et al. (2017) or Corbett-Davies and Goel (2018)

Prior and Posterior Insurance Pricing (1)



Prior premium for period $t+1$ is a function of features available $\boldsymbol{x}_t$

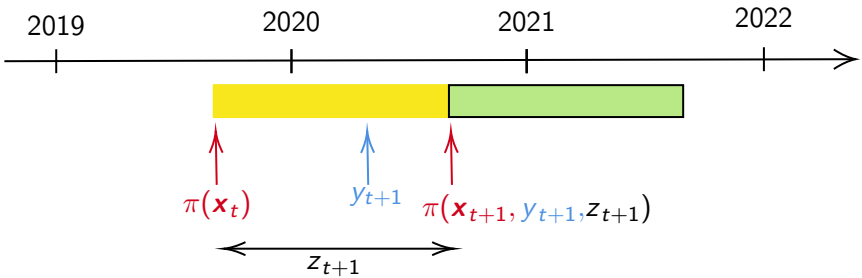
Prior and Posterior Insurance Pricing (2)



Prior premium for period $t+1$ is a function of features available $\boldsymbol{x}_t$

Posterior premium for period $t+2$ is a function of $\boldsymbol{x}_{t+1}$ and claims history y_{t+1}

Prior and Posterior Insurance Pricing (3)



Prior premium for period $t+1$ is a function of features available $\boldsymbol{x}_t$

Posterior premium for period $t+2$ is a function of $\boldsymbol{x}_{t+1}$, y_{t+1} and driving experience z_{t+1}

Wrap-Up & Open Challenges

See Charpentier and Denuit (2020) for a recent state-of-the-art

- predictive models (econometrics or ML) are everywhere in insurance
- impossible to use black boxes for ratemaking
- more and more regulation towards transparency, fairness, equity
- so far, only actuarial modeling, not much about economics of insurance (moral hazard, competition, price elasticity, etc)
- how to include *ex-post* observations, e.g. telematics (endogeneity issues)?
- how to include competition? (learning games)

- to go further freakonometrics.hypotheses.org or charpentier.arthur@uqam.ca

References I

Arneson, R. J. (2011). Luck egalitarianism–a primer. In Knight, C. and Stemplowska, Z., editors, *Responsibility and Distributive Justice*, pages 24–50. Oxford University Press.

Bonchi, F., Hajian, S., Mishra, B., and Ramazzotti, D. (2017). Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, 3(1):1–21.

Charpentier, A. and Denuit, M. (2020). On limits for machine learning algorithms in insurance. In *Insurance data analytics : some case studies of advanced algorithms and applications*, pages 210–235. Economica.

Charpentier, A., Élie, R., and Remlinger, C. (2020). Reinforcement learning in economics and finance. *arXiv*, 2003.10014.

Charpentier, A., Flachaire, E., and Ly, A. (2018). Econometrics and machine learning. *Economics and Statistics*, 505-506:147–169.

Cohen, G. A. (1989). On the currency of egalitarian justice. *Ethics*, 99(4):906–944.

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv*, 1808.00023.

de Leeuw, J., Vrieling, A., Shee, A., Atzberger, C., Hadgu, K., Biradar, C., Keah, H., and Turvey, C. (2014). The potential and uptake of remote sensing in insurance: A review. *Remote Sensing*, 6(11):10888–10912.

References II

Denuit, M., Hainault, D., and Trufin, J. (2019a). *Effective Statistical Learning Methods for Actuaries I (GLMs and Extensions)*. Springer Verlag.

Denuit, M., Hainault, D., and Trufin, J. (2019b). *Effective Statistical Learning Methods for Actuaries III (Neural Networks and Extensions)*. Springer Verlag.

Denuit, M., Hainault, D., and Trufin, J. (2020). *Effective Statistical Learning Methods for Actuaries II (Tree-based methods and Extensions)*. Springer Verlag.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Commununications of the ACM*, 59(2):56–62.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA. Association for Computing Machinery.

Frees, E. W. J., Meyers, G., and Cummings, A. D. (2014). Insurance ratemaking and a gini index. *The Journal of Risk and Insurance*, 81(2):335–366.

References III

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Jørgensen, B. (1997). *The Theory of Dispersion Models*. Chapman & Hall.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman & Hall.

O'Neill, O. (1997). Genetic information and insurance: Some ethical issues. *Philosophical Transactions: Biological Sciences*, 352(1357):1087–1093.

Óskarsdóttir, M., Ahmed, W., Antonio, K., Bart Baesens, Rémi Dendievel, T. D., and Reynkens, T. (2019). Social network analytics for supervised fraud detection in insurance. *KUL Working Paper*.

Pasquale, F. (2015). *The black box society: the secret algorithms that control money and information*. Harvard University Press.

Schauer, F. (2006). *Profiles, Probabilities and Stereotypes*. Harvard University Press.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. *Statistics: applications and new directions (Calcutta, 1981)*, pages 579–604.

Wüthrich, M. V. (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, 2018(6):465–480.