Machine Learning in Actuarial Science & Insurance

Arthur Charpentier

Machine learning for economists and applied social scientists, 2020
Arthur Charpentier

Université du Québec à Montréal

@freakonometrics
freakonometrics
freakonometrics.hypotheses.org
What would be a “fair contribution”? see O’Neill (1997)

- pure actuarial fairness contributions for individual policyholders should perfectly reflect their predicted risk levels → 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

From Econometric Techniques to Machine Learning

Goodness of Fit & Uncertainty

Model Interpretation

Price Discrimination & Fairness

To Go Further
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, \( \pi = \mathbb{E}_P(S) \)

(possible) losses \( S = Y_1 + \cdots + Y_N \)

Premium \( \pi = \mathbb{E}_P(S) \)

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

Under standard assumptions,

\[
\pi(x) = \mathbb{E}[S|X] = \mathbb{E}[N|X] \cdot \mathbb{E}[Y|X] = \mathbb{E}[1_{S>0}|X] \cdot \mathbb{E}[S|X, S > 0]
\]

- frequency
- average cost
- occurrence
- individual cost
Merging claims & underwriters databases (per policy), e.g. \((n_i, e_i, x_i), i = 1, \cdots, n_p\)

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

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

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

or possibly GAMs (for nonlinearities, or spatial component)

\[ \theta_i = \alpha_0 + s_1(x_{1,i}) + \alpha_2 x_{2,i} + \cdots + \alpha_p x_{p,i}, \quad s_1(x) = \sum_j \gamma_j \varphi_j(x) \]
From Econometric Techniques to Machine Learning (2)

or \((y_i x_i), i = 1, \cdots, n_c\)

<table>
<thead>
<tr>
<th>id</th>
<th>no cover</th>
<th>cost</th>
<th>zone</th>
<th>pwr</th>
<th>agecar</th>
<th>agedrv</th>
<th>model</th>
<th>gas</th>
<th>dsty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1870</td>
<td>17219</td>
<td>1TP</td>
<td>C</td>
<td>5</td>
<td>0</td>
<td>52</td>
<td>12</td>
<td>R</td>
</tr>
<tr>
<td>2</td>
<td>1963</td>
<td>16336</td>
<td>1TP</td>
<td>E</td>
<td>9</td>
<td>0</td>
<td>78</td>
<td>12</td>
<td>R</td>
</tr>
<tr>
<td>3</td>
<td>4263</td>
<td>17089</td>
<td>2MT</td>
<td>C</td>
<td>10</td>
<td>7</td>
<td>27</td>
<td>17</td>
<td>D</td>
</tr>
<tr>
<td>4</td>
<td>5181</td>
<td>17801</td>
<td>1TP</td>
<td>D</td>
<td>5</td>
<td>2</td>
<td>26</td>
<td>3</td>
<td>D</td>
</tr>
<tr>
<td>5</td>
<td>6375</td>
<td>17485</td>
<td>1TP</td>
<td>B</td>
<td>7</td>
<td>4</td>
<td>46</td>
<td>6</td>
<td>R</td>
</tr>
</tbody>
</table>

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

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

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

\[
\mathbb{E}[Y|X] = \mathbb{P}[Y \leq s|X] \cdot \mathbb{E}[Y|X, Y \leq s] + \mathbb{P}[Y > s|X] \cdot \mathbb{E}[Y|X, Y > s]
\]

or some Tweedie regression on \(S\) (but less flexible), Tweedie (1984).
Probabilistic interpretations: estimation using maximum likelihood - Generalized (Mixed) Linear Models (see McCullagh and Nelder (1989) or Jørgensen (1997))

- Individual model: \( S = 1_{S > 0} \cdot \tilde{S} \)

- 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 \( k \in [1, 2] \)

\[
\log \mathcal{L} = \sum_{i=1}^{n} y_i \frac{\mu_i^{1-k}}{1-k} - \frac{\mu_i^{2-k}}{2-k} - \ell(y_i, \mu_i), \quad \text{where} \quad \mu_i = \exp[x_i^T \beta]
\]

Machine learning: interpret \( -\log \mathcal{L} \) as a loss
One can also include some penalty if the number of possible covariates is too large...
Consider a simple model: number of claims $N \in \{0, 1\}$ and fixed cost, say $1,000$. Simple classification problem.

Assume that $N \sim B(\theta)$ and we have some covariate $x$.

Two important features:

- the dispersion of the heterogeneity, i.e. the variance of $\theta$
  
  e.g. assume that $\theta$ has a Beta distribution on $[0, 1]$

- the dependence between heterogeneity $\theta$ and covariate $x$

\[ 0.0 \quad 0.1 \quad 0.2 \quad 0.3 \quad 0.4 \quad 0.5 \]

\[ 0% \quad 10% \quad 20% \quad 40% \]

\[ \text{False Positive Rate} \quad \text{True Positive Rate} \]

\[ d=10\%, \text{low correlation} \]

\[ d=10\%, \text{medium correlation} \]

\[ d=10\%, \text{high correlation} \]
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 \( \theta \)
  
e.g. assume that \( \theta \) has a Beta distribution on \([0, 1]\)
- ▶ the dependence between heterogeneity \( \theta \) and covariate \( x \)

\[\begin{align*}
\text{Low correlation} & \quad \text{High correlation} \\
\text{False Positive Rate} & \quad \text{False Positive Rate} \\
\text{True Positive Rate} & \quad \text{True Positive Rate}
\end{align*}\]
Goodness of Fit & Uncertainty (2)

- the dispersion of the heterogeneity, $d = 10\%$ or $40\%$
- the dependence between heterogeneity $\theta$ 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 $\hat{\pi}(x_i)$, policyholders ordered by premiums, $\hat{\pi}(x_1) \geq \hat{\pi}(x_2) \geq \cdots \geq \hat{\pi}(x_n)$, plot

$$\{F_i, L_i\} \quad \text{with} \quad F_i = \frac{i}{n} \quad \text{and} \quad L_i = \frac{\sum_{j=1}^{i} s_j}{\sum_{j=1}^{n} s_j}$$

(proportion of insured) (proportion of losses)
Goodness of Fit & Uncertainty (3)

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

\[
\text{Var}[S] = \mathbb{E}\left[\text{Var}[S|\mathbf{X}]\right] + \text{Var}\left[\mathbb{E}[S|\mathbf{X}]\right]
\]

\[
\mathbb{E}\left[\text{Var}[S|\mathbf{X}]\right] = \mathbb{E}\left[\text{Var}[S|\Theta]\right] + \mathbb{E}\left\{\text{Var}\left[\mathbb{E}[S|\Theta]|\mathbf{X}\right]\right\}
\]

(1) \((\theta_i, s_i)\) and \(\mathbb{E}(S_i|\mathbf{X}_i)\)
(2) \((\theta_i, s_i - \mathbb{E}(S_i|\mathbf{X}_i))\)
(3) distribution of \(S - \mathbb{E}(S|\mathbf{X})\)
Model Interpretation & Explainability (1)

Most machine learning 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)
Model Interpretation & Explainability (2)

Various tools:

- Local Interpretable Model-Agnostic Explanations (LIME) for general models,
- and for regression types, partial dependence plots (PDP), Individual Conditional Expectation (ICE), or Accumulated Local Effects (ALE)

\[
pdp(x_1) = \mathbb{E}[\pi(x_1, X_2)] \quad \text{and} \quad \widehat{pdp}(x_1) = \frac{1}{n} \sum_{i=1}^{n} \widehat{\pi}(x_1, x_{i,2})
\]

Or define \( ale(x_1) = \int_{-\infty}^{x_1} \mathbb{E} \left[ \frac{\partial \pi(z_1, X_2)}{\partial x_1} \right] dz_1 \), as in Apley and Zhu (2016).

For more details, see https://christophm.github.io/interpretable-ml-book/, or Lakkaraju et al. (2019), Samek et al. (2017), Ribeiro et al. (2016), Lipton (2016), Lundberg and Lee (2017), Gilpin et al. (2018), etc.
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. Formally, $\mathbf{x} = (x_1, x_2)$ where only variables $x_2$ can be used.
Algorithmic Fairness (for classifiers)

Feldman et al. (2015), Bonchi et al. (2017) or Corbett-Davies and Goel (2018)

Let $s$ denote the (underlying) score of a classifier $s(x) = \mathbb{P}[Y = 1 | X = x]$, and $p$ the classifier $p(x) = 1_{[t, \infty)}(s(x))$.

- **anti-classification**: some attributes $x_1$ cannot be used

\[ p(x_1, x_2) = p(x'_1, x_2) \text{ whenever } x_1 \neq x'_1 \]

- **classification parity**: standard metric do not depend on $x_1$,
  e.g. for a classifier, the false positive rate should not depend on $x_1$

\[ \mathbb{P}[p(X_1, X_2) = 1 | Y = 0, X_1 = x_1] = \mathbb{P}[p(X_1, X_2) = 1 | Y = 0], \quad \forall x_1. \]

- **calibration equity**:

\[ \mathbb{P}[Y = 1 | S(X_1, X_2) = s, X_1 = x_1] = \mathbb{P}[Y = 1 | S(X_1, X_2) = s] \]
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, nothing 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


References II


References III


