Loss Reserving Analytics

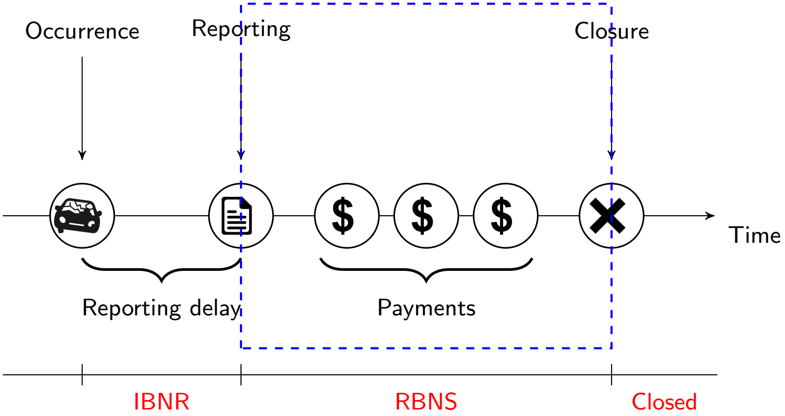
Micro-level RBNS reserving

Katrien Antonio
LRisk - KU Leuven and ASE - University of Amsterdam

April, 2019

Introduction

Development of a single claim



Research focus

Development of an RBNS claim

- Research questions with focus on RBNS:
  - What is the number of payments for an RBNS claim?
  - What is the size of these future payments?
  - When do we make these payments?
  - When will the claim settle or close?
  - Combining the model for IBNR claim counts and the development of RBNS claims, can we allocate an IBNR reserve?

Research contributions

- Antonio & Plat (2014, SAJ):

  development of individual claims in continuous time, parametric

  inspired by Norberg (1993,1999), Haastrup & Arjas (1996), Cook & Lawless (2007).

- Pigeon, Antonio & Denuit (2013, ASTIN Bulletin):

  development of individual claims in discrete time, parametric, complex, rigid dependence structure for individual development factors

  inspired by chain–ladder method.

Research contributions

- Pigeon, Antonio & Denuit (2014, IME):

  development of individual claims in discrete time, claim payments and incurred losses, parametric with complex, rigid dependence structure

  inspired by chain–ladder method, inspired by PIC method of Wüthrich & Merz (2010).

- Antonio & Godecharle (2015, NAAJ):

  development of individual claims in discrete time, historical simulation from empirical data, include claim markers

  inspired by Drieskens, Henry, Walhin & Wielandts (SAJ, 2012) and Rosenlund (ASTIN, 2012).

Research contributions

- Nielsen, Miranda Martinez, Verrall, Hiabu et al.:

  Double Chain Ladder (DCL, see R package) (2012)

  Continuous Chain Ladder (2013, 2019).

- (Many) Other (recent) contributions in this field including

  Zhao et al. (2009, 2010), Huang et al. (2015a,b, 2016),

  Larsen (2007), Denuit and Trufin (2017, 2018), Wuthrich (2018a,b), ASTIN Working Party (2017, 2018).

!Highly scattered literature!

My (short to mid-term) mission statement

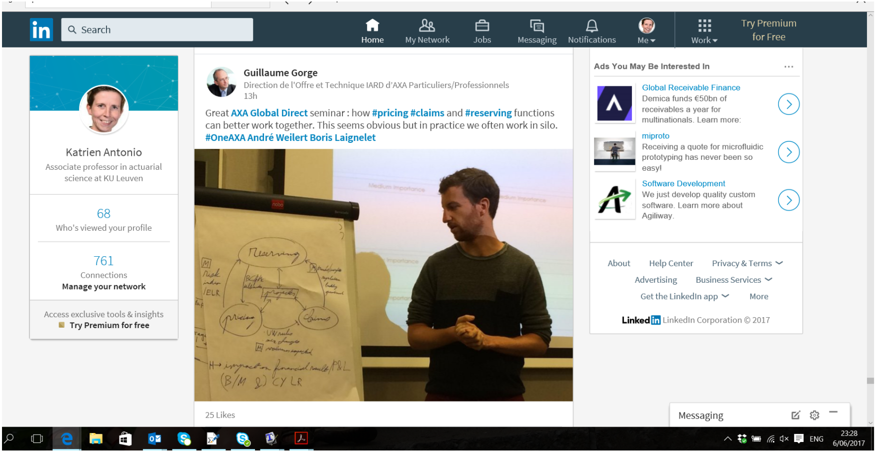
1. Structure the (highly) scattered literature on analytics for loss reserving.
2. Hybrid strategy, take data (or: feature) driven position between micro- and macro-level.
3. *Less is more*, unify pricing and reserving methodology, while preserving the advantages of the macro approach.
4. Lessons to learn from the machine learning literature.
5. Use multiple evaluation dates, instead of single out-of-time evaluation.
6. Use multiple portfolios, *no free lunch*.

My (short to mid-term) mission statement

Mr M. H. Tripp, F.I.A.: Why do we throw away information? This question has already been hinted at, and needs reinforcing in the domain for thinking about in the future. I have never been keen on silos, and it is important to learn between disciplines. Looking at the life side of our profession, you realise that work like this takes place at policy level detail. If you look within the general insurance part of the actuarial profession, there is a body of thinking that has grown up around premium rating and a body of thinking that has grown up around reserving. Are we getting 'over-siloed'? Could aspects of the methodology and the thinking that has gone into using GLMs for premium rating be brought more into play when it comes to reserving, where, at present, we tend to use aggregated claims data? I wonder whether we are missing out on using information that is available from exposure descriptions and from the circumstances of individual claims. I know that the traditional response to this is that there is all too much variability, but, in attempts to remove heterogeneity from data and to try to find better for the future, I look for support in thinking this through.

M. Tripp, F.I.A., Stochastic claims reserving - Abstract of the discussion, British Actuarial Journal, 2002.

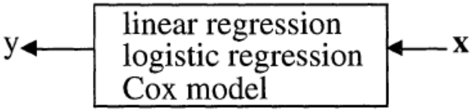
My (short to mid-term) mission statement



Katrien's LinkedIn timeline, June 2017.

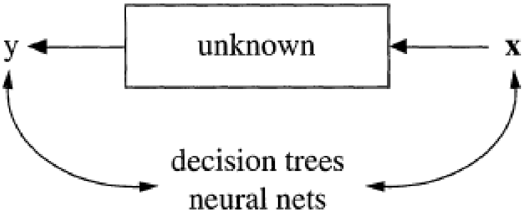
Analytics: 'the two cultures'

- Read the Breiman (2001, Stat Science) paper on *Statistical modeling: the two cultures*.
- Data modeling culture
  - assume stochastic data model, estimate parameter values
  - validate with goodness-of-fit tests and residual inspection.



Analytics: ‘the two cultures’

- Read the Breiman (2001, Stat Science) paper on *Statistical modeling: the two cultures*.
- Algorithmic modeling culture
    - inside of the box is complex and unknown
    - find algorithm $f(\boldsymbol{x})$ to predict $\boldsymbol{y}$
    - measure by predictive accuracy.



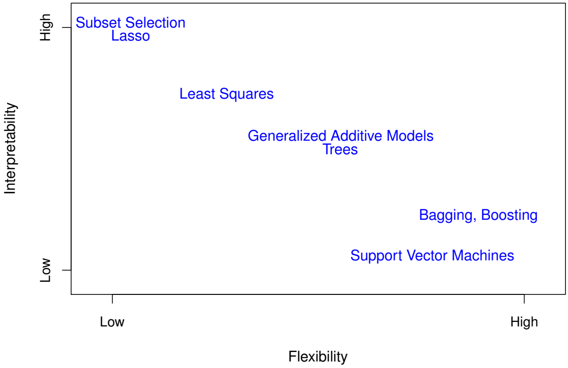
Analytics: 'the two cultures'

	Statistical Learning	Machine Learning
origin	statistics	computer science
$f(X)$	model	algorithm
emphasis	interpretability, precision and uncertainty	large scale applicability, prediction accuracy
jargon	parameters, estimation	weights, learning
CI	uncertainty of parameters	no notion of uncertainty
assumptions	explicit a priori assumption	no prior assumption, learn from the data

(Taken from Why a mathematician, statistician and machine learner solve the same problem differently.)

Analytics: (some) challenges

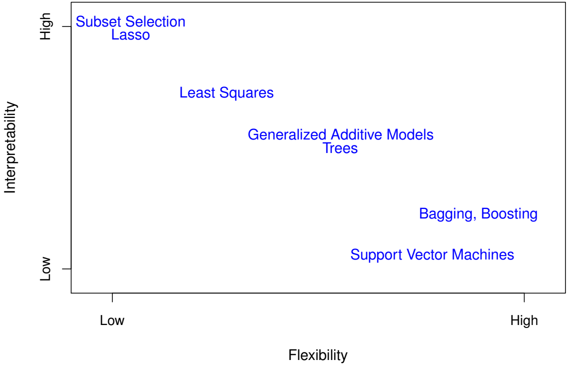
Trade off between flexibility and model interpretability!



(Picture taken from James et al., An introduction to statistical learning, 2017.)

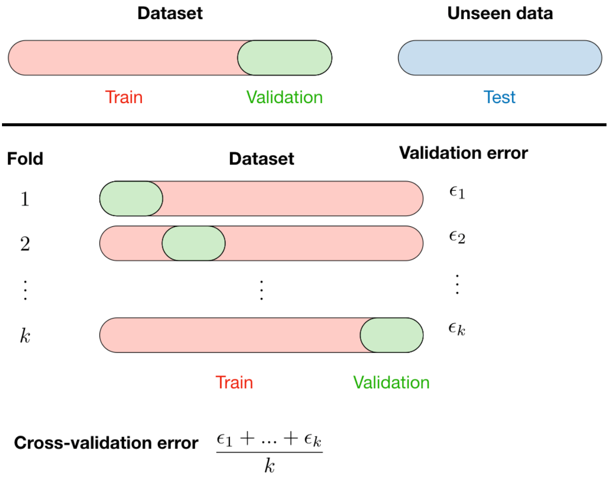
Analytics: (some) challenges

The 'black box' versus 'white box' discussion!



(Picture taken from James et al., An introduction to statistical learning, 2017.)

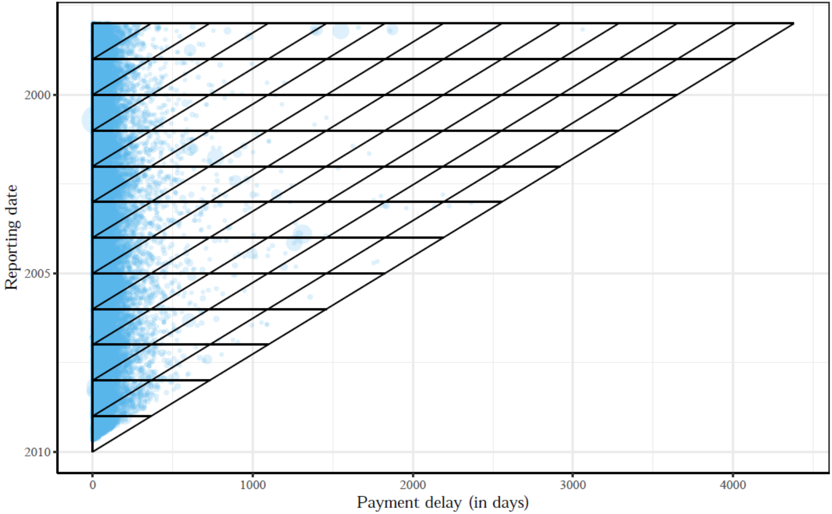
Analytics: model building strategy



(Picture taken from https://developer.ibm.com/recipes/tutorials/machine-learning-and-ibm-watson-studio/.)

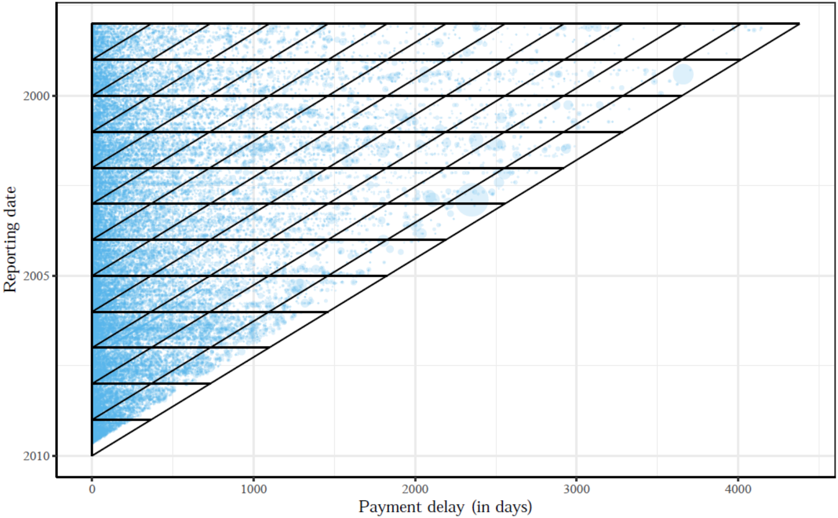
Development of an RBNS claim

Continuous time



Development of an RBNS claim

Continuous time



Development of an RBNS claim

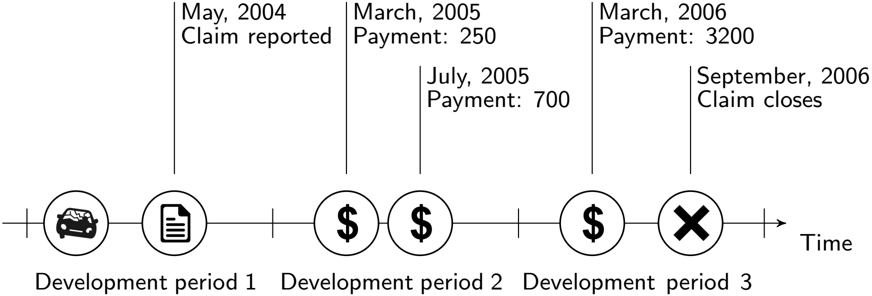
Triangular view

meldings jaar	afwikkelingsjaar 1	2	3	4	5	6	7	8	9	10	11
1998	4 531 234	831 294	38 768	24 946	79 026	21 622	690	18 311	8893	6782	12 305
1999	5 757 864	1 031 688	42 510	117 949	108 993	13 819	476	0	0	0	0
2000	6 124 614	965 253	150 217	17 848	6753	3193	0	12 388	0	0	
2001	6 522 440	1 455 000	105 690	121 644	15 159	3722	446	0	0		
2002	8 020 500	1 482 132	59 294	57 939	1994	5512	11 844	815			
2003	8 316 119	1 599 504	90 951	34 840	30 319	9551	12 992				
2004	9 111 343	1 383 280	73 007	36 161	24 432	8047					
2005	9 628 820	1 563 299	189 738	3367	19 982						
2006	10 048 711	2 082 880	65 442	139 400							
2007	10 780 382	1 722 627	76 651								
2008	10 669 883	1 594 990									
2009	6 625 840										

Mind the use of reporting year instead of occurrence year!

Development of an RBNS claim

Discrete time



- Claim reported
- Payment: 0

- Payment: 950

- Payment: 3200
- Claim closed

Development of an RBNS claim

Notations

- Index the individual claims by k and the development periods by j.
- Per development period:
  - C_{kj}: closure indicator
  - P_{kj}: payment indicator
  - Y_{kj}: payment size.
- The three essential building blocks (cfr. Antonio & Plat, 2014)!
- Mind the presence of covariates!

Development of an RBNS claim

Type of covariates

- Deterministic covariates

  Known for future development periods.

- Dynamic covariates

  Functions of past development information, unknown for future development periods.

- Feature engineering matters!

Development of an RBNS claim

Hierarchical likelihood

The likelihood of the observed development process for a single claim is:

$$
\begin{aligned}
f(C_{1,\ldots,T}, P_{1,\ldots,T}, Y_{1,\ldots,T}) = & \prod_{j=1}^{T} f(C_j \mid C_{1,\ldots,j-1}, P_{1,\ldots,j-1}, Y_{1,\ldots,j-1}) \times \\
& \prod_{j=1}^{T} f(P_j \mid C_{1,\ldots,j}, P_{1,\ldots,j-1}, Y_{1,\ldots,j-1}) \times \\
& \prod_{j=1}^{T} f(Y_j \mid C_{1,\ldots,j}, P_{1,\ldots,j}, Y_{1,\ldots,j-1}),
\end{aligned}
$$

where T is the number of observed development years.

We model the building blocks C, P and Y in this likelihood with a Generalized Linear Model (GLM).

The statistical model

GLMs

- Closure indicator:

  Binomial GLM with complementary log-log link

  $$\Pr(C_j = 1 \mid C_{1,\ldots,j-1} = 0, P_{1,\ldots,j-1}, Y_{1,\ldots,j-1}) = 1 - \exp(-\exp(\boldsymbol{y}' \cdot \boldsymbol{\beta})).$$

- Payment indicator:

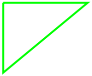
  Binomial GLM with logit link

  $$\Pr(P_j = 1 \mid C_{1,\ldots,j-1} = 0, C_j, P_{1,\ldots,j-1}, Y_{1,\ldots,j-1}) = \frac{\exp(\boldsymbol{y}' \cdot \gamma)}{1 + \exp(\boldsymbol{y}' \cdot \gamma)}$$

- Payment size: Gamma GLM, below a high threshold (to exclude very large payments).

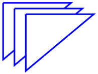
The statistical model

RBNS claim development



Chain ladder (Poisson GLM)

Classical chain ladder applied to the run-off triangle with payments aggregated per `reporting` and `development year`.



Aggregate GLM

GLM fitted for each of $(C, P$ and $Y)$ using only the covariates `reporting` and `development year`.

Granular GLM

GLM fitted for each of $(C, P$ and $Y)$ with selected (static and dynamic) covariates.

Model selection

Imbalance in-sample and out-of-sample data

Run-off triangle with number of open claims per reporting and development year

reporting year	development year 1	2	3	4	5	6
1998	14 507	2256	51	11	6	5
1999	15 936	2325	75	24	11	4
2000	15 818	2224	73	18	6	3
2001	17 079	2895	103	29	14	3
2002	19 656	2929	112	31	12	6
2003	18 342	2713	137	25	12	3

In-sample and out-of-sample distribution of the development year:

	1	2	3	4	5	6
in-sample (%)	88.626	11.045	0.264	0.046	0.015	0.004
out-of-sample (%)	0	87.151	7.999	2.730	1.413	0.707

Model selection

Imbalance in-sample and out-of-sample data

Divide the run-off triangle in training, validation and evaluation cells

reporting year	development year 1	2	3	4	5	6
1998	14 507	2256	51	11	6	5
1999	15 936	2325	75	24	11	4
2000	15 818	2224	73	18	6	3
2001	17 079	2895	103	29	14	3
2002	19 656	2929	112	31	12	6
2003	18 342	2713	137	25	12	3

... and verify the distribution of development year again:

	1	2	3	4	5	6
training (%)	91.063	8.717	0.179	0.031	0.005	0.004
validation (%)	0	95.688	3.365	0.588	0.359	0
evaluation (%)	0	87.151	7.999	2.730	1.413	0.707

Model selection

Imbalance in-sample and out-of-sample data

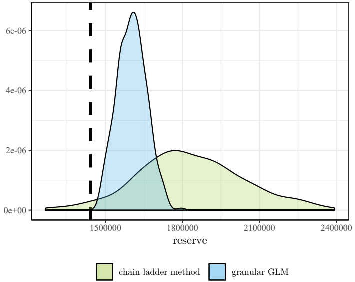
- We apply a model building strategy in line with machine learning literature:
  - Calibrate on training cells.
  - Select covariates on validation cells.
  - Recalibrate on training and validation cells, predict the evaluation cells.

Model evaluation

Traditional one day view

Fit and evaluate the model on 31 December 2003:

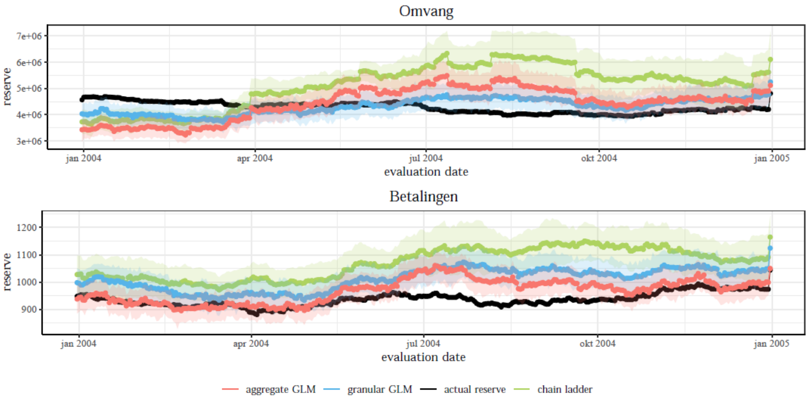
dev. year	actual	granular GLM	chain ladder
2	1 110 556	1 140 453	1 281 761
3	126 417	119 937	125 258
4	130 200	184 242	71 107
5	44 753	102 647	249 168
6	29 633	55 475	129 629
total	1 441 560	1 602 757	1 856 926



Model evaluation

Dynamic view

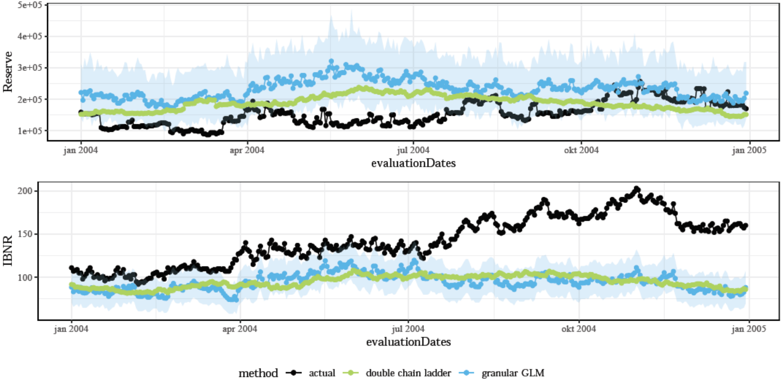
Moving window, fit and evaluate the reserve over an extended period of time (RBNS, LTS).



Model evaluation

Dynamic view

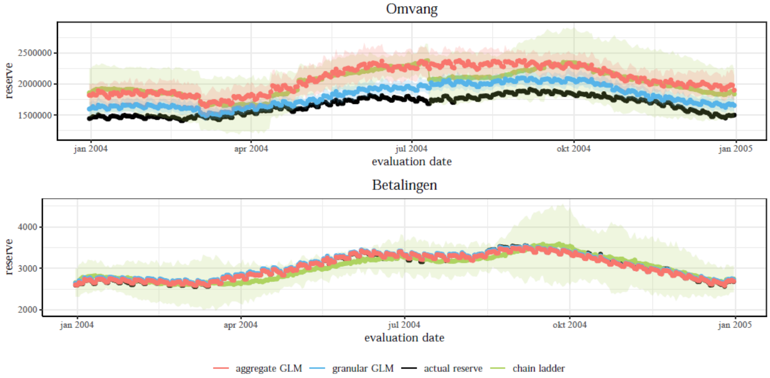
Moving window, fit and evaluate the reserve over an extended period of time (IBNR, LTS).



Model evaluation

Dynamic view

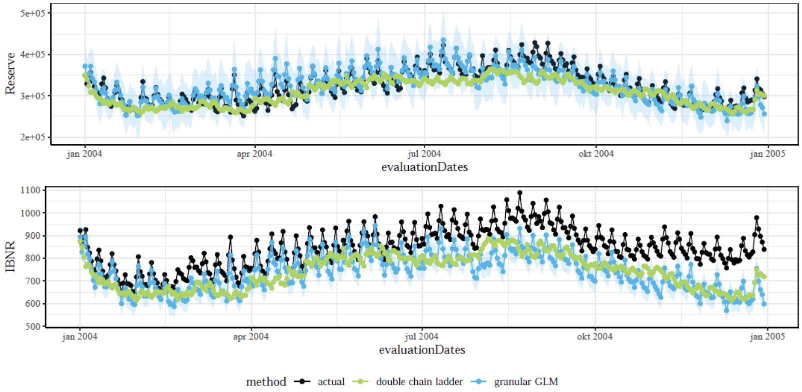
Moving window, fit and evaluate the reserve over an extended period of time (RBNS, MAT).



Model evaluation

Dynamic view

Moving window, fit and evaluate the reserve over an extended period of time (IBNR, MAT).



Structuring the scattered literature

Denuit & Trufin (2018)

Reserving by combining multiple aggregated triangles for closure, payment and size.

Wüthrich (2018)

Machine learning method (regression tree) for the payment and closure indicator.

Larsen (2007)

GLM fitted for each of the components (C, P and Y) in the development process.

Analytics

Going beyond GLMs in loss reserving?

Table A.1: A summary of models and some of their characteristics

Model	Allows $n < p$	Pre-processing	Interpretable	Automatic feature selection	# Tuning parameters	Robust to predictor noise	Computation time
Linear regression†	×	CS, NZV, Corr	✓	×	0	×	✓
Partial least squares	✓	CS	✓	∘	1	×	✓
Ridge regression	×	CS, NZV	✓	×	1	×	✓
Elastic net/lasso	×	CS, NZV	✓	✓	1–2	×	✓
Neural networks	✓	CS, NZV, Corr	×	×	2	×	×
Support vector machines	✓	CS	×	×	1–3	×	×
MARS/FDA	✓		∘	✓	1–2	∘	∘
K-nearest neighbors	✓	CS, NZV	×	×	1	∘	✓
Single trees	✓		∘	✓	1	✓	✓
Model trees/rules†	✓		∘	✓	1–2	✓	✓
Bagged trees	✓		×	✓	0	✓	∘
Random forest	✓		×	∘	0–1	✓	×
Boosted trees	✓		×	✓	3	✓	×
Cubist†	✓		×	∘	2	✓	×
Logistic regression*	×	CS, NZV, Corr	✓	×	0	×	✓
{LQRM}DA*	×	NZV	∘	×	0–2	×	✓
Nearest shrunken centroids*	✓	NZV	∘	✓	1	×	✓
Naïve Bayes*	✓	NZV	×	×	0–1	∘	∘
C5.0*	✓		∘	✓	0–3	✓	×

†regression only *classification only

Symbols represent affirmative (✓), negative (×), and somewhere in between (∘)

(Picture taken from Kuhn & Johnson, Applied predictive modeling, 2013.)

Analytics

Going beyond GLMs in loss reserving?

- Wüthrich (2018, SAJ) builds a classifier (with `rpart`)

$$\Pr[Y^{(\nu)}_{ij|k+1} = y, Z^{(\nu)}_{ij|k+1} = z | \mathcal{F}_{i+j+k}] = p^{(y,z)}_{j+k}(\boldsymbol{x}^{(\nu)}_{i,j|k}),$$

  where

  - Y is indicator 'payment/no payment'
  - Z is indicator 'open/closed'
  - ν is a claim that occurred in year i, was reported with j years of delay
  - k refers to run-off
  - $\boldsymbol{x} \in \mathcal{X}$ refers to available (static, dynamic) features.

Analytics

Going beyond GLMs in loss reserving?

- Wüthrich (2018, EAJ) refines CL structure to

$$C_{i,j}(\boldsymbol{x}) \approx f_{j-1}(\boldsymbol{x}) \cdot C_{i,j-1}(\boldsymbol{x}),$$

  and uses feature $\boldsymbol{x} \in \mathcal{X}$ individual CL factors.

- These CL factors $f_{j-1} : \mathcal{X} \to \mathbb{R}_+$ minimize a loss function

$$\mathcal{L}_j = \sum_i \sum_{\boldsymbol{x}} \frac{(C_{i,j}(\boldsymbol{x}) - f_{j-1}(\boldsymbol{x}) \cdot C_{i,j-1}(\boldsymbol{x}))^2}{\sigma^2_{j-1} \cdot C_{i,j-1}(\boldsymbol{x})},$$

  and are built using a feed-forward neural network with one hidden layer having $q = 20$ neurons.

Wrap-up

- The message is not that chain-ladder should disappear!
- Take home messages:
  - the presented methods increase insight in the available data and the dynamics in claim development patterns;

    (fits within the increasing interest in data analytics);

    (claim and policy characteristics can be taken into account).
  - caution: many choices involved, should be done with care!