Insurance Data Science Conference 15 - 16 June 2023

Programme and Abstract Booklet

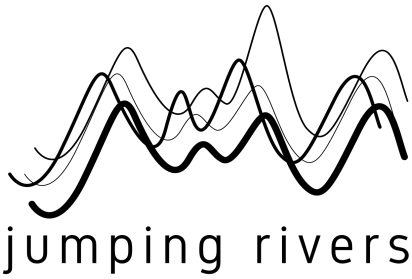
Scientific Committee

2023-05-23

Contents

Conference sponsors

Gold sponsors



Jumping Rivers combines data science consultancy and knowledge transfer with provision of managed software to support businesses in gaining invaluable insights from their data.

Silver sponsors

Posit: The open source data science company.

Mirai Solutions: Smarter analytics - better decisions

Markel: Bold ideas. Honest actions.

Carl H. Lindner III Center for Insurance and Risk Management

Scientific Committee

Academic	Industry
Katrien Antonio (KU Leuven and University of Amsterdam)	Markus Gesmann (co-organiser, Insurance Capital Markets Research, London)
Arthur Charpentier (Université du Québec à Montréal)	Davide de March (Markel, London)
Gian Paolo Clemente (Università Cattolica del Sacro Cuore, Milano)	Grainne McGuire (Taylor Fry)
Christophe Dutang (LJK, Université Grenoble Alpes)	Wui Hua (John) Ng (Reinsurance Group of America)
Montserrat Guillén (Universitat de Barcelona)	Ronald Richman (Old Mutual Insure, South Africa)
Ioannis Kyriakou (co-organiser, Bayes Business School, City, University of London)	Markus Senn (Partner Re, Zurich)
Pietro Millossovich (co-organiser, Bayes Business School, City, University of London)	Jürg Schelldorfer (Swiss Re, Zurich)
Silvana Pesenti (University of Toronto)	Giorgio Alfredo Spedicato (Unipol Group, Milan)
Andreas Tsanakas (co-organiser, Bayes Business School, City, University of London)	
Mario Wüthrich (ETH Zurich)	
Diego Zappa (Università Cattolica del Sacro Cuore, Milano)	
Rui Zhu (co-organiser, Bayes Business School, City, University of London)	
Johanna Ziegel, (University of Bern)	

Keynotes

- **Luca Baldassarre** *(Lead Data Scientist, Swiss Re):* Responsible AI trade-offs in Insurance
- **Rosalba Radice** *(Professor of Statistics, City, University of London):* A Unifying Copula Regression Framework with Applications in Insurance and Health Economics
- **Mark Sellors** *(Board member, Data Orchard):* APIs and the future of data science

Programme

Venue

- City, University of London, **Northampton Square**, London **EC1V 0HB**

15 June 2023

08:00 - 09:00 Registration, tea & coffee

09:00 - 09:15 Room B200: Opening remarks

09:15 - 10:15 Room B200: Keynote 1 (Chair: Andreas Tsanakas)

- **Luca Baldassarre** (Swiss Re): Responsible AI trade-offs in Insurance

10:15 - 11:15 Regular Session 1

Room B200: Machine learning and predictive modelling (Chair: Ron Richman)

- **Munir Hiabu** (University of Copenhagen): On functional decompositions, post-hoc machine learning explanations and fairness
- **Mario V. Wüthrich** (RiskLab, ETH Zurich): Isotonic recalibration under a low signal-to-noise ratio
- **Can Baysal** (Munich Re): Two-step Bayesian hyperparameter optimisation to efficiently build insurance market price models

Room B103: Advances in mortality modelling (Chair: Pietro Millossovich)

- **Salvatore Scognamiglio** (University of Naples): Accurate and Explainable Mortality Forecasting with the LocalGLMnet
- **Mike Ludkovski** (University of California Santa Barbara): Expressive Mortality Models through Gaussian Process Compositional Kernels
- **Asmik Nalmpatian** (Department of Statistics - LMU Munich): Modern Machine Learning approaches in mortality modeling considering the impact of COVID-19

11:15 - 11:40 Coffee break

11:40 - 12:40 Lightning Session 1

Room B200: Stream 1 (Chair: Jürg Schelldorfer)

- **Aurelien Couloumy** (Novaa-Tech): Neural generative techniques for synthetic data creation in insurance: context and use case
- **Thao Nguyen** & **Davide de March** (Markel International): Sentence similarity models to develop a new risk appetite tool
- **Navarun Jain** (Lux Actuaries & Consultants): Saving the World: Predictive Early Warning Systems for Conflict Risk using Neural Networks
- **Yafei (Patricia) Wang** (Lloyd's of London): Machine Learning and XAI for underwriting
- **Nuzhat Jabinh** (FRSA Ethics in AI consultant): Data is not neutral: ethics and AI
- **Valerie du Preez** (Dupro Advisory) & **Paul King** (University of Leicster): AI Risk: How much should we care?

Room B103: Stream 2 (Chair: Diego Zappa)

- **Mark Shoun** (Ledger Investing): Fitting Development and Tail Models Jointly via Mixture Modeling
- **Sebastian Calcetero-Vanegas** (University of Toronto): A Credibility Index Approach for Effective a Posteriori Ratemaking with Large Insurance Portfolios
- **Chris Halliwell** & **Cynon Sonkkila** (Markel): Log-Normal-Pareto: A Case Study
- **Gabriele Pittarello** (University of Rome): Chain Ladder Plus: a versatile approach for claims reserving
- **Mick Cooney** (Describe Data): A Bayesian Approach to Customer Lifetime Value
- **Tim Edwards** (Howden Tiger): Estimating return periods for extreme events - a frequenstist and Bayesian perspective

12:40 - 13:40 Lunch

13:40 - 14:40 Regular Session 2

Room B200: NLP case studies (Chair: Davide de March)

- **Jürg Schelldorfer** (Swiss Re): Actuarial Applications of Natural Language Processing Using Transformers: Case Studies for Using Text Features in an Actuarial Context
- **Paola Gasparini** (Bupa): Advanced analytics and machine learning to identify fraudulent health insurance claims
- **Bavo D.C. Campo** (KU Leuven): Insurance fraud network data simulation machine: Generating synthetic fraud network data sets to develop and to evaluate insurance fraud detection strategies

Room B103: Fairness & explainability (Chair: Andreas Tsanakas)

- **Olivier Côté** (Université Laval): Causal Inference and Fairness in Insurance Pricing
- **James Ng** (Trinity College Dublin): Generalized Bayesian Inference with Fairness Constraints
- **Deniz Günaydin-Bulut** (Swiss Re): Why this claim? Incorporating local model explainability in a reinsurance setting

14:40 - 15:40 Panel discussion (Room B200)

Algorithmic underwriting: the future of specialty insurance? (Chair: John Ng, RGA)

- **Davide Burlon** (Principal, Insurance Solutions. Munich Re)
- **Mick Cooney** (CTO, Describe Data)
- **Dana Cullen** (Senior Associate, SCOR Ventures)
- **Melanie Zhang** (Head of Algorithmic Pricing at Ki)

15:40 - 16:00 Coffee break

16:00 - 17:00 Regular Session 3

Room B200: Challenges in modelling insurance data (Chair: Mick Cooney)

- **Sindre Henriksen** (Eika Forsikring): Implementing ML Ops in insurance: a case study using a complex, multi-model Customer Lifetime Value system
- **Zhiyu Quan** (University of Illinois at Urbana-Champaign): Imbalanced learning for insurance using modified loss functions in tree-based models
- **Freek Holvoet** (KU Leuven): Neural networks for insurance pricing with frequency and severity data: a benchmark study from data preprocessing steps to technical tariff

Room B103: Telematics & graphs (Chair: Mario Wüthrich)

- **Xenxo Vidal-Llana** (Universitat de Barcelona): Non-crossing neural network quantile regression estimation for driving data with telematics
- **Marco De Virgilis** (Arch Insurance): Territorial Ratemaking and Graph Theory

- **Diego Zappa** (Università Cattolica del Sacro Cuore): Estimating the road accident risk of a road network

17:00 - 18:00 Keynote 2 (Room B200) (Chair: Ioannis Kyriakou)

- **Mark Sellors** (Data Orchard): APIs and the future of data science

19:00 Conference dinner

- **Ironmongers' Hall**, Shaftesbury Place, Barbican, London EC2Y 8AA

16 June 2023

09:15 - 10:15 Room B200: Keynote 3 (Chair: Mathias Lindholm)

- **Rosalba Radice** (Bayes Business School, City, University of London): A unifying and flexible bivariate copula regression framework

10:15 - 11:15 Lightning Session 2

Room B200: Stream 1 (Chair: Grainne McGuire)

- **Patrick Hogan** (PartnerRe, Zürich): …whatever remains, however improbable, must be a bug
- **Roland Schmid** (Mirai Solutions): Py-shiny for Reinsurance: ready or not, here we come
- **Priyank Shah** (Lane Clark & Peacock LLP): Embedding data science in reserving
- **William Mesquita** (TROVADORES D´EQUAÇÕES LDA): Fully automated ETL Process Using Azure
- **Bence Zaupper** (Finalyse): Optimisation and automation of capital projections in insurance
- **Amin Karbassi** (Axa): Application of Information Retrieval (IR) for automation of Risk Engineering research within unstructured data

Room B103: Stream 2 (Chair: Mike Ludkovski)

- **Despoina Makariou** (University of St. Gallen): A bivariate mixed Poisson claim count regression model with varying dispersion and shape
- **Lina Palmborg** (Stockholm University): Reinforcement learning in search of optimal premium rules
- **Bernard Wong** (University of New South Wales): Machine Learning with High-Cardinality Categorical Features in Actuarial Applications
- **Guillaume Biessy** (LinkPact & Sorbonne Université): Revisiting Whittaker-Henderson Smoothing
- **Agni Orfanoudaki** (Saïd Business School, University of Oxford): Algorithmic Insurance
- **Sukrita Singh** (Saïd Business School, University of Oxford): Algorithmic Insurance: A Conformal Prediction Framework

11:15 - 11:45 Coffe break

11:45 - 12:45 Regular Session 4

Room B200: Risk modelling (Chair: Markus Gesmann)

- **Claudio Giorgio Giancaterino** (Intesa SanPaolo Vita): Earthquakes Risk Modelling with Quantile Approach
- **Nicholas Robert** (DeNexus): Vine Copulas for Systemic Cyber Risk Modelling
- **Jacky Poon** (nib Travel Insurance): From Chain Ladder to Probabilistic Neural Networks for Claims Reserving

Room B103: Tree based models in insurance (Chair: Munir Hiabu)

- **Henning Zakrisson** (Stockholm University): Multi-Parametric Gradient Boosting Machines with Non-Life Insurance Applications
- **Mathias Lindholm** (Stockholm University): Local bias adjustment, duration-weighted probabilities, and automatic construction of tariff cells
- **Arthur Maillart** (Detralytics): Distill knowledge of additive tree models into GAMs

12:45 - 13:45 Lunch

13:45 - 14:45 Lightning Session 3

Room B200: Stream 1 (Chair: Salvatore Scognamiglio)

- **Karol Maciejewski** (Milliman): The fast and the fabulous. Harnessing GPU power for high-performance life insurance computations

- **Markus Gesmann** (Insurance Capital Markets Research): Measuring daily value creation of global specialty (re)insurance
- **Guillaume Attard** (Ageoce Solutions): Evolution of the Soil Wetness Index (SWI) in France: Analysis with Google Earth Engine
- **Shirley Ng** (Vantage Risk): A Practitioner Guide to Marginal Pricing - Pricing with Portfolio Impact in Mind
- **George Wright** (Vounder Analytics): Is the Lloyd's insurance market ready for an open-source capital modelling framework?
- **Annette Hoffmann** (University of Cincinnati): Modeling Underwriting Cycles in Property-Casualty Insurance: The Impact of Catastrophic Events

Room B103: Stream 2 (Chair: Rui Zhu)

- **Philipp Ratz** (Université du Québec à Montréal): Solving censored regression problems using a multitask approach
- **Matteo Crisafulli** (Università di Roma, La Sapienza): A neural network approach for selecting efficient reinsurance strategies
- **Michelle Dong** (The Australian National University): Forecasting Mortality by cause with Zero Death Counts
- **Valeria D'Amato** (University of Salerno): Explore latent factors of longevity trends with frailty-based stochastic models
- **Marie Michaelides** (Université du Québec à Montréal): Individual claims reserving with dependent censored data

14:45 - 14:55 Room B200: Closing comments (Markus Gesmann)

Keynotes

Responsible AI trade-offs in Insurance

Luca Baldassarre, Lead Data Scientist, Swiss Re

Abstract: AI is now widely impacting many industries, including insurance. The combination of larger, more diverse and granular datasets with advanced machine learning techniques has tremendous potential to revolutionize pricing, underwriting, claims handling, marketing and customer interactions.

However, like any new technology, AI's benefits come with significant risks, not only financial. Despite increasing regulatory proposals and civil society scrutiny, comprehensive safety measures are not yet widely adopted.

In this talk, I will unpack the several facets of Responsible AI using concrete examples, highlight the emerging trade-offs, and provide some recommendations on how to reap the full benefits of AI, while mitigating its risks.

Contact details

- Web: https://www.swissre.com/profile/Luca_Baldassarre/ip_f66e2b
- LinkedIn: https://www.linkedin.com/in/lucabaldassarre/

APIs and the future of data science

Mark Sellors, Board Member, Data Orchard

Abstract: Data science rarely thrives in isolation and there are lots of ways for a data scientist to integrate their work into an organisations decision making processes. Creating network APIs and integrating with machine-to-machine communication is one of the most powerful, but least well understood techniques.

Imagine an insurance company in which the data, models and assumptions are connected via APIs, so that information can be trusted and flow seamlessness from one function to another, from business planning and capital setting, to pricing and underwriting, and claims handling and reserving, while at the same time each team might use different tools and applications.

In this talk, Mark will introduce the concept of network APIs and make the case that you probably already have everything you need to start creating and using them. How these APIs can be leveraged can often be a source of confusion and frustration, so we'll be shining a light on use cases as well as diving into how and when to start using them. Along the way, we'll look at what it might take to bring the rest of your organisation along with you on this journey and some of the challenges you might face making your APIs ready for the much-feared "production deployment".

Contact details

- Web: https://www.dataorchard.org.uk/people/mark-sellors
- Linkedin: https://www.linkedin.com/in/msellors

A unifying and flexible bivariate copula regression framework

Rosalba Radice, Professor of Statistics, City, University of London

Abstract: Motivated by the need in many fields to model joint outcomes, we present a unifying and flexible bivariate copula regression framework that is capable of handling peculiar shapes of response data via a vast range of distributions, allows for a wide variety of copula dependence structures, and permits to specify all model parameters (including the dependence parameters) as functions of additive predictors.

Fitting the models within this framework can be a challenging task in practice. To this end, parameter estimation is carried out via a carefully structured algorithm based on a computationally efficient and stable penalised maximum likelihood estimation approach.

The framework has been made available via the R package GJRM which is very easy and intuitive to use. The methodology will be illustrated by discussing insurance and health economics-related case studies which have motivated some of the developments incorporated in GJRM.

Contact details

- Web: https://www.bayes.city.ac.uk/faculties-and-research/experts/rosalba-radice

Abstracts of contributed talks

Algorithmic Insurance

Agni Orfanoudaki, Saïd Business School, University of Oxford (Presenter)

Dimitris Bertsimas, Sloan School of Management, Massachusetts Institute of Technology

Abstract: As machine learning algorithms start to get integrated into the decision-making process of companies and organizations, insurance products are being developed to protect their owners from liability risk. Algorithmic liability differs from human liability since it is based on a single model compared to multiple heterogeneous decision-makers and its performance is known a priori for a given set of data.

Traditional actuarial tools for human liability do not take these properties into consideration, primarily focusing on the distribution of historical claims. We propose, for the first time, a quantitative framework to estimate the risk exposure of insurance contracts for machine-driven liability, introducing the concept of *Algorithmic Insurance*.

Specifically, we present an optimization formulation to estimate the risk exposure of a binary classification model given a pre-defined range of premiums. We adjust the formulation to account for uncertainty in the resulting losses using robust optimization. Our approach outlines how properties of the model, such as accuracy, interpretability, and generalizability, can influence the insurance contract evaluation.

To showcase a practical implementation of the proposed framework, we present a case study of medical malpractice in the context of breast cancer detection. Our analysis focuses on measuring the effect of the model parameters on the expected financial loss and identifying the aspects of algorithmic performance that predominantly affect the risk of the contract.

Keywords: Algorithmic Insurance, Machine Learning, Algorithmic Risk, Insurance Contracts

References

1. Bertsimas, D., Orfanoudaki, A. (2021). Algorithmic Insurance. *arXiv preprint arXiv:2106.00839*.

Contact details

- Email: agni.orfanoudaki@sbs.ox.ac.uk
- Homepage: https://www.sbs.ox.ac.uk/about-us/people/agni-orfanoudaki
- Twitter: @AgniOrfanoudaki

Application of Information Retrieval (IR) for automation of Risk Engineering research within unstructured data

Amin Karbassi, Casualty Risk Consultant, AXA XL

Abstract: In the insurance industry, Risk Engineering's task to identify and highlight Property & Casualty ("P&C") exposure trends to support the underwriting and pricing of risks, often involves investigation of information from various publications or online resources that are considered unstructured data and can often make this a tedious task. To this end, risk engineers are required to review each report/webpage manually and decide if the information is relevant to a specific exposure of interest (e.g., Product Liability, Professional Indemnity, etc.) before going deeper and studying those selected documents in more detail.

In this presentation, a technical solution for automating and analyzing unstructured data - scraping unstructured data from the web and sorting the information based on its relevance - is demonstrated.

Nowadays, websites often use frontend frameworks which render dynamic content by loading a JSON or XML file from their backend to populate the user-facing site. Examples of this can be seen on the publicly available accident reports from the National Transportation Safety Board's website [1]. For this reason, the scraping methodology leverages calls to the API endpoint to efficiently gather a large corpus of accident reports. Consequently, for the Information Retrieval (IR), a 'word2vec' based Vector Space Model (VSM) [2] pre-trained on Google News [3] is implemented in Python to represent the accident reports in a high-dimensional vector space.

The semantic similarity between the accident reports and pre-defined user's query is then calculated through cosine similarity, enabling effective filtering of relevant reports. In this way, the user is able to quickly and efficiently gather large amounts of un-structured data from the web and filter them based on relevance to specific exposures, resulting in time and resources savings compared to manual data gathering and analysis. Additionally, the process can be scheduled and executed periodically, to gather new reports for a specific query providing up-to-date data to stay informed about industry trends and emerging incidents.

Keywords: Information Retrieval, word2vec, unstructured data, Gensim

References

1. National Transportation Safety Board. https://www.ntsb.gov/Pages/home.aspx
2. Radim Rehurek (2022) Word2Vec Model on Gensim. https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html#sphx-glr-auto-examples-tutorials-run-word2vec-py
3. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, (2013). https://arxiv.org/pdf/1301.3781.pdf

Contact details

- Email: amin.karbassi@axaxl.com
- Repository: https://github.com/aminkarbassi
- Social media: https://ch.linkedin.com/in/aminkarbassi

Modeling Underwriting Cycles in Property-Casualty Insurance: The Impact of Catastrophic Events

Annette Hofmann, University of Cincinnati (presenter)

Cristina Sattarhoff, University of Kiel

Abstract:
This paper challenges the question of existence and predictability of underwriting cycles in the U.S. property and casualty insurance industry. Using a long time series of underwriting data, we demonstrate the existence of a hidden periodic component in annual aggregated loss ratios, supporting an underwriting cycle length of 8-9 years. Going beyond previous research and studying almost thirty years of quarterly underwriting data, we improve forecasting performance by (dis-)connecting cycles and catastrophic events. Superior out-of-sample forecasting performance in models with intervention variables flagging the time point of catastrophic outbreaks is achieved in terms of mean squared/absolute forecast errors. Following Hansen et al. (2011), we evaluate model confidence sets containing the most accurate model with a certain confidence level. The analysis suggests that reliable forecasts can be achieved net of the irregular major peaks in loss distributions that arise from natural and other catastrophes as well as big 'unusual' black swan events.

Keywords: modeling cycles, property insurance, catastrophic events

References

Grace, M. F. and Hotchkiss, J. L. (1995). External impacts on the property-liability insurance cycle. Journal of Risk and Insurance, 62(4):738 96754.

Gron, A. (1994a). Capacity constraints and cycles in property-casualty insurance markets. The RAND Journal of Economics, 110 96127.

Jawadi, F., Bruneau, C., and Sghaier, N. (2009). Nonlinear cointegration relationships between non-life insurance premiums and financial markets. Journal of Risk and Insurance, 76(3):753 96783.

Jiang, S.-j. and Nieh, C.-C. (2012). Dynamics of underwriting profits: Evidence from the u.s. insurance market. International Review of Economics & Finance, 21(1):1 9615.

Owadally, M., Zhou, F., and Wright, I. (2018). The insurance industry as a complex social system: competition, cycles, and crises. Journal of Artificial Societies and Social Simulation, 21(4):2.

Contact details

- Email: annette.hofmann@uc.edu
- Homepage: https://business.uc.edu/lcirm
- Social media: https://www.linkedin.com/in/annette-hofmann-phd-43904652

Distill knowledge of additive tree models into GAMs

Arthur Maillart, 1. Detralytics 2. Université de Lyon, Université Lyon 1, ISFA (presenter)

Christian Y. Robert, 1. Center for Research in Economics and Statistics, ENSAE 2. Université de Lyon, Université Lyon 1, ISFA

Abstract: Generalized additive models with pairwise interactions (GA2Ms) are a leading model class for interpretable machine learning. GA2Ms were originally trained using smoothing splines.

Recently, tree-based GAMs where shape functions are gradient-boosted ensembles of bagged trees were proposed. EBM (Explainable Boosting Machine) or its sparse version EBM-BF, that greedily grows the next tree on the best and most informative feature to reduce error as much as possible at each step, are typical examples.

In this paper, we propose a competing two-step GLM approach where we combine an additive tree based one-hot encoding of the continuous features with a penalized GLM learning task. The penalization is known as binarsity (Alaya et al. (2019)) and penalizes the model weights learned from feature one-hot encodings to avoid in particular collinearity.

Numerical experiments illustrate the very good performances of our approach on several datasets compared to EBM and EBM-BF. A case-study in trade credit insurance is also provided.

Keywords: GAM, EBM, XAI, binarsity

References

1. Alaya, M.Z., Bussy, S., Gaïffas, S., Guilloux, A. (2019). Binarsity: a penalization for one-hot encoded features in linear supervised learning. *JMLR* **20(118)**, 1-34.

Contact details

- Email: a.maillart@detralytics.eu

Modern Machine Learning approaches in mortality modeling considering the impact of COVID-19

Asmik Nalmpatian, Department of Statistics - LMU Munich (presenter)

Prof. Dr. Christian Heumann, Department of Statistics - LMU Munich

Dr. Stefan Pilz, Data and Analytics Department - Munich Re

Abstract: The last two centuries have seen a significant increase in life expectancy. Although past trends suggest that mortality will continue to decline in the future, uncertainty and instability about the development is greatly increased due to the ongoing COVID-19 pandemic. It is therefore of essential interest, particularly to annuity and life insurers, to predict the mortality of their members or policyholders with reliable accuracy.

The goal of this study is to improve the state-of-the-art stochastic mortality models using machine learning techniques and generalize them to a multi-population model. Detailed cross-country results conducted for Finland, Germany, Italy, the Netherlands, and the United States show that the best forecasting performance is achieved by a generalized additive model that uses the framework of APC analysis. Based on this finding, trend forecasts of mortality rates as a measure of longevity are fulfilled for the future, given a range of COVID-19 scenarios, from mild to severe.

Discussing and evaluating the plausibility of these scenarios, this study is useful for preparation, planning and informed decision-making especially in the current epidemiological uncertainty.

Keywords: Mortality modeling, covid impact, generalized additive models, APC method, excess mortality, partial APC plots, machine learning

References

1. Lee, R.D., Carter, L.: Modeling and forecasting US mortality. howpublished of the American Statistical Association (1992)
2. Wood, S.N.: Generalized additive models: an introduction with R. CRC press (2017)
3. Hastie, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. New York: Springer (2016)
4. Reither, E., Hauser, R., Yang, Y.: Do birth cohorts matter? Age-periodcohort analyses of the obesity epidemic in the United States. Social science and medicine (2009)
5. Girosi, F., King, G.: Demographic Forecasting. Princeton University Press (2008)

Contact details

- Email: asmik.nalmpatian@gmx.de
- Social media: Linkedin (https://www.linkedin.com/in/asmik-nalmpatian/), Twitter (https://twitter.com/AsmikNalmpatian)

Neural generative techniques for synthetic data creation in insurance: context and use case

Aurélien Couloumy, Novaa-Tech, ISFA Lyon

Akli KAIS, CCR Re

Abstract: The availability of data in insurance is often a complex topic: lack of data, limited quality, bias in tools or models, regulatory constraints, price, ethics, etc. The creation of synthetic data may allow to overcome these problems.

In this presentation, we first review what synthetic data are, and how the creation of such data can be helpful in insurance (data quality, model and tool trustworthiness, confidentiality aspects).

We propose an overview of sampling based neural generative techniques (GAN, CTGAN, TVAE). Then, we dive into a motor pricing case study. We run sensitivity tests (methods, epoch, sampling size). We analyse models results from a data quality perspective and compare them to common approaches (simple imputation, MissForest). Finally, we extend the discussion to adversarial scenarios (uncertainty) with BNNs usage and other case studies experimented.

Keywords: Insurance, Synthetic Data, CTGAN, TVAE, BNN, Uncertainty

References

L. Xu et al, (2019) Modeling tabular data using conditional GAN. https://arxiv.org/abs/1907.00503

K. Kuo (2019) Generative Synthesis of Insurance Datasets https://arxiv.org/abs/1912.02423v2

Y Gal, (2016) Uncertainty in Deep Learning, http://www.cs.ox.ac.uk/people/yarin.gal/website//thesis/thesis.pdf

Contact details

- Email: acouloumy@novaa-tech.com
- Social media: https://www.linkedin.com/in/aur%C3%A9lien-couloumy-5aa778a8/

Insurance fraud network data simulation machine: Generating synthetic fraud network data sets to develop and to evaluate insurance fraud detection strategies

Bavo D.C. Campo, KU Leuven (presenter)

Katrien Antonio, KU Leuven

Abstract: Traditionally, the detection of fraudulent insurance claims relies on business rules and expert judgement which makes it a time-consuming and expensive process [1].

Consequently, researchers have been examining ways to develop an efficient and accurate fraud detection model. The use of features engineered from the social network of parties involved in a claim is a particularly promising strategy to flag potentially fraudulent claims (see for example [1, 2]). When developing a fraud detection model, however, we are confronted with several challenges. The uncommon nature of fraud, for example, creates a high class imbalance which complicates the development of analytic classification models.

In addition, only a small number of claims are investigated and get a label, which results in a large corpus of unlabelled data. We design a simulation machine that is inspired by the real non-life motor insurance data used in Óskarsdóttir et al. [1]. This data contains both traditional claims characteristics as well as social network features.

When generating the synthetic data, the user has control over several data-generating mechanisms. We can specify the total number of policyholders and parties, the desired level of imbalance and the (effect size of the) features in the fraud generating model. Hereby, it enables researchers and practitioners to examine several methodological challenges as well as to back-test their (development strategy of) insurance fraud detection models in a range of different settings.

Keywords: social networks, simulation machine, fraud detection, class imbalance, missing data

References

[1] Óskarsdóttir, M., Ahmed, W., Antonio, K., Baesens, B., Dendievel, R., Donas, T., Reynkens, T. (2022). Social Network Analytics for Supervised Fraud Detection in Insurance. *Risk analysis* **42 (8)**, 1872–1890
[2] Van Vlasselaer, V., Eliassi-Rad, T. Akoglu, L. Snoeck, M., Baesens, B. (2016). Gotcha! Network-based fraud detection for social security fraud. *Management Science* **63 (9)**, 3090–3110

Contact details

- Homepage: https://bavodc.github.io/
- Repository: https://github.com/BavoDC
- Social media: https://www.linkedin.com/in/bavodccampo/

Optimisation and automation of capital projections in insurance

Bence Zaupper, Finalyse (presenter)

Abstract: The projection of regulatory and economic capital is a key component in the financial planning of an insurance company. Due to implications on regulatory compliance, profitability and dividend payment capacity it involves senior stakeholders up to board level.

Decision making is supported by sensitivity analysis and scenario testing over the planning horizon (typically 3-5 years). However, due to heavy reliance on spreadsheet models and archaic software infrastructure this is often limited to the impact assessment of changes in single factors (such as interest or mortality rates) on capital.

A case study is presented to show how the advanced modelling capabilities of R can be used capital optimisation and process automation. Furthermore, the power of interactive Shiny apps and dashboards to enhance decision making will be demonstrated.

High-performance cashflow and capital modelling enables multi-factor sensitivity analysis and reverse scenario testing, asset portfolio optimisation (using FRAPO and ROI) and impact assessment of management actions.

Report and workflow automation is achieved by structured Rmarkdown documents which are easy to update and include enhanced visualisation with ggplot2 (2D), plotly and rgl (3D).

Keywords: Capital management, Risk management, Optimisation, Process automation

References

1. Dedu, S., Serban F. (2015) Multiobjective Mean-Risk Models for Optimization in Finance and Insurance
2. Pfaff B. (2013). *Financial Risk Modelling and Portfolio Optimization with R - Second edition.* Wiley.
3. Peikert A, Brandmaier A. (2020). A Reproducible Data Analysis Workflow With R Markdown, Git, Make, and Docker

Contact details

- Email: bence.zaupper@finalyse.com
- Homepage: https://www.finalyse.com/
- Repository: https://github.com/bencezaupper
- Linkedin: https://ie.linkedin.com/in/zaupperbence

Machine Learning with High-Cardinality Categorical Features in Actuarial Applications

Benjamin Avanzi, University of Melbourne

Gregary Taylor, University of New South Wales

Melantha Wang, University of New South Wales

Bernard Wong, University of New South Wales (presenter)

Abstract: High-cardinality categorical features are pervasive in actuarial data (e.g. occupation in commercial property insurance). Standard categorical encoding methods like one-hot encoding are inadequate in these settings.

In this work, we present a novel *Generalised Linear Mixed Model Neural Network* ("GLMMNet' ') approach to the modelling of high-cardinality categorical features. The GLMMNet integrates a generalised linear mixed model in a deep learning framework, offering the predictive power of neural networks and the transparency of random effects estimates, the latter of which cannot be obtained from the entity embedding models. Further, its flexibility to deal with any distribution in the exponential dispersion (ED) family makes it widely applicable to many actuarial contexts and beyond.

We illustrate and compare the GLMMNet against existing approaches in a range of simulation experiments as well as in a real-life insurance case study. Notably, we find that the GLMMNet often outperforms or at least performs comparably with an entity embedded neural network, while providing the additional benefit of transparency, which is particularly valuable in practical applications.

Importantly, while our model was motivated by actuarial applications, it can have wider applicability. The GLMMNet would suit any applications that involve high-cardinality categorical variables and where the response cannot be sufficiently modelled by a Gaussian distribution.

Keywords: Categorical features; Generalised linear mixed models; Neural networks; Categorical embedding; Random effects; Variational inference; Insurance analytics

Contact details

- Email: bernard.wong@unsw.edu.au
- Article link: https://doi.org/10.48550/arXiv.2301.12710

Two-step Bayesian hyperparameter optimisation to efficiently build insurance market price models

Can Baysal, Munich RE (presenter)

Abstract: Strong competition in personal lines business forces primary insurers to set their prices not only based on risk, but also on market prices. Players in the market are on a constant pursuit to minimise the money-left-on-table and identify the segments that need discounting in order to maximise their profits and market share. To achieve this goal, aggregator (price comparison website) data is used to reverse-engineer market prices with models similar to those used for technical risk pricing. The resulting market price models allow primary insurers to predict the market prices for new business quotes. However, the highly dynamic nature of the general insurance market (especially motor and home) and the rapid change in market prices require these models to be updated frequently to keep market price predictions accurate over time.

Pricing actuaries and data scientists have been challenging the traditional GLM models with popular machine learning algorithms such gradient boosting machine and XGBoost to build more accurate models. Nevertheless, these algorithms require tuning a set of hyperparameters prior to model training, making this process as time-consuming as the size of the data that insurers are dealing with. In this talk, I propose to use Bayesian approach for hyperparameter optimisation in two main steps to leverage the trade-off between exploration and exploitation. The aim is to obtain the appropriate hyperparameter combination that can be run in each model update, requiring minimal manual work and time. We will see the evolution of model accuracy at each step, and then discuss the benefits of training models with this approach in the day-to-day processes of an insurance pricing team.

Keywords: Non-life insurance pricing, Bayesian hyperparameter optimisation, machine learning, gradient boosting models

References

1. Brochu, E., Cora, V. M., & De Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint* **arXiv**, 1012.2599.
2. Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology* **17(1)**, 26-40.

Contact details

- Email: cbaysal@munichre.com
- Linkedin: https://www.linkedin.com/in/can-baysal/

Log-Normal-Pareto: A Case Study

Cynon Sonkkila, Actuarial Consultant, Markel (presenter)

Chris Halliwell, Senior Data Scientist, Markel (presenter)

Abstract: In the insurance industry, predicting portfolio performance accurately has been a persistent challenge for long tail lines. Traditionally, actuaries have relied on data manipulation including manual adjustments to treat for data quality issues, this comes with scalability limitations. In this study, we propose a Bayesian hierarchical model that employs transactional level data to address statistical challenges associated with these models.

We apply this model to a moderately sized dataset of triangulated individual claim data for several commercial long tail lines, and use it to generate posterior distributions for the parameters of the ultimate severity distribution. This model is designed to produce on-demand forecasts. We also demonstrate that our data does not follow a log-normal distribution, and show that the MLP function provides an excellent fit for commercial line severity in this setting[1][2].

To account for heteroskedasticity and negative IBNR, we explore growth curve models[3][4] and incremental models[5] alongside generalized additive models provided by the lme4 framework under BRMS.

In the future, we plan to include a frequency model for loss ratio and profitability forecasting to complement our approach. Our study highlights the potential of advanced statistical models and data-driven techniques to overcome longstanding challenges in the insurance industry.

Keywords: Frequency, Severity, BRMS, Stan, Loss Ratio, Profit Study

References

1. Basu S., Jones C.E.. , MNRAS , 2004 , vol. 347 pg. L47
2. Shantanu Basu, M. Gil, Sayantan Auddy, The MLP distribution: a modified lognormal power-law model for the stellar initial mass function, Monthly Notices of the Royal Astronomical Society, Volume 449, Issue 3, 21 May 2015, Pages 2413–2420, https://doi.org/10.1093/mnras/stv445
3. Markus Gesmann (Jul 15, 2018) Hierarchical loss reserving with growth curves using brms. Retrieved from https://magesblog.com/post/2018-07-15-hierarchical-loss-reserving-with-growth-cruves-using-brms/
4. Mick Cooney (Nov 30, 2017) Modelling Loss Curves in Insurance with RStan. Retrieved from https://mc-stan.org/users/documentation/case-studies/losscurves_casestudy.html
5. Greg McNulty (July 15, 2017) Severity Curve Fitting for Long Tailed Lines: An Application of Stochastic Processes and Bayesian Models Retrieved from https://www.casact.org/sites/default/files/2021-07/Severity-Curve-Fitting-McNulty.pdf

Contact details

- Email:
  - cynon@sonkkila.co.uk / cynon.sonkkila@markel.com
  - chrisjhalliwell@gmail.com / chris.halliwell@markel.com

Earthquakes Risk Modelling with Quantile Approach

Claudio Giorgio Giancaterino, Intesa SanPaolo Vita (presenter)

Abstract: Forecasting earthquakes is one of the most challenging job because they don't show specific patterns resulting by predictions. An earthquake is a natural disaster based on a shaking of Earth's surface, and caused by a sudden slip on a fault. It releases energy in waves that travel through the Earth's crust. A big earthquake can inflict massive death and huge infrastructural damages with also huge losses for Insurance/Reinsurance Companies.

Predicting the occurrences of an earthquake, losses can be reduced. Current scientific studies linked to earthquake forecasting focus on three key points:

1. when the event will happen,
2. where it will happen, and
3. how large it will be.

The purpose of this job is to predict the magnitude of the next earthquakes in the following period by the help of Supervised Learning models.

In order to manage the uncertainity linked with the prediction is used the quantile loss function and in this way there will be a prediction interval for each estimation. Instead of modeling the expected value of the conditional distribution of the outcome, like the least squares, the goal is to try to estimate quantiles of that conditional distribution.

Given the earthquake hazard, likely the magnitude cannot be completely accurately predicted from the available features, and it can be influenced by other variables.

For this aleatoric uncertainty, quantile regression makes possible to give a finer description of that distribution without making strong assumptions on its shape. Indeed, predicting the median, lower and upper quantiles of the target will be able to have a confidence region in between which the true value is likely to belong.

Some models have been explored looking at the performance and features importance.

Keywords: K-Nearest Neighbors, Random Forest, Gradient Boosting Machine, Neural Networks, Deep Learning, Permutation Feature Importance, Partial Dependence Plot, Quantile Regression, Earthquakes forecasting

References

1. P.J. Brockwell, R.A. Davis (2016). *Introduction to Time Series and Forecasting*, Springer
2. M. H. A. Banna et al. (2020). *Application of Artificial Intelligence in Predicting Earthquakes: State-of-the-Art and Future Challenges*, IEEE
3. J. Pai, L. Yunxian, A. Yang, C. Li (2022). *Earthquake parametric insurance with Bayesian spatial quantile regression*, Insurance: Mathematics and Economics

Contact details

- Email: c.giancaterino@gmail.com
- Repository: https://github.com/claudio1975/Earthquakes_Risk_Modelling
- Social media: https://www.linkedin.com/in/claudiods/

Why this claim? Incorporating local model explainability in a reinsurance setting

Deniz Gunaydin-Bulut, Swiss Reinsurance Company Ltd. (presenter)

Nora Leonardi, Swiss Reinsurance Company Ltd.

Abstract: Claims handlers at a large reinsurer only have the capacity to review a fraction of claims. What if these skilled employees could focus on only the complex cases where they can add true value, e.g., by identifying an error or proposing an alternative action?

To enhance our claims management, claim experts are enabled with a claim triaging application, which has an operationalized machine learning (ML) model at its core. The decision remains with the human. The ML model leverages 10 years of structured claims data to predict the likelihood that a human assessor adds value to an incoming claim. The ML model utilizes feature elimination and ensemble of tree-based models to solve the imbalanced classification with the best accuracy. To understand the general mechanisms in the model, we used global model-agnostic methods for explainability. While they were deemed insightful by domain experts, individual claim predictions were uninterpretable. To increase the trustworthiness, we added local explanations via Shapley values.

We observed two main categories of benefits from adding local explainability. First, from an implementation perspective, it improved the original model and feature engineering by uncovering hidden patterns. Second, from the users' perspective, they can interpret the model's suggestion for each individual claim, and fully engage their critical thinking to ensure that they don't blindly trust the prediction. This in turn enhances how we gather user feedback. Even though Shapley values provided additional insight on individual model predictions, we also observed different results between local and global explainability methods. Thus, not one approach is always right. Hybrid implementation of model explainability methods helps to increase transparency and understanding of the model for both data scientists and users.

Keywords: Responsible Artificial Intelligence, Explainable Artificial Intelligence (XAI), Global Model Explainability, Local Model Explainability, Shapley Values, Human-in-the-loop System, Transparency, Interpretable Models

References

1. Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. (2018). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *arXiv*:1801.01489.
2. Shapley, Lloyd S. (1953). A value of n-person games. *Contributions to the Theory of Games*, 307-317.
3. Sundararajan, Mukund, and Amir Najmi. (2019). The many Shapley values for model explanation. *arXiv*:1908.08474.

Contact details

- Email: Deniz_Guenaydin@swissre.com, Nora_Leonardi@swissre.com
- Social media: https://www.linkedin.com/in/denizgunaydin/, https://www.linkedin.com/in/noraleonardi/

A bivariate mixed Poisson claim count regression model with varying dispersion and shape

Dr. George Tzougas, Heriot Watt University

Dr. Despoina Makariou, University of St. Gallen (presenter)

Abstract: We consider a family of bivariate mixed Poisson regression models with varying dispersion and shape for approximating the different types of claims and their associated counts in nonlife insurance.

Our main contribution is that we develop an Expectation Maximization (EM) algorithm for estimating the parameters of the models.

We exemplify our approach by fitting the bivariate Poisson generalised regression model with varying dispersion and shape parameters to property damage and bodily injury count data from a European motor insurer.

Keywords: Multivariate Poisson-Generalized Inverse Gaussian Distribution; EM Algorithm; Non-life Insurance

References

1. Tzougas, G. and Makariou, D., 2022. The multivariate Poisson-Generalized Inverse Gaussian claim count regression model with varying dispersion and shape parameters. Risk Management and Insurance Review.
2. Jeong, H., Tzougas, G. and Fung, T.C., 2023. Multivariate claim count regression model with varying dispersion and dependence parameters. Journal of the Royal Statistical Society Series A: Statistics in Society, p.qnac010.

Contact details

- Email: despoina.makariou@unisg.ch
- Homepage: https://www.ivw.unisg.ch/research-area-catastrophe-risk-and-machine-learning-in-insurance/
- Repository: myrepository.com
- Social media: https://www.linkedin.com/in/dr-despoina-makariou-71667959/

Estimating the road accident risk of a road network

Diego Zappa (*presenter*) (1), Gabriele Cantaluppi (1), Gian Paolo Clemente (2), Francesco della Corte (2), Nino Savelli (2)

(1) Università Cattolica del Sacro Cuore, Department of Statistical Sciences

(2) Università Cattolica del Sacro Cuore, Department of Mathematics for Economics, Finance and Actuarial Science

Abstract: Estimating the risk of motor vehicle accidents in road networks is a relevant topic for both socio-political decisions and insurance companies. To this end, we show how information on road structure and related traffic volumes can be used to map the risk related to each link of the road network.

Because of computational burdens, we do not follow standard Bayesian methods to estimate spatially structured and unstructured risk components, but we include spatial dependence by dividing the domain into non-disjoint subregions and applying spatially lagged (SLX) models. A comparison of the two approaches is provided.

A key goal of the project is to validate the hypothesis that the riskier the trajectory covered by insurance customers, the higher the risk of having an accident. To this end, using proprietary databases, we analyse the trajectories of customers who had black boxes installed and compare the estimated car occurrences with the estimated accident frequency using only standard customer profile databases.

Results are so far limited to the only frequency component but they demonstrate the possibility of accurately adjusting the premiums for driving a car using the estimated risk map at the street level.

Keywords: Car crashes, Spatial models, Black Boxes

References

1. Guillen, M., Nielsen, J.P., Peréz-Marìn A.M. & Elpidorou, V. (2020) Can Automobile Insurance Telematics Predict the Risk of Near-Miss Events?, *North American Actuarial Journal*, **24:1, 141-152**, DOI: 10.1080/10920277.2019.1627221.
2. Wüthrich, M. V. and Buser, C. (2023) *Data Analytics for Non-Life Insurance Pricing*. Swiss Finance Institute Research Paper No. 16-68, Available at SSRN: https://ssrn.com/abstract=2870308 or http://dx.doi.org/10.2139/ssrn.2870308

Contact details

- Email: diego.zappa@unicatt.it

Neural networks for insurance pricing with frequency and severity data: a benchmark study from data preprocessing steps to technical tariff

Freek Holvoet, KU Leuven (presenter)

Katrien Antonio, KU Leuven & University of Amsterdam

Roel Henckaerts, KU Leuven & Prophecy Labs

Abstract: Insurers usually turn to generalized linear models for modelling claim frequency and severity data. Due to their success in other fields, machine learning techniques are gaining popularity within the actuarial toolbox. Our paper contributes to the literature on frequency-severity insurance pricing with machine learning ([1]) via deep learning structures.

We present a benchmark study on four insurance data sets with frequency and severity targets in the presence of multiple types of input features. We compare in detail the performance of: a generalized linear model on binned input data, a gradient-boosted tree model, a feed-forward neural network (FFNN), and the combined actuarial neural network (CANN) proposed by [2].

Our CANNs combine on the one hand a neural network with a GLM and on the other hand a neural network with a gradient-boosting model (GBM). We explain extensively the data preprocessing steps with specific focus on the multiple types of input features typically present in tabular insurance data sets, such as postal codes, numeric and categorical covariates. Autoencoders ([3]) are used to embed the categorical variables into the neural networks and we explore their potential advantages in a frequency-severity setting.

Finally, we construct global surrogate models ([4]) for the neural nets' frequency-severity models. These surrogates enable the translation of the essential insights captured by the FFNNs or CANNs to GLMs. As such, an interpretable tariff table results that can easily be deployed in practice.

Keywords: property and casualty insurance, pricing, neural networks, embeddings, interpretable machine learning

References

1. Henckaerts, R., Côté, M-P., Antonio, K., Verbelen, R. (2021). Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods. *North American Actuarial Journal* **25(2)**:255-285.
2. Wüthrich, M., Merz, M. (2018). Editorial: Yes we CANN!. *ASTIN Bulletin*.
3. Delong, Ł., Kozak, A. (2021). The use of autoencoders for training neural networks with mixed categorical and numerical features. *SSRN Electronic Journal*.
4. Henckaerts, R., Antonio, K., Côté, M.-P. (2022). When stakes are high: balancing accuracy and transparency with model-agnostic interpretable data-driven surrogates. *Expert Systems with Applications*, **202**:117230.

Contact details

- Email: freek.holvoet@kuleuven.be

Chain Ladder Plus: a versatile approach for claims reserving

Gabriele Pittarello, 'La Sapienza' University of Rome (presenter)

Munir Hiabu, University of Copenhagen

Andrés Villegas, University of New South Wale

Abstract: This paper introduces yet another stochastic model replicating chain ladder estimates and furthermore considers extensions that add flexibility to the modelling.

We show that there is a one-to-one correspondence between chain-ladder's individual development factors and averaged hazard rates in reversed development time. By exploiting this relationship, we introduce a new model that is able to replicate chain ladder's development factors.

The replicating model is a GLM model with averaged hazard rates as response. This is in contrast to the existing reserving literature within the GLM framework where claim amounts are modelled as response. Modelling the averaged hazard rate corresponds to modelling the claim development and is arguably closer to the actual chain ladder algorithm.

Furthermore, the resulting model only has half the number of parameters compared to when modelling the claim amounts; because exposure is not modeled. The lesser complexity can be used to easily introduce model extensions that may better fit the data. We provide a new R-package, `clmplus`, where the models are implemented and can be fed with run-off triangles.

We conduct an empirical study on 30 publicly available run-off triangles making a case for the benefit of having `clmplus` in the actuary's toolbox.

Keywords: chain ladder, hazard function, generalized linear model, claims reserving

References

1. Bischofberger, S. M., Hiabu, M., & Isakson, A. (2020). Continuous chain-ladder with paid data. Scandinavian Actuarial Journal , 2020 (6), 477–502.
2. Hiabu, M. (2017). On the relationship between classical chain ladder and granular reserving. Scandinavian Actuarial Journal , 2017 (8), 708–729.

Contact details

- Email: gabriele.pittarello@uniroma1.it
- Repository: https://github.com/gpitt71
- Social media: https://www.linkedin.com/in/gabrielepittarello/

Is the Lloyd's insurance market ready for an open-source capital modelling framework?

George Wright, Vounder Analytics

Abstract: Capital modelling is one of the more software driven parts of insurance actuarial skills. Capital modelling requirements have moved on significantly since the implementation of Solvency II, and we've seen the incumbent software providers disrupted by quicker and more powerful new entrants.

As models are tested, developed and improved one factor that holds them back is the small resource pool of sufficiently experienced individuals who can develop using the propriety software tools available.

While there are a handful of examples of capital models being developed using open-source languages, there still lacks a volume of tested examples in this area.

In this presentation I put forward a case for the development of more tools using more widely used languages (such as Python, R or Julia) will:

- increase the pool of developers available to support the capital modelling teams;
- lower the barrier to entry for other insurance professionals to start developing software solutions; and
- result in higher quality capital modelling and better data-driven business decisions.

Keywords: Capital modelling, Lloyd's, Python

Contact details

- Email: george@vounder.co.uk
- Homepage: www.vounder.co.uk
- Social media: ttps://www.linkedin.com/in/george-wright-86824575

Evolution of the Soil Wetness Index (SWI) in France: Analysis with Google Earth Engine

Guillaume Attard, Ageoce Solutions

Aurélien Couloumy, Novaa-Tech

Guillaume Attard, Ageoce Solutions

Abstract: Between 1989 and 2021 in France, the damages caused by the shrinkage and swelling of clay soils represented more than 15 billion euros [1]. This phenomenon mostly occurs after severe drought events that can be identified by low values of the soil wetness index (SWI), which is the meteorological criteria used by authorities in the national disaster compensation model.

If the return period of the SWI value is higher than 25 years, then municipalities can be recognized and supported by the compensation program. In the current context of climate change, the following questions arise: - How did the SWI evolve in France since the beginning of its calculation at a national scale in 1969? - Are there eventually some locations more impacted than others in terms of soil wetness evolution? - What are the implications considering the strong relationship between this meteorological index and the compensation program?

In this communication, we provide an insight of the SWI trend over France. The SWI image collection provided by meteo-France has been ingested in Google Earth Engine [2] and a multi-seasonal analysis has been performed to identify the linear trend between 1970 to 2021.

The results allow us to identify the areas where the soil moisture loss is the most severe. Additionally, in some areas,the amplitude of the SWI trend indicates that the current compensation model is under threat.

Keywords: Drought, Remote, Earth Engine

References

1. France Assureurs (2022) Le risque sécheresse et son impact sur les habitations. Web article*
2. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment* **202**, 19-27.

Contact details

- Email: g.attard@ageoce.com
- Homepage: www.ageoce.com
- Social media: https://www.linkedin.com/company/ageoce/

Revisiting Whittaker-Henderson Smoothing

Guillaume Biessy, LinkPact & Sorbonne Université

Abstract: Introduced a century ago, Whittaker-Henderson (WH) smoothing remains as of today widely used among actuaries to build both unidimensional and bidimensional decrement tables for mortality and other biometric risks. This presentation aims at revisiting WH smoothing from a modern statistical perspective and tackles 5 related questions of special practical relevance.

First, enlighten the choice of the observation and weight vectors to which WH smoothing should be applied when building a decrement table, by connecting it to a maximum likelihood estimator introduced in a survival analysis framework. Second, adopt a bayesian interpetation of WH smoothing to obtain credibility intervals around its results. Third, cover selection of smoothing parameters, relying on maximisation of the restricted maximum likelihood. Fourth, improve numerical performances in cases where the number of observations, and therefore parameters, becomes overwhelming. Finally, extrapolate the results of WH smoothing while maintaining the consistency between fitted and extrapolated values.

Keywords: smoothing, survival analysis, mortality tables, restricted maximum likelihood, extrapolation

Contact details

- Email: guillaume.biessy78@gmail.com
- Repository: https://github.com/GuillaumeBiessy/WH

Multi-Parametric Gradient Boosting Machines with Non-Life Insurance Applications

Henning Zakrisson, Stockholm University (presenter)

Łukasz Delong, Warsaw School of Economics

Mathias Lindholm, Stockholm University

Abstract: A general multi-parametric gradient boosting machine (GBM) approach is introduced. The starting point is a standard univariate GBM, which is generalised to higher dimensions by using cyclic coordinate descent. This allows for different covariate dependencies in different dimensions. The suggested approach is also easily extended to, e.g., multi-parametric versions of XGBoost.

Given weak assumptions the method can be shown to converge for convex negative log-likelihood loss functions, which is the case, e.g., for d-parameter exponential families. Further, when having d-parametric distribution functions, it is important to design appropriate early stopping schemes. A simple alternative is introduced and more advanced schemes are discussed.

The flexibility of the method is illustrated both on simulated and real insurance data examples using different multi-parametric distributions, with both convex and non-convex losses.

Keywords: gradient boosting machines, cyclic coordinate descent, multi-parameter exponential family, early stopping

References

1. Delong, Łukasz, Mathias Lindholm, and Henning Zakrisson. "A Note on Multi-Parametric Gradient Boosting Machines with Non-Life Insurance Applications." *Available at SSRN 4352505 (2023).*
2. Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." *Annals of statistics (2001): 1189-1232.*

Contact details

- Email: zakrisson@math.su.se

From Chain Ladder to Probabilistic Neural Networks for Claims Reserving

Jacky Poon, Head of Finance - nib Travel. Member of the Machine Learning in Reserving Working Party

Abstract: The concept of an individual claims reserving model using proabilistic neural networks may sound intimidating to many actuaries. So in this presentation we take a simple chain ladder, and make incremental improvements to reach a crescendo with a probabilistic mixture density network on individual claims.

The session is intended to appeal for all levels of experience, from newbies to the field looking for an introductory overview and step-by-step explanations to experienced researchers, with the the model including concepts believed to be novel for use in reserving by the author, such as a customized neural network architecture, and log-normal mixture density networks.

An accompanying Python notebook with full workings in the data will be provided for further reading and to allow attendees to fully replicate the analysis.

This will be a virtual session - unfortunately the author is based in Sydney and will not be able to travel to London at the conference dates.

Keywords: Machine Learning, Reserving, Mixture Density Network

References

The research references papers including the following:

[1] Poon, J.H., 2019. Penalising unexplainability in neural networks for predicting payments per claim incurred. Risks 7, 95.

[2] MT Al-Mudafer, 2020, Probabilistic Forecasting with Neural Networks Applied to Loss Reserving,

Contact details

- Email: jacky.poon@nibtravel.com
- Homepage: https://actuariesinstitute.github.io/cookbook/docs/index.html
- Repository: https://github.com/JackyP
- Social media: https://www.linkedin.com/in/jacky-poon/

Generalized Bayesian Inference with Fairness Constraints

Tin Lok James Ng, School of Computer Science and Statistics, Trinity College Dublin, Ireland (presenter)

Abstract Algorithmic fairness is receiving significant attention in the academic and broader literature. Most machine learning algorithms are trained using data. If the underlying training data contains biases, algorithms trained on them will reflect and even amplify and perpetuate existing bias in the data.

Recent research has shown that machine learning models can result in potential biases when making decisions for people in different subgroups, which can lead to detrimental effects on specific demographic groups such as vulnerable ethnic minorities.

While much of the fair machine learning literature has focused on fairness definitions and designing fairness-aware learning algorithms, uncertainty quantification of fair machine learning algorithms is an important yet under-studied aspect.

We adopt a Bayesian constrained inference approach to incorporate fairness constraints in machine learning models. We further generalize the framework to Gibbs posterior inference by placing a model-based likelihood with a general loss function.

Asymptotic properties of the proposed framework are studied and applications to regression and classification settings are illustrated.

Keywords: Machine Learning Fairness, Gibbs Posterior, Constrained Inference

References

1. Liu, X., Tong, X., and Liu, Q. (2021), "Sampling with Trusthworthy Con- straints: A Variational Gradient Framework," in *Advances in Neural Information Processing Systems*, Vol. 34, pp. 23557–23568.
2. Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016), "A general framework for updating belief distributions," *Journal of the Royal Statistical Society Series B*, 78, 1103–1130
3. Sen, D., Patra, S., and Dunson, D. (2018), "Constrained inference through posterior projections," *ArXiv*

Contact details

- Email: ngja@tcd.ie

Actuarial Applications of Natural Language Processing Using Transformers: Case Studies for Using Text Features in an Actuarial Context

Andreas Troxler, AT Analytics

Jürg Schelldorfer, Swiss Re (presenter)

Abstract:
This talk demonstrates workflows to incorporate text data into actuarial classification and regression tasks. The main focus is on methods employing transformer-based models. The talk provides practical approaches to handle classification tasks in situations with no or only few labeled data. A dataset with short property insurance claims descriptions is used to demonstrate the techniques. The results achieved by using the language-understanding skills of off-the-shelf natural language processing (NLP) models with only minimal pre-processing and fine-tuning clearly demonstrate the power of transfer learning for practical applications.

This case study has been done as part of the "Data Science" working group of the Swiss Association of Actuaries (SAA). The group publishes tutorials that discuss the use of machine learning techniques for actuarial applications. The tutorials are self-explanatory and its code and data is publicly available on the website www.actuarialdatascience.org.

Keywords: Natural language processing, NLP, transformer, multi-lingual models, domain-specific fine-tuning, integrated gradients, extractive question answering, zero-shot classification, topic modeling

References

1. Troxler, A., Schelldorfer, J. (2022) Actuarial Applications of Natural Language Processing Using Transformers: Case Studies for Using Text Features in an Actuarial Context, *arXiv*, 2206.02014, https://arxiv.org/abs/2206.02014

Contact details

- Email: Juerg_Schelldorfer@swissre.com, andreas.troxler@atanalytics.ch
- Homepage: www.actuarialdatascience.org
- Repository: https://github.com/JSchelldorfer/ActuarialDataScience
- Linkedin: https://www.linkedin.com/company/actuarial-data-science-in-the-swiss-association-of-actuaries/

The fast and the fabulous. Harnessing GPU power for high-performance life insurance computations

Karol Maciejewski and Mehdi Echchelh, Milliman (presenters)

Abstract Data science most often relies on heavy computations on data, either due to complexity of calculations or volume of the data, or both. In this practical technical presentation we want to discuss our experiences working on design and implementation of high-performance large-scale computations required in the context of life insurance ALM projections using Python, GPU and compute clusters.

Several of the commonly used Python data science and machine learning libraries provide simple entry points to the GPU computations that don't require GPU knowledge to use. However, if one wants to go beyond that and develop fully custom models, at least some understanding of the GPU architecture and packages allowing its utilization is necessary to truly take advantage of this technology and avoid common pitfalls. We will present the difference in the computation model of the CPU and GPU, types of problems that can be efficiently tackled using massive parallelization on GPU and clusters, selected Python libraries that can be used by modellers with different levels of experience and some hints on what to do and what not to do when implementing GPU computations. We will also show benchmarks of our implementations to show that speed-ups in order of magnitude of up to 1000x are possible with an appropriately chosen model and implementation architecture.

We believe our considerations and examples can be very useful to anyone working with life insurance data. While we focus on multidimensional liability and asset projections, these methods can be applied to any data engineering or data science problems equally well. We also use this opportunity to highlight key differences between developing sand-boxed models for ad-hoc studies versus IT-approved production models.

Keywords: life insurance, projection model, model design, ALM, Python, HPC, CUDA, GPU, Numba, Numpy, Cupy, Dask

References

1. Maciejewski, K., Echchelh, M., Sznajder D. (2023). Building a high-performance in-house projection and ALM model. Architecture and implementation considerations in Python. *pending publication on Milliman website*

Contact details

- Email: karol.maciejewski@milliman.com
- LinkedIn: https://www.linkedin.com/in/karolmaciejewski/
- Email: mehdi.echchelh@milliman.com
- LinkedIn: https://www.linkedin.com/in/mehdi-echchelh/

Reinforcement learning in search of optimal premium rules

Lina Palmborg, Stockholm University (presenter)

Filip Lindskog, Stockholm University

Abstract: In simplified settings, optimal premium rules can be derived by classical approaches such as dynamic programming methods providing solutions to a Bellman equation. Realistic insurance settings involve features such as reporting/payment delays and fluctuations in the number of policyholders, partly in response to varying premium levels. In such settings, classical approaches are not applicable due to the size of the state space and lack of explicit expressions for transition probabilities.

I will discuss how to design efficient algorithms in search of optimal premium rules in realistic insurance settings requiring reinforcement learning combined with function approximation. Both theoretical properties and practical aspects of such algorithms will be presented.

The talk is based on Palmborg & Lindskog (2023) and current work.

Keywords: Reinforcement learning, premium control

References

1. Palmborg, L., Lindskog, F. (2023). Premium control with reinforcement learning. *ASTIN Bulletin: The Journal of the IAA*, 1-25. Open access https://doi.org/10.1017/asb.2023.13

Contact details

- Email: lina.palmborg@math.su.se

Territorial Ratemaking and Graph Theory

Marco De Virgilis, Arch Insurance (presenter)

Abstract: In non-life insurance, territory-based risk classification is useful for various insurance operations including marketing, underwriting and ratemaking. This presentation will show how to develop a modelling framework to produce territorial risk scores that employs aggregate insurance claims alongside geographical information and graph theory.

Actuarial methodologies based on GIS (Geographic Information Systems) have been implemented and used predominately in the industry for many years; nowadays, in fact, almost all ratemaking methodologies for the pricing of personal lines of automobile and homeowners insurance in the United States include a geographical component.

Alongside new developments in the realm of spatial data science, in recent years, graph theory has established itself as an important mathematical tool in a wide variety of subjects ranging from geography to linguistic and chemistry.
In this presentation we will show how to combine data and modeling approaches from both fields, spatial and graph related in order to produce accurate risk scores.

The presentation will:

a. Briefly review the theoretical foundations behind Graph Theory and Spatial Analysis.
b. Show how to apply suitable algorithms on typical geographical insurance data sets.
c. Compare different approaches in terms of predictive performance and practical usability.

Learning Objectives:

a. The audience will be knowledgeable on the theoretical foundation of graph theory and spatial modeling techniques, with emphasis on the practical side. Participants will appreciate how to employ such techniques in the context of territorial ratemaking.
b. The audience will gain insights on how to implement graph analysis in ratemaking. The literature on this topic is scarce and coming from other disciplines. This presentation will give clear explanations of the technicalities involved in defining and manipulating graphs built on spatial data.
c. The attendees will gain a working knowledge of how to evaluate several algorithms that will be presented. They will be able to assess the effectiveness of such techniques from several standpoints. In particular, areas such as goodness of fit and implementation hurdles and requirements will also be addressed.

All the techniques and the analysis presented will be based on open-source softwares and packages to ensure reproducibility and easy implementation by the audience.

Keywords: Ratemaking, Graph, Spacial Data Science, Geographic Risk

References

1. Begher, F. et al. (2011), *Territorial Analysis for Ratemaking.* University of California at Santa Barbara.
2. Pebesma, E., Bivand, A. (2022). *Spatial Data Science: With Applications in R.* Chapman & Hall/CRC.
3. Taylor, G. (2001), *Geographic Premium Rating by Whittaker Spatial Smoothing.* CAS E-Forum.
4. Werner G. (1999), *The United States Postal Service's New Role: Territorial Ratemaking.* CAS E-Forum.
5. Wilson, R. (1996). *Introduction to Graph Theory.* Longman.

Contact details

- Email: devirgilis.marco@gmail.com
- Repository: https://github.com/marcopark90
- Social media: http://www.linkedin.com/in/marco-de-virgilis

Individual claims reserving with dependent censored data

Marie Michaelides, Université du Québec à Montréal (presenter)

Mathieu Pigeon, Université du Québec à Montréal

Hélène Cossette, Université Laval

Abstract: We study the dependence between the different insurance coverages offered within a single policy when a claim occurs and the impact that this dependence has on the total reserves amount.

We propose an individual claims reserving model that allows to estimate the development of each claim through the different insurance coverages that it impacts. More specifically, for a single claim, we jointly model the activation delays of the coverages. We then complete the model by estimating not only the correlated activation delays of the different coverages but also the subsequent development of each claim, namely the payments that might occur after activation of the coverages and their corresponding severities. We finally propose an illustration of our model using a recent automobile dataset from a Canadian insurance company.

Keywords: claims reserving, dependence, censored data

References

1. Denuit, M., Purcaru, O., Van Keilegom, I. (2006). Bivariate Archimedean Copula Models for Censored Data in Non-Life Insurance. *Journal of Actuarial Practice* **13(2006)**, 5-32.
2. Akritas, M., Van Keilegom, I. (2003). Estimation of bivariate and marginal distributions with censored data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **65(2)**, 457-471.
3. Genest. C., Rivest, L-P. (1993). Statistical Inference Procedures for Bivariate Archimedean Copulas. *Journal of the American Statistical Association* **88(423)**, 1034-1043.
4. Beran. R. (1981). Nonparametric Regression with Randomly Censored Survival Data. *Technical Report. University of California, Berkeley*, 19.

Contact details

- Email: michaelides.marie@uqam.ca
- Social media: Linkedin

Isotonic recalibration under a low signal-to-noise ratio

Mario V. Wüthrich, RiskLab, ETH Zurich

Abstract: Insurance pricing systems should fulfill the auto-calibration property to ensure that there is no systematic cross-financing between different price cohorts in the pricing system. Often, regression models are not auto-calibrated. We present the method of isotonic recalibration to a given regression model which gives us the auto-calibration property. As a nice side result we see that under a low signal-to-noise ratio, this isotonic recalibration step leads to explainable pricing systems because the resulting isotonically recalibrated regression functions have a low complexity.

This is joint work with Johanna Ziegel.

Keywords: auto-calibration, isotonic regression, isotonic recalibration, low signal-to-noise ration, neural networks, algorithmic regression models.

References

1. Wüthrich, M.V., Ziegel, J. (2023). Isotonic recalibration under a low signal-to-noise ratio. arXiv:2301.0269.
2. Wüthrich, M.V. (2023). Model selection with Gini indices under auto-calibration. European Actuarial Journal, in press.
3. Wüthrich, M.V., Merz, M. (2023). Statistical Foundations of Actuarial Learning and its Applications. Springer Actuarial. Open Access. https://link.springer.com/book/10.1007/978-3-031-12409-9

Contact details

- Email: mario.wuethrich@math.ethz.ch
- Homepage: https://people.math.ethz.ch/~wmario/

Fitting Development and Tail Models Jointly via Mixture Modeling

Mark Shoun, Ledger Investing (presenter)

Abstract: Within the context of loss development for casualty insurance, tail factor estimation is often regarded as an afterthought, and tail factor selection is poorly integrated with the rest of the loss development process. When tail factors are based on extrapolation from a loss triangle, the tail factor model is typically a parametric curve fitted to observed development factors. By analogy to the tail model, we refer to the primary loss development model as the "body" model. When the tail model is combined with the body model, this effectively creates a sharp discontinuity at the transition point between the two models.

We propose a framework for loss development where the model is a mixture of a traditional loss development model for the body and a parametric extrapolation model for the tail. The two models and the mixture weights are estimated simultaneously. The component weights in the mixture model are a function of development lag, allowing for a natural, monotonic transition between the body and tail models. We illustrate the performance of the blended model on data from US insurers' statutory filings, and demonstrate that the blended model improves development factor estimation for development lags covered by the triangle, not just the tail factor.

Keywords: Loss Reserving, Tail Factors, Mixture Modeling, Bayesian Modeling

Contact details

- Email: mark@ledgerinvesting.com
- Linkedin: https://www.linkedin.com/in/mark-shoun/

Measuring daily value creation of global specialty (re)insurance

Markus Gesmann, Insurance Capital Markets Research

Abstract: Capital markets have a growing appetite for insurance as an asset class. Investment in insurance risk, e.g. via insurance linked security structures or Lloyd's participation is seen as a portfolio diversifier. However, the lack of regular valuation for these instruments inhibits liquidity and makes them less fungible.

We have developed an equity index that aims to provide a daily capital markets perspective, based on the publicly listed companies that participate in Lloyd's, the leading marketplace for global specialty (re)insurance risks.

The innovative weighting methodology of the 'RISX' index ensures the index correlates more with Lloyd's risk profile than wider equity markets, and hence gives real time insight into capital market's perspective and confidence in global specialty (re)insurance as an asset class.

The talk will discuss the methodology of the index, which is based on premium rather than market capitalisation, and its wider application for daily valuation and forecasting Lloyd's underwriting results.

Keywords: RISX, Lloyd's, equity index, valuation, time series, ILS

References

- About RISX Index: https://risxindex.com/downloads/Case_for_RISX.pdf
- Factsheet: https://risxindex.com/downloads/RISX_Factsheet.pdf
- Methodology: https://moorgatebenchmarks.com/wp-content/uploads/2021/04/ICMR-ReInsurance-Specialty-Index-Methodology-v1.2.pdf

Contact details

- Email: markus.gesmann@insurancecapitalmarkets.com
- Web site: https://risxindex.com
- Social media: https://www.linkedin.com/in/markus-gesmann/

Local bias adjustment, duration-weighted probabilities, and automatic construction of tariff cells

Mathias Lindholm, Stockholm University

Abstract: We study non-life insurance pricing and present a general procedure for constructing a distribution-free locally unbiased predictor of the risk premium based on any initially suggested predictor. The resulting predictor is piecewise constant, corresponding to a partition of the covariate space, and by construction auto-calibrated.

Two key issues are the appropriate partitioning of the covariate space and the handling of randomly varying durations, acknowledging possible early termination of contracts. A basic idea in the present paper is to partition the predictions from the initial predictor, which as a by-product defines a partition of the covariate space. Two different approaches to create partitions are discussed in detail using (i) duration-weighted equal-probability binning, and (ii) binning by duration-weighted regression trees.

Given a partitioning procedure, the size of the partition to be used is obtained using cross-validation. In this way we obtain an automatic data-driven tariffication procedure, where the number of tariff cells corresponds to the size of the partition. We illustrate the procedure based on both simulated and real insurance data, using both simple GLMs and GBMs as initial predictors.

The resulting tariffs are shown to have a rather small number of tariff cells while maintaining or improving the predictive performance compared to the initial predictors.

Keywords: Local bias adjustment, duration-weighted probabilities, non-life pricing, automatic tariffication

References

1. Lindholm, M., Lindskog, P., Palmquist, J. (2023). Local bias adjustment, duration-weighted probabilities, and automatic construction of tariff cells. *Scandinavian Actuarial Journal (available online)*.

Contact details

- Email: lindholm@math.su.se
- Homepage: https://staff.math.su.se/lindholm/

A neural network approach for selecting efficient reinsurance strategies

Matteo Crisafulli, Università di Roma, La Sapienza

Abstract: Reinsurance treaties are one of the main instruments used by insurance companies for reducing their risks and balancing technical performance. The choice of optimal reinsurance is a topic object of extensive academic and professional studies, which consists in finding the combinations of reinsurance strategies which maximize the objective of the insurance company. The complexity of this analysis increases when we are interested in the joint optimization of more than one metric (i.e. a multi-objective optimization problem), since there is typically a trade-off between the multiple objectives which move in different directions.

We propose the application of a neural network model for finding the efficient frontier in this multi-objective optimization problem, requiring limited data and preserving the possibility of deriving the strategies which generate the Pareto front. Numerical application is performed assuming a multi-line non-life insurer, with parameters from the Italian market, which is interested in finding the optimal combinations of reinsurance treaties maximizing its Solvency Ratio (SR) and Return on Equity (RoE).

We show how this approach offers another perspective for determining efficient reinsurance strategies, which can be especially useful in case of high number of potential combinations defining each strategy.

Keywords: neural network, efficient frontier, reinsurance

References

1. Cheng, X., Jin, Z., and Yang, H. (2020). Optimal insurance strategies: A hybrid deep learning markov chain approximation approach. *ASTIN Bulletin: The Journal of the IAA*, 50(2):449–477.
2. Navon, A., Shamsian, A., Chechik, G., and Fetaya, E. (2020). Learning the pareto front with hypernetworks. *arXiv preprint*
3. Ruchte, M. and Grabocka, J. (2021). Scalable pareto front approximation for deep multi-objective learning. *IEEE International Conference on Data Mining (ICDM)*, 1306–1311.
4. Zanotto, A. and Clemente, G. P. (2021). An optimal reinsurance simulation model for non-life insurance in the solvency II framework. *European Actuarial Journal*, 1–35

Contact details

- Email: matteo.crisafulli@uniroma1.it

Forecasting Mortality by cause with Zero Death Counts

Michelle Dong, The Australian National University (presenter)

Han Lin Shang, Macquarie University

Aaron Bruhn, The Australian National University

Francis Hui, The Australian National University

Abstract: We consider a compositional power transformation, known as alpha transformation, to model and forecast a time series of cause-specific life-table death counts. As a generalisation of the isometric log-ratio transformation (alpha = 0), the alpha transformation relies on the tuning parameter alpha, determined in a data-driven manner. Using the human cause-of-death database, the alpha-transformation produces more accurate forecasts than the log-ratio transformation.

In addition, the alpha transformation can address the problem of zero counts, commonly occurring at higher ages. The improved forecast accuracy of cause-specific life-table death counts is important to predict future mortality incidence rates. It can provide insight into the impact of external drivers of change, including the emergence of climate-related risks over time.

Keywords: Cause of Death, Compositional Data Analysis, Alpha-Transformation, Mortality Forecast

References

1. Kjaergaard, S., Ergemen, Y. E., Kallestrup-Lamb, M., Oeppen, J., Lindahl-Jacobsen, R. et al. (2019), Forecasting causes of death using compositional data analysis: the case of cancer deaths, Department of Economics and Business Economics, Aarhus University.
2. Oeppen, J. et al. (2008), 'Coherent forecasting of multiple-decrement life tables: a test using Japanese cause of death data'.
3. Tsagris, M. T., Preston, S. and Wood, A. T. (2011), 'A data-based power transformation for compositional data'

Contact details

- Email: zhe.dong@anu.edu.au

A Bayesian Approach to Customer Lifetime Value

Mick Cooney, Describe Data

Abstract: Calculating customer lifetime value in a non-contractual setting can be challenging due to the arbitrary nature of customer behaviour and the lack of knowledge or precision for lifetime measurement.

This talks covers attempts to implement a Bayesian version of some standard Buy-Till-You-Die (BTYD) models common in retail analytics. In particular, we discuss the challenges in setting priors and calculating likelihoods and cover this use of Simulation-Based Calibration to diagnose issues with these models.

We conclude with some discussion around uses of this model in insurance, such as for modelling transactional claims development.

Keywords: Bayesian, customer life time value, buy till you die, Stan

References

1. Schmittlein, David C., Donald G. Morrison, and Richard Colombo (1987) Counting Your Customers: Who They Are and What Will They Do Next? Management Science, 33 (January), 1–24

Contact details

- Email: mcooney@describedata.com
- Homepage: http://kaybenleroll.github.io/data_workshops/
- Repository: https://github.com/kaybenleroll/data_workshops.git
- Social media: https://www.linkedin.com/in/mick-cooney/

Expressive Mortality Models through Gaussian Process Compositional Kernels

Mike Ludkovski, University of California Santa Barbara (presenter)

Jimmy Risk, Cal Poly Pomona

Abstract: We develop a flexible Gaussian Process (GP) framework for learning the covariance structure of Age- and Year-specific mortality surfaces. Our compositional search builds off the Age-Period-Cohort (APC) paradigm to construct a covariance prior best matching the spatio-temporal dynamics of a given mortality dataset.

Utilizing the additive and multiplicative structure of GP kernels, we design a genetic programming algorithm to identify the fittest kernels according to the Bayesian Information Criterion. Several synthetic case studies demonstrate the ability of our Genetic Algorithm (GA) to recover various APC structures. We then apply the GA on several national-level mortality datasets from the Human Mortality Database.

Our machine-learning based analysis provides novel insight into the presence/absence of Birth Cohort effects in different populations, the relative smoothness of mortality surfaces along the Age and Year dimensions, and the nonstationarity of the mortality surface covariance structure. Our modelling work is done with the PyTorch libraries in Python.

Keywords: Mortality modeling, Human Mortality Database, Gaussian process models, cohort effect

References

1. D. Duvenaud, J. Lloyd, R. Grosse, J. Tenenbaum, and G. Zoubin (2013). Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning (ICML)*, PMLR, 1166–1174.
2. A. Hunt and D. Blake (2020). Identifiability in age/period/cohort mortality models. *Annals of Actuarial Science*, 14(2):500–536
3. M. Ludkovski, J. Risk, and H. Zail (2018). Gaussian process models for mortality rates and improvement factors. *ASTIN Bulletin*, 48(3):1307–1347
4. M. Ludkovski, J. Risk (2023). Expressive Mortality Models through Gaussian Process Compositional Kernels, *in preparation*.

Contact details

- Email: ludkovski@pstat.ucsb.edu
- Homepage: ludkovski.faculty.pstat.ucsb.edu
- Twitter: @MLudkovski

On functional decompositions, post-hoc machine learning explanations and fairness

Munir Hiabu, University of Copenhagen (presenter)

Joseph T. Meyer, Heidelberg University

Marvin N. Wright, Leibniz Institute for Prevention Research & Epidemiology, University of Bremen

Abstract: In the last decade machine learning algorithms have shown unprecedented accuracy in a variety of applications and tasks. However, current state-of-the-art machine learning algorithms are black-box models. As such, they make it seemingly hard to understand the relationship between predictors and response.

Current post-hoc machine learning explanations usually only focus on explaining an approximation of the fitted model or merge interaction effects into local explanations.

In this talk I will discuss that if predictions are the composition of low dimensional structures, then interpretation of the exact model is possible via a functional decomposition of the output function. A functional decomposition unifies the notion of local explanations, global explanations, and causal effects. The latter can be used for individual fairness considerations and discrimination-free pricing. Examples of machine learning predictors that are compositions of low dimensional structures are gradient boosting machines and random planted forests.

An accompanying R-package is available at https://github.com/PlantedML/glex.

Keywords: interpretable machine learning, local explanation, global explanation, causality, fairness

References

1. Hiabu, M., Meyer, J. T., Wright, M. N. (2023). Unifying local and global model explanations by functional decomposition of low dimensional structures. [https://arxiv.org/abs/2208.06151](https://arxiv.org/abs/2208.06151

Contact details

- Email: mh@math.ku.dk
- Homepage: https://mhiabu.github.io/
- Repository:https://github.com/PlantedML
- Twitter: https://twitter.com/hiabumunir

Saving the World: Predictive Early Warning Systems for Conflict Risk using Neural Networks

Navarun Jain, Lux Actuaries & Consultants (presenter)

Sergey Trofimov FIA, Lux Actuaries & Consultants (co-presenter)

Abstract: Civil war and organised violence continue to erupt throughout the world into the 21st century, with thousands of victims annually. In addition to the direct damage inflicted such conflicts also have a negative impact on the socio-economic environment, including but not limited to economic growth, willingness to do business and demographics. Armed conflict appears to be more likely in some countries than in others, and we use a data-driven approach to understand this likelihood and its main drivers. Retrospectively, for each conflict, the root cause or a complex of such causes can be assessed. It therefore can be useful to create early warning predictors for armed conflicts and potentially use them to understand what drives conflicts and wars. Such early warning indicators can play a key role in P&C ratemaking, especially in the Marine, Construction and Property LOBs.

Our model is designed not only to determine the degree of risk of conflict for each country in the near future but also to show the main objectively measurable indicators that can be linked to the emergence of conflicts that have already occurred and may affect the development of conflict in the future. The key engine of the Model is a Multi-layer Perceptron. This is a Supervised Feed-forward Neural Network with multiple layers which is trained to predict whether the country is in the conflict status or not, based on several input features sourced from reliable platforms such as the World Bank, International Labour Organisation, UN Population Program and the REIGN (Rulers, Elections and Irregular Governance) Databse among others.

The key production output from the Model is a world map where countries are ranked in accordance with the risk profile assessed by the model based on the likelihood of the conflict in each month in the next 3 years (up to the end of 2024):

- Low Risk: Likelihood of conflict does not exceed 25%
- Low to Moderate Risk: Likelihood of conflict is between 25% and 50%
- Moderate to High Risk: Likelihood of conflict is between 50% and 75%
- High Risk: Likelihood of conflict exceeds 75%.

The Model outputs for each country the assigned likelihood of being in each state (in conflict / not in conflict) which allows explicit ranking.

Keywords: Supervised Learning, Neural networks, War risk, Predictive analytics

References

Contact details

- Email: navarun.jain@luxactuaries.com
- Email: sergey.trofimov@luxactuaries.com

Vine Copulas for Systemic Cyber Risk Modelling

Nicholas Robert, DeNexus (presenter)

Diana Carrera, DeNexus (presenter)

Romy R. Ravines, DeNexus

Abstract: Systemic cyber risk is an increasing threat to businesses and society, characterised by complex and cascading consequences that are hard to predict[1]. Reliable actuarial models for risk accumulation are lacking in the insurance industry, leaving a gap of unmet demand for coverage worth up to 1 trillion dollars. (Re)insurance firms seek models that describe the uncertainty around the patterns of portfolio co-exposure, and the distribution of losses arising as a result of the event. The challenge of building these models is compounded by the complex dependencies in the digital world, the dynamic nature of cyber threats, and the scarcity of data.

We present a novel approach for modelling cyber accumulation risk using vine copulas, which are a technique for specifying complex joint distributions by composing bivariate conditional copulas[2]. Vine copulas have seen significant usage in financial and insurance contexts due to their ability to model complex relationships between different factors[3], such as cyber asset characteristics like shared service, technological, or operational dependencies. We discuss approaches for constructing vines which capture asymmetric dependence and tail risk, or emphasise specific knowledge of the cyber portfolio risk structure [2, 3]. We utilise a proprietary catalog of cyber incidents to estimate the optimal structure and parameters of the vines, complemented by a Bayesian unit-risk modeling system to estimate financial losses.

To address computational challenges with high dimensional copulas[4], we utilise a custom simulation library to construct and fit large vines, enabling modelling of large portfolio accumulation. We demonstrate the effectiveness of this method on synthetic scenarios based on real-world cyber events, and compare it with alternative approaches such as point-process or epidemic network models[5]. We find that this vine copula based approach yields significant advantages in robustness and accuracy of VaR estimates, as well as interpretability of portfolio exposure outcomes.

Keywords: systemic cyber risk, vine copulas, loss accumulation, simulations, dependency uncertainty modelling

References

1. World Economic Forum. (2022). Systemic cybersecurity risk and role of the global community. Retrieved from https://www3.weforum.org/docs/WEF_GFC_Cybersecurity_2022.pdf
2. Aas, K., Czado, C., Frigessi, A., & Bakken, H. (2009). Pair-copula constructions of multiple dependence. Insurance: Mathematics and Economics, 44(2), 182-198.
3. Denuit, M., Guillén, M., Trufin, J., & Vanasse, C. (2017). Regular vines and insurance risk analysis. Insurance: Mathematics and Economics, 75(1), 43-56.
4. Nagler T., Bumann C., & Czado C. (2019). Model selection in sparse high-dimensional vine copula models with an application to portfolio risk. Journal of Multivariate Analysis 172(1), 180-192.
5. Hillairet C., & Lopez O. (2021). Propagation of cyber incidents in an insurance portfolio: counting processes combined with compartmental epidemiological models. Scandinavian Actuarial Journal.

Contact details

- Email: nr@denexus.io, dc@denexus.io, rr@denexus.io

Data is not neutral: ethics and AI

Nuzhat Jabinh FRSA Ethics in AI consultant

Abstract: Data is not neutral: ethics and AI

There is often an assumption that data is neutral, that it doesn't carry values or meanings other than that it represents. All too often bias in AI systems is caused by existing bias in data sets that reflect wider societal attitudes.

This talk will give a quick overview of an ethical approach to AI and its benefits. This will include how AI can benefit the insurance industry.

My main points are: while ethics are important in their own right, there is a business case for them. (This will be discussed earlier in the talk). There is also regulatory pressure to build ethical AI. It is likely to lead to more robust products.

Keywords: Ethics, AI, collaboration, London Insurance Market, business value, business case, ROI

References

1. Gartner hype cycle

Contact details

- Email: N@nuzhat.net
- Homepage: Nuzhat.net
- Social media: https://uk.linkedin.com › nuzhat-jabinh-frsa

Causal Inference and Fairness in Insurance Pricing

Olivier Côté, Université Laval (presenter)

Arthur Charpentier, Université du Québec à Montréal

Marie-Pier Côté, Université Laval

Abstract: The insurance industry has long been wary of including sensitive variables in pricing, whether to comply with laws, ethical considerations, or to avoid reputational risk [5].

The ubiquity of big data and the growing complexity of machine learning algorithms raise concerns regarding the possibility of indirect discrimination on sensitive variables. [2].

Fairness is an essential subject in insurance, as insurers use personal information and interact with a large part of the population. The unfair bias correction problem in insurance is similar to the general goal of causal inference: using causal assumptions to avoid (unfair) biases against (prohibited) confounders [4]. Causal inference is an already-available science that allows us to reformulate our fairness in modelling problem [1].

In this talk, we pursue the discussion on discrimination of [3] and expand the causal perspective on fairness with the goal of limiting indirect discrimination.

Keywords: Fairness, Discrimination, Causal Inference, Actuarial Science, Insurance, Modeling, Ethics

References

1. Araiza Iturria, C. A., Hardy, M., & Marriott, P. (2022). A discrimination-free premium under a causal framework. Available at SSRN 4079068.
2. Embrechts, P., & Wüthrich, M. V. (2022). Recent challenges in actuarial science. Annual Review of Statistics and Its Application, 9, 119-140.
3. Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2022). A Discussion of Discrimination and Fairness in Insurance Pricing. arXiv preprint arXiv:2209.00858.
4. Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2022). Discrimination-free insurance pricing. ASTIN Bulletin: The Journal of the IAA, 52(1), 55-89.
5. Xin, X., & Huang, F. (2022). Anti-discrimination insurance pricing: regulations, fairness criteria, and models. Fairness Criteria, and Models (May 1, 2022).

Contact details

- Email: olivier.cote.12@ulaval.ca
- Repository: github.com/OliCoSide
- Social media: https://www.linkedin.com/in/olivier-cote-act/

Advanced analytics and machine learning to identify fraudulent health insurance claims

Paola Gasparini, BUPA (presenter)

Paula Ramon Armas, BUPA (presenter)

Abstract: Healthcare insurance fraud is a significant problem that can result in higher premiums for policyholders and decreased access to quality care [1].

In this work, we present a tool that we developed utilising data science techniques to detect fraudulent activity in healthcare insurance claims. The tool employs cosine similarity to identify providers with similar characteristics, scatter plots to visualize patterns in the data, outlier trees [2] to identify unusual claims, and network analysis to uncover connections between suspicious consultants. By applying these methods to a large dataset of insurance claims, we were able to identify a number of fraudulent claims that would have otherwise gone undetected.

Our results demonstrate the effectiveness of using data science tools for fraud detection in the healthcare insurance industry, and we believe the tool can significantly reduce fraudulent activity and improve the efficiency of the insurance claims process.

Keywords: Healthcare insurance, Fraud detection, Advanced Analytics, Machine Learning

References

1. Gee J., Button M. (2015). *The financial cost of healthcare fraud. In: what data from around the world shows.* London: PFK Littlejohn LLP
2. David Cortes. *Package 'outliertree'; Explainable Outlier Detection Through Decision Tree Conditioning*

Contact details

- LinkedIn: linkedin.com/in/paola-gasparini-723a6a34

. . . whatever remains, however improbable, must be a bug

Patrick Hogan, Senior Data Scientist at PartnerRe, Zürich

Abstract: Open-source modelling relies on a multitude of interacting elements. The underlying technical pieces generally work seamlessly together, and when things go wrong the cause usually lies with the modeller.

However very occasionally these components don't quite play as nicely together as expected, and in our case produced a Bayesian mortality model which was subtly yet profoundly flawed.

We present a detective story in which rigorous model validation, a systematic approach to debugging and fruitful collaboration with our system admins & the open-source community allowed us to identify, isolate and guard against a pernicious bug lurking deep within our technical stack.

Keywords: open-source software, model validation, mortality modelling, generalised additive models

References

1. Hogan, P. M. (2023). Spline model prediction inconsistencies when changing BLAS/LAPACK implementation. *brms GitHub issue #1465*: https://github.com/paul-buerkner/brms/issues/1465.
2. R Core Team (2022). *R Installation and Administration*, Section A.3: Linear algebra.

Contact details

- Email: patrick.hogan@partnerre.com
- Repository: https://github.com/pmhogan
- Social media: https://www.linkedin.com/in/patrick-hogan-ch/

Solving censored regression problems using a multitask approach

Philipp Ratz, Université du Québec à Montréal (UQAM)

Abstract: There exists a stack of well-established parametric estimators that work well on regular and aggregated survival data but less so on individual prediction problems. More recently, algorithms for predictive survival problems were proposed, but they often favour predictive accuracy over interpretability and calibration.

To bridge the gap, we show that common survival estimators can be reformulated into problems that involve solving multiple related "tasks". This reformulation then allows to leverage techniques from the field of machine learning which work well in high dimensions, without sacrificing the interpretability of the results completely.

Deriving these new estimation techniques directly from well-known survival models such as the Kaplan-Meier further permits the use of existing techniques to deal with common issues such as dependent censoring. We present two applications where we estimate individual survival curves from censored data and predict survival quantiles under censoring.

Keywords: Survival Analysis, Neural Networks, Quantiles

References

1. Kvamme, Håvard, and Ørnulf Borgan. "Continuous and discrete-time survival prediction with neural networks." *Lifetime data analysis 27* (2021): 710-736.
2. Lee, Changhee, et al. "Deephit: A deep learning approach to survival analysis with competing risks." *Proceedings of the AAAI conference on artificial intelligence. Vol. 32.* No. 1. 2018.
3. Chen, George. "Nearest neighbor and kernel survival analysis: Nonasymptotic error bounds and strong consistency rates." *International Conference on Machine Learning. PMLR, 2019*.

Contact details

- Email: ratz.philipp@courrier.uqam.ca
- Homepage: phi-ra.github.io
- Repository: github.io/phi-ra
- Social media: https://www.linkedin.com/in/philippratz/

Embedding data science in non-life reserving

Priyank Shah, LCP (presenter)

Charlie Stone, LCP

Abstract: A lot is needed from non-life insurers' reserving processes. Insurers require reserves for financial reporting to be delivered in short timescales, whilst also needing reserving to provide detailed insight and analysis of insurers' performance. Data science has the potential to help actuaries automate significant parts of the reserving process and generate valuable insights from their data. However it can be challenging to embed techniques and ideas from data science into the reserving process in practice.

We will share our experience of embedding data science in reserving over the last four years, where we have worked with insurers and regulators to develop upon standard approaches and introduce novel visualisation tools. In particular, we will look at three examples:

- Automated identification of trends in reserving diagnostics.
- Automating the selection of traditional reserving methods and assumptions, and priortising assumptions for actuarial review.
- Improving the setting of IEULR assumptions with time series forecasting.

Keywords: Reserving, Practical, Trends, Actuary, Automation

References

1. Richman, R., Balona, C. (2020.) The Actuary and IBNR Techniques: A Machine Learning Approach. *Available at SSRN: https://ssrn.com/abstract=3697256*

Contact details

- Email: charlie.stone@lcp.uk.com
- LinkedIn: https://uk.linkedin.com/in/charlie-stone-a24b5125
- Email: priyank.shah@lcp.uk.com
- LinkedIn: https://www.linkedin.com/in/priyankshah7/

Py-shiny for Reinsurance: ready or not, here we come

Roland A. Schmid, Mirai Solutions (presenter)

Paula K. Ando, Mirai Solutions

Abstract: Shiny for Python has been launched, and while it is officially still in **alpha** stage, many people are experimenting with it and we can expect rapid progress.

With corporate reinsurance clients interested in extending and modernizing solutions to **explore, accumulate and manage risks in interactive and responsive ways**, we decided to develop a freely available open-source service through a Py-Shiny app. This could serve as a basis and hopefully provide some common functionality out-of-the-box, while raising and solving current shortcomings of Py-Shiny along the way and thus helping to accelerate its development and maturing process. Python is typically the first-choice language for scalable enterprise-grade solutions in the GIS-domain as well as when it comes to working with large simulation data in a framework like (Py-)Spark.

We show that even now Shiny for Python is already in a very useable state, although there is still a lot of more specific functionality and tooling available in Shiny for R, which hasn't been provided to Python yet.

The core functionality of our initial **Risk Explorer** app / service shall help to visualize, assess, communicate and understand risks and exposure, covering the following features:

- interactively create and modify (reshape) custom (ad-hoc) hazard footprints, download footprint data
- upload scenario (footprint) data and visualize
- upload exposure (location) data and visualize
- calculate and download distances to center of footprints / scenarios
- associate / amend details: define intensities, vulnerabilities / damage ratios, probabilities

In addition we aim to provide and integrate a direct connection to the data of the United States Geological Survey (USGS).

During this talk we introduce **Risk Explorer** illustrating an **effective use-case based on the US terrorism scenarios relevant under ORSA**.

Keywords: Reinsurance, Exposure, Scenario, Risk Insight, SolvencyII, ORSA, Shiny, Python, Leaflet

References

1. Posit Software, PBC (RStudio). https://shiny.rstudio.com/py/.
2. Mirai Solutions GmbH. https://github.com/miraisolutions/PublicTalks/tree/master/InsuranceDataScience2023.
   Abstract plus screenshots of current work-in-progress version of the app.

Contact details

- Email: roland.schmid@mirai-solutions.com
- Homepage: https://mirai-solutions.ch
- Repository: https://github.com/miraisolutions
- Social media: Linkedin.com/in/Roland-A-Schmid

Accurate and Explainable Mortality Forecasting with the LocalGLMnet

Francesca Perla, Department of Management and Quantitative Sciences, University of Naples, Parthenope,

Ronald Richman, Old Mutual Insure and University of the Witwatersrand,

Salvatore Scognamiglio, Department of Management and Quantitative Sciences, University of Naples, Parthenope (presenter),

Mario V. Wüthrich, RiskLab, Department of Mathematics, ETH Zurich.

Abstract: Recently, accurate forecasting of mortality rates with deep learning models has been investigated in several papers in the actuarial literature [1,2]. Most of the models proposed to date are not explainable, making it difficult to communicate the basis on which mortality forecasts have been made.

We adapt the LocalGLMnet proposed in [3] to produce explainable forecasts of mortality rates using locally connected neural networks, and we show that these can be interpreted as autoregressive time-series models of mortality rates. These forecasts are shown to be highly accurate on the Human Mortality Database and the United States Mortality Database.

Finally, we show how regularizing the LocalGLMnet can produce improved forecasts, and that by applying auto-encoders, observations of mortality rates can be denoised to improve forecasts even further.

Keywords: Mortality forecasting, explainable deep learning, Lee–Carter model, LocalGLMnet

References

1. Richman, R., Wüthrich, M.V. (2021). A neural network extension of the Lee–Carter model to multiple populations *Annals of Actuarial Science* **15(2)**, 346-366.
2. Perla, F., Richman, R., Scognamiglio, S., Wüthrich, M.V. (2021).Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal* **2021(7)**, 572-598.
3. Richman, R., Wüthrich, M.V. (2023). LocalGLMnet: interpretable deep learning for tabular data. *Scandinavian Actuarial Journal* **2023(1)**, 71-95.

Contact details

- Email: salvatore.scognamiglio@uniparthenope.it

A Credibility Index Approach for Effective *a Posteriori* Ratemaking with Large Insurance Portfolios

Sebastian Calcetero-Vanegas, University of Toronto (presenter)

Andrei L. Badescu, University of Toronto

X. Sheldon Lin, University of Toronto

Abstract: A posteriori ratemaking in insurance involves determining premiums that take into account both policyholders' attributes and their claim history using a Bayesian model, known as a credibility model. Most data-driven models used for this task are mathematically intractable, and thus, credibility premiums must be obtained through numerical methods, such as simulation via MCMC. However, these methods can be computationally expensive and prohibitive for large portfolios when applied at the policyholder level. Additionally, these computations are often considered "black-box" procedures as there is no clear expression showing how the claim history of policyholders is used to upgrade their premiums.

In this talk, a methodology is proposed to derive a closed-form expression to compute credibility premiums for any given Bayesian model by introducing a credibility index that serves as an efficient summary statistic of a policyholder's claim history. By using this closed-form solution, the computational burden of a posteriori ratemaking for large portfolios can be reduced through the use of surrogate modeling, and it also provides a transparent and interpretable way of computing premiums.

Keywords: Credibility theory, Experience rating, Mixed models, Bayesian computation, Large Portfolios

References

1. Calcetero-Vanegas, S. F. , Badescu, A. L., & Lin, X. S. (2022). A Credibility Index Approach for Effective a Posteriori Ratemaking with Large Insurance Portfolios. arXiv preprint arXiv:2211.06568. Also available at SSRN.

Contact details

- Email: sebastian.calcetero@mail.utoronto.ca
- Homepage: https://www.sfcalceterov.com/
- Repository: https://actsci.utstat.utoronto.ca/
- Social media: https://www.linkedin.com/in/sebastiancalcetero/

A Practitioner Guide to Marginal Pricing - Pricing with Portfolio Impact in Mind

Shirley Ng, Vantage Risk

Abstract: When faced with multiple submissions, underwriters must make decisions that optimise their aggregate exposure, maximise their premium, and minimise impact from shock loss events and volatility. In an underwriting cycle with reduced capacity, knowing the aggregate impact at the underwriting stage is more critical than ever. Individual case pricing does not serve the purpose well. This is where pricing with portfolio impact comes in.

This concept is not alien to property catastrophe reinsurance underwriters. A "bang for bucks" report illustrates the impact of each risk contribution to the portfolio TVaR and Var by region peril. The same principle can be applied to non-property catastrophe risks.

Ideally, you want to link each case pricing output to an exposure and capital model to determine the optimal strategy. However, this is not always possible. The level of data granularity in submissions varies greatly.

Marginal pricing typically requires high compute power, however, there are techniques that an analyst can use to approximate the impact of each risk at the underwriting stage on their desktop computer.

By the end of this presentation, pricing analysts will learn how to provide additional insights to underwriters and, in turn, incentivise underwriters to refer more risks to analysts. Management need to recognise the importance of marginal pricing and enforce the data collection. Marginal pricing is only achievable with support from management.

Keywords: Collective Risk Model, Portfolio Management, Risk Management, Pricing, Reinsurance, Marine, Energy, Casualty, Cyber, Liability, Correlation

References

1. Casella, G. and Berger, R.L. (2002) *Statistical Inference. 2nd Edition* Duxbury Press, Pacific Grove
2. Kevin P. Murphy (2012) *Machine Learning - A Probabilistic Perspective* MIT Press
3. S Wang(1999) *Aggregation of correlated risk portfolios: models and algorithms* Proceedings of the Casualty Actuarial society 85 (163), 848-939

Contact details

- Email: shirley.ng21@imperial.ac.uk

Implementing ML Ops in insurance: a case study using a complex, multi-model Customer Lifetime Value system

Sindre Henriksen, Eika Forsikring, Hamar, Norway (presenter)

Øyvind Klåpbakken, Eika Forsikring, Hamar, Norway

Fredrik Wollert Hansen, Eika Forsikring, Hamar, Norway

Abstract: Machine learning (ML) systems offer significant and idiosyncratic challenges in their development, productionisation, and maintenance. This has led to the emergence of a set of technologies and practices clustered together under the umbrella term of "ML Ops" (Alla & Adari, 2021).

We chronicle our journey of implementing a complex, multi-model Customer Lifetime Value (CLV) system in a Norwegian insurance company and showcase the ML Ops required to support the resulting system complexity. A common challenge in this field is that the operational model that may work for 1-3 "proof-of-concept" machine learning systems often does not scale to a scenario where a company has 10+ ML systems in production. Indeed, our CLV model alone is composed of multiple constituent ML subsystems, necessitating the use of modern ML Ops practices.

We detail the development of an in-house feature store, the use of a cloud-based machine learning platform, and the deployment of a data and model monitoring framework which supports an overarching goal of *observability* of the machine learning process. We also show how the various components of the modern ML Ops stack facilitate the scaling of the entire machine learning life cycle. While a well-chosen technology stack is paramount to success, our focus is squarely on the emerging consensus around principles and practices required to operationalise machine learning (Shankar et al., 2022), and on how we have successfully applied these principles in the insurance industry.

Keywords: ML Ops, Customer Lifetime Value, Machine Learning

References

1. Alla, S., & Adari, S. K. (2021). What is MLOps?. In Beginning MLOps with MLFlow (pp. 79-124). Apress, Berkeley, CA.
2. Shankar, S., Garcia, R., Hellerstein, J.M., & Parameswaran, A.G. (2022). Operationalizing Machine Learning: An Interview Study. arXiv:2209.09125.

Contact details

- Email: sid@eika.no
- Social media: https://www.linkedin.com/in/drhenriksen

Algorithmic Insurance: A Conformal Prediction Framework

Sukrita Singh, Saïd Business School, University of Oxford (presenter)

Agni Orfanoudaki, Saïd Business School, University of Oxford

Abstract: Machine learning algorithms have grown in sophistication and importance over the years but their use in high-risk applications has been impeded by factors such as the absence of a vehicle to absorb losses from algorithmic errors.

Products that transfer this risk from algorithm users to insurance providers are being developed in a new field called "Algorithmic Insurance", encouraging the uptake of machine learning for practical purposes. While there has been significant past work in the domain highlighting the need for algorithmic insurance, and qualitative research into product design, there has been limited work outlining quantitative approaches to do so in practice.

Bertsimas and Orfanoudaki [1] presented the first quantitative framework to enable risk estimation of derived insurance contracts for the binary classification case. This paper builds on their work and aims to develop a generalizable framework that is applicable to both regression and classification learners, using techniques from the conformal prediction literature.

Conformal prediction is a black-box uncertainty quantification method that can be used to generate prediction intervals with statistical coverage guarantees.[2] We show how conformal prediction can quantify the performance risk associated with a predictive algorithm.

Specifically, we propose a pricing regime for regression problems using the conformal prediction error assessment technique [3] and for classification problems using conformal risk control [4]. We further extend this framework to discuss how pricing may be considered at a portfolio level that includes a set of machine learning models instead of individual learners.

Keywords: Algorithmic Insurance, Uncertainty quantification, Machine Learning, Conformal Prediction

References

1. Bertsimas, Dimitris, Agni Orfanoudaki. (2021). Algorithmic insurance.
2. Angelopoulos, Anastasios N., Stephen Bates. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification.
3. Prinster, Drew, Anqi Liu, Suchi Saria. (2022) Jaws: Predictive inference under covariate shift.
4. Angelopoulos, Anastasios N., Stephen Bates, Adam Fisch, Lihua Lei, Tal Schuster. (2022) Conformal Risk Control

Contact details

- Email: sukrita.singh@sbs.ox.ac.uk
- Social media: https://uk.linkedin.com/in/sukrita-singh-412631b4

Sentence similarity models to develop a new risk appetite tool

Davide De March (presenter), Thao Nguyen, Kevin O'Donovan (presenter), Markel International

Abstract: Natural Language Processing (NLP) is a branch of artificial intelligence that deals with the interaction between computers and human languages. It involves developing algorithms and models to process, understand, and generate human-like language.

NLP is widely used in the tech industry, but it is still a greenfield in the insurance sector; while not widely used yet, some interesting examples start to appear in the insurance market. Examples where NLP has been proven successful in the insurance sector are mainly related to improve customer service, automate workflows, and enhance risk analysis with technique such as chatbots, OCR and fraud detection. (Goossens et al. 2022, Ly et al. 2020).

Markel has developed an Azure portal for underwriters and brokers able to ingest free text with the description of customer's business activity and assesses the risk appetite of this new business using NLP models.

The solution uses the "allMiniLM-L6-v2" sentence-BERT (sBERT) pre-trained model from Hugging Face (Reimers and Gurevych 2019) which has been enhanced with internal data. A cosine similarity score is then applied to the encoded sentences to verify concordance between the incoming text and internal risk appetite.

Results of the similarity score will be presented in a live demo (by omitting the internal risk appetite comparison).

Keywords: NLP, Hugging Face, Sentence similarity

References

1. Goossens, A., Berth, L., Decoene, E., Van Veldhoven, Z., Vanthienen, J. (2022). *Automatically Extracting Insurance Contract Knowledge Using NLP*. In: Abramowicz, W., Auer, S., Str.
2. Ly, A., Uthayasooriyar, B., Wang, T. (2020) A survey on natural language processing (nlp) and applications in insurance, *https://doi.org/10.48550/arxiv.2010.00462*.
3. Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *https://doi.org/10.48550/arxiv.1908.10084*.

Contact details

- Email: davide.demarch@markel.com
- Linkedin: https://www.linkedin.com/in/davidedemarch/

Estimating return periods for extreme events - a frequenstist and Bayesian perspective

Tim Edwards, Howden Tiger

Abstract: How frequentist and Bayesian techniques may be used to assess the likelihood of historically experienced event and annual aggregate loss levels. Bayesian techniques may extend available loss statistics by using hazard indices to augment the information available.

Some specific case studies and questions to address:

- Was the 2021/22 drought event for a La Niña event unusual?
- How likely should we assume the 2021/22 drought is from an historical perspective comparing frequentist and Bayesian techniques?
- What caveats do we need to consider when comparing these results?

Keywords: Extreme events, catastrophe risks, frequentist, Bayesian

Contact details

- Email: Tim.Edwards@howdentiger.com

Explore latent factors of longevity trends with frailty-based stochastic models

Maria Carannante, University of Salerno

Valeria D'Amato, University of Salerno (presenter)

Steven Haberman, Bayes Business School, City University of London

Massimiliano Menzietti, University of Salerno

Abstract: Mortality improvement trends have been mainly studied by means of stochastic mortality models and affine models. Nevertheless, the analysis of the changes in the mortality trend as the underlying risk factors vary remains relatively unexplored.

Systematic mortality improvement trends vary with the risk characteristics of individuals in a population (including by age and gender) and this variation determines the degree of mortality heterogeneity within a population [4][3]. In the actuarial literature, frailty is represented by an unobservable risk factor which affects the mortality rate of an individual. The unobserved frailty factor encompasses all the factors affecting human mortality other than age. Conversely, in the medical literature, frailty is defined in terms of increased vulnerability (even to small stress factors) and decreased physiological reserve in older people [1].

In this research, we study the prevalence of frailty and explore the relationship between frailty and mortality, in order to obtain more accurate projections of the longevity trend. The idea behind the research consists in identifying the main latent factors explaining the frailty component, and detecting the covariates to determine its heterogeneity, by means of the Variable Importance of Machine Learning method. Variable importance allows selecting the better variables without assuming a functional form of the model.

We point out that frailty is mainly due to comorbidities that impact on the process of deterioration in terms of the human body's physiological capacity. In this setting, we provide a range of frailty-based stochastic models for modelling and projecting mortality rates based on the seminal Lee Carter model [2].

Keywords: Mortality trends, Frailty, Variable Importance, Lee-Carter family model

References

1. Clegg A, Young J, Iliffe S, Rikkert M.O., Rockwood K. (2013) Frailty in elderly people, Lancet 381(9868), 752–62
2. Lee, R.D., Carter, L.R. (1992) Modeling and forecasting u. s. mortality. Journal of the American Statistical Association 87(419), 659-71
3. Meyricke, R., Sherris M. (2013) The determinants of mortality heterogeneity and implications for pricing annuities, Insurance: Mathematics and Economics 53 (2), 379–87
4. Vaupel, J. W., K. G. Manton, and E. Stallard (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality, Demography 16(3), 439–54

Contact details

- Email: mcarannante@unisa.it

AI Risk: How much should we care?

Valerie du Preez FIA, Dupro Advisory (presenter)

Paul King PhD FIA, University of Leicster (presenter)

Abstract: We will explore key issues & challenges in AI risk management based on research from a transregional actuarial focus group. We will define the gaps and challenges faced when it comes to implementing and utilising modern modelling techniques from a risk-perspective, in the context of regulation. We will also discuss best practice guidelines and how to take a professional approach to AI risk management, based on key themes identified such as bias and explainability. With examples, we will identify the regulatory, professional and industry resources available that could support the professional adoption of data science and AI techniques. We explore where gaps might still exist.

We will explore:

- How to manage key risks associated with AI including minimising bias and improving explainability and transparency
- Industry examples of modelling issues in traditional and non-traditional actuarial work
- Best practice professional guidelines for data science and AI risk management
- How to adopt AI techniques in a professional, risk-controlled and ethical way

This discussion is aimed at participants who are looking to improve their AI and data science risk management and governance frameworks. No prior technical knowledge on the topic is required.

Keywords: Risk management, AI best practice, regulation, bias, explainability

References

1. Institute and Faculty of Actuaries and Royal Statistics Society. (2019). A Guide for Ethical Data Science. Available at: https://www.actuaries.org.uk/system/files/field/document/An%20Ethical%20Charter%20for%20Date%20Science%20WEB%20FINAL.PDF
2. Institute and Faculty of Actuaries. (2021). Ethical and professional guidance on Data Science: A Guide for Members. Available at: https://www.actuaries.org.uk/system/files/field/document/IFoA_Ethical_Professional_Guidance_Data_Science_Feb_2021.pdf
3. Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M.V. (2022). Discrimination-free insurance pricing. *ASTIN Bulletin: The Journal of the IAA*, 52(1), 55-89. doi:10.1017/asb.2021.23

Contact details

- Email: paul.king@leicester.ac.uk

Fully automated ETL Process Using Azure

William Mesquita (presenter), TROVADORES D´EQUAÇÕES LDA

Jonathan Sales, Universidade Federal do Ceará – UFC

Anna Riepin, Markel International

Abstract: In today's data-driven world, the need for efficient and automated data integration is more critical than ever. The Extract, Transform, and Load (ETL) process is a fundamental part of data integration that involves extracting data from multiple sources, cleaning, transforming it into a unified format, and loading it into a target database, data warehouse or lakehouse[1].

The presentation will highlight the benefit for the insurance sector of the extensive use of the Azure technologies and show the importance of a reliable ETL process for data science teams. We will discuss the automation of ETL processes using Azure, focusing on scalability, fault-tolerance, and robustness. We will showcase how Azure Data Factory[2] and Azure Databricks[3] has been leveraged to create Markel's lakehouse ETL process, highlighting the benefits of their combined use. We will also present how the solution has been engineered to integrate data from various sources, including on-premise systems, cloud-based services, and third-party applications, and to deploy modeling in production.

Lastly, we would demonstrate how the Azure end-to-end data integration solution is designed to enable collaboration among data scientists on a cloud environment.

Keywords: ETL, Azure, Cloud environment, Data warehousing, Data lake, Lakehouse architecture, Data quality, MLops

References

1. Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In Proceedings of CIDR (p. 8).
2. Microsoft (2022) Azure Data Factory: Data Integration in the Cloud. Available at: https://azure.microsoft.com/mediahandler/files/resourcefiles/azure-data-factory-data-integration-in-the-cloud/Azure_Data_Factory_Data_Integration_in_the_Cloud.pdf
3. Databricks (2023) Azure Databricks. Available at: https://pages.databricks.com/AzureDatabricks-DE.html.

Contact details

- Email: william.m.mesquita@gmail.com
- Social media: linkedin.com/in/williammesquita

Non-crossing neural network quantile regression estimation for driving data with telematics

Xenxo Vidal-Llana, Universitat de Barcelona (presenter)

Montserrat Guillen, Universitat de Barcelona

Abstract: The state-of-the-art methodologies to estimate Value at Risk (VaR) and Conditional Tail Expectation (CTE) controlled by covariates are mainly based on quantile regression and do not consider explicit constraints to guarantee that non-crossing conditions across VaRs and their associated CTEs always hold. We implement a non-crossing neural network that:

a) estimates VaRs and CTE simultaneously,
b) is conditional on covariates, and
c) preserves the natural quantile level order.

We implement a Non-Crossing Dual Neural Network, a deep learning model capable of handling driving data using a telematics dataset from 2015 for quantile levels 0.9, 0.925, 0.95, 0.975 and 0.99. Improvements compared to quantile regression using lineal optimization and CTE estimation of one quantile level at a time are discussed. We also conclude that our method improves a Monotone Composite Quantile Regression Neural Network approximation and that it can be implemented in many areas of risk analysis.

Keywords: risk evaluation, telematics, quantile regression, motor insurance, value at risk, conditional tail expectation

Contact details

- Email: juanjose.vidal@ub.edu

Machine Learning and XAI for underwriting

Yafei (Patricia) Wang, Lloyd's of London (presenter)

Abstract: Underwriting is a risk assessment process that classifies insurance applications into different categories. Traditional underwriting is costly, time-consuming and perceived as a barrier for the underserved population. Despite many attempts in automating life and health insurance underwriting, the dominant approach is a mix between prescriptive rule-based engines and manual underwriting. Only a third of all applications are processed by these prescriptive rule-based engines; the remaining complicated applications are underwritten manually, so this process is lengthy and costly. Therefore, a more efficient improvement to the current approach is predictive machine learning models.

This research aims to construct predictive machine learning models to predict underwriting decisions for life and health insurance applications, using reinsurer data that are predominantly applications with complex medical conditions and large sum insured. The models are designed to provide an end-to-end solution, so machine learning techniques such as natural language processing and clustering analysis are used to process real-world data; in particular, free-text descriptions of impairments and occupations, which the traditional statistical models cannot process. Machine learning algorithms such as XGB, Random Forest and bagging are used to predict the underwriting decisions. Various feature selection methods recursive feature elimination and Boruta are used to improve prediction accuracies.

Ten machine learning algorithms are run, and their performances are compared using various performance metrics. The extreme gradient boost algorithm performs the best, with 99.5% accuracy on the training set and 81.0% accuracy on the held-out testing set. In particular, the model achieved 94.4% predication accuracy for the standard class in the held-out testing set. Explainable AI such as SHAP and LIME plots give insights on how various features contribute to the predictions, both at global, cohort and local level.

Keywords: underwriting, Machine Learning, XAI, NLP, XGBoost, LIME, SHAP

References

1. Chen, T. & Guestrin, C., 2016, XGBoost: A Scalable Tree Boosting System, KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016 Pages 785–794, https://doi.org/10.1145/2939672.2939785
2. Ribeiro, M. T., Singh, S., Guestrin, C., 2016, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016 Pages 1135-1144, https://arxiv.org/abs/1602.04938
3. Lundberg, S. M.& Lee, S., 2017, A Unified Approach to Interpreting Model Predictions, NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, December 2017 Pages 4768–4777, https://doi.org/10.48550/arXiv.1705.07874

Contact details

- Email: patriciawangyafei@gmail.com
- LinkedIn: https://www.linkedin.com/in/patricia-wang-1446894a/

Imbalanced learning for insurance using modified loss functions in tree-based models

Changyue Hu, University of Illinois at Urbana-Champaign

Zhiyu Quan, University of Illinois at Urbana-Champaign (presenter)

Wing Fung Chong, Heriot-Watt University

Abstract: Tree-based models have gained momentum in insurance claim loss modeling; however, the point mass at zero and the heavy tail of insurance loss distribution pose the challenge to apply conventional methods directly to claim loss modeling. With a simple illustrative dataset, we first demonstrate how the traditional tree-based algorithm's splitting function fails to cope with a large proportion of data with zero responses. To address the imbalance issue presented in such loss modeling, this paper aims to modify the traditional splitting function of Classification and Regression Tree (CART). In particular, we propose two novel modified loss functions, namely, the weighted sum of squared error and the sum of squared Canberra error. These modified loss functions impose a significant penalty on grouping observations of non-zero response with those of zero response at the splitting procedure, and thus significantly enhance their separation. Finally, we examine and compare the predictive performance of such modified tree-based models to the traditional model on synthetic datasets that imitate insurance loss. The results show that such modification leads to substantially different tree structures and improved prediction performance.

Keywords: Predictive model of insurance claims, Imbalanced learning, Custom loss, Canberra distance, Regression tree, Tree-based algorithms

References

1. Smith, A., Wang, Z. (2015). The art of modeling article. *Journal of XY* **3/4**, 120-122.
2. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees.* Taylor & Francis Group, LLC, Boca Raton, FL.
3. Guelman, L., Guillén, M., Pérez-Marín, A.M. (2015). Uplift random forests. *Cybernetics and Systems* **46 (3–4)**, 230–248.
4. Henckaerts, R., Côté, M.-P., Antonio, K., Verbelen, R. (2021). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal* **25 (2)**, 255–285.
5. Wüthrich, M.V. (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal* **2018 (6)**, 465–480.

Contact details

- Email: zquan@illinois.edu
- Homepage: https://www.linkedin.com/in/zhiyufrankquan/

Index of presenters