Insurance Data Science Conference

Abstracts

RiskLab, ETH Zurich

June 14, 2019, Rooms HG E7 & HG F7

Keynote Talks:

Talk 1: 9:15–10:00

Talk 2: 16:00–16:45

Room HG E7

Robust Algorithmics: A Foundation for Science?!

Joachim M. Buhmann

ETH Zurich, Department of Computer Science
jbuhmann@inf.ethz.ch

Abstract.
The ALGORITHM is the idiom of modern science, as Bernard Chazelle phrazed it. I like to go a step further in this talk by claiming that algorithmics lays the foundation of modern science. The scientific method of "systematic observation, measurements, and experiments, as well as the formulation, testing, and modification of hypotheses" requires algorithms for knowledge discovery in complex experimental situations. Algorithms in data science map data spaces to hypothesis classes. Beside running time and memory consumption, such algorithms should be characterized by their sensitivity to the signal in the input and their robustness to input fluctuations. The achievable precision of an algorithm, i.e., the attainable resolution in output space, is determined by its capability to extract predictive information. I will advocate an information theoretic framework for algorithm analysis where an algorithm is characterized as a computational evolution of a posterior distribution on the output space.
The method allows us to investigate complex data analysis pipelines as they occur in computational neuroscience and neurology as well as in molecular biology. I will demonstrate this design concept for algorithm validation with a statistical analysis of diffusion tensor imaging data. A theoretical result for sparse minimum bisection yields statistical hints why random combinatorial optimization problems are hard to solve.

AI in Actuarial Science

Ronald Richman

AIG
ronald.richman@aig.com

Abstract.
Deep learning techniques have begun to appear in the actuarial literature, and have produced promising results in several areas of actuarial practice, from pricing and reserving for non-life insurance to forecasting mortality rates and analyzing telematics data. Building on a recent introductory paper [1], this talk has as an aim to review the state of the art in applying deep learning techniques within actuarial science and, in doing so, highlight the potential that these techniques hold for solving actuarial problems. Furthermore, as the application of these techniques begins to mature, new challenges are coming to light that we believe should be considered by the actuarial community. Therefore, there are three goals for this talk – first, to provide a high-level overview of deep learning and show how neural networks are a generalization of the regression models that actuaries currently apply, secondly, to provide a snapshot of recent successes of applying deep learning to actuarial problems and thirdly, to examine the challenges and emerging solutions to the problems actuaries might face when applying deep learning techniques.

References.

[1] Richman, R. (2018). AI in Actuarial Science. *SSRN Manuscript* ID 3218082. Version July 24, 2018.

Session 1:
Non-Life Insurance Pricing

Time: 10:15–11:00

Room HG E7

Modelling Multiple Guarantees on a Household Level in Motor Insurance using Multivariate Credibility

Florian Pechon

UCLouvain
florian.pechon@uclouvain.be

Abstract.
Actuarial risk classification studies are typically confined to univariate, policy-based analyses: individual claim frequencies are modelled for a single guarantee in isolation, without accounting for the interactions between the different coverages bought by the same policyholder. Moreover, independence between the policyholders is generally assumed. However, some events may trigger multiple guarantees of an insurance product at the same time. Moreover, some latent but important risk factors may be shared across multiple guarantees and across a household, inducing dependence between guarantees and policyholders from the same household.
In this talk, we present some ideas using multivariate credibility theory on how the dependence between the different guarantees can be accounted for, while also taking into account the dependence between policyholder from the same household. The analysis is performed on a motor insurance portfolio, using the two most common guarantees: Third-Party Liability insurance and Material Damage insurance.
The results show that the dependence between the claim frequencies of the two considered guarantees is strong, even for different policyholders from the same household. The model in turn allows obtaining better predictions of the claim frequencies and possible cross-selling opportunities can be identified.

References.
[1] Pechon, F., Denuit, M., Trufin, J. (2018). Multivariate modelling of multiple guarantees in motor insurance of a household. *Working paper.*
[2] Pechon, F., Trufin, J., Denuit, M. (2018). Multivariate modelling of household claim frequencies in motor third-party liability insurance. *ASTIN Bulletin* **48/3**, 969-993.

Optimizing Insurance Pricing Models in Collison

Yves Staudt

University of Lausanne
yves.staudt@unil.ch

Abstract.
Adequate pricing of car insurance contracts is an important challenge for many non-life insurance companies. When actuaries rely on models using covariates to explain the claims loss exposure, they often model the severity and the frequency separately. The aim of this paper is to compare the performance of machine learning methods to the traditional regression model approaches in severity modeling by including the residence location of the policyholder. In a first step, building on generalized linear models and following [1], we follow a data-driven procedure to build a regression model including the most significant factors. Thereby, we consider generalized additive models, since continuous covariates are often related to the response in a nonlinear way. In a second step, we compare the results of these models to those of random forests. In our applications we rely on a data set to cover the loss exposure of a collision insurance portfolio of a Swiss insurer from 2011 to 2015. The data contains an exposure of around 500 thousand policyholder-years including observations from about 81 000 settled claims. The data from 2011 to 2014 are used for training the data and the year 2015 will serve as a test dataset. The validation set will be created with the help of the cross-validation.

References.
[1] Henckaerts, R., Antonio, K., Clijsters, M., Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal* **2018/8**, 1-25.
[2] Staudt, Y., Wagner, J. (2018). Comparison of machine learning and traditional severity-frequency regression models for car insurance pricing. *Working paper, University of Lausanne.*
[3] Klein, N., Denuit, M., Lang, S., Kneib, T. (2014). Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape. *Insurance: Mathematics and Economics* **55**, 225-249.
[4] Denuit, M., Lang, S. (2004). Non-life rate-making with Bayesian GAMs. *Insurance: Mathematics and Economics* **35/3**, 627-647.

Generative Adversarial Networks (GANs) for Claim Frequency Prediction Models

Armando Vieira[1] **and Javier Rodriguez Zaurin**[2]

[1]Data Scientist @Hazy. armandosairmais@gmail.com
[2]Head of Data Science @Simply Business. jrzaurin@gmaul.com

Abstract.
Risk assessment and underwriting is the core of the insurance industry. Gaining advantage in a competitive market by offering tailored products to the customers relies on adequate risk models. To build such models a substantial amount of data and a sizeable number of claims records is required. However, when expanding to new markets or relatively unexplored territories such level of information is not always available. In order to build risk models, Insurance companies have to buy data from some local provider or rely on extrapolations from better known markets. In this presentation we will show how Generative Adversarial Networks (GANs)[1] can alleviate this problem by generating good quality data from very small samples, even from different domains and without alignment. GANs have been overwhelmingly used in computer vision[2], natural language processing[3] or e-commerce[4] with remarkable results, but their application to the insurance industry remains still unexplored.
Through the presentation, we will briefly explain what GANs are and how they can be used to synthesize data in order to enhance very infrequent events and create better claim frequency prediction models.

References.
[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems* **2014**, 2672–2680.
[2] Karras, T., Laine, S., Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *arxiv* 1812.04948
[3] Nie, W., Narodytska, N., Patel, A. (2019). RelGAN: relational generative adversarial networks for text generation. *International Conference on Learning Representations.*
[4] Kumar, A., Biswas, A., Sanya, S. (2018). eCommerceGAN : a generative adversarial network for E-commerce. *arxiv* 1801.03244

Session 2:
Workflows and Satellite Data

Time: 10:15–11:00

Room HG F7

Principles and Best Practices for Applied Machine Learning Models [in Industry]

Jürg Schelldorfer

Swiss Re
Juerg_Schelldorfer@swissre.com

Abstract.
In insurance, there is an established best practice, governance and regulation around the actuarial models used in pricing, reserving and solvency. On the other hand, data science is still maturing as a field, since interest increased rapidly in recent years. Many organizations are beginning to consolidate their activities around data analytics, machine learning, model building and productive deployment, and are in the process of professionalizing these activities. Still other organizations are taking the first steps on their journeys into data science, and would like to learn from those who have gone before.
At Swiss Re, we brought together the collective expertise and experience of numerous expert practitioners and managers from data science and risk management to create a definitive set of principles and best practices that guides all our data science activities.

TensorFlow Probability - Why We Should Care

Roland A. Schmid

Mirai Solutions
roland.schmid@mirai-solutions.com

Abstract.
TensorFlow Probability is an open source Python library built using TensorFlow, which makes it easy to combine deep learning with probabilistic models on modern hardware [3]. Efficiency and numerical stability are key goals of TensorFlow Probability, which aims to offer R-like richness in capabilities for statisticians and data scientists, provides an API similar to SciPy (like TensorFlow's API has been modelled after NumPy) and supports automatic differentiation like for instance the probabilistic programming language Stan does [1]. In contrast to the previous, TensorFlow Probability inherits the TensorFlow ecosystem benefits such as distributed computing, hardware acceleration (GPUs/TPUs besides CPUs), batching and vectorization, device-specific kernel optimizations (also for random number generation), graph operations and TensorBoard visualization. TensorFlow Probability also features several innovations, e.g. with bijectors or higher-order distributions. In this talk we give an introduction and lay out why we think that this might be the one new technology / library you need to hear about this year. We explain how to use TensorFlow Probability following an example from *Bayesian Methods for Hackers* [2] and finish with a use-case that shows how we use TensorFlow Probability for "real-time" ad-hoc assessment of tail events in parameterized risk scenarios.

References.
[1] Dillon, J., Langmore, I., Tran, D., et al. (2017). TensorFlow Distributions. *Available at* `arXiv:1711.10604`.
[2] Davidson-Pilon, C., McAteer, M., Seybold, B., et al. (2018-2019). Probabilistic Programming and Bayesian Methods for Hackers: Chapter1, Introduction, TensorFlow Probability. *Jupyter notebook on CamDavidsonPilon's GitHub account.*
[3] Dillon, J., Bhimji, W. (2018). Frontiers of TensorFlow: Space, statistics, and probabilistic ML. Proceedings of *O'Reilly Artificial Intelligence Conference.*

How to Use Satellite Data with Machine Learning in Insurance

Damian Rodziewicz

Appsilon Data Science
damian@appsilon.com

Abstract.
The talk is about new possibilities arising from analyzing satellite imagery in the insurance industry. Satellite data changes the game, as it allows to travel in time and reach information not available to business. Combined with the advances in image recognition and computing power, satellite data analysis offers Insurers possibilities to automate or streamline processes and design better products.
Satellite data is huge and non-obvious. Thanks to currently available technologies you can access it, build forecasts and observe events that were undetectable before. We will show you what possibilities can be offered with the use of deep learning on satellite images and how our data science department has been successfully working with satellite data to build decision support systems for business.
Last but not least, we'll explore three emerging use cases: (1) validating claims in agriculture, (2) quantifying the spread of disaster events and (3) asset valuation.

Lightning Talks 1

Time: 11:30–12:30

Room HG E7

Hybrid Tree-Based Models for Insurance Claims

Zhiyu Quan

University of Connecticut
zhiyu.quan@uconn.edu

Abstract.
Modeling loss costs for short-term insurance contracts has conventionally been based on claim frequency and claim severity. While it is not uncommon to use a two-part framework with frequency and severity as components, there has been an interest in the use of Tweedie Generalized Linear Model (GLM) as a direct approach. For most insurance claims datasets, there is typically a large proportion of zero claims that leads to imbalances that cause inferior prediction accuracy of these traditional approaches. As an alternative approach, we propose to use tree-based models with a hybrid structure that involves a two-step algorithm. The first step is the construction of a classification tree to build the probability model for frequency. In the second step, we employ elastic net regression model at each terminal node from the classification tree to build the distribution model for severity. This hybrid structure captures the benefits of tuning hyperparameters at each step of the algorithm thereby allowing for an improved prediction accuracy. We examined the performance of this model vis-à-vis the Tweedie GLM using the LGPIF and simulated datasets. Our empirical results indicate that this hybrid tree-based model produces more accurate predictions without loss of intuitive interpretation.

References.
[1] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees.* Taylor Francis Group, LLC.
[2] Frees, E.W., Derrig, R.A., Meyers, G. (2014). *Predictive Modeling Applications in Actuarial Science.* Cambridge University Press.
[3] Hothorn, T., Hornik, K., Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* **15/3**, 651-674.
[4] Xacur, O.A.Q., Garrido, J. (2015). Generalised linear models for aggregate claims: to Tweedie or not? *European Actuarial Journal* **5/1**, 181-202.
[5] Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67/2**, 301-320.

An R Package for Simulating Stochastic Individual-Based Population Models

Daphné Giorgi

Sorbonne Université
daphne.giorgi@sorbonne-universite.fr

Abstract.
Recent advances in probability have contributed to the development of a general framework for the stochastic modelling of heterogeneous and age-structured population dynamics. These so-called Individual-Based Models (IBMs) have applications in several fields, including mathematical biology, ecology, demography, or actuarial sciences. For instance, IBMs can be used for simulating the evolution of an insurance portfolio, assessing the basis and demographic risk, or validating mortality forecasts by studying consistency between subnational and national mortality forecasts.
The simulation of stochastic IBMs is often limited by high computational costs and time. This package implements efficient algorithms based on thinning methods, in which computational costs and runtimes are significantly reduced. Algorithms are implemented in C++ for efficiency. The user-friendly R interface allow users to define the parameters of the model (mortality rates, birth rates, ...) straightforwardly, and is consistent with the package StMoMo for stochastic mortality modelling. In addition, the package provides good graphical representations and analysis of outputs, such as age-pyramids or life tables obtained from the simulated data.

Joint work with S. Kaakai (Le Mans Université) and V. Lemaire (Sorbonne Université).

Distribution Fitting in Insurance with the Shiny App DistrFit

Kornelius Rohmeyer, Ralph Scherer

University of Oldenburg
kornelius.rohmeyer@uni-oldenburg.de

Abstract.
Distribution fitting from loss data is an important part of insurance pricing, risk management and solvancy calculations. We present an R package with accompanying Shiny app that supports the estimation of a large set of loss and count distributions for individual claims and their frequency. Emphasis is put on tailored fitting and ranking of modeled distributions according to user given criteria and weights. Incorporation of expert judgement is supported and robust methodology available by utilizing the distrMod package.
DistrFit allows for basic data preparation steps before the fitting process as well as fitting based simulation procedures afterwards to estimate VaR, TVaR and further parameters. In addition, we will show how the DistrFit app could be integrated into a greater framework of further data preparation and simulation tools.

References.
Kohl, M., Ruckdeschel, P. (2010). R Package distrMod: S4 Classes and Methods for Probability Models. *Journal of Statistical Software*, **35/10**, 1-27.
Delignette-Muller, M.L., Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, **64/4**, 1-34.
Chang, W., Cheng, J., Allaire J., Xie Y. McPherson, J. (2018). shiny: Web Application Framework for R. *R package version 1.2.0*, https://CRAN.R-project.org/package=shiny.

Impact of NLP Techniques on Risk Management Practices in Reinsurance: a Business Case Approach

Aurélien Couloumy, Loris Chiapparo

Reacfin - Institut de Science Financière et d'Assurances (ISFA)
aurelien.couloumy@reacfin.com

Abstract.
Of all the great applications of Data Science in (re)insurance, the understanding and the use of textual information is probably one of the most promising in the short term. In risk management more particularly, there are plenty of data available such as legal documents, regulatory texts, technical notes, or claims reports that can be used to save time, to improve risk assessments or just to enrich market understanding.
In this presentation, we propose to dive deep in this subject by setting up a business case that aims at using Natural Language Processing (NLP) techniques and reinsurance wordings to improve pricing and risk understanding.
In a first part, we will propose a quick recap to introduce the business case and the data. Then we will come back to the main techniques developed including word embedding approaches and the use of recurrent neural networks. Finally, we will demonstrate with a Python demo how and why such an approach can improve the tasks of actuaries and risk managers.

Exploring Patterns of Convergence and Divergence among Insurers and Banks: An Analysis Using Text Similarly Measures

Lei Fang

Cass Business School, City, University of London
Lei.Fang@cass.city.ac.uk

Abstract.
Building on a database with the full history of the 'Risk Factors' section of the SEC's 10-K reports filed by US banks and insurance companies from 2006 to 2018 (banks = 770; insurance = 159), we perform a systematic analysis of convergence between banks and insurers, using topic modelling (Bao and Datta 2014) and cosine similarity methods (Mihalcea et al 2006). We define 'convergence' as banks and insurers using increasingly similar language in the text of their Risk Factor returns.

Our preliminary results show an emerging trend: text-based measures of convergence between banks and insurers show a consistent pattern of increase over time, irrespective of the similarity measure used. Our results further show that this trend is consistent for all sub-categories of insurers against banks – e.g. Property and Casualty, Life, Reinsurance etc. Yet, our analysis also shows that when focusing on the insurers' sub-categories, some types of insurers appear to be diverging over time: for instance, preliminary results show that the Risk Factor reports of Property & Casualty insurers and of Reinsurers are becoming less similar. The results described here are in line with the management and the insurance literatures, which highlight that Industry boundaries are becoming more blurred (Kim et al 2015) and risk factors more common (Turk 2015). This is a joint work with Gianvito Lanzolla and Andreas Tsanakas.

References.

[1] Bao, Y., Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science* **60/6**, 1371-1391.

[2] Kim, N., Kim, W., Lee, H., Lee, H., Suh, J.H. (2015). Dynamic patterns of industry convergence: evidence from a large amount of unstructured data. *Research Policy* **44/9**, 1734-1748.

[3] Mihalcea, R., Corley, C., Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *AAAI* **6/6**, 775-780.

[4] Turk, M.C. (2015). The convergence of insurance with banking and securities industries, and the limits of regulatory arbitrage in finance. *Columbia Business Law Review* **2015/3**, 967-1073.

Improving Verification of Car Insurance Claims by Processing Photos with Machine Learning

Marek Rogala

Appsilon Data Science
marek@appsilon.com

Abstract.
Deep learning challenge accuracy of existing algorithms in a variety of fields. In insurance, we can use it to automate and improve damage classification, risk prediction, market analytics and many other tasks.
In this talk we'll focus on car insurance claims. Processing such claims requires gathering information like licence number, vehicle identification number (VIN), mileage and other dashboard indicators. They have to be written down and verified. Doing this manually is not cost-effective. We'll show how this process can be streamlined by introducing Machine Learning models for recognizing and verifying numbers present in client-submitted photos. We'll not only look at the models for numbers detection and recognition using deep learning networks with LSTM and Convolutional Networks. We'll also look the broader business perspective, and propose a pipeline for processing these image data from the moment they arrive from client to providing final decision.

How Context Influences Model Selection

Benjamin Abt

PartnerRe
benjamin.abt@partnerre.com

Abstract.
When selecting a statistical model, emphasis is often on the theoretical properties of model candidates. In this talk, we will to show how the ultimate usage of the model, business constraints, and software can influence the choice.
Two teams at PartnerRe independently modelled the same subject matter: the life cycle of disability claims (i.e., the developments from the moment a disability is incurred until the insured returns to work). Both teams operated in their own unique context – they had different goals, tools and markets. For example, one team aimed at estimating total cost of a claim for pricing, the other wanted to support claim managers' prioritization. One team ended up fitting survival regressions, the other used binary classification models at various constant horizons.
We present a comparison of the two approaches on identical data and feature sets. Observed properties of the two approaches are linked back to the context in which the models were originally developed.

Towards Open Collaboration in Insurance Innovation

Kevin Kuo

RStudio
kevinykuo@gmail.com

Abstract.
In the past year, we saw big data and distributed computing maturing and artificial intelligence coming into the mainstream. These technologies were merely buzzwords not too long ago, but have disrupted a wide range of industries, including insurance. One factor that is often credited for the rapid advances in these technologies is the open and collaborative nature of the computer science and machine learning communities. With a few exceptions, the insurance industry hasn't always embraced open source software and sharing research, but the trend has been shifting recently as actuaries refine and align their roles in an environment driven by data science. We discuss benefits, challenges, and ramifications of conducting actuarial research in the open, and introduce Kasa AI [1], a recent initiative for community-driven research and software development for insurance. We demonstrate ongoing projects in life and nonlife insurance, and walk through workflows for participating.

References.
[1] `https://kasa.ai`

Lightning Talks 2

Time: 11:30–12:30

Room HG F7

Automobile Insurance Fraud Detection using the Evidential Reasoning Approach and Data-Driven Inferential Modelling

Xi Liu

The University of Manchester
xi.liu-2@manchester.ac.uk

Abstract.
Automobile insurance fraud detection has become critically important for reducing the costs of insurance companies. Majority of insurance companies adopts expert knowledge to detect fraud. Experience-based knowledge are interpretable and re-usable but such knowledge is simply used in practice, which leads to some degrees of misjudgement. This paper aims to establish a unique Evidential Reasoning (ER) rule to combine multiple pieces of independent evidence from both experience based indicators and probabilities of fraud obtained from historical data. Each piece of evidence is weighted and then combined conjunctively with the weights optimised by using a maximum likelihood evidential reasoning (MAKER) framework for data-driven inferential modelling. Based on a real-world insurance claim dataset, our experimental results reveal that the proposed approach preserves the interpretability and usability of expert detection system, and anticipates the changes in fraud practices by tracking the trend of the weights of experience-based indicators. Furthermore, the experimental results show that the proposed approach outperforms a number of widely used machine learning models, such as random forests and support vector machine.

References.
[1] Yang, J.B., Xu, D.L. (2013). Evidential reasoning rule for evidence combination. *Artificial Intelligence* **205**, 1-29.
[2] Yang, J.B., Xu, D.L. (2017). Inferential modelling and decision making with data. In *Automation and Computing (ICAC), 2017 23rd International Conference* (pp. 1-6). IEEE.

Fraud Detection in Insurance with Social Network Analytics

María Óskarsdóttir

KU Leuven
maria.oskarsdottir@kuleuven.be

Abstract.
Fraud is encountered in various domains in the form of tax evasion, money laundering and credit card fraud. Fraud also occurs in the insurance industry, for example, when policyholders file claims that are exaggerated or based on intentional damage. Such fraudulent cases are often the result of organized schemes carried out by committed groups of collaborating fraudsters that go to great lengths in order to hide their tracks while maximizing their gain. We strive to detect such groups of fraudsters by linking together claims and the involved parties in a massive social network. As such, we are able to look beyond the classical properties of the claim, the policyholder and the policy, and study the social structures of collaborating fraudsters in insurance fraud detection tools and models. A first attempt of using social network analysis for fraud detection in insurance was made by Šubelj et al. (2011). However, their approach is not time-dependent which is unrealistic as fraudsters adapt their methods. In our study, we leverage the entire network of claims and policyholders of an insurance company and apply a fraud propagation algorithm, based on personalized PageRank (Page et al., 1999). Thereby, we obtain scores that quantify the claims' exposure to fraud, relative to known fraudulent claims. These scores are then combined with the intrinsic features of the claim, policyholder and policy in an analytical fraud detection model to flag highly suspicions claims that need to be investigated further. This approach, developed by Van Vlasselaer et al. (2016) to detect social security fraud, deals with the time evolving nature of fraud, by weighting relationships in time, as well as the high class imbalance which is typical in fraud datasets. We discuss how it can be used to detect fraudulent insurance claims and illustrate it on a real life dataset.

References.
[1] Šubelj, L., Furlan, Š., Bajec, M. (2011). An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications* **38/1**, 1039-1052.
[2] Page, L., Brin, S., Motwani, R., Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web.* Stanford InfoLab.
[3] 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.

How to Build a Data Driven Insurance Business

Nicolas Attalides

Mango Solutions
nattalides@mango-solutions.com

Abstract.
Much has been made in recent years of business' increasing reliance on data to remain competitive, but for the insurance sector, data analysis is in the DNA. With survival based purely on accurate, risk-based pricing, the sector has been 100% dependent on data analysis for years. However, with competition intensifying in the industry, organisations wanting to remain ahead of the curve need to make a step change in the way they use data, understanding its value as a strategic asset that can inform better decision-making. In this presentation, we will explain how to proactively use data and advanced analytics to build a data driven insurance business. Drawing on a recent case study with a global insurance company, he will cover four areas that must be addressed to transform the business model:

1. Identifying the decisions that have the biggest business impact, in order to create an analytics framework;
2. Ensuring that the internal and external data being collected and stored is good quality to support those decisions;
3. Provision of expertise to identify ways to improve these and run analytics regularly;
4. Creating the right technology infrastructure to support data and analytics.

Solvency Contagion Modeling

Tobias Baltensperger

Mirai Solutions
tobias.baltensperger@mirai-solutions.com

Abstract.
The quantification of systemic risk and solvency contagion in financial networks has gained increasing attention since the financial crisis. Many researchers developed extensions to the basic network valuation framework developed by Eisenberg and Noe [1], including costs of default [2], obligations with different seniorities [3], and credit valuation adjustment [4]. Scientific studies are mostly focused on systems of interconnected banks [1-5], which have limited information about each other's balance sheets. In contrast, Mirai Solutions, in close cooperation with a globally-active Zurich-based insurer, adopted the framework to the company's network of internal (legally separated) entities. One of the main challenges of this task was to include special types of financial instruments into the model without violating the fundamental mathematical properties of the framework. In this talk, I will outline the basic solvency contagion framework, and extend it by internal default guarantees, a key tool for multinational companies to protect entities with a high value for the group from defaulting. Moreover, I will touch on the technical aspects of using a Monte Carlo approach with the proposed framework, based on an R / C++ implementation using RcppArmadillo and RcppParallel.

References.
[1] Eisenberg, L., Noe, T.H. (2001). Systemic risk in financial systems. *Management Science* **47/2**, 236-249.
[2] Rogers, L.C.G., Veraart, L.A.M. (2013). Failure and rescue in an interbank network. *Management Science*, **59/4**, 882-898.
[3] Fischer, T. (2014). No-arbitrage pricing under systemic risk: accounting for cross-ownership. *Mathematical Finance* **24/1**, 97-124.
[4] Barucca, P., Bardoscia, M., et al. (2016). Network valuation in financial systems. *Available at* `arXiv:1606.05164v2`.
[5] Bardoscia, M., Barucca, P., Codd, A.B., Hill, J. (2017). The decline of solvency contagion risk. *Bank of England Staff Working Paper No. 662.*

Scenario Weights for Importance Measurement - An R Package for Sensitivity Analysis

Silvana M. Pesenti

Cass Business School, City, University of London
Silvana.Pesenti@cass.city.ac.uk

Abstract.
When modelling portfolios of risks, it is of central importance to analyse the propagation of changes in model assumptions. As is typical in applications, we view a model as a random vector of input factors that is mapped, via an aggregation function, to a random output. Performing sensitivity testing includes stressing the inputs and observing the impact on the output, as well as stressing the output and monitoring the impact on different inputs (*reverse sensitivity testing*). We propose an approach to sensitivity analysis, based on [1], that circumvents the need for additional simulation runs and thus does not require time-consuming re-evaluations of the aggregation function. The approach is implemented via the **R**-package SWIM.

Specifically, we define a *stress* on a random variable as a probabilistic modification, resulting from an increase or decrease in e.g. moments or risk measures such as VaR and ES. The distribution of the stressed random variable is chosen such that, subject to the constraints, the Kullback-Leibler divergence is minimised. In a Monte Carlo setting, the **R**-package calculates the importance weights of the resulting change of probability measure. Thus, using the weighting of simulated scenarios, the entire probabilistic characterisation of the stressed model is provided. Calculation of the stressed model including usual common sensitivity metrics and plotting facilities are implemented in the **R**-package.

This is joint work with Alberto Bettini, Pietro Millossovich and Andreas Tsanakas.

References.
[1] Pesenti, S.M., Millossovich, M., Tsanakas, A. (2018). Reverse sensitivity testing: What does it take to break the model? *European Journal of Operational Research* **274/2**, 654–670.

Accuracy and Robustness of Machine Learning Methods to Estimate the SCR

Salvatore Scognamiglio

University of Naples "Parthenope"
salvatore.scognamiglio@uniparthenope.it

Abstract.
Solvency II, the European Directive [2] for insurance and reinsurance companies, induced revolutionary changes in the logic of control and management of risks. It introduces new fundamental measures, such as the *Probability Distribution Forecast* (PDF) and the *Solvency Capital Requirement* (SCR), whose estimation involves a *market-consistent evaluation* of assets and liabilities. An expression in closed form of all components of assets and liabilities is unfeasible, due the complexity of their payoff. Numerical techniques such as Monte Carlo (MC) simulations are therefore used.
In Solvency II evaluation framework, insurance undertakings must estimate the *Net Asset Value* (NAV), defined as the difference between the total amount of assets and liabilities, on a one-year time horizon. The NAV at one-year depends on the future Economic and Actuarial Scenario, and an estimation of its probability distribution – the PDF – requires *nested Monte Carlo* simulation. Nested MC method yields accurate results (details about the trade-off between the inner and outer simulations can be found in [1]) but it can present unacceptable computational costs. One possible solution is the well known Least Square Monte Carlo (LSMC) method based on orthogonal polynomials [3]. We propose an alternative approach based on *Machine Learning* techniques. In particular, the performance of *Deep Learning Network* (DLN) and *Support Vector Regression (SVR)*, integrated in the Disar system – an asset-liability computational system for monitoring life insurance policies [1] – is investigated in terms of accuracy and robustness in the evaluation of PDF and SCR. The results are compared against the traditional LSMC methodology.

References.
[1] Casarano, G., Castellani,G., Passalacqua, L., Perla, F., Zanetti, P. (2017). Relevant applications of Monte Carlo simulation in Solvency II. *Soft Computing* **21**, 1187-1192.
[2] Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II), Official Journal of the European Union, L335/1, 17.12.2009.
[3] Longstaff, F.A., Schwartz, E.S.(2001). Valuing american options by simulation: a simple leastsquares approach. *The Review of Financial Studies* **14/1**, 113-147.

Terrorism as a Virus: Epidemiological Techniques for Outbreak Detection

Mick Cooney

Agrippa Data Consulting
mcooney@agrippadataconsulting.com

Abstract.
Terrorism-related risks is a specialty risk around for over forty years, covering multiple hazards such as kidnap, loss of life, property damage and business interruption.
In this talk we discuss the use of the Global Terrorism Database, maintained by the University of Maryland in the US, to analyse patterns of behaviour in global terrorism.
In particular we discuss how global patterns in terrorist violence behave like a disease. Positive feedback loops cause upwards spirals of violence and law-enforcement interventions act like vaccine programs and quarantines.
We discuss the use of epidemiological, time series and anomaly detection techniques to analyse these signals in the data, and show how multiple approaches usually result in robust analyses.

References.
[1] The Global Terrorism Database. https://www.start.umd.edu/gtd/
[2] Brennan, P. (2016). A machine learning approach to the analysis of terrorism. MSc Thesis.

Market Segmentation in Life Insurance: A Case Study

Kshitij Srivastava

Milliman
kshitij.srivastava@milliman.com

Abstract.
Most life insurance companies have historically faced a key problem in the inability to distinguish policyholders who are likely to behave quite differently from one another. This has led to overall inefficiencies and challenges in the marketing and development of new products. Market segmentation has been used in many other industries already to help alleviate those inefficiencies. Segmentation can be used to understand policyholder behavior and simultaneously improve company profitability as well as provide better value to customers based on their unique needs. Kshitij Srivastava, a data scientist at Milliman's Life & Annuity Predictive Analytics team will discuss how the team uses machine learning models to better understand policyholder behavior and their drivers. Better understanding of policyholder behavior can help insurers design products that are more targeted and more able to suit specific customer needs. For example, a customer with immediate liquidity needs is more likely to place value on the ability to get money out now than they are on a product feature that may provide them with more money down the road. This talk will start with an overview of market segmentation methods (including clustering) followed by applications in life insurance. At the end, Kshitij will also discuss how he used RStudio's Shiny platform to convert this study into a data product.

References.
[1] Burns, E., Wang, D. (2017). Predictive analytics in policyholder behavior. SOA Malaysia presentation.
[2] Burns, E., Kullowatz, M. (2018). Milliman values 2018 GLWB industry utilization study. Milliman.

Session 3:
Loss Reserving

Time: 13:30–14:30

Room HG E7

From Chain Ladder to Individual Claims Reserving using Machine Learning Techniques

Alessandro Carrato, Michele Visintin

Allianz SE
alessandro.carrato@gmail.com

Abstract.
Recent years have seen the emergence of a lot of research on the application of machine learning techniques in actuarial science. In particular, there has been a noticeable amount of papers regarding machine learning applied to P&C Loss Reserving. Overall, there is a common understanding that machine learning techniques provide better prediction accuracy of the outstanding liabilities compared to traditional methods. Nevertheless, the greater accuracy is offset by a difficult interpretations of results. This makes them in line of principle not suitable in an increasingly regulated world, as it is the insurance business. Our objective is to show how we can introduce elements of machine learning into the traditional actuarial reserving methods in a gradual way. We strive to achieve a balance between predictive power and interpretability by introducing step-by-step new machine learning elements, with the possibility to simply start from the legacy paid/incurred datasets underlying the loss claim triangles without introducing any cumbersome data requirement or significant IT investment.

References.
[1] Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin* **23**, 213-221.
[2] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning.* Springer, Second Edition.
[3] Halliwell, L.J. (2007). Chain-ladder bias: its reason and meaning. *Variance* **1/2**, 214-247.
[4] Wüthrich, M.V., Merz, M. (2008). *Stochastic Claims Reserving Methods in Insurance.* Wiley.
[5] Wüthrich, M.V. (2018). Neural networks applied to chain-ladder reserving. *European Actuarial Journal* **8/2**, 407-436.

Individual Reserving with Claim Specific Covariates

Jonas Crevecoeur

KU Leuven
jonas.crevecoeur@kuleuven.be

Abstract.
Holding sufficient capital is essential for an insurance company to ensure its solvability. Hence, predicting the amount of capital needed to fulfil future liabilities in an accurate way is an important actuarial task. Insurers record detailed information related to claims (e.g. cause of the claim) and policies (e.g. value of the insured risk) for pricing insurance contracts. However, this same information is largely neglected when estimating the reserve. We present a flexible framework for including these claim specific covariates. Our framework focuses on three building blocks in the development process: the time to settlement, the number of payments and the size of each payment. We present a well-chosen generalized linear model (GLM) for each of these stochastic building blocks. Standard model selection techniques for GLMs allows us to determine the appropriate covariates in these models. We demonstrate how these covariates determine the granularity of our reserving model. On the one extreme, including many covariates leads to large differences in the development process of individual claims. On the other extreme, including no covariates, corresponds to specifying a model for data aggregated in a triangle. The set of selected covariates then naturally determines the position the actuary should take in between those two extremes. Moreover, since similar generalized linear models are applied in the pricing of insurance contracts, our project bridges the gap between pricing and reserving methodology. We illustrate our method on a case study of a real life insurance dataset.

This research is a joint work with Katrien Antonio (KU Leuven, University of Amsterdam) and Roel Verbelen (KU Leuven).

Deep Representation Learning using Stacked Autoencoder in General Insurance Reserving

Phani Krishna Kandala

Swiss Re
phani_kandala@swissre.com

Abstract.
We propose a novel approach for loss reserving based on blended unsupervised and supervised deep neural networks which requires minimal feature engineering. This approach involves the following two steps:
1) Non-linear latent features or deep representation extraction using unsupervised stacked autoencoder.
2) Loss reserving is performed using the extracted latent features from step 1 as a starting point of multilayer perceptron.
We evaluates the value of using Stacked Denoising Autoencoder[2], a variant of Stacked Autoencoder,as a robust feature extraction method when noisy data exists. [1] has shown that by using the classical actuarial regression model as a starting point of the neural network calibration can minimise the loss (subject to overfitting) if the actuarial model misses important model structure. In our work, we used deep representation learning techniques for non-linear latent feature extractionand used it as a good initialization to multilayer perceptron for loss reserving. In [1], simultaneous learning and prediction using a common Neural Network structure across all Lines of Business (LoBs) was explored.In our work, we demonstrated, how generic features across multiple LoBs can be learnt in the first step with Stacked Autoencoders. Also we illustrated how loss reserve predictions can be done using a combination of generic features extracted from Step 1 and individual LoB specific learning in Step 2 resulting in reducing bias further. Similar work done in other industries such as prediction of future patients[3] and wind power prediction[4] was also considered prior to arriving at this novel approach.

References.
[1] Gabrielli, A., Richman, R., Wüthrich, M.V. (2018). Neural network embedding of the over-dispersed Poisson reserving model. *SSRN Preprint*.
[2] Pascal V., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 3371-3408.
[3] Miotto, R., Li L., Kidd, B.A., Dudley, J.T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*.
[4] Tasnim, S., Ashfaqur, R., Oo, A.M.T., Haque, M.E. (2017). Autoencoder for wind power prediction. In: *Renewables: Wind, Water, and Solar*, Springer.

Session 4:

NatCat, Risk Management and Cyber Risk

Time: 13:30–14:30

Room HG F7

Deep Learning of Remote Sensing Images to Predict Wildfires

Christian Klose

Swiss Re
christian_klose@swissre.com

Abstract.
Societal and economic risks of wildfires have become more concerning in recent years due to a changing climate. Wildfires remain one of the least predictable perils due to uncertainties reflecting the lack of knowledge about fire triggers, fire fuel availability, physical setting, and weather. Over the past decades, many novel data sources, models, and predictive techniques became available, including satellite imaginary, social media data, weather forecasting models, and artificial intelligence methods [1]. However, most of the prediction models focus on short term prediction of up to several weeks. In addition, these models tend to only rely on a few features, including temperature, precipitation [3,4] or drought conditions [5]. But to what extend can forecasts of wildfires be extended to several months or even years prior to future wildfire seasons? This study presents research results on how monthly remote sensing over several years (2001-2018) coupled with LSTM driven deep networks [2] can be utilized in the insurance industry to reduce uncertainties in forecasting wildfire occurrence and severity for portfolio steering and underwriting purposes.

References.
[1] Abatzoglou, J.T., Williams, A.P. (2016). Impact of anthropogenic climate change on wildfire across western US forests. *PNAS* **113/42**, 11770-11775.
[2] Hochreiter, S., Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation* **9/8**, 1735-1780.
[3] Littell, J. S., Oneil, E.E., McKenzie, D., Hicke, J.A. Lutz, J.A., Norheim, R.A., Elsner M.M. (2010). Forest ecosystems, disturbance, and climatic change in Washington State, USA. *Climatic Change* **102/1-2**, 129-158.
[4] Westerling, A.L. (2006). Warming and earlier spring increase western U.S. forest wildfire activity. *Science* **313/5789**, 940-943.
[5] Westerling, A.L., Gershunov, A., Brown, T.J., Cayan, D.R., Dettinger, M.D. (2003). Climate and Wildfire in the Western United States. *Dx.Doi.org* **84/5**, 595-604.

Neural Networks for Life Insurance Risk Management

Lucio Fernandez-Arjona

University of Zurich and Zurich Insurance Group
lucio.fernandez.arjona@business.uzh.ch

Abstract.
Insurance companies make extensive use of Monte Carlo simulations in their capital and solvency models. In life insurance, however, these models face the additional complexity of requiring two nested layers of Monte Carlo simulations. The inner layer is used for pricing the liabilities and the outer layer for calculating the capital requirements.
To overcome the computational challenge of such nested simulation approach, most large life insurance companies use a proxy model to avoid the inner simulation layer. These proxy models are mainly of two types, known as replicating portfolios and least-squares-monte-carlo (LSMC). They differ in the basis functions used for the approximation, the former using financial instruments and the latter using polynomials.
In this talk, we will present an example based on a variable annuity guarantee, showing the main challenges faced by practitioners in the construction of replicating portfolios: the feature engineering step and subsequent basis function selection problem.
We will then describe how neural networks can be used as a proxy model and how to apply risk-neutral pricing on a neural network to integrate such a model into a market risk framework. The proposed model naturally solves the feature engineering and feature selection problems of replicating portfolios. We apply the neural network model to the variable annuity example mentioned above, showing how risk and capital calculations compare for neural networks, replicating portfolios and polynomial approximations.
All datasets will be made public to encourage academics and practitioners in the field to replicate our results and compare them against additional algorithms.

References.
[1] Beutner, E., Pelsser, A., Schweizer, J. (2016). Theory and validation of replicating portfolios in insurance risk management SSRN Manuscript ID 2557368.
[2] Pelsser, A., Schweizer, J. (2016). The difference between LSMC and replicating portfolio in insurance liability modeling. *European Actuarial Journal* **6/2**, 441–494.
[3] Cambou, M., Filipović, D. (2018). Replicating portfolio approach to capital calculation. *Finance and Stochastics* **22/1**, 181–203.

Predicting Cyber-Attacks using the 'Hawkes' R Package

Alexandre Boumezoued
Caroline Hillairet

Milliman Paris & ENSAE-ParisTech, CREST
alexandre.boumezoued@milliman.com
caroline.hillairet@ensae.fr

Abstract.
Among the several features of cyber-attacks one wants to reproduce, those related to the memory of events and self-exciting behavior is of major importance, as it underlies the clustering and auto-correlation of times of cyber-attacks.
In this talk, we will describe the step by step specification, calibration and simulation using R of a multivariate Hawkes model for modelling and predicting cyber-attacks frequency, based on a public US dataset containing features of cyber-attacks targeting the Healthcare industry.
After a short description of the challenges related to the pricing of cyber-insurance products, we will first detail some statistical tests invalidating the classical Poisson modelling. Then, we will discuss the multivariate Hawkes model specification based on data segmentation and the interpretations associated with the simulated parameters, using the 'Hawkes' R package - particular focus will be given to the challenges related to calibration. We will finally illustrate prediction results providing the full distribution of future cyber-attacks times of occurrence.

Session 5:
Life and Mortality Modeling

Time: 14:30–15:30

Room HG E7

Medium Data and Socio-Economic Mortality

Andrew Cairns

Heriot-Watt University
a.j.g.cairns@hw.ac.uk

Abstract.
In this session we will use highly-detailed geographical population data from the UK's Office for National Statistics to assess what socio-economic and other factors are most strongly associated with high or low mortality rates. We will investigate what types of information about where you live affects mortality and life expectancy through consideration of the socio-economic mix and location of your neighbourhood.
Data:

- ONS population and deaths data at the level of small geographical areas (LSOA's)
- socio-economic covariates for each LSOA
- geographical covariates for each LSOA.

We will discuss first which combination of covariates have the strongest predictive power. Second we will investigate how much regional variation there is in our resulting models: is region or locality genuinely a significant factor in the level of mortality or is observed regional variation simply reflecting differences in the socio-economic makeup of local populations?
The use of non-parametric statistical methods will allow us to investigate how much variation there is across England in mortality rates and life expectancy. We can then use that to inform mortality assumption setting in pension scheme valuations.

A Data Analytics Paradigm for the Construction, Selection, and Evaluation of Mortality Models

Andrés M. Villegas

School of Risk and Actuarial Studies and ARC Centre of Excellence in Population Ageing Research (CEPAR), UNSW Sydney
a.villegas@unsw.edu.au

Abstract.
Humanity has made, and continues to make, significant progress in averting and delaying death, which burdens society with increased longevity costs. This has brought to the fore the critical importance of mortality forecasting for actuaries and demographers. Consequently, numerous mortality models have been proposed, with the most popular and commonly-referenced models belonging to a generalised age-period-cohort framework. These models decompose observed historical mortality rates across the dimensions of age, period, and cohort (or year-of-birth), which can then be extrapolated to forecast future outcomes. Recently, a large number of models have been proposed within this framework, many of which are over-parameterised and produce spurious forecasts, particularly over long horizons and for noisy data sets.
In this paper we exploit data analytics techniques to provide a comprehensive framework to construct, select, and evaluate discrete-time age-period-cohort mortality models. To devise this robust framework, we leverage two key statistical learning tools –cross validation and regularisation – to draw as much insight as possible from limited data sets. We first propose a cross validation framework for model selection, which can be tailored to determine the features of mortality models that are desired for different actuarial applications, including period and cohort-based forecasting. This enables the answering of questions regarding the effects of population size and structure, age, and forecasting basis and horizon on the preferred model selection. We also present a regularisation approach to construct bespoke mortality models by automatically selecting the most appropriate parametric forms to best describe and forecast particular data sets, using a trade-off between complexity and parsimony. We illustrate this using empirical data from the Human Mortality Database and simulated data sets.

Joint work with Dilan SriDaran, Michael Sherris and Jonathan Ziveyi.

Use of Administrative Health Databases for Modelling Longevity Improvement

Elena Kulinskaya, Ilyas Bakbergenuly and Lisanne Gitsels

University of East Anglia
e.kulinskaya@uea.ac.uk

Abstract.
In the past decade, the longevity improvement has slowed down in a number of developed countries including UK. This slowdown is not uniform, but mainly concerns the older, sicker and/or more deprived parts of the population. The effects of the major drivers of longevity need to be quantified to explain these changes and to predict future trends. It is well recognised [1] that the longevity improvement of the past 30 years is explained by declining cardiovascular disease, due in a large extent to wide implementation of preventive statin therapy. In our previous work [2], we quantified the effects of statin prescription at particular ages on population life expectancy. In the current study we implemented a more realistic longitudinal design and novel landmark survival analysis methods to be able to provide a more detailed answer. We have used primary care medical records from The Health Improvement Network (THIN), an UK primary care database, to select a cohort of $110,000$ healthy individuals residing in England or Wales who turned 60 between 1990 and 2000, and were neither diagnosed with cardiovascular disease nor prescribed statins. This cohort was followed up until January 2017, and their medical history was updated every 6 months. Landmark analyses were carried out by fitting Cox's proportional hazards models for all-cause mortality at each landmark age, adjusted for statin therapy, cardiac risk and other medical history. Statin therapy was measured as current prescription, cumulative proportion, and age at first prescription. Cardiac risk was based on QRISK2 score and diagnosis of cardiovascular disease. The initial results, after adjustment for cardiac risk and other medical history, show that statin therapy is associated with increasing survival benefits at older ages.
This research is supported by the Actuarial Research Centre.

References.
[1] Palin J. (2017) (on behalf of the Continuous Mortality Investigation – CMI) Mortality improvements in decline. *The Actuary*, August 2017.
[2] Gitsels, L.A., Kulinskaya, E., Wright, N. (2019). How medical advances and health interventions will shape future longevity. *British Actuarial Journal* **24**. DOI: 10.1017/S1357321719000059

Session 6:
Industry Case Studies

Time: 14:30–15:30

Room HG F7

Using Boosted Regression Trees in Insurance

Jakob L.K. Gerstenlauer

Allianz Lebensversicherung-AG
jakob.gerstenlauer@allianz.de

Abstract.
Boosted regression tree implementations are increasingly popular off-the-shelf machine learning algorithms. They are credited with high accuracy based on their versatility in estimating non-linear interactions among both numeric and categorical inputs [1]. They lack, however, global model coefficients, which complicates model interpretation and the estimation of effect sizes. As a result, making boosted regression trees interpretable, is currently an active field of research [2]. Another obstacle to the successful application of boosted regression tree algorithms are the high number of hyperparameters and the associated effort of tuning them.
Here, I describe practical experiences which we collected with the xgboost algorithm [3] throughout several data science projects. First, some practical tips for data preprocessing (e.g. one hot encoding using sparse matrices) will be given. Second, I will discuss efficient hyperparameter tuning using a combination of automated tuning based on the *mlr* framework and manual tuning based on some rules of thumb. Third, a practical approach for estimating global coefficients based on SHAP values [4] will be presented. Estimating effects is crucial, because it allows us to efficiently communicate the key findings of our data analysis to stakeholders. I will conclude with more general findings and recommendations, regarding the suitability of boosted regression tree algorithms for the analysis of insurance data.

References.
[1] Friedman, J., Hastie, T., Tibshirani, R. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition. *Springer*.
[2] Lundberg, S.M., Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765-4774.
[3] Chen, T., Guestrin, C. (2016). XGBoost: a scalable tree boosting system. *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 785-794.
[4] Lundberg, S.M., Erion, G.G., Lee, S.I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv:1802.03888.

Feature Synthesis Using t-SNE and Clustering

Mark Lee

Insight Risk Consulting
mark.lee@insightriskconsulting.co.uk

Abstract.
While modern supervised machine learning algorithms can often be used to improve the predictive power of models based on insurance data, their implementation and use in the industry can face barriers. These range from implementation issues due to legacy systems, to reticence from underwriters, regulators, and other stakeholders to trust "black-box" algorithms which cannot be easily and transparently explained or adjusted. Rather than use the most powerful supervised learning model, in some cases it can be advantageous to look instead at feature synthesis using unsupervised learning algorithms. Here we use a case study to show how a combination of t-distributed Stochastic Neighbour Embedding (t-SNE) [1] and clustering algorithms can be used to tailor features that fit into a more "traditional" multiplicative tabular model, are reasonably transparent and readily understood by stakeholders, and yet boost the predictive power compared to older models.

References.
[1] van der Maaten, L.J.P., Hinton, G.E. (2008). Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605.

Medical Underwriting Triage: An End-To-End Machine Learning Case Study

Markus Senn

PartnerRe
markus.senn@partnerre.com

Abstract.
PartnerRe rapidly grows its presence on the Canadian market for individual life insurance. This growth means that our local medical underwriting team has to handle 25 percent more cases every year. To manage these additional cases without increasing headcount at the same pace, the team has to operate more and more efficiently. In a market where response time is a key differentiator, we improve efficiency by using machine learning to triage and prioritize the underwriting team's case work.
In this talk, we describe the model pipeline end to end. That is, we cover the entire implementation chain from accessing data to the eventual evaluation of the added value. In particular, we discuss often neglected topics such as model risk mitigation, technical deployment of the model and integration into business processes.