Neural Networks and Random Forest in Insurance Risk Modelling

SDS2020, Online

26^{th} June 2020

Dr. Jürg Schelldorfer

Senior Data Scientist, Swiss Re

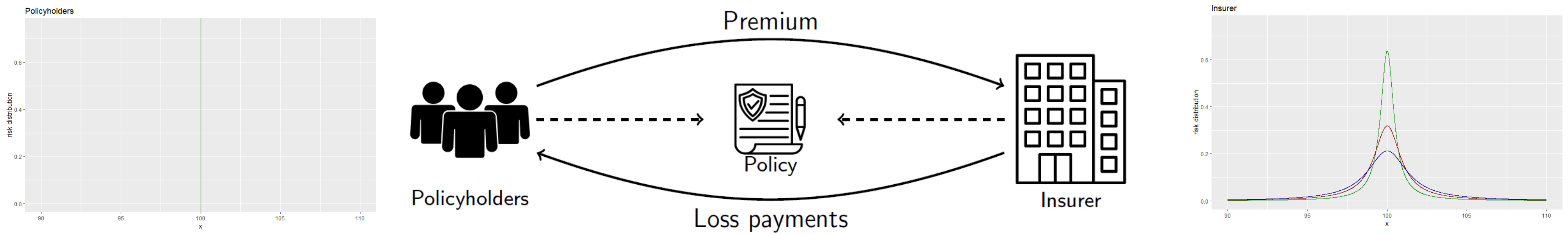
Chair of the «Data Science» working group, Swiss Association of Actuaries (SAA)

Disclaimer

The opinions expressed in this presentation are those of the author only. They are inspired by the work that the author is doing for both Swiss Re and the SAA, but they do not necessarily reflect any official view of either Swiss Re or the SAA.

ML and insurance risk modeling (1/2)

- What is **insurance risk modeling**?
  - Insurance risk modeling consists of the quantitative modeling of insurance risks, i.e. the risks that an insurance company takes through ist underwriting activities [a].
  - Insurance companies offer **protection against financial losses**. They allow individuals to **trade uncertainty for certainty, by transferring the risk to the insurer in exchange for a fixed premium**. An insurer sets the price for an insurance policy **before its actual cost is revealed** [b].
- What is **insurance pricing**?
  - Insurance pricing is one part of insurance risk modeling, determining the **fixed premium**.



[a],[b] Paper(s): [Financial report Swiss Re](), [Tree-based machine learning for insurance pricing]()

ML and insurance risk modeling (2/2)

- Generally speaking, why is insurance risk modeling more than solving a «standard» machine learning problem?

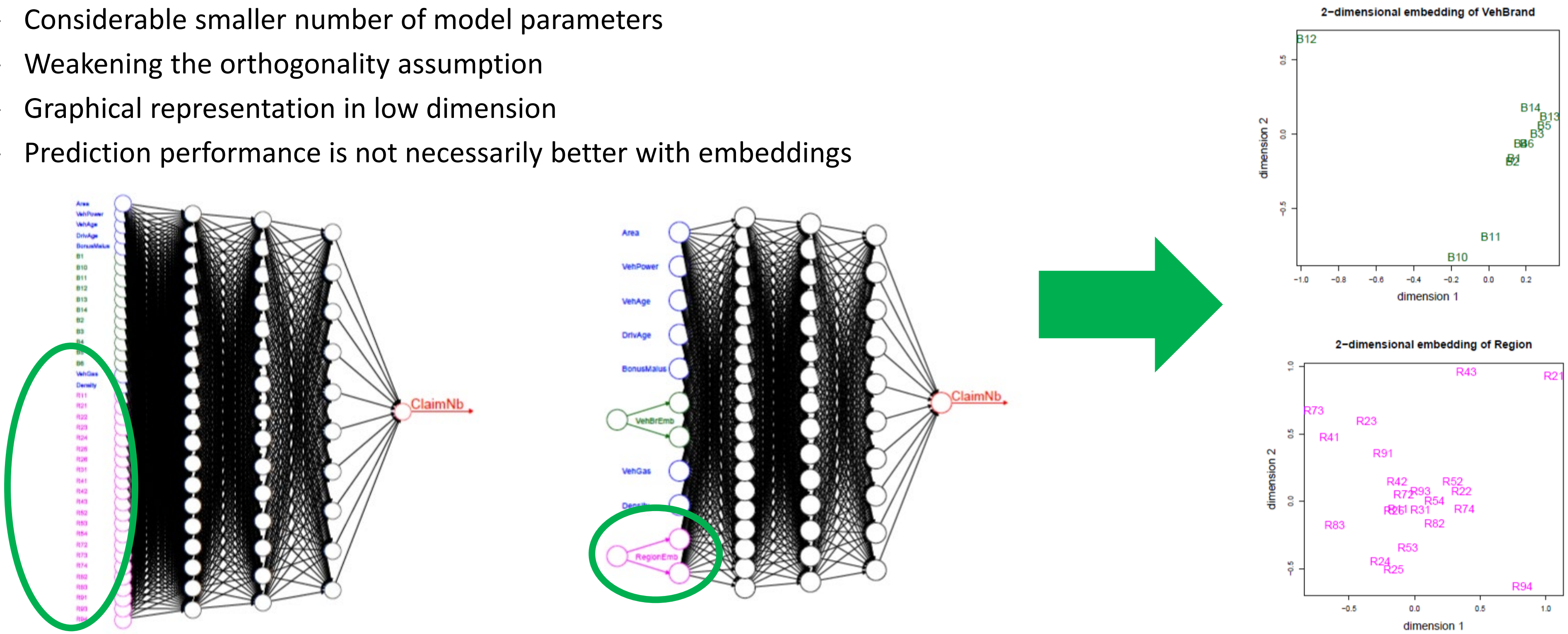
	Insurance Risk Modeling	«Standard» Machine Learning
Modelling target	Probabilistic Forecast*	Point Forecast
Statistical Distributions	Non-Gaussian (asymmetric, skewed)	Gaussian (symmetric)
Focus of distribution	Tail (Extreme Value Theory)	Bulk (Central Limit Theorem)
Model selection «criteria»	• Prediction performance • Stability over time • Smoothness • Interpretability / Explainability • Inclusion of expert knowledge • Uncertainty quantification • Regulatory framework • Political and social aspects	• Prediction performance • Computation performance
Size of data	From no to small to big data	Medium to big data
Aggregation of individual data	Weightening (e.g. case weights, exposure)	Equal weightening of observations
Domain knowledge	(Re-)insurance and insurance product knowledge	Subject-specific

* due to the capital requirements to absorbe negative financial losses.

Table of Content

1 – Embeddings for categorical features[1]

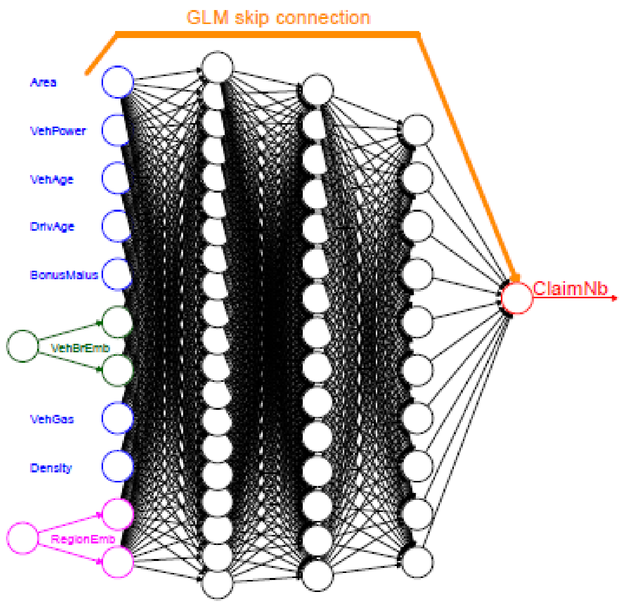
- In insurance pricing, there are often many categorical features (i.e. vehicle brand, region, age group,…) which consist of many levels.
- Usually, the categorical features are encoded as dummy variables (or one-hot encoding), i.e. the levels are orthogonal in the feature space.
- With neural networks, one should use (feature) embeddings:
  - Considerable smaller number of model parameters
  - Weakening the orthogonality assumption
  - Graphical representation in low dimension
  - Prediction performance is not necessarily better with embeddings



[1] Paper(s): https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3320525

2 – Combined Actuarial Neural Networks (CANN)[2]

- In most cases, there already exists a productive GLM.
- As the GLM has been developed and fine-tuned over years, there is a very good understanding of the pros and cons of the GLM.
- Instead of building a neural network from scratch, the actuaries are looking for an approach of using neural networks to improve the current GLM.
- The idea is to nest the GLM into a network architecture using a **skip connection** that directly links the input layer to the output layer.
- This approach is called CANN.



- We start the gradient descent algorithm for fitting the CANN model with the GLM solution.
- By that, the **algorithm explores the network architecture for additional model structure that is not present in the GLM**.
- Analyzing the results can hence be used to identify where the GLM needs to be improved, e.g. which interactions are missing.
- In this way we obtain an improvement of the GLM by network features. This provides a more systematic way of using network architectures to improve the GLM.
- CANN allow for **uncertainty quantification** due to its low computational effort.
- This approach is not restricted to GLM's, any regression algorithm can be chosen for the skip connection. However, this is not possible for all regression models.

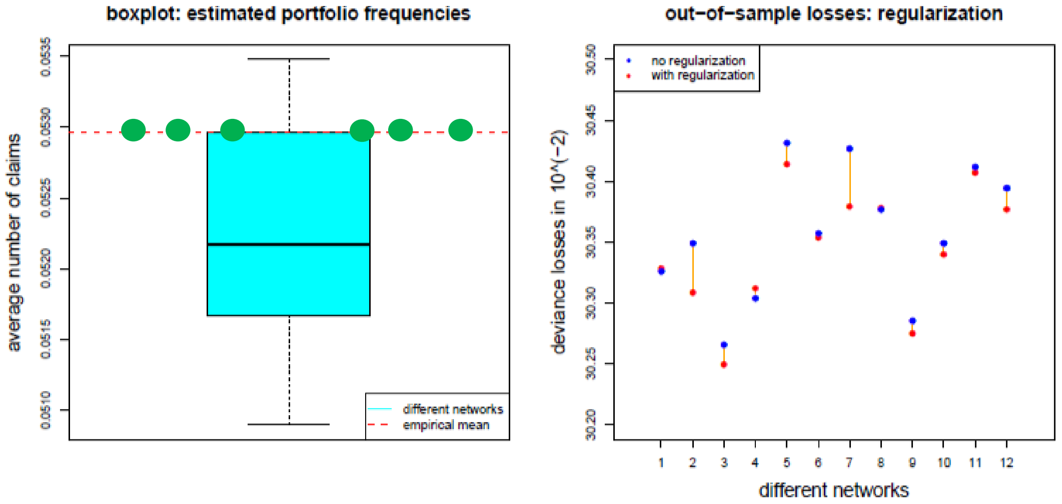
[2] Paper(s): https://doi.org/10.1017/asb.2018.42; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3320525

3 – Portfolio bias in neural networks[3]

Let us exmaine various pricing models:

- The same price μ_0 (empirical portfolio average) for all policyholders provides the empirical **portfolio premium ϕ_{GLM}**.
- Using a GLM, the price gets differentiated between the policyholders according to their risk characteristics (=features). Overall, the sum of all predicted prices μ_i gives again the empirical portfolio premium ϕ_{GLM}.
- Mathematically, GLM provide unbiased estimates on a portfolio level under the canonical link.
- Using a neural network for determining the individual prices μ_i provides price differentiation, but the neural network provides a portfolio average ϕ_{NN} which is different than ϕ_{GLM}.
- Is this an issue? YES, **network calibrations have a bias** and one needs to correct for these biases, the insurance company does not earn the price it needs to cover its liabilities!

- The neural networks show a bias (mainly negative) whereas the green dots illustrate the level of various GLMs (●).

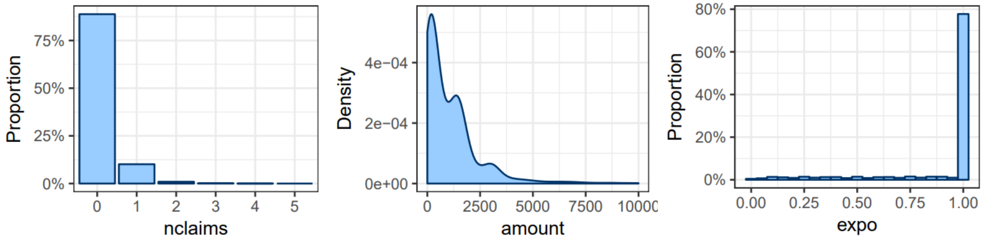


- Reason for the bias of neural networks: **early stopping criteria**.
- Solution: If we work in a GLM with canonical link function, this can simply be achieved by an additional MLE step using the **neuron activations in the last hidden layer as new covariates in the GLM**.

[3] Paper(s): https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3347177; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3226852

4 – Random forest for insurance data[4]

- Typical insurance data (MTPL from the CASdatasets R package):



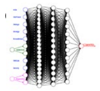
- The plots are characteristic for (non-life) insurance data: highly unbalanced count data with excess zeros (left) and varying exposure on the frequency side (right) combined with scarce, but potentially **long- or even heavy-tailed continuous data** on the severity side (middle).
- The default random forest implementation in R (e.g. randomForest, ranger) or Python (e.g. sklearn) are based on the standard squared error loss function.
- The squared error loss function is not necessarily a good choice when modeling integer-valued frequency data or right-skewed severity data.

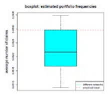
- The loss function used in the algorithm needs to be adjusted **such that the specific characteristics of insurance data are carefully considered**.
- Claim frequency modeling involves count data, typically assumed to be **Poisson** distributed. Therefore, an appropriate loss function is the Poisson deviance.
- **The exposure needs to be taken into account** in the expected number of claims. The Poisson deviance loss function can account for different policy durations.
- Right-skewed and long-tailed severity data is typically assumed to be **gamma** or **log-normally** distributed.
- What is the issue using gaussian-based random forest? → The tails of the distribution are not modelled accurately.
- Use of the **distRforest** R package.

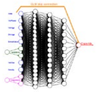
[4] Paper(s): https://arxiv.org/abs/1904.10890 ; https://github.com/henckr/distRforest

Conclusions

Neural networks and random forests may substantially improve classical insurance risk models, if appropriately applied.


Embeddings of categorical features reduce the neural network size and allow for visualisations of the catogorical feature levels in a low-dimensional space.


Neural networks need to be corrected for its bias to determine the correct technical price. The bias stems from the using early stopping criteria.


CANN provide the framework for extending the GLM’s, allowing to improve the accuracy of the model as well as providing a framework to **assess the uncertainties**.

Random forest (and also neural network) loss function needs to be aligned with the characteristcs of insurance data.

And yet, a complex and very well calibrated GLM may still be as good as an advanced machine learning model in terms of accuracy.

Appendix

References

- www.actuarialdatascience.org
- Nesting Classical Actuarial Models into Neural Networks, Schelldorfer J. and Wüthrich M.V. (2019), SSRN
- Editorial: Yes, we CANN!, Wüthrich, M.V., Merz, M. (2019). ASTIN Bulletin 49/1
- Bias Regularization in Neural Network Models for General Insurance Pricing, Wüthrich M.V. (2019), SSRN
- Insights from Inside Neural Networks, Ferrario A., Noll A., Wüthrich M.V. (2018), SSRN
- Boosting insights in insurance tariff plans with tree-based machine learning methods, Henckaerts R. et al. (2019), arXiv
- distRforest, Henckaerts R. et al. (2019), GitHub

Acknowledgements

People:

- All members of the SAA working group
- Roel Henckaerts
- Mara Nägelin
- Dr. Alexander Noll
- Dr. Simon Renzmann
- Ron Richman

Insititutions:

- Swiss Association of Actuaries (SAA)
- RiskLab at ETH Zurich
- MobiLab for Analytics at ETH Zurich

Companies:

- Swiss Re

Visit

www.actuarialdatascience.org

Articles, data and code on applications of machine learning to insurance risk modelling (aka actuarial data science)

Machine Learning in insurance risk modeling

The SAA working group «Data Science» has published the following tutorials in Actuarial Data Science (ADS):

1. French Motor Third-Party Liability Claims
2. Insights from Inside Neural Networks
3. Nesting Classical Actuarial Models into Neural Networks
4. On Boosting: Theory and Applications
5. Unsupervised Learning: What is a Sports Car?
6. Lee and Carter go Machine Learning: Recurrent Neural Networks
7. The Art of Natural Language Processing: Classical, Modern and Contemporary Approaches to Text Document Classification
8. Peeking into the Black Box: An Actuarial Case Study for Interpretable Machine Learning

Further tutorials will follow!