Machine Learning with High-Cardinality Categorical Features in Actuarial Applications

Insurance Data Science, 2023
Bayes Business School,
City, University of London, UK.

Presented by Bernard Wong[1]
Joint work with Benjamin Avanzi[2], Greg Taylor[1], Melantha Wang[1]

[1] School of Risk and Actuarial Studies, UNSW Business School, UNSW Sydney
[2] Centre for Actuarial Studies, Department of Economics, University of Melbourne

Outline of Talk

1. Context: High Cardinality Features in Actuarial Modelling
2. Proposed Approach
3. Case Study - Insurance Application
4. Conclusion

1. Context: High Cardinality Features in Actuarial Modelling

- Presence of multiple categorical features, some with **a large number of categories (e.g. 300+)**

Claim ID	Occupation (ANZSIC4)	Sum Insured	...	Total Incurred
1	Supermarket and Grocery Stores	736,673		2,919.61
2	Cafes and Restaurants	239,858		705.27
3	Fruit and Vegetable Retailing	174,661		108.88
4	Tiling and Carpeting Services	5,355,696		1,002.61
5	Other Specialised Machinery/Equipment Manufacturing	271,402		3,234.89
6	Clothing Retailing	1,157,769		634.61

Table: Example SME building insurance data
(numbers are randomised, for illustrative purposes only)

- ML models **cannot read categorical inputs** on their own
- Standard approach is one-hot encoding (e.g. Henckaerts et al. 2018), which **fails as cardinality grows**

Current Modelling Options

$$Y = f(X, Z) + \epsilon$$

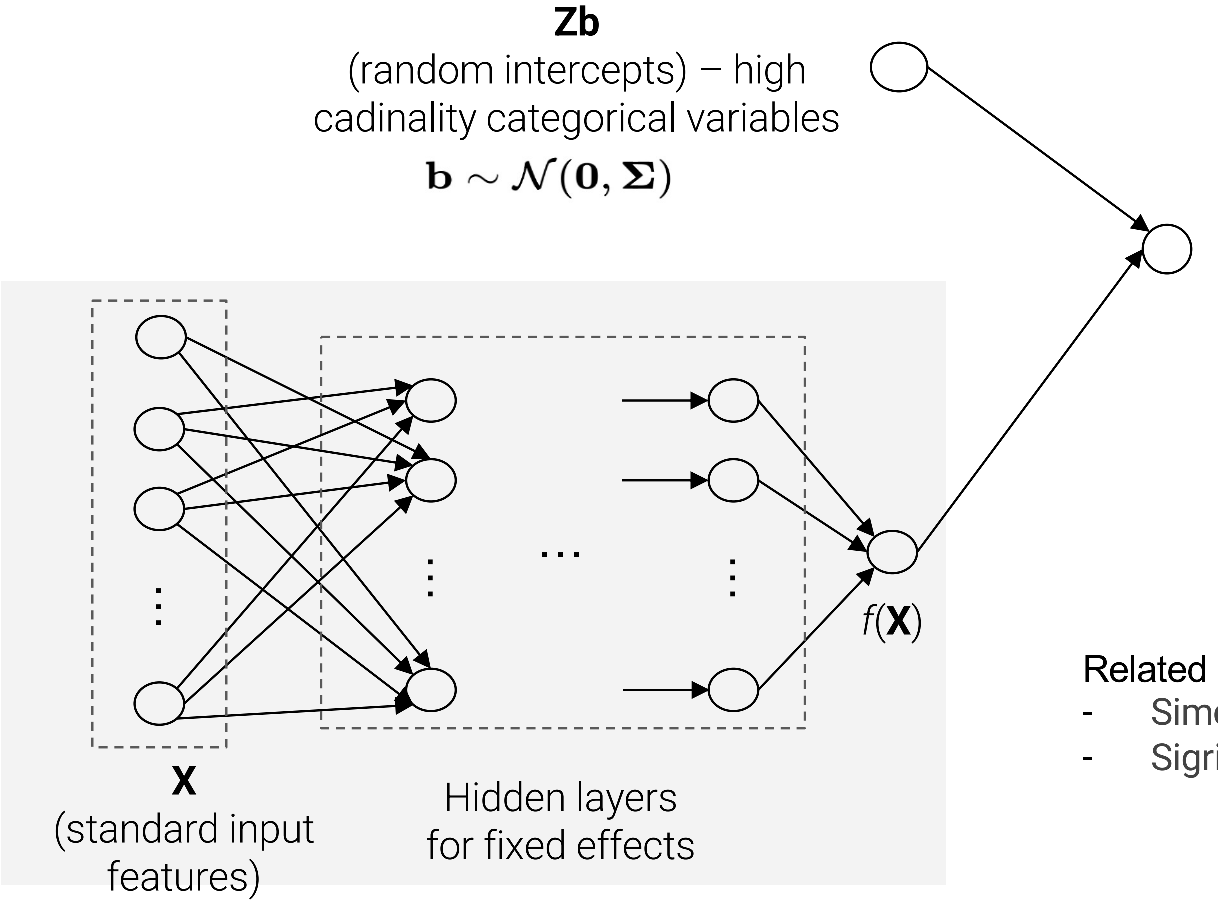
An illustrative example (State of New York, 2022):

Claim Identifier	Accident Date	Cause of Injury
4095286	08/10/2015	Lifting
4464102	12/14/2016	(Caught in) Object Handled
5193732	05/04/2019	Holding or Carrying
5444778	02/11/2020	Lifting
5809180	09/09/2021	Falling Or Flying Object

The Challenge: Too many categories in Z to learn the effect of each individual category well.

1. Make Z smaller in dimension (regrouping of "similar" categories)
2. Make Z look more like X before learning f (representation learning)
3. Pool the effects of categories in Z (generalised linear mixed models, or GLMMs)

2. Proposed Solution – GLMMNet



Output: $p(\mathbf{y} \mid \mathbf{b})$

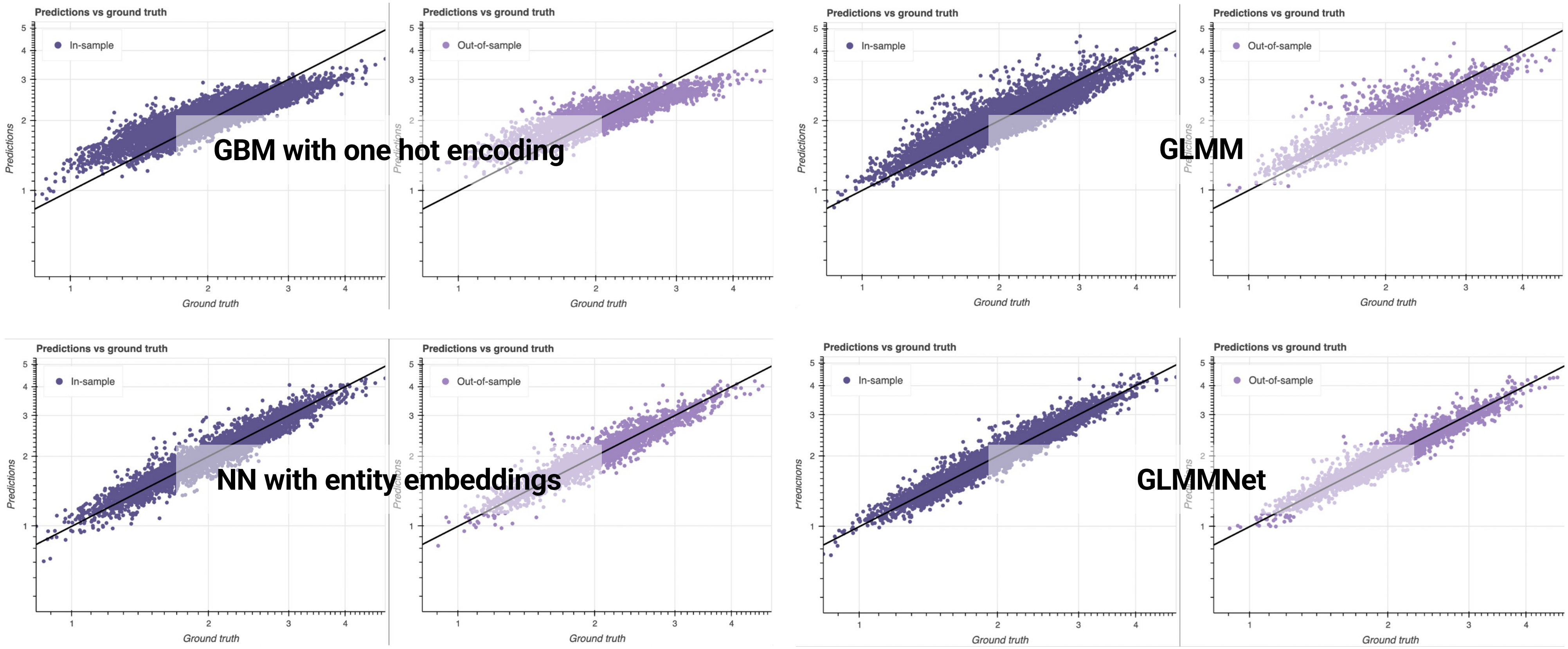
$$\mathbf{y}|\mathbf{b} \sim \mathcal{D}(\boldsymbol{\theta}) \text{ with } \boldsymbol{\mu} = \mathbb{E}(\mathbf{y}|\mathbf{b})$$

$$g(\boldsymbol{\mu}) = f(X) + Z\mathbf{b}$$

Related References:

- Simchoni & Rosset (2021, 2023), Mandel et al (2021)
- Sigrist (2021, 2022)

3. Simulation Example: Predictions vs Ground Truth (Left: IS, Right: OOS)



4. Insurance Case Study: SME Building Insurance Data

Columns

27

- **Response variable**: `total_incurred` (claim severity)
- **Features**: individual claim characteristics
  - `sum_insured`, `state_risk`, `roof_type`, `years_insured` ...
  - `ANZSIC4` occupation code (over 300 levels)

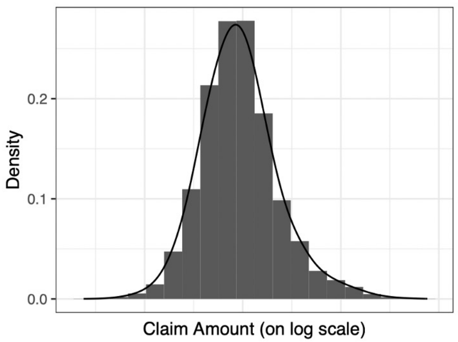
Rows

~27,000

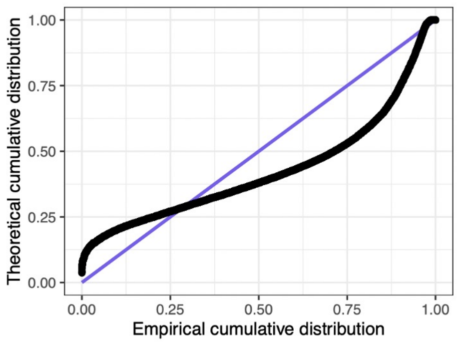
- SME building and contents insurance claims, over the period of 2010-2015
- ~1% of the rows were removed due to negative or <1 incurred amounts

Overview of Data

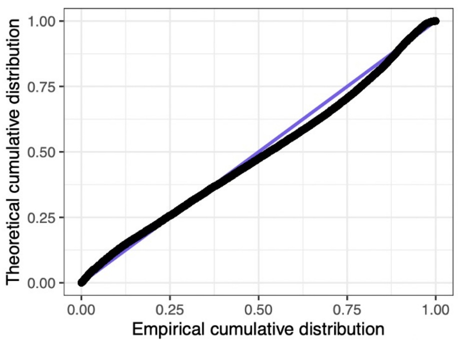
(axis removed due to commercial confidentiality)



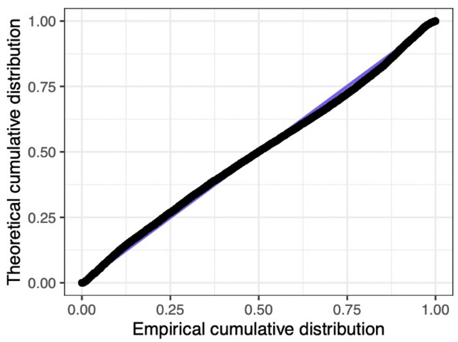
(a) Histogram of claim amounts



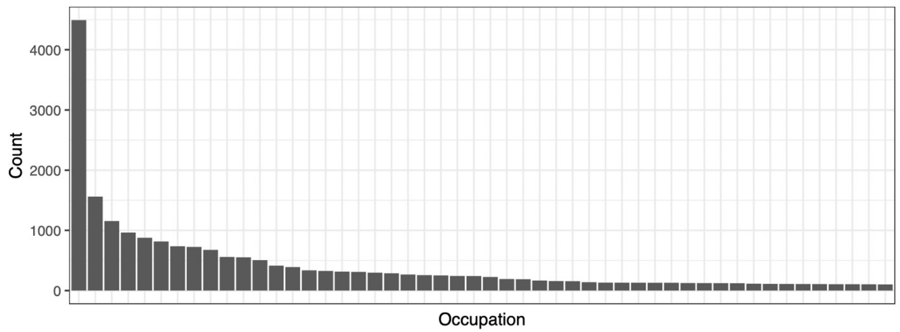
(b) P-P plot (Gamma)



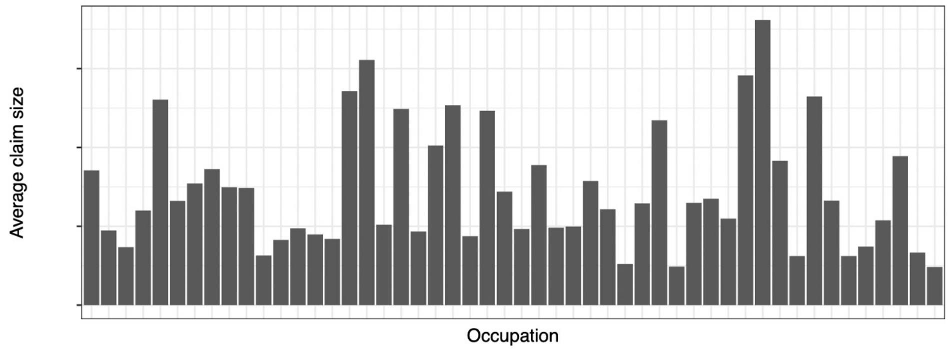
(c) P-P plot (lognormal)



(d) P-P plot (loggamma)



(a) Number of claims from the top 100 common occupation classes



(b) Average claim size for the (same) top 100 common occupation classes

-> skewed, high noise, unbalanced, variable average claim sizes.

Results – Model Comparison

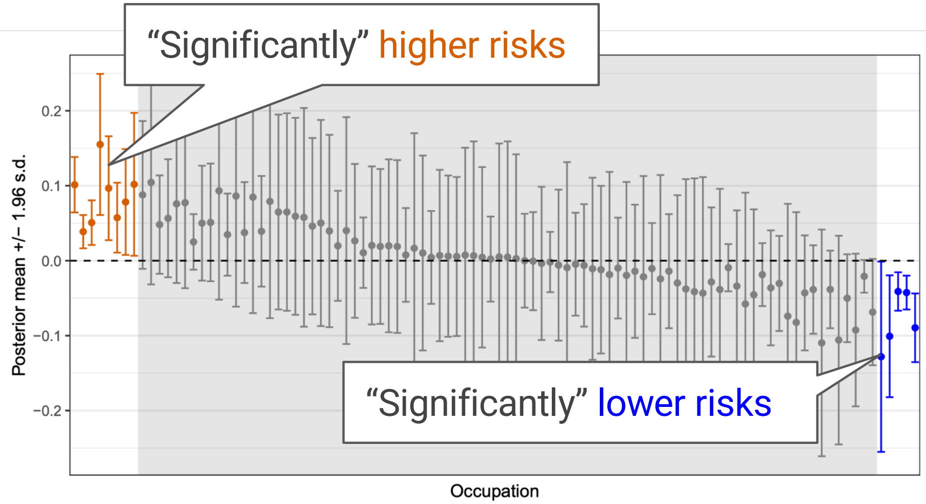
	Lognormal (out-of-sample)			Loggamma (out-of-sample)		
	MedAE	CRPS	NLL	MedAE	CRPS	NLL
GLM_one_hot	4108	0.7931	9.623	1946	0.8557	9.751
GBM_one_hot	3903	0.7682	9.586	1545	0.7643	9.580
NN_ee	4086	0.7666	9.584	1606	0.7612	9.578
GLMM	3864	0.7666	9.584	1570	0.7629	9.577
GLMMNet	3783	0.7751	9.595	1633	0.7662	9.583
GLMMNet_l2	3549	0.7634	9.580	1618	0.7626	9.577

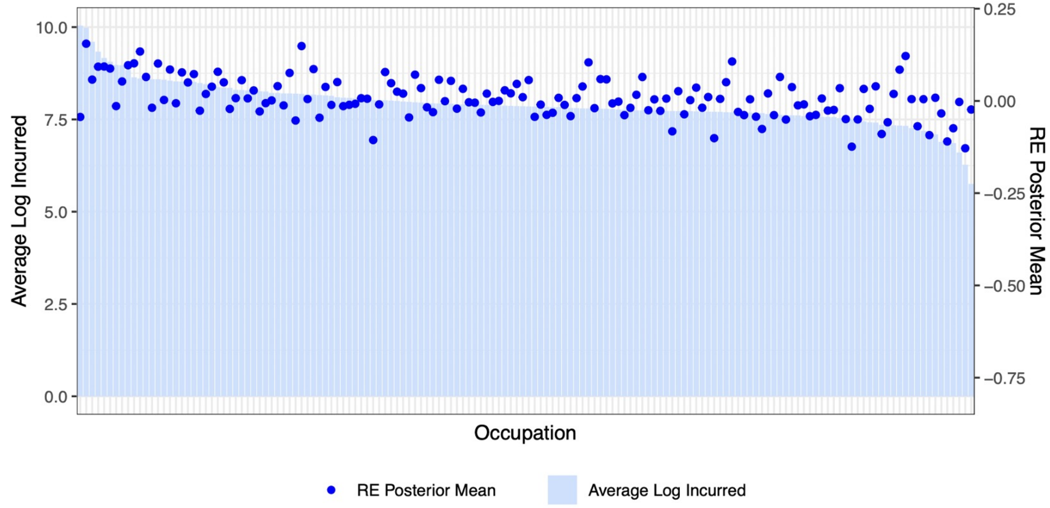
Comparison of lognormal and loggamma model performance on the out-of-sample set.

- In the family of lognormal models, **the regularised GLMMNet outperforms** all other models
- Among the loggamma models, the regularised GLMMNet comes as a close second to NN_ee
- Importantly, **regularisation is required to reduce overfitting** and helps the model generalise

Looking into the Model: Transparency of random effects

- Insights into how belonging to a certain category changes one's risk profile.





Left: Posterior predictions of the random effects in 95% confidence intervals
Right: Average log incurred by occupations, overlaid with RE predictions

4. Summary

The challenge: Too many categories in Z to learn the effect of each individual category well.

In this work, we:

1. Reviewed the existing approaches to insurance modelling with high-cardinality categorical features.
2. Developed GLMMNet, a flexible, implementable model that combines the statistical strength and transparency of mixed effects models and the predictive power of neural networks for insurance settings.
3. Compared the performance of the various modelling options using both simulated and real data.

Code is available on github: agi-lab/glmmnet. Current paper is available on arXiv:2301.12710*.

Appendix - A How-To Guide to Using GLMMNet in Practice

Ingredients

- A dataset with some high-cardinality categorical variable you want to model
- Our code on GitHub: agi-lab/glmmnet (in Python)

Method

1. **Import** functions for building and making predictions from GLMMNet.
2. **Tidy up** the data: train-val-test split & feature preprocessing.
3. **Train** the GLMMNet and experiment with the hyperparameters.
4. **Evaluate** model performance on OOS set.
5. Extract the random effect predictions and **interpret** the findings.