Convention A: Application of the NLP's models in Loss Modeling in the actuarial field

Manuel Caccone

March 19, 2024

Preface

Classic Loss Models vs New Loss Models

Loss modeling in actuarial science is a classic problem. The most common models are the following:

- Frequency-Severity models
- Compound models

Normally these models are available for two sets of personalization:

- **A priori** models;
- A **posteriori** models.

Both methods usually are based on the use of the classical GLM (Generalized Linear Models) framework, moreover the Poisson distribution for the frequency and the Gamma distribution for the severity.

ISOA

Classic Loss Models vs New Loss Models

Anyway, giving **personalization to the models is not enough**. The specific ratemaking process is not able to capture the complexity of loss data, acquiring a lot of noise and not being able to capture the real behavior of the policyholders.

Classic ratemaking will provide coefficients linking the policyholder's characteristics to the **CONTEXT** of loss data.

Coefficients and Context

We are looking for **CONTEXT**. In particular:

- The **frequency** of the claims related to a specific cluster of policyholders;
- The **severity** of the claims related to a specific cluster of policyholders.

In general let X_j be the vector of the loss distribution of policyholder's cluster j for $j \in J$ in omogegeous portfolio. An example of frequency-severity model can be:

- $N_j \sim Poisson(\lambda_j)$
- $Y_j \sim Gamma(\alpha_j, \beta_j)$

Coefficients and Context

A PERFECT model cannot provide something like that, that is risks besides others with "cross-existence":

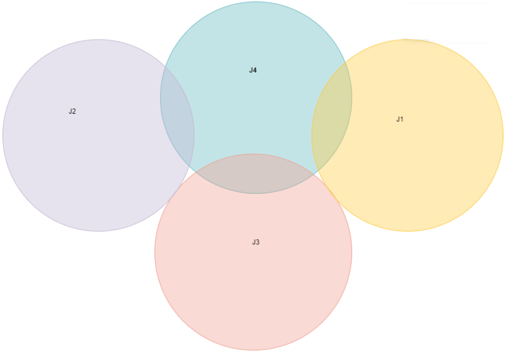


Figure 1: Eulero-Venn coefficients context

Coefficients and Context

- It's clear that the coefficients are not able to capture the complexity of the loss data, due to the noise component. Noise is the result of the **randomness** of the loss data and the weakness of the ratemaking process due to the **lack of information** (or maybe **lack of valid work hypothesis**).
- GLMs can be a limitation, a **weak flexibility** dealing with **dependence** between the policyholders. It's known the concept of **independence** between the policyholders (*indipendent and identically distributed random variables*).

Coefficients and Context

An efficient clusterization can be the solution to the problem. The clusterization is the process of grouping the policyholders in homogeneous clusters. The clusterization is the result of the **context** of the loss data and the **dependence** between the policyholders. In classic theory, the clusterization is based on:

- *Chi-squared* distance based methods;
- *Pitkanen's* method[1];
- *K-means* algorithm et similia.

[1] Pitkänen (1975)

Coefficients and Context

- **Misclassification** of loss event can be another great problem, expecially in multi-peril policies. In general the loss event classification is guided by the **policyholder's declaration** and the **loss adjuster's evaluation**.
- A loss event enters into the insurer database if the policyholder declares the loss event and the loss adjuster evaluates the loss event. The loss adjuster's evaluation is the result of the **loss adjuster's expertise** and the **loss adjuster's evaluation process**. It can be a blind process respect to the insurer's database or modeling needs.

Coefficients and Context

Let:

- $X^F_{i,j}$ loss deriving from i policyholder with j cluster in F peril (the misspecified peril);
- $X^T_{i,j}$ loss deriving from i policyholder with j cluster in T peril (the true one);

Missclassification leads to the following problem:

$$\mu^F_{i,j} = \mathbb{E}[X^F_{i,j}] \neq \mu^T_{i,j} = \mathbb{E}[X^T_{i,j}]$$

ISOA

Core problem

Multi-peril policies and pre-clustering via NLP

- Pre-clustering via NLP can be a solution to the problem. The pre-clustering via NLP is the process of grouping the policyholders in homogeneous clusters using the **policyholder's declaration** and the **loss adjuster's evaluation**. The pre-clustering via NLP is the result of the **context** of the loss data and the **dependence** between the policyholders.
- Loss adjuster's evaluation normally it's not included in actuaries' database. The loss adjuster's evaluation is the result of the **loss adjuster's expertise** and the **loss adjuster's evaluation process**. It can be a blind process respect to the insurer's database or modeling needs.

Fraud detection via NLP

- The fraud detection via NLP it's another great application of the NLP's models in the actuarial field. The fraud detection via NLP is the process of identifying the **fraudulent claims** using the **policyholder's declaration** and the **loss adjuster's evaluation**.
- "pre-clustering via NLP" can help to prevent misclassification of loss event, **including fraudulent claims**[2].

[2] Boulieris et al. (2023)

How to?

Collecting the data

- first of all, we need to collect the data. The data can be the following:
  - the **policyholder's declaration**;
  - the **loss adjuster's evaluation**;
  - the **loss data**.

Figure 2: Loss Documents

Collecting the data

Let I_k for $k \in \{1, \cdots, N\}$ with N claims, each I_k works with:

a) training new model starting from general **pretrained** models[3] with "signed-data";
b) **pre-clustering** of the policyholders through the **embeddings** of a "big" pretrained model (OPENai e.g.).

In a) case, the **pretrained** model will be developed with specific purpose in order to prevent, for example, fraud claims. In b) case, a "big" pretrained model will be used to **embed** I_k collapsing with very specific **context** of the claim's information.

[3] Devlin et al. (2018)

Zero-Shot Classification

- **Zero-shot classification** (ZCS) is a machine learning technique that enables the classification of data into categories not seen during the model's training phase. This approach can be extremely useful in fields where categories may change over time or where it's not feasible to collect a large labeled dataset for every possible category.
- fixed categories of loss normally are adopted by the modelers responding to the **business needs** (Line of business) and the **regulatory needs** (Solvency II, IFRS 17, etc.). The ZCS can be a solution to the problem of the **misclassification** of the loss event, giving flexibility and fitting claims to categories with right **context**.

ISOA

Zero-Shot Classification

- **ZCS** it's a **difficult** task for a *transformer* model, anyway it's possible to use big models with a lot of parameters and a lot of data[4]. A **big** model will be able to **embed** the claim's information into the **fitting context**.
- *OPENai* models nowaday are the best models for the task, gold standard in the "NLP Models" field.

[4] HugginFace (2024)

ISOA

Zero-Shot Classification

- **Embedding** is "roughly" the process of transforming the claim's information in a **vector** of numbers. The vector of numbers will be the **context** of the claim's information.
- **ZCS** involves the use of a **classifier** that has been trained on a set of categories, working onto any others for which it was trained at all.

ISOA

Zero-Shot Classification

- Here a simple example of **ZCS architecture**[5]:

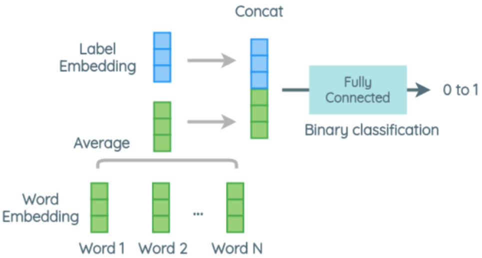


Figure 3: ZCS architecture

[5] Chaudhary (2020)

Data

Data presentation

The data used in the analysis come from Health Multi-Peril policies, in particular:

- Dentistry
- Visits
- Mammography
- Analysis
- Diagnostics.

There are 78 claims in the dataset, generating claims amount for all perils Dentistry,Visits,Mammography,Analysis,Diagnostics.

ISOA

Data presentation

Here the average claims amount for each peril:

Peril	Abbreviation	AverageClaim
Dentistry	Dnts	85.55
Visits	Vsts	40.88
Mammography	Mmmg	40.38
Analysis	Anly	27.99
Diagnostics	Dgns	35.29

Application

Confusion Matrix

Operating the ZCS given the same perils as categories onto the *Loss Adjuster Relations* , we can obtain the following confusion matrix (Before and After the ZSC) in order to measure the mispecification of the perils:

	Anly	Dgns	Dnts	Mmmg	Vsts	Totale
Anly	5	3	3	1	4	15
Dgns	4	1	1	1	2	14
Dnts	4	1	3	1	3	12
Mmmg	0	2	1	1	6	8
Vsts	2	7	4	4	14	29
—-	—	—	—	—	—	—
Totale	16	9	12	10	31	78

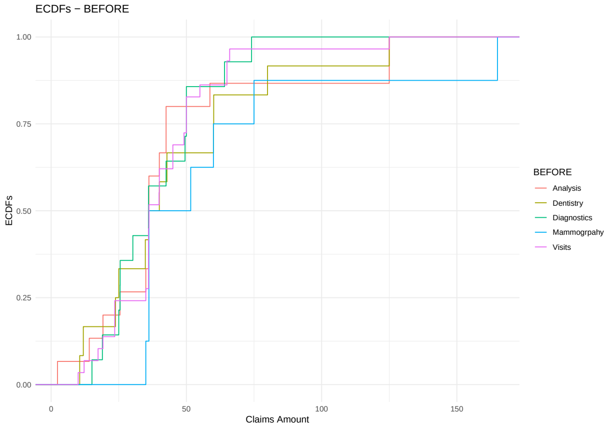
Distributions Before and After

We have a **69.23 % value of mispecification ratio**. Here the new average claims amount before and after the ZSC:

Peril	Abbrev.	AvCl AFTER	AvCl BEFORE
Dentistry	Dnts	85.55	45.83
Visits	Vsts	40.88	41.43
Mammogrpahy	Mmmg	40.38	61.88
Analysis	Anly	27.99	44.94
Diagnostics	Dgns	35.29	38.75

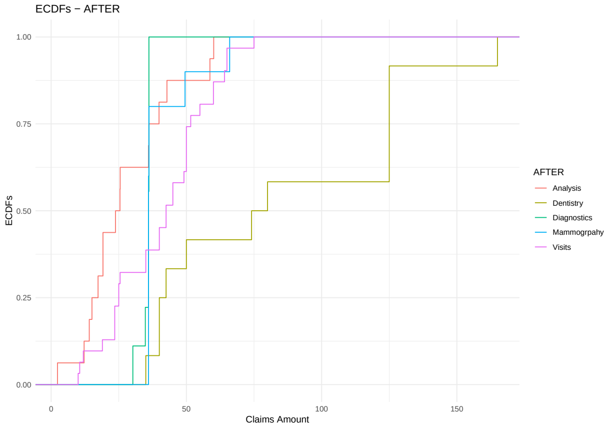
Distributions Before and After

Here the empirical distributions function values **BEFORE** graphically represented:



Distributions Before and After

Here the empirical distributions function values **AFTER** graphically represented:



Correlation matrix

Correlation Matrix **BEFORE**:

	Anly	Dnts	Dgns	Mmmg	Vsts
Anly	1	-0.16	0.22	0.29	-0.13
Dnts	-0.16	1	0.13	-0.07	-0.16
Dgns	0.22	0.13	1	0.18	0.3
Mmmg	0.29	-0.07	0.18	1	0.35
Vsts	-0.13	-0.16	0.3	0.35	1

Correlation matrix

Correlation Matrix **AFTER**:

	Anly	Dnts	Dgns	Mmmg	Vsts
Anly	1	-0.16	0.38	0.26	-0.32
Dnts	-0.16	1	-0.97	-0.55	-0.03
Dgns	0.38	-0.97	1	0.59	-0.13
Mmmg	0.26	-0.55	0.59	1	0.38
Vsts	-0.32	-0.03	-0.13	0.38	1

Conclusions

Conclusions

- The **ZCS** is a great tool to prevent the **misclassification** of the loss event, giving flexibility and fitting claims to categories with right **context**.
- Incorrect classification of a loss event results in a 'narrowing distribution,' consolidating the context (same for all) of the claim's information . Recovering the **context** we can see a "spreading distribution" instead.
- The **ZCS** can restablish reasonable **correlation** between the claims amount, due to a better catched **context**.

ISOA

Future findings

- By employing cosine similarity, we can assess the **proximity** between the **embeddings** of claims to gauge their similarity. 'Real correlations' serve as valuable inputs for a copula aggregation framework, where the weights of the claims are adjusted according to the similarity context of the claims.

Thank you

Thank you for your attention.

Manuel Caccone
Gruppo Unipol
manuel.caccone@unipolsai.it

References I

Boulieris, Petros, John Pavlopoulos, Alexandros Xenos, and Vasilis Vassalos. 2023. "Fraud Detection with Natural Language Processing." *Machine Learning*, 1–22.

Chaudhary, Amit. 2020. "Zero Shot Learning for Text Classification." 2020. https://amitness.com/2020/05/zero-shot-text-classification.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *CoRR* abs/1810.04805. http://arxiv.org/abs/1810.04805.

HugginFace. 2024. "Zero-Shot Classification." 2024. https://huggingface.co/tasks/zero-shot-classification.

Pitkänen, Paavo. 1975. "Tariff Theory." *ASTIN Bulletin* 8 (2): 204–28. https://doi.org/10.1017/S0515036100009338.

ISOA