Adjusting Manual Rates to Own Experience: Comparing the Credibility Approach to Machine Learning A Preprint

Christophe Dutang, Quentin Guibert, Giorgio Alfredo Spedicato

Adjusting Manual Rates to Own Experience: Comparing the Credibility Approach to Machine Learning

A Preprint

Giorgio A. Spedicato, Ph.D FCAS FSA CSPA C.Stat
Leitha SRL, Unipol Group

Milan, Italy
spedicato_giorgio@yahoo.it

Christophe Dutang, Ph.D
LJK, CNRS
Grenoble INP - UGA
Bat. IMAG, 700 avenue Centrale, 38401 St Martin d'Hères
christophe.dutang@grenoble-inp.fr

Quentin Guibert, Ph.D
CEREMADE, CNRS
Univ. Paris-Dauphine - PSL
Pl. du Ml de Lattre de Tassigny, 75016 Paris, France
quentin.guibert@dauphine.psl.eu

March 15, 2023

Abstract

Credibility theory is the usual framework in actuarial science when it comes to reinforcing individual experience by transfering rates estimated from collective information. Based on the paradigm of transfer learning, this article presents the idea that a machine learning (ML) model pre-trained using a rich market data porfolio can improve the prediction of rates for an individual insurance portfolio. This framework consists first in training several ML models on a market portfolio of insurance data. Pre-trained models provide valuable information on relations between features and predicted rates. Furthermore, features shared with the company dataset are used to predict rates better than the same ML models trained on the insurer's dataset alone. Our approach is illustrated with classical ML models on an anonymized dataset including both market data and data from an European non–life insurance company, and is compared with a hierarchical Bühlmann-Straub credibility model. We observe the transfert learning stragegy combining company data with external market data significantly improves the prediction accuracy compared to a ML model only trained on the insurer's data and provides competitive results compared to hierarchical credibility models.

***K*eywords** Transfer learning · Hierarchical credibility theory · Bühlmann credibility theory · Boosting · Deep Learning

1 Introduction

The use of market data as an aid for setting own company rates has been a common practice in the insurance industry. External data, as provided by insurance rating bureaus (cf. Appendix A.1), reinsurers or advisory organizations, may supplement internal company's data that may be scarce or unreliable because of a non-representative and/or a too short history, or, non-existing at all, e.g., when entering in a new business line or a new territory. According to Porter and CPCU (2008), pools of insurers existed already in the second part of the 19th century in the US that supported their members in setting raters thought data collection and standardized policy forms. In the first part of the 20th century, the McCarran Fergusson Act partially exempted the Insurance industry from the US federal antitrust regulation, thus NAIC laws explicitly allowed cooperation in setting rates specifying the role of rating organization. The importance of external data has been historically recognized by regulators to support adequate rates that preserve the company's solvency, to avoid an excessive competition and to ease the entrance of new players, e.g., granting a partial antitrust – law exception in the US jurisdiction (Danzon 1983).

When the insurer takes into account its own experience in order to enhance the credibility of its rates, it needs to benchmark its portfolio experience compared to the market one. The actuarial profession traditionally used techniques based on Bayesian statistics and non-parametric credibility to optimally combine the market and insurer's portfolio experience in the technical rates. From this point of view, the contribution of market data makes it possible to satisfy the two classic approaches addressed by the credibility theory: the limited fluctuation credibility theory and the greatest accuracy credibility theory, e.g., Norberg (2004). The former refers to the need of incorporating individual experience into the rate calculation in order to stabilize the level of individual rates. The second approach corresponds to the application of modern credibility theory and consists in combining both individual and collective experiences to predict individual rates by mean square error minimization.

Credibility theory is extensively used in non-life insurance. Early models were not based on policyholders' rate-making variables, see, e.g., Bühlmann and Gisler (2006) for a comprehensive presentation. Some advanced regression credibility models have been proposed in the actuarial literature, such as the so-called Hachemeister model (Hachemeister et al. 1975). On the contrary, rates based on Generalized Linear Models (GLMs), the current gold standard in personal rates pricing (Goldburd, Khare, and Tevet 2016), are only based on the impact of ratemaking factors, giving no credit to the individual policy experience.
Nevertheless, mixed effects GLMs allow to incorporate policyholders' experience within the GLM tariff structure (Xacur and Garrido 2018; Antonio and Beirlant 2007) but they are not widespread used. All these regression approaches enable insurers to incorporate individual risk profile covariates into a credibility model. The structure of insurance data, notably the distinction between own experience and market experience, is dealt with the use of the hierarchical credibility model of Bühlmann and Straub (1970). In some situations, the use of company data is not possible at all and only a tariff at market level is reliable.

For instance, in France, a two-level Bühlmann-Straub rating model is used for fire and business interruption insurance (Douvillé 2004), with data collected the French association of private and mutual insurers, FA. In other countries, as far as the authors' knowledge, public insurance bureaus exist certainly in Italy, Germany, UK and Brazil. The Italian Association of Insurers, ANIA, aggregates data from the Motor lines (but only pure premiums with few covariates), the Long-Term Care insurance for Health, while it collects and extensive statistical plans (with many covariates) for Crop insurance. In Germany, the German Insurance Association (GDV) provides data similar to the Italian ones for many lines. The Industry data and subscription section of the Association of British Insurers (ABI) provides (at least) yearly aggregate data for many P&C, Health and Life LOBs. Finally, the Brazilian Insurance Regulator SUSEP provides aggregated losses and exposures for the Motor Liability insurance aggregated by key rating variables.

Furthermore, credibility theory is also largely used in life insurance applications for modeling mortality risks. A first attempt for stabilizing mortality rates by combining the mortality data of a small population with the average mortality of the neighboring populations is proposed by Ahcan et al. (2014). Regarding this issue of limited mortality data (small population or short historical period of observations), Li and Lu (2018) introduces a Bayesian non-parametric model for benchmarking a small population compared to a reference population. Bozikas and Pitselis (2019) focus on a credible regression framework to efficiently forecast populations with a short-base-period. In order to improve mortality forecasting, some recent contributions have been done in the literature for combining usual mortality models, such as the Lee and Carter (1992) model and the Bühlmann credibility theory, see Tsai and Lin (2017), Tsai and Zhang (2019) and Tsai and Wu (2020) among others.

The recent widespread/massive usage of Machine Learning (ML) has provided many more techniques to the practitioner actuaries. Gradient Boosting Models (GBM) and Deep Learning (DL) models for motor third-part liability (MTPL) pricing are presented e.g. in Noll, Salzmann, and Wuthrich (2020), Ferrario, Noll, and Wuthrich (2020), Schelldorfer and Wuthrich (2019), and Ferrario and Hämmerli (2019). More recently, Hanafy and Ming (2021) show that random forest (RF) is more efficient (in terms of accuracy, kappa, and AUC values) than logistic regression, XGBoost, decision trees, naive Bayes, and KNN to predict claim occurrence. Matthews and Hartman (2022) compare RF, GBM and DL against GLMs to predict the claim amount and the claim frequency on a commercial auto insurance and demonstrate the efficiency and the accuracy for future ratemaking models. Henckaerts et al. (2021) also show that GBM outperform the classical GLMs and allow the insurer to form profitable portfolios and to guard against potential adverse risk selection. Furthermore, non-pricing applications have been carried out, e.g., Spedicato, Dutang, and Petrini (2018) model the policyholder behavior; Rentzmann and Wuthrich (2019) present recent advances in unsupervised learning for vehicle classification as DL autoencoders; Kuo (2019) shows how neural networks with embedding may offer a sensibly better prediction on tabular loss development triangles using the NAIC reserving dataset. We refer to the comprehensive review of ML in P&C studies by Blier-Wong, Cossette, et al. (2021).

On the life insurance side, the application of DL on lapse modeling (Kuo, Crompton, and Logan 2019) as well as the DL version (Richman and Wuthrich 2019; Nigri et al. 2019) of the classical Lee-Carter model are worth mentioning. For a more comprehensive review, see Richman (2021a) or Richman (2021b). Recently, Diao and Weng (2019) combine the use of credibility and regression tree models. In these publications, the ultimate goal of the use of ML is to improve the usual regression setup in actuarial science based on the GLM. However, these techniques, such as the Gradient Boosting Models (GBM) and the Deep Learning (DL), can also be used in a manner that permits to "transfer" what the model has learned on a much bigger dataset (as the market data, MKT) to a smaller set (the portfolio data of the company, CPN). For that, "Transfer learning" (henceforth TRF) reuses knowledge learned in different data sources to improve performance of learners. This area in machine learning has become particularly popular in recent years, in particular in computer vision DL modes to fine tune standard architectures on specific recognition tasks, see Zhuang et al. (2021) for a comprehensive review. In our experience, such approaches tend to develop in the insurance industry for ratemaking models with the incorporation of new data sources (Blier-Wong, Baillargeon, et al. 2021), but they can be used in other areas, for instance to train life insurance valuation models (Cheng et al. 2019).

In this paper, our aim is to take advantages of ML for easily handling complex non–linear relationships compared to standard credibility based approaches to assess the policyholders' risk more accurately. Hence, our work contrasts with traditional methods to ML ones in the task of blending market data to individual portfolio experience. First, we anticipate a difference between the credibility approach and the ML used in this paper: the credibility approach naturally uses the longitudinal structure to calibrate its parameters, while this is not a prerequisite for ML models which only need to share some variables. We apply our approach on a (properly anonymized) dataset comprised both market and own portfolio experience coming from an European non–life insurance pool. Final comparisons will focus not only on predictive performance, but also practical applicability in terms of computational request, ease of understanding and interpretability of results.

The rest of this paper is organized as follows. We recall the hierarchical Bühlmann-Straub (HBS) credibility model in Section 2. Section 3 presents the main ML algorithms used in this paper. Section 4 compares the performance between ML and HBS models based on a market dataset and a company dataset, and Section 6 concludes this paper.

2 Hierarchical credibility model

In this section, we briefly describe the hierarchical credibility theory of Bühlmann and Gisler (2006) used in this paper for modeling the claim frequency. We also refer to Goulet (1998) for a general introduction.

Consider a large portfolio of I individual risks which includes heterogeneous risk profiles, as well as market and company data. The model considered is defined as an unbalanced claim model since different claim histories are available across individual risks. Intrinsically, the credibility approach is based on a longitudinal data structure where individuals/policyholders' clusters are repeatedly observed in a given time period. Generally, the company data experience is often shorter than that of the market. In addition, we assume that market and company datasets share the same features, which makes it possible to fit into a framework compatible with homogeneous transfer learning approaches (Zhuang et al. 2021).

For the ease of this presentation and without a loss of generality, we use a five-level model structured in a hierarchical tree, as presented in Figure 1 with usual notation.

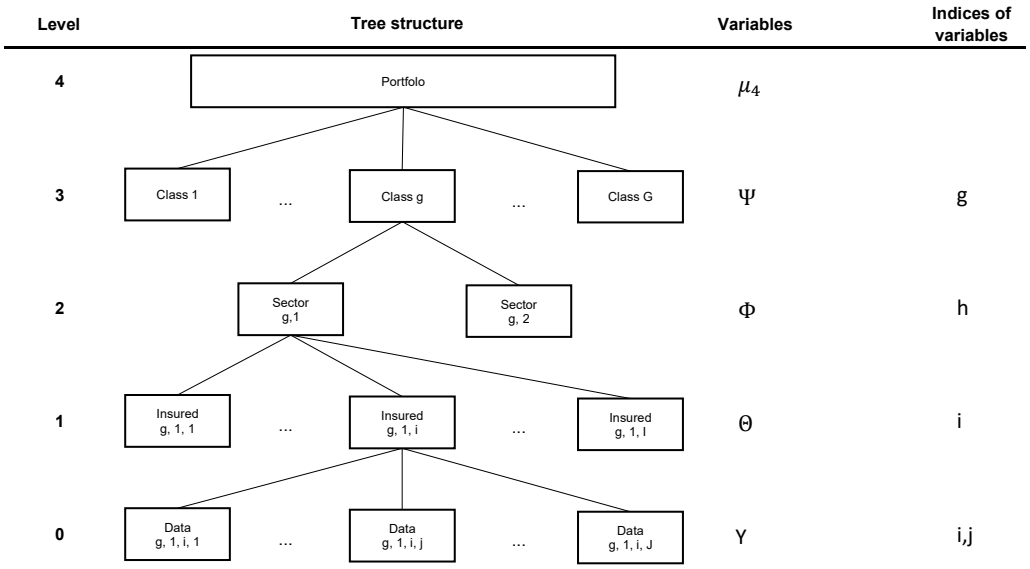


Figure 1: Representation of a five-level hierarchical tree structure

The five levels are given as follows from the top to the bottom, based on the classical assumptions of hierarchical credibility theory:

- Level 4: This is the entire portfolio with market and company information.
- Level 3: The portfolio is divided into risk classes. We introduce parameters related to this risk level, Ψ_g, $g = 1, \ldots, G$, which are independent and identically distributed.
- Level 2: Each risk class is divided into sectors. Given Ψ_g, we denote by $\Phi_{g,h}$, $h = 1, \ldots, H$ the class risk parameters which are assumed to be conditionally independent and identically distributed.
- Level 1: Given $\Phi_{g,h}$, $\Theta_{g,h,i}$, $i = 1, \ldots, I$ are the individual risk parameters which are conditionally independent and identically distributed.
- Level 0: Given $\Theta_{g,h,i}$, data are available on the study period $[1, J_{g,h,i}]$. We denote by $\mathbf{Y}_{g,h,i} = (Y_{g,h,i,1}, \ldots, Y_{g,h,i,J_{g,h,i}})$, the vector of observations over years which are conditionally independent, identically distributed and have a finite variance. We also introduce the vector of the relative known weights $\mathbf{w}_{g,h,i} = (w_{g,h,i,1}, \ldots, w_{g,h,i,J_{g,h,i}})$ over the same observation period.

In Section 5.2, the class variable related to Level 2 results from the combination from several categorical variables. These variables comprise unobservable risk factors allowing to partition the data space. Seven variables are used to build up the credibility tree in the numerical application. That is by adding intermediary levels in Figure 1, we consider 10-level hierarchical trees later in this paper.

In order to estimate credibility rates, we define the following notations and structural parameters for $i = 1, \ldots, I$ and $j = 1, \ldots, J_{g,h,i}$:

- Level 4: Define $\mu_4 = E[Y_{g,h,i,j}]$ the collective rates.

- Level 3: Define $\mu_3(\Psi_g) = E[Y_{g,h,i,j} \mid \Psi_g]$ for observations $Y_{g,h,i,j}$ that stem from Ψ_g, $\sigma_3^2(\Psi_g) = Var[\mu_2(\Phi_{g,h}) \mid \Psi_g]$ and $\sigma_3^2 = Var[\mu_3(\Psi_g)]$.
- Level 2: Define $\mu_2(\Phi_{g,h}) = E[Y_{g,h,i,j} \mid \Phi_{g,h}]$ for observations $Y_{g,h,i,j}$ that stem from $\Phi_{g,h}$, $\sigma_2^2(\Phi_{g,h}) = Var[\mu_1(\Theta_{g,h,i}) \mid \Phi_{g,h}]$ and $\sigma_2^2 = E[\sigma_3^2(\Psi_g)]$.
- Level 1: Define $\mu_1(\Theta_{g,h,i}) = E[Y_{g,h,i,j} \mid \Theta_{g,h,i}]$ for observations $Y_{g,h,i,j}$ that stem from $\Theta_{g,h,i}$, $\sigma_1^2(\Theta_{g,h,i})) = Var[Y_{g,h,i,j} \mid \Theta_{g,h,i}]w_{g,h,i,j}$ and $\sigma_1^2 = E[\sigma_2^2(\Phi_{g,h}))]$.
- Level 0: Define $\sigma_0^2 = E[\sigma_1^2(\Theta_{g,h,i}))]$.

Similarly to the Bühlmann-Straub model, the credibility estimates for these parameters are based on the Hilbert projection theorem, see Chapter 6 of Bühlmann and Gisler (2006). With the above notations, we obtain the following classical results for hierarchical (inhomogenenous) credibility estimators

$$\widehat{\mu(\Psi_g)} = \widehat{\alpha_g^{(3)}}\,\widehat{B_g^{(3)}} + (1 - \widehat{\alpha_g^{(3)}})\widehat{\mu_4},$$
$$\widehat{\mu(\Phi_{g,h})} = \widehat{\alpha_{g,h}^{(2)}}\,\widehat{B_{g,h}^{(2)}} + (1 - \widehat{\alpha_{g,h}^{(2)}})\widehat{\mu(\Psi_g)},$$
$$\widehat{\mu(\Theta_{g,h,i})} = \widehat{\alpha_{g,h,i}^{(1)}}\,\widehat{B_{g,h,i}^{(1)}} + (1 - \widehat{\alpha_{g,h,i}^{(1)}})\widehat{\mu(\Phi_{g,h})},$$

where formula of credibility factors $\widehat{\alpha_g^{(3)}}$, $\widehat{\alpha_{g,h}^{(2)}}$, $\widehat{\alpha_{g,h,i}}^{(1)}$ and weighted means $\widehat{B_g^{(3)}}$, $\widehat{B_{g,h}^{(2)}}$ and $\widehat{B_{g,h,i}^{(1)}}$ are given in Appendix A.6. They depend on structural parameters which can easily be estimated non-parametrically. Therefore, a HBS model provides a recursive computation of weighted empirical means whose parameters minimize quadratic losses. There are no distribution assumption when deriving estimators and thus HBS models are full non-parametric models.

3 Modeling approach with Transfer Learning

In this section, we present the transfer-learning (TRF) based framework, as well as ML models used in this paper. This research aims to compare the predictive power of traditional credibility and ML methods that use an initial estimate of loss costs, e.g., from MKT experience, to predict those of a smaller portion (the CPN one) in a subsequent period (the test set). Therefore, following the idea of the greatest accuracy credibility theory, our modeling process aims to predict the losses on the last available year (the test set) by training models on the experience of the previous years, eventually split into a train and validation test.

3.1 Transfer Learning

Figure 2 describes the main steps of our TRF approach compared to ML models fully trained on a market train dataset, called "MKT" approach, or on a company dataset, called "CPN" approach. To fit a ML model via the MKT approach (resp. the CPN approach), we split the historical data from a benchmark (resp. a company) dataset solely between a train and validation subsets. Then, the performance is assessed based on a test data from the company dataset.

The TRF approach relies on experience from the MKT approach and uses the corresponding pre-trained model as a starting point. Next, we fine-tune the MKT model based on experience from the CPN dataset. After this step, the model contains both information from the market and the company, and should offer better predictions.

3.2 Machine learning models

Now, we focus on ML models that permit to use an initial estimate of losses performed on another set via Transfer Learning. While the paper explores the use of such approaches applying ML methods, traditional GLMs may be used as well, see Appendix A.7 for a brief introduction. In GLMs, one can perform log-linear regressions to estimate both the frequency and the severity of the claim. These outputs can used as offsets in subsequent models. For instance, under a log-linear regression framework and initial log-estimate of either the frequency, the severity of the pure - premium may be set as an offset for a subsequent model (Yan et al. 2009).

ML methods used in insurance pricing are strongly non - linear and are able to automatically find interactions among ratemaking factors and exclude non relevant features. In particular two techniques are acquiring

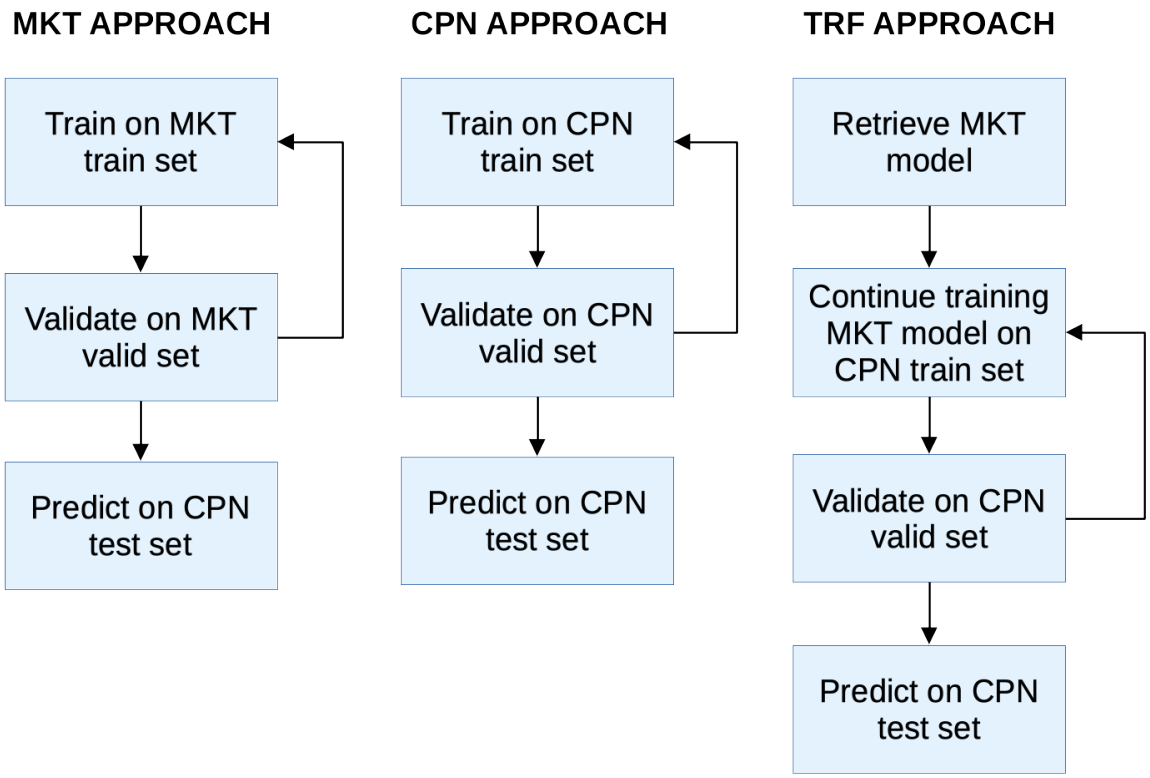


Figure 2: ML models training diagram

widespread importance: boosting and deep learning. Both techniques allow the use of an initial estimate of loss / exposure to risk to train the model on last observations.

All the ML models used in this work hold the Poisson assumption. That is each i-th insurance policy is described by independent claim count N_i such that

$$N_i \sim \mathcal{P}oi\left(\lambda(x_i) \times v_i\right), i \in 1, \ldots, n,$$

being x_i the covariates' vector and v_i the exposure related to the i-th policy for a sample of size n. Thus, ML models try to find the best functional form for $\lambda\left(.\right)$ by minimizing the Poisson loss function, typically on the test set

$$L(\lambda(.)) = \frac{1}{n}\sum_{i=1}^{n} 2n_i \left[\frac{\lambda\left(x_i\right) v_i}{n_i} - 1 - \ln\left(\frac{\lambda\left(x_i\right) v_i}{n_i}\right)\right],$$

where n_i are observed claim counts.

3.2.1 Boosting techniques

The boosting approach (Friedman 2001) can be synthesized by the following formula

$$F_t\left(x\right) = F_{t-1}\left(x\right) + \eta \times h_t\left(x\right),$$

that is, the prediction at the t-it step is given by the contribution, to the prediction of the previous step, of a weak predictor $h_t\left(x\right)$, properly weighted by a learning (shinkage) factor η, being x the covariate vector. The most common choice for the weak predictor $h_t\left(x\right)$ lies in the classification and regression trees (CART) family (Breiman 2017), from which the Gradient Boosted Tree (GBT) models take the name. CARTs partitions the feature space in an optimal way to receive (more) homogeneity on the resulting subsets (in terms of the

modeled outcome). Such optimal partition is determined by recursively searching for the stage-wise optimal split among all standardized binary splits (SBS). At first stage, given an optimal partition of size $K > 0$ of the feature space, $(X_k^{(1)})$ with $k = 1, 2, \dots K$, the estimated frequency is constant in each element of the partition and determined by the MLE estimate

$$\hat{\lambda}_k = \frac{\sum_{i=1}^{n} \mathbf{1}_{\{x_i \in X_k^{(1)}\}} n_i}{\sum_{i=1}^{n} \mathbf{1}_{\{x_i \in X_k^{(1)}\}} v_i}.$$

As well presented by Noll, Salzmann, and Wuthrich (2020), a "weak learner" is a SBS with just one split (e.g. $K = 2$ leaves) such that the estimated frequency is

$$\lambda^{(1)}(x_i) = \hat{\lambda}_1 \mathbf{1}_{\{x_i \in X_1^{(1)}\}} + \hat{\lambda}_2 \mathbf{1}_{\{x_i \in X_2^{(1)}\}}.$$

The boosting approach starts from an initial estimate given by the above formula. We can define "working weights" as $w_i = \hat{\lambda}^{(1)(x_i)} v_i$ so that N_i follows a Poisson distribution $\mathcal{P}oi\left(\mu(x_i) \times w_i\right)$. With a new SBS partition set $X^{(2)}$, we can recursively estimate $\mu(x_i)$ using a supplementary SBS such that

$$\hat{\mu}^{(2)}(x_i) = \left(\hat{\mu}_1^{(2)} \mathbf{1}_{\{x_i \in X_1^{(2)}\}} + \hat{\mu}_2^{(2)} \mathbf{1}_{\{x_i \in X_2^{(2)}\}}\right)^{\eta},$$

where $\hat{\mu}_1$ and $\hat{\mu}_2$ are estimated using formulas analogue as the first step. We obtain an improved regression function $\lambda^{(2)}(x) = \lambda^{(1)}(x) \times \hat{\mu}^{(2)}(x)$. The $\eta \in (0, 1]$ parameter is the learning (shinkage) factor and it is used to make the learner even more weaker, as values close to zero move the learner towards one. The estimation can be iterated for M times and as it is performed in log-scale, this reduces to the formula exposed at the beginning of the paragraph.

It can be shown that “boosting” weak predictors lead to very strong predictive models (Elith, Leathwick, and Hastie 2008). Almost all winning solutions of data science competitions held by Kaggle are at least partially based on the eXtreme Gradient Boosting (XGBoost) algorithm (Chen and Guestrin 2016), the most famous GBT model. More recent and interesting alternatives to be tested are: LightGBM (Ke et al. 2017), which is particularly renowned for its speed, and Catboost (Prokhorenkova et al. 2017), which has introduced an efficient solution for handling categorical data.

The structural difference between XGBoost and LightGBM lies in the approach used to find trees' splits. XGBoost uses a histogram-based approach: features are organized in discrete bins on which the candidate split values of the trees are determined. LightGBM focuses the attention on instances characterized by large error gradients, the ones where growing a further tree would be more beneficial (leaf-wise tree growth). In addition, a dedicated treatment is given to categorical features. In general, none of these algorithms systematically outperforms the others on any given use case, depending by the specific dataset and by the chosen hyperparameters (Gursky 2020; Nahon 2019). We choose LightGBM mainly as significantly faster than XGBoost, a definitive benefit when there is need to iterate the training through different combinations of hyperparameters. CatBoost is not considered in this stage as less mature compared to the other two algorithms.

A set of hyperparameters defines a boosted model and even more defines a GBT one. The core hyperparameters that influence the boosting part are the number of models (trees), $t = 1, 2, \dots, T$ (typically between 100 and 1000) and the learning (shrinkage) rate η, whose typical values lies between 0.001 and 0.2. $h_t\left(x\right)$ can be, when it belongs to the CART family, the maximum depth, the minimum number of observations in final leafs, the fraction of observations (rows or columns) that are considered when growing each tree. The optimal combination of hyperparameters is learned using either a grid search approach or a more refined one (e.g. bayesian optimization).

When applied to claim frequency prediction, they are fit to optimize a Poisson log-loss function. In addition, to handle uneven risk exposure, the log - measure of exposure risk is given (in log scale) as an init-score $\left(F_t\left(x\right)\right)$ to initialize the learning process. The init-score (or base margin) in the boosting approach has the same role of the traditional GLM offset term (Goldburd, Khare, and Tevet 2016).

3.2.2 Deep Learning

An artificial neuron is a mathematical structure that applies a (non linear) activation function to a linear combination of inputs, i.e.

$$\phi\left(z\right) = \phi\left(< x_i, \bar{w} > + \beta\right),$$

being $\bar{w}$ and β the weights and the intercept, respectively. Popular choices of activation functions are: the sigmoid $\phi(z) = 1/(1+\exp(-z))$, the hyperbolic tangent $\tanh(z)$ and the REctifier Linear Unit $\phi(z) = z\mathbf{1}_{\{z\geq 0\}}$.

A neural network consists in one or more layers of interconnected neurons, that receives a (possibly multivariate) input set and retrieves output set (Goodfellow, Bengio, and Courville 2016). Modern Deep Neural Networks (DNN) are constructed by many (deep) layers of neurons. Deep Learning has been knowing a hype in interest for a decade, thanks to the availability of huge amount of data, computing power (in particular GPU computing) and the development of newer approaches to reduce the overfitting that had halted the widespread adoption of such techniques in previous decades. Different architectures have reached state of the art performances in many fields; e.g., convolutionary neural networks achieved top performance in computer vision (e.g. image classification and object detection) (Meel 2021), while recurrent neural networks, see, e.g., Hochreiter and Schmidhuber (1997) for Long Short Term Memory ones, provides excellent results in Natural language processing tasks like sequence-to-sequence modeling (translation) and text classification (sentiment analysis). For applications in actuarial science, we refer to the recent review of Blier-Wong, Cossette, et al. (2021), and to the work of Richman (2021a) and Richman (2021b) for DNN.

Simpler structures are needed for a claim frequency regression, the multi-layer perception (MLP) architectures that basically consist in stacked simple neurons layers, from the input one to the single output cell one. This structure is dealt to handle the relation between the ratemaking factors and the frequency (the structural part). Thus, holding the Poisson assumption $N_i \sim \mathcal{P}oi(\lambda(x_i) \times v_i)$, the structural part is modeled as $\lambda(x_i) = \beta_0 + \sum_{j=1}^{Q} \beta_j \phi(z_i)$ being Q the number of neurons of the preceding hidden layer. To handle different exposures, the proposed architecture is based on the solution presented by Ferrario, Noll, and Wuthrich (2020) and Schelldorfer and Wuthrich (2019). A separate branch collects the exposure v_i, applies a log-transformation, then this exposure is added in a specific layer just before the final one (that has a dimension of one).

Training a DL model consists in providing batches of data to the network, evaluating the loss performance and updating the weights in the direction that minimizes the training (back-propagation). The whole dataset is provided to the fitting algorithms many times (epochs) split in batch. One of the common practices to avoid over-fitting is to use a validation set where the loss is scored at each epoch. When it starts to systematically diverge, the training process is stopped (early stopping).

4 Numerical illustrations

In this section, we compare the prediction performance between our ML and credibility models. In this study, the analysis is performed on two real and anonymized datasets, the CPN and MKT, pre-processed and split into train, validation and test set as discussed in Section 4.1. In particular, we recall that the predicting performance of the fitted models are assessed on the company test dataset, even if models have been calibrated on the company or the market datasets or both. Then, the models are fitted on the train set and the predictive performance is assessed on the test set. The validation set is used in DL and LightGBM models to avoid overfitting. Finally, the different models are compared in terms of predictive accuracy, using the actual / predicted ratio, and risk classification performance, using the Normalized Gini Index (NGI) (Frees, Meyers, and Cummings 2014). The latter index has become quite popular in the actuarial academia and among practitioners to compare competing risk models. Let y_i be the actual number of claims ranked by their modeled score $v_i \times \hat{\lambda}(x_i)$. The NGI is defined as

$$NGI = \frac{2\sum_{i=1}^{n} i y_i}{n\sum_{i=1}^{n} y_i} - \frac{n+1}{n}.$$

In addition to NGI, we also compute the mean absolute error (MAE) and the root mean square error (RMSE), which are also popular metrics for comparing the predicting performance, see Appendix A.2.

4.1 The structure of the dataset

Two (anonymized) datasets are provided, one for the market (`"mkt_anonymized_data.csv"`) and one for the company (`"cpn_anonymized_data.csv"`), henceforth referred as MKT and CPN datasets, see Appendix A.4. These datasets share the same structure, as each company provides its data in the same format to the pool, that aggregates individual filings into a marketwide file, that is provided back to the companies. In particular, it is important to note that the MKT dataset includes CPN data. The dataset contains the year to year

exposures and claim numbers, aggregated by some categorical variables. More precisely, the losses are the number of damaged units, while the exposure are the number of insured units. Therefore, only the frequency component has been modeled as the ratio between the claim number and the exposure. Henceforth, losses in this paper shall be considered as a synonym for claim numbers. Our aggregated dataset contains variables listed bellow:

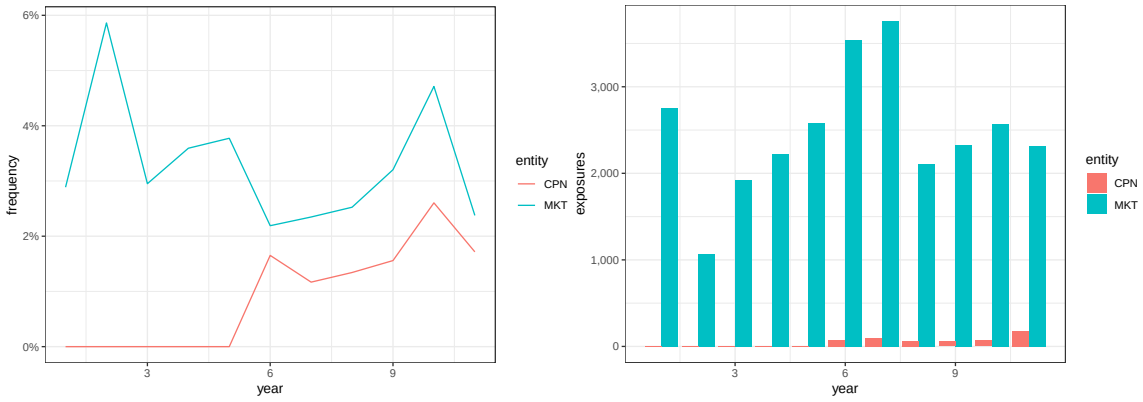
- `exposure`: the insurance exposure measure, on which the rate is filled (aggregated outcomes).
- `claims`: the number of claims by classification group (aggregated outcomes).
- `zone_id`: territory (aggregating variable).
- `year`: filing year (aggregating variable).
- `group`: random partition of the dataset into train, valid and test sets.
- `cat1`: categorical variable 1, available in the original file (aggregating variable). It can be considered as a risk classification, and, possibly, the most important predictors. The number of exposures insured strongly depends on this variable. Also, the `cat1` distribution may vary significantly between companies.
- `cat2`: categorical variable 2, available in the original file (aggregating variable).
- `cat3`: categorical variable 3, available in the original file (aggregating variable).
- `cat4-cat8`: categorical variables related to the territory (joined to the original file by `zone_id`).
- `cont1-cont13`: numeric variables related to the territory (joined to the original file by `zone_id`).
- `entity`: a categorical variable either `"CPN"` or `"MKT"`.

Variable names, levels and numeric variable distributions are masked and anonymized for privacy and confidentiality purposes. Categorical and continuous variables are anonymized by label encoding and scaling (calibrated on market data).

Figure 3 displays exposures and claim frequencies by year for each entity (MKT, CPN). Furthemore, the last available year (2008) is used as test set, while data from precedent years is randomly split between train and validation sets on a 80/20 basis, see Table 1. Market data is available for eleven years, while company data for the last five ones. Also, the number of exposures is widely dependent on the `cat1` variable.

Table 1: Dataset sizes

	Test	Training	Validation	Total
CPN	19124	50430	12519	82073
MKT	89805	527388	130995	748188



(a) Claim frequency

(b) Exposure

Figure 3: Claim frequencies and exposures

Tables 2 and 3 compare explanatory variables by domain. The frequency distribution is reported for categorical one, while summary statistics are computed for continuous ones (mean, standard deviation, minimum and maximum). The `zone_id` and `cat1` statistics have been put in Appendix A.5 for the sake of synthesis. Note also that the variable `year` is not taken as an explanatory variable neither for ML nor credibility models. We implicitly assume that the claim process is stationary.

Table 2: Descriptive statistics for continuous variables.

	Market statistics				Company statistics			
	Min.	Mean	Max.	Std.Dev.	Min.	Mean	Max.	Std.Dev.
cont1	-0.6984692	0	21.5702212	1.000001	-0.6984692	-0.0693805	19.9167608	0.7930418
cont2	-3.3849066	0	6.8058838	1.000001	-3.3257812	-0.0104225	6.6250557	0.9615486
cont3	-3.8761156	0	6.1180898	1.000001	-3.8761156	-0.0207215	4.5260526	0.9734832
cont4	-0.9210659	0	6.9693657	1.000001	-0.9210659	-0.0957618	5.4379990	0.9418158
cont5	-7.7412741	0	3.3530692	1.000001	-7.7204408	0.1036776	3.1162034	0.9816364
cont6	-5.3338005	0	3.7045000	1.000001	-5.3338005	-0.0748241	3.6727833	1.0077024
cont7	-1.4725984	0	3.1211930	1.000001	-1.4725984	-0.0774874	2.8880116	1.0032119
cont8	-1.4039038	0	6.2454848	1.000001	-1.4039038	-0.1138889	6.2454848	1.0239311
cont9	-1.7815236	0	4.3161718	1.000001	-1.7815236	-0.0722550	4.3161718	1.0185563
cont10	-4.0784342	0	3.8562753	1.000001	-3.7914358	0.0398642	3.8562753	0.9809852
cont11	-2.1552892	0	0.9704065	1.000001	-2.1552892	0.0381052	0.9704065	0.9571870
cont12	-0.9924332	0	2.4714341	1.000001	-0.9924332	-0.0136351	2.4714341	0.9298889
cont13	-0.4892582	0	4.8858112	1.000001	-0.4892582	-0.0433888	4.8858112	0.9750647

Table 3: Frequency tables for categorical variables.

	Market statistics for each level												Company statistics for each level											
	0	1	2	3	4	5	6	7	8	9	10	11	0	1	2	3	4	5	6	7	8	9	10	11
cat2	0.008	0.021	0.04	0.004	0.158	0.057	0.348	0.043	0.026	0.239	0.024	0.033	0.02	0.058	0.002	0.21	0.051	0.341	0.027	0.013	0.271	0.005	0.003	
cat3	0.969	0.014	0.017										0.973	0.011	0.016									
cat4	0.788	0.089	0.123										0.811	0.086	0.103									
cat5	0.047	0.632	0.321										0.052	0.628	0.321									
cat6	0.944	0.056											0.939	0.061										
cat7	0.133	0.867											0.124	0.876										
cat8	0.02	0.05	0.086	0.058	0.025	0.081	0.617	0.006	0.02	0.037			0.014	0.063	0.091	0.046	0.024	0.084	0.631	0.001	0.02	0.027		

4.2 Implementation details

In this section, we present the operations performed on the data and the implementation of the different models. The dataset preprocessing is performed in a `Python` 3.8 environment, using the *Pandas* and *Scikit-Learn* libraries (Reback et al. 2020; Pedregosa et al. 2011) for Extraction Transformation and Loading (ETL) stages. The `R` Software (R Core Team 2022) and the `Python` programming language are the environments used for the analysis.

4.2.1 Boosting approach

The LGB model is used to apply boosted trees on the provided datasets, minimizing the Poisson deviance. As for most modern ML methods, a LGB model is fully defined by a set of many hyperparameters for which default values may not be optimal for the given data. Indeed, there is no closed formula to identify the best combination for the given data.

Therefore, an hyperparameter optimization stage is performed. For each hyperparameter, a range of variations is set, then a 100-run trial is performed using a Bayesian Optimization (BO) approach performed by the `hyperopt` Python library (Bergstra, Yamins, and Cox 2013). Under the BO approach, each subsequent iteration is performed toward the point that minimizes the loss to be optimized, being the loss distribution by hyperparameter updated for each iteration using a bayesian approach. As suggested by boosting trees practitioners (Zhang and Yu 2005), the number of boosted models is not estimated under the BO approach, but determined by early stopping. The loss is scored on the validation set and the number of trees chosen is the value beyond which the loss stops decreasing and starts diverging up.

The CPN and MKT models use the standard exposure (in logarithm base) as `init score`. The TRF model instead uses as init score the a priori prediction of the MKT model on the CPN data. The *LightGBM* Python library is used for the boosted models (Ke et al. 2017). Computation is performed on an AMD-FX 9450 processor with 32 GB RAM. In general, fitting one model takes in average a minute on this environment.

4.2.2 Deep Learning

Several approaches may be considered for building a DL architecture. Since the hyperparameters space of a DL architecture is very vast, comprising not only fitting level degrees of freedom (the optimizer, the number of epochs, the batch size), designing the best search strategy and network architecture (the number of layers, the number of neurons within, search etc.) is challenging. At this regard, it is common among practitioners to start with a knowingly working architecture for a similar task and to perform moderate changes. It is also worth mentioning that more sophisticated approaches of DL architecture optimization are being developed (e.g. the Neural Architecture Search (Elsken, Metzen, and Hutter 2019)), but the presentation of such techniques is beyond the scope of this paper.

In this paper, the chosen DL architecture is set by several trials, based on previous experiments practitioners architecture found in the literature for tabular data analysis (Schelldorfer and Wuthrich 2019; Kuo, Crompton, and Logan 2019). Our approach consists of introducing a dense layer to collect the inputs and handling categorical variables using embedding. Three hidden layers perform the feature engineering and knowledge extraction from the input. Dropout layers is added to increase the robustness of the process. As anticipated in the methodological section, the exposure part is separately handled in another branch and then merged in the final layer. The same model's structure is used for both the CPN, MKT and TRF models. The TRF model is build first using the pre-trained weights calculated on the market data and continuing the training process on the CPN data in a second step. Figure 7 in Appendix A.8 displays the model structure as exported by the Keras-Tensorflow routine.

Overfitting is controlled using an early-stopping callback scoring the loss on the validation test and stopping the learning procedure (Zhang and Yu 2005) if the loss is not improved for more than 20 epochs. DL models are trained in *Keras-Tensorflow* 2.4 (Chollet et al. 2018), taking on average 40s per epoch.

4.2.3 Credibility models

Regarding the credibility approach, the original datasets that were in longitudinal format have been processed into a wide format (also called unbalanced) needed by the `R` package *actuar* (Dutang, Goulet, and Pigeon 2008). Furthermore, as required by the hierarchical Bühlmann-Straub model, continuous variables are discretized using the entire dataset (in order to have the widest ranges) based on the Random-Forest algorithm. Using the R package *ForestDisc* (Maïssae 2020) which proposes a random-forest discretization approach, we discretize continuous variables into 3 or 4 levels by group of variables (`cont1-cont2`, `cont3-cont6`, `cont7-cont8`, `cont9-cont10`, `cont11-cont13`) based on their (undisclosed) meaning.

The fitting process of hierarchical credibility models is performed by the `cm` function of the R package *actuar* (Dutang, Goulet, and Pigeon 2008) which allows to fit various forms of credibility models, see, e.g., Goulet et al. (2021). The response variable used for credibility models is the claim frequency (and not the number of claims). Therefore, predicted claim frequencies are multiplied by exposure to obtain the number of predicted claims.

Several credibility models are compared in terms of performance. We first carry out a simple Bühlmann-Straub model using only the `zone_id` variable for the three approaches CPN, MKT and TRF. Note that for TRF, a new variable `entity` is created to distinguish the company and the market data. This base Bühlmann-Straub model is called `BSbase` in the following.

Then, we select the most appropriate hierarchical Bühlmann-Straub (HBS) model by the most appropriate permutation of categorical explanatory variables `cat1-cat8` since there is no particular order among them, except `cat1-cat3`. More precisely, we consider hierarchical credibility structures as follows `cat1`, `cat2`, `cat3`, then a permutation of `cat4`, `cat5`, `cat6`, `cat8`, and finally `zone_id` (and eventually `entity` for TRF) is done. There are 4!=24 possible HBS models. The best categorical HBS model that minimizes the mean squared error when fitting models is called `HBScateg` in the following.

Finally, we apply the same procedure to select another HBS model using categorical explanatory variables `cat1-cat8` and (discretized) continuous variables `cont1-cont13`. As there are too many (17!) HBS models, we restrict to the following hierarchical credibility structures as follows `cat1`, `cat2`, then a permutation of `cont5`, `cont7-10` variables (the most significant continuous variables). There are 5!=120 possible HBS models. This best HBS model is called `HBScont` in the following.

Due to the high number of HBS fitted and used for prediction on the validation dataset, we use parallel computation using the `R` (core) package *parallel*, while models' comparisons are performed in an `R` environment. The running times are summarized in Table 4 and show that MKT and TRF approaches are particularly long

to validate. Indeed the fitting time contains only a call to `cm()` to every HBS models while the validation time makes the prediction for every policies of the validation set, see Table 1. The prediction computation is particularly long, but it requires for each policy the exact location in the credibility tree structure starting from the top.

Table 4: Best HBS models and running times (hours)

Used variables	Approach	Best model	Fitting time	Validation time	Testing time
categorical	**cpn**	cat1:cat2:cat3:cat4:cat6:cat8:cat5:zone_id	0.0076	0.74	0.411
	mkt	cat1:cat2:cat3:cat5:cat8:cat4:cat6:zone_id	0.0647	49.93	1.285
	trf	cat1:cat2:cat3:cat4:cat6:cat8:cat5:zone_id:entity	0.1857	104.39	2.566
all	**cpn**	cat1:cat2:cont7:cont8:cont10:cont9:cont5:zone_id	0.0494	6.99	0.399
	mkt	cat1:cat2:cont10:cont9:cont5:cont8:cont7:zone_id	0.3551	212.52	2.214
	trf	cat1:cat2:cont10:cont5:cont9:cont7:cont8:zone_id:entity	0.4331	244.66	2.560
none	**cpn**	zone_id	0.0000	0.00	0.001
	mkt	zone_id	0.0000	0.00	0.001
	trf	zone_id:entity	0.0001	0.00	0.054

As explained above, the fitting of HBS models is carried out on the training dataset, the best model (in terms of RMSE) is selected on the validation dataset, and the overall comparison is done on the test dataset.

4.2.4 Assessment of performance

The empirical data available for the study faces a risk for which year to year loss cost may materially fluctuate due to external conditions (systematic variability) much more than the portfolios' risks heterogeneity composition. In this regard, the performance assessment has considered not only the discrepancy between the actual and predicted losses, but the ability of the model to rank risks, namely providing a sensible order of which policies are most prone to suffer a loss in the coverage period. This can be achieved even in contexts where getting an acceptable estimate of the pure premium is more challenging, e.g., due to a systematic unmodeled social or environmental trend either in the frequency or in the severity. The ability of ML to identify non-linear patterns and interactions is useful both to model the pure premium and to rank risks.

In order to compare credibility and ML models within or between model classes, we use the following metrics: the NGI, the ratio between the sum of observed claims and the sum of expected ones (denoted by *actual to predicted ratio*), as well as the mean absolute error *MAE* and the root mean squared error *RMSE*. The Gini is a metric of discriminancy and ranks models according to their ability to predict, while the *actual to predicted ratio* is used to check if the model is generally unbiased on a total basis. For both metrics the closer to one the metric is, the better the model is. *MAE* and *RMSE* measure the overall distance between observations and predictions. Best models are identified by the lowest values.

The choice of models deserves a final consideration. The RMSE and NGI indices typically move in the same direction, so minimizing the prediction error, which is the pivotal objective of risk-pricing, also means maximizing the models' discriminating ability, which may be of greater underwriting or marketing interest. If this is not possible, the analyst will rely on either the first or the second metric depending on the business context. Finally, the availability of tools to interpret models should be taken into account; indeed, it may become an essential selection criteria in some contexts where the explicability of a model is essential for regulatory or marketing purposes.

4.3 Models' interpretation and predictive performance results

In this section, we focus on interpreting the ML and credibility models. For that, we examine variables importance for ML models and analyze credibility factor densities related to the best HBS models. In a second step, we assess performance of the both approaches.

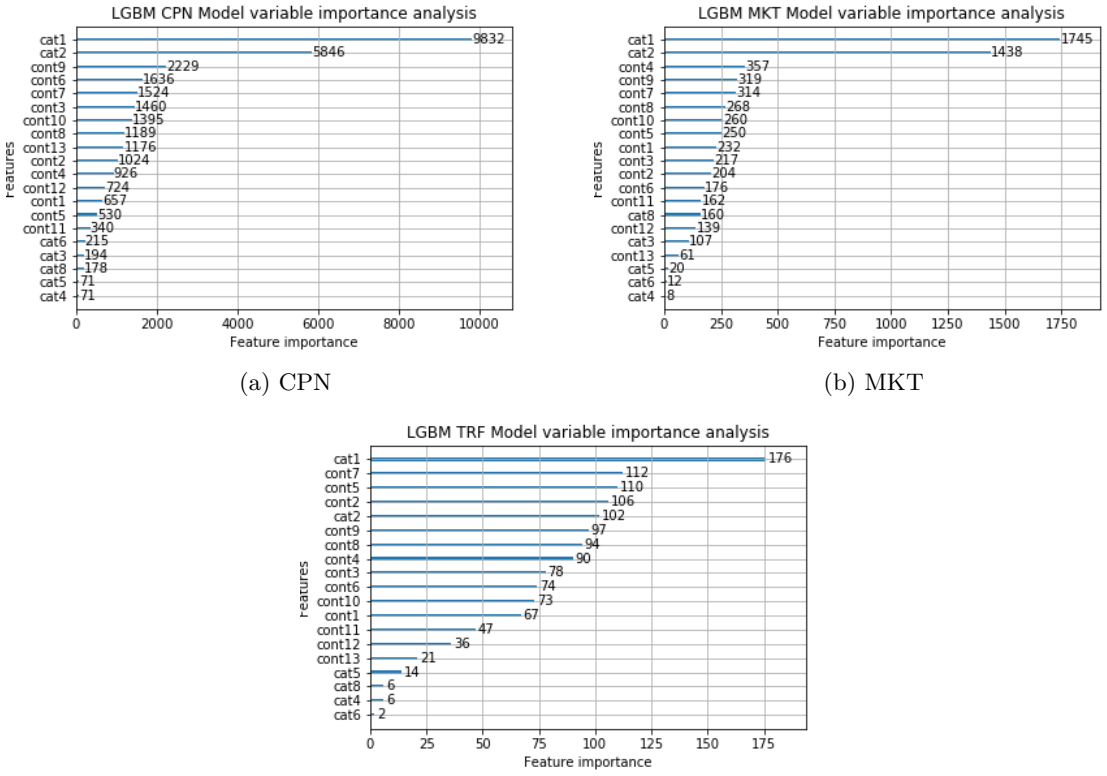
4.3.1 Models interpretation for ML

ML models have been longly considered black boxes, but methods to provide explanations of the models' structure and provide outputs have been developed, and even in actuarial science, see, e.g., Lorentzen and Mayer (2020). In our application, we simply focus on the variable importance analysis internally calculated by the LightGBM model. That measure of variable importance broadly reflects the gain of using that features in LightGBM trees to reduce the training losses. Variable importance analysis in DL models is not automatically

calculated during the training stage and would require the use of a separate algorithm, e.g., the one with Shapley Values (SHAP) (Lundberg and Lee 2017) that is outside the scope of this paper. Another possible approach would be the use of permutation importance for which however there are no readily available routines for Tensorflow Datasets. However, it is reasonable to assume that variables' ranking are similar between the two ML approaches.

Figure 4 displays LightGBM for the CPN, MKT and TRF approaches. The following considerations can be drawn:

1. `cat1` and `cat2` are consistently ranked as the most important predictors both for MKT and CPN approaches. Subfigures 4a and 4b, whose relative importance is markedly superior to that of the other variables that immediately following; the ranking of the remaining variable is indeed very similar,
2. The TRF plot, Subfigure 4c, is more difficult to interpret. It indicates which variables mostly shall be used to correct the difference between the MKT model and the CPN one. While `cat1` keeps the first place, the relative importance of other variables is higher than in the MKT and CPN plots.



(a) CPN

(b) MKT

(c) TRF

Figure 4: LightBM Variable importance for CPN, MKT and TRF approaches

By construction, the ML models used here are black box and require post-hoc interpretability tools to analyze the effect of features. Given the anonymous nature of the data, we limit ourselves to an illustration of the overall interpretability of the variables which is relevant for an actuary to understand the overall effect of variable on the tariff. Depending on the audience involved in the interpretability analysis (Delcaillau et al. 2022) (e.g. an actuary or a policyholder interested by its tariff), it may however be necessary to discuss in depth the local interpretabily and variables interaction issues.

4.3.2 Models interpretation for credibility

The approach we use to select the best HBS model is based on permutations, which implicitly leads to taking into account the importance of variable when building the hierarchical tree structure. Therefore, the structure

Table 5: Models comparison on the test company set

		Normalized Gini		Actual predicted ratio		MAE		RMSE	
Model	Approach	Metric	Rank	Metric	Rank	Metric	Rank	Metric	Rank
DL	cpn	0.9087134	7	0.7754874	15	269.2162	14	4976.958	13
	mkt	0.9213489	6	0.9244061	8	207.7253	5	3194.502	3
	trf	0.9247027	4	0.9665417	6	201.8016	4	3368.277	5
BST	cpn	0.9242127	5	0.8408846	13	249.4886	9	4912.252	12
	mkt	0.9389341	2	0.9745253	5	187.3897	2	3066.079	1
	trf	0.9401530	1	1.0524500	7	179.0908	1	3198.866	4
HBScateg	cpn	0.8912439	9	0.8540033	12	266.5052	13	5723.639	15
	mkt	0.9343191	3	0.9097275	9	199.3581	3	3154.323	2
	trf	0.8966275	8	0.9874937	2	246.1944	8	5125.650	14
HBScont	cpn	0.8878352	11	0.9751283	4	253.0847	10	4192.023	6
	mkt	0.8883086	10	1.1743643	14	240.3153	6	4583.056	10
	trf	0.8856011	15	1.1436449	11	241.2772	7	4577.653	9
BSbase	cpn	0.8862437	14	0.9782633	3	259.1008	12	4560.274	7
	mkt	0.8871308	13	1.0107743	1	255.5869	11	4565.516	8
	trf	0.8875098	12	0.8598436	10	275.3586	15	4584.378	11

of best HBS model selected in Table 4 can be compared with variable importance results depicted previously. We note in particular the role of the variables `cat1` and `cat2`, whose importance remains unchanged for MKT and CPN approaches. The `cat2` variable also stands out significantly for the TRF approach, which is not the case with the LightGBM model.

Additionally, Figure 5 displays the empirical distributions of fitted credibility factors for the best HBS model with categorical variables for the three approaches (CPN, MKT, TRF). Recall that the higher the probability of the coefficient being close to 1, the more significant the variable is in the construction of the hierarchical tree. For both CPN and MKT, Subfigures 5a and 5b, we observe higher credibility factors for the same variables: `cat2`, `cat3` and the third variable in the hierarchical structure. Whereas for TRF, Subfigure 5c, lower credibility factors are fitted even for `cat2` and `cat3`.

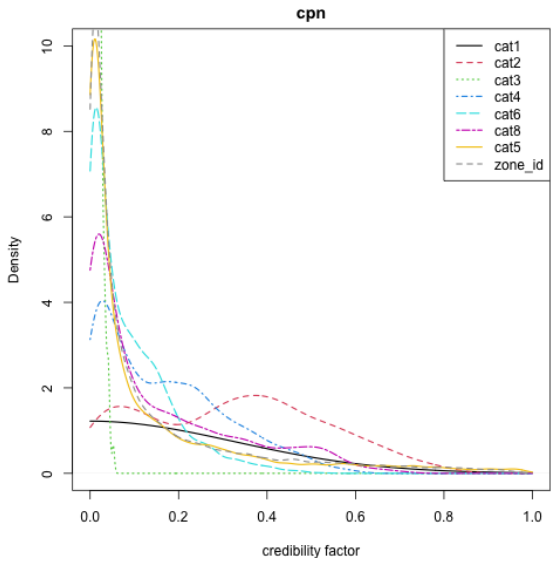
These analysis provide an empirical approach to globally measure the importance of variables on the tariff. Unlike a black box model, these analyses are directly derived from the structure of the model. In addition, its hierarchical structure and the value of the credibility coefficients allow to visualize the decision process of the algorithm and the resulting local predictions similar to a decision tree. The HBS model is therefore easily interpretable and transparent.

This approach to interpret the HBS model is however constrained by the choice of a credibility-based approach, which therefore depends on the claims history of the policyholder. From this point of view, the predictions of the model do not necessarily depend only on the importance of a variable, but also on the experience accumulated on the claims history. In some situations, the seniority of the claim is not important or recent information may better represent the current nature of the risk. Further research is needed to develop indicators that would distinguish the relationship between variables, their effect on the rate and the importance of past experience in a credibility framework.

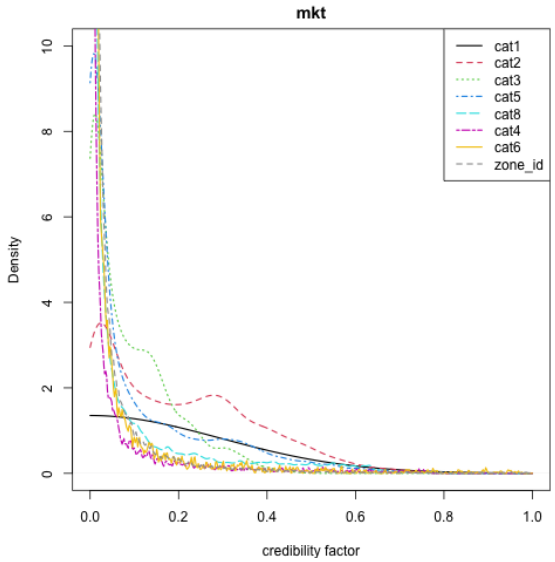
4.3.3 Predictive performance analysis

Table 5 reports the predictive performance, evaluated on the company test set, for the deep learning (DL), Light Gradient Boosting (BST) and credibility models, whereas Figure 6 displays the normalized Gini against other metrics for each model point. The columns *Approach* indicates whether the model is trained on market-only (MKT), company-only (CPN) or company data using a transfer learning approach (TRF). Again, we stress out that the predictive performance of the different approaches are carried out on the (same) test company dataset in order to be comparable.

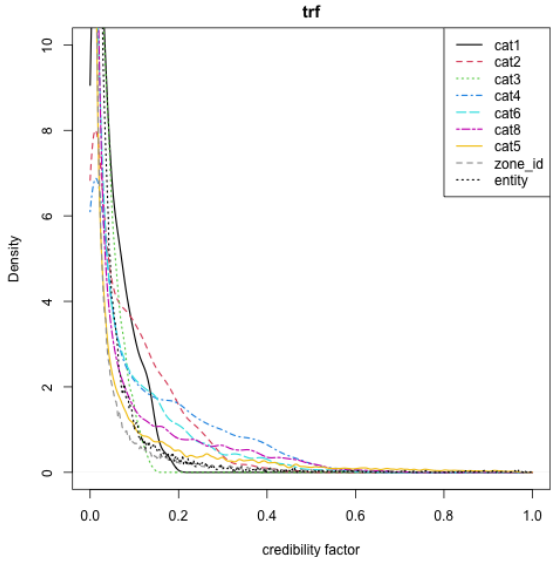
First, we see that the actual/predicted ratio is between 0.9 - 1.1 for all models, but as expected company ones are the worse. This result was indeed expected since the MKT dataset includes the company's data. We anticipate that as the test set considers a year different from the train and validation pools, the predictions



(a) CPN



(b) MKT

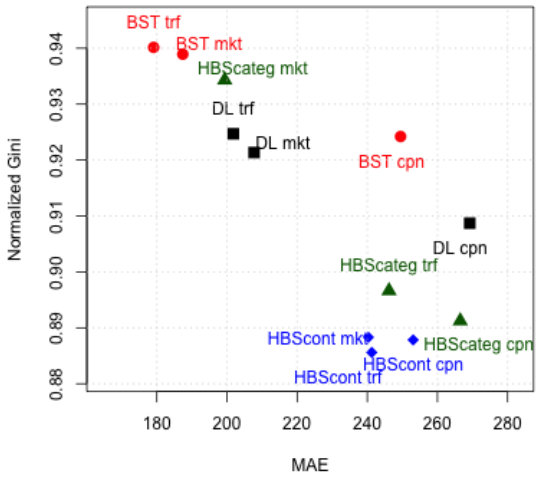


(c) TRF

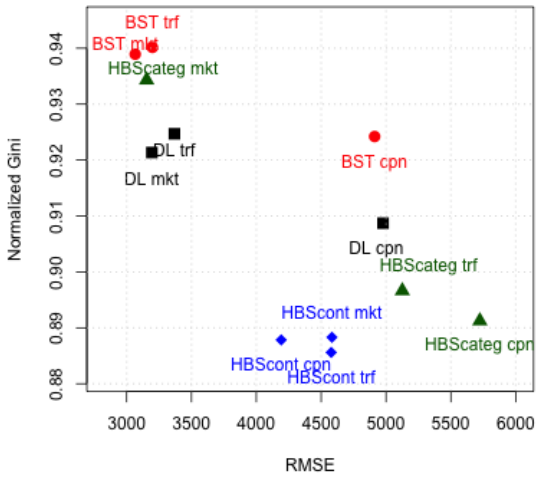
Figure 5: Empirical densities of credibility factors of best HBS models

may be structurally biased as the insured risk strongly depends by the year's context and that frequency trending is not consider in the modeling framework at all. Nevertheless, it shows that the MKT data brings in this case a superior experience than the use of only the CPN data.

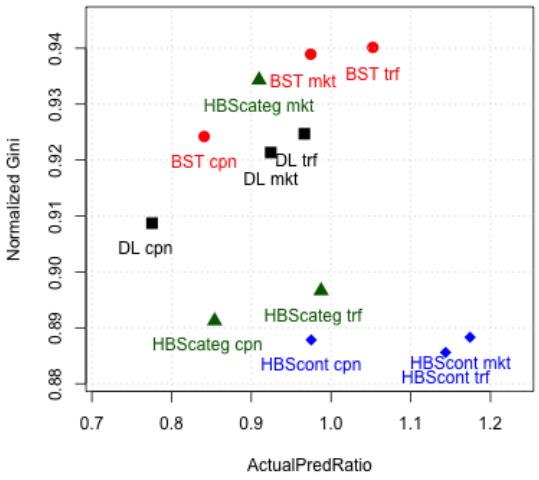
The results obtained by the LightGBM model have the best performance with the TRF and MKT approaches when measured by the Normalized Gini index and the MAE. We also note that the LightGBM with the TRF approach is the best model in terms of RMSE. The HBS model built with categorical explanatory variables performs well with the MKT approach and is the second or the third best model depending on the metric considered. Due to its non-parametric nature, this model is very flexible to adjust to different feature effects. We generally observe that combining company data with external market data will give a significant advantage in predictive performance both for ML and credibility models. However, only the LightGBM model seems to be able to exploit the TRF approach in an appropriate way. Indeed, it seems that the TRF approach penalizes the credibility methods, which can be explained by a more important weight given to the information related to the company.



(a) CPN



(b) MKT



(c) TRF

Figure 6: Normalized Gini against other metrics

Regarding the predictive accuracy, on the other hand, and especially for the DL models, we cannot rule out that the superiority of TRF approaches holds for all possible ML architectures.

5 Conclusion

Credibility theory is widely used in actuarial science to enhance an insurer's rating experience. In particular, hierarchical models allow to take into account the effect of different covariates on the premium by splitting the portfolio into different levels. They are easily interpretable and provide to actuaries a clear picture of the pricing process by classifying policyholders according to their risk and claim history. However, they are not very flexible and make it difficult to capture non-linearities or interaction effects between variables.

In this paper, we present an application of ML methods, namely the LightGBM model and a Deep-learning model, that can be compared to the hierarchical credibility approach to transfer the experience applied on a

different, but similar, book of business to a newer one. Two approaches for each model are examined: the first one consists in using directly a ML model pre-trained on market data, while the second one relies on a transfer learning logic where the pre-trained model is fitted on the insurer's data. We perform our empirical analysis transferring loss experience from an external insurance bureau to a specific company portfolio. We focus on the global predictive performance and not individual features or cluster of exposures (e.g. `zone_id`) due to the anonymized format of the data. Our approaches allow to significantly improve the prediction performance of ML models compared to a model only trained on the insurer's data. Our results show the interest and the efficiency of pre-training a ML model on a reference dataset. We also observe that HBS models perform well on the market data or the company data alone in our application, so that the transfer does not improve the prediction power compared to the MKT or the CPN approaches depending on the chosen metrics (MAE or RMSE). Finally, ML approaches obtain more competitive results compared to credibility models on this dataset. However, it can be reasonable to expect that as far as the company data increases, the advantage of the MKT and TRF approaches decreases with respect to a model trained only on company data.

Hierarchical credibility and ML models are flexible enough to deal with other types of data or business in insurance applications when reference data is available. The only disadvantage is the training-validation-test computation time which might be too high for big datasets. However applying MKT or TRF approaches should be transposed to specific context by replacing a "market"/"company" situation to e.g. "holding group"/"entity" or "company"/"line of business" situations, as in practice the loss experience of competitors remains unknown. ML models have also an practical advantage in their implementation which is relatively automated, while HBS model implementation may require a manual and expensive selection phase to derive the best features combination. Moreover, the code to train the ML models shown in this study or similar ones is readily available, cf. Appendix A.3 and also can be replicated on PCs without too much effort, should enough computational resource be provided.

Several uses of this technique seem possible for the rating of any insurance products. In fact, while our exercise is applied to agricultural insurance, in theory it can be applied to every insurance industry context where the set of ratemaking variables shared between two distinct portfolios is non-empty, holding the common ratemaking variables the same domain between the two portfolios. First, the "transfer of experience" may be performed within the same company for example when new products, tailored for niche books of business, are created. Initial losses estimates may be performed on the initial product and then applied as initial scores on the newer portfolio. A second application can be considered in reinsurance. The nature of their business allows reinsurance companies to underwrite similar risks from different primary insurers. Often, a small proportional treaty is the way to fully overview the loss experience of a new underwritten portfolio. When setting the reinsurance cover or when assisting their clients to set rates for new covers or new markets, the need of blending individual and market experiences emerges, so that reinsurers can make benchmark datasets for training ML models. However, in non-life insurance, it will be necessary to ensure that these benchmarks contain characteristics comparable to those of the insurance product to be priced in order to properly extrapolate the results, as previously anticipated.

Nevertheless, the models applied in this paper can be improved in different ways. HBS models used need categorical variables, which led us to categorize the continuous variables (and to lose information). It is an opened question if regression credibility so-called Hachemeister models could improve predictions. In addition, the computational performance of HBS models is challenging on large insurance portfolios for actuaries. For example, future research could focus on improving the variable selection process which is currently cumbersome although the model is based on explicit formula. Finally, future work can explore in how we interpret the marginal effect of explanatory variables of credibility models. A possible direction may consist in developing summary indicators based on credibility model to assess the feature importance and the role of policyholder's experience.

These connections between credibility theory and machine learning techniques open some pathways for future research. We use an empirical approach for building the hierarchical tree structure of the credibility model. A first way of improvement consists in defining the tree structure through different partitioning tree models, similarly to Diao and Weng (2019) where the partitioning algorithm directly includes credibility theory. From there, it is natural to consider that such a credibility regression tree can be applied to other ensemble decision tree algorithms, such as boosted trees. It will be interesting to measure the interest of an approach based on transfer learning on this type of models.

6 Acknowledgments

This work has been sponsored by the Casualty Actuarial Society (CAS) and the Society of Actuaries (SOA) Individual Grants Competition for 2020. The authors wish to give a special thanks to CAS research and publication staff for their support.

The authors are also very grateful for the useful suggestions of the two anonymous referees, which led to significant improvements of this article. The remaining errors, of course, should be attributed to the authors alone.

This paper also benefits from fruitful discussions with members of the French chairs DIALog (Digital Insurance And Long-term risks) and RE2A (Emerging or atypical risks in Insurance), two joint initiatives under the aegis of the Fondation du Risque.

A Appendix

A.1 Rating bureaus

According to IRMI (2022), a Rating Bureau is "an organization that collects statistical data (such as premiums, exposure units, and losses), computes advisory rating information, develops standard policy forms, and files information with regulators on behalf of insurance companies that purchase its services". The use of their services has become progressively less compulsory in recent decades, and their activities have become increasingly consultative: single carries may purchase their data collection service and decide whether and how to use them. In the US market, the most relevant rating bureaus are: NCCI (for Worker Compensation, WC), the Insurance Service Office (ISO), that serves most personal and commercial lines, the Surety Association of America (SAA) that operates in the surety and crime insurance, and the American Association of Insurance Services (AAIS) specialized in many commercial lines different from WC.

A.2 Usual metrics

Consider a set of observations y_i and its corresponding predictions $\hat{y}_i$ for $i = 1, \ldots, n$. The MAE and RMSE metrics used are

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \; RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

These values represent the absolute, the squared norms of residual vectors.

A.3 Code

The modeling is performed using both `R` (R Core Team 2022) and `Python` (Van Rossum and Drake 2009).

The GitHub Repo provides the full code used for the computations as well as an extract of the datasets used (with 150 `zone_id` randomly chosen).

A.4 Data Preparation and Anonymization

The market and company data files are loaded. An initial renaming of the variable is performed, conventionally naming the continuous one as `cont_x` while the categorical one as `cat_x`, being `x` a number from one up to the number of variables of such category. The following criterion is used to filter out anomalous observations: presence of missing values in any of the observations, zero exposures.

Then, the available data is split threefold: the last available year has been set to the test set, while the remaining years have been split into a train / validation set using a 80/20 ratio. Therefore we have available three dataset for the marked data, and another three for the company one.

A.5 Other descriptive statistics

Tables 6 and 7 give descriptive statistics for `cat1` and `zone_id` which have a large number of levels.

Table 6: Frequency tables for zone id variable.

	Market most frequent levels							Company Market most frequent levels					
Nb. levels	2036	2835	2779	2812	2805	2839	Nb. levels	2779	2852	2835	2812	2805	2839
2710	0.00484	0.0057	0.00587	0.00608	0.00831	0.0085	5207	0.00252	0.00276	0.003	0.00331	0.00372	0.00373

Table 7: Frequency tables for cat1 variable.

	Market most frequent levels							Company Market most frequent levels					
Nb. levels	158	245	167	119	154	273	Nb. levels	120	245	167	119	154	273
170	0.03604	0.04879	0.05574	0.06342	0.08577	0.19843	281	0.04085	0.04758	0.04824	0.07063	0.08456	0.11705

A.6 Parameter estimation in HBS

Let $g \in \{1,\dots,G\}$, $h \in \{1,\dots,H\}$ $i \in \{1,\dots,I\}$, $j \in \{1,\dots,J_{g,h,i}\}$. We define index subsets I_h and H_g given the father index (h and g resp.) by $I_h = \{i, \Theta_i \in \Theta(\Phi_h)\}$ and $H_g = \{h, \Phi_h \in \Phi(\Psi_g)\}$. The parameters $\widehat{\alpha_g^{(3)}}$, $\widehat{\alpha_{g,h}^{(2)}}$, $\widehat{\alpha_{g,h,i}^{(1)}}$, $\widehat{B_g^{(3)}}$, $\widehat{B_{g,h}^{(2)}}$ and $\widehat{B_{g,h,i}^{(1)}}$ of the HBS presented in Section 2 are given in Table 8, see Theorem 6.4 of Bühlmann and Gisler (2006) for details. We refer to Section 6.6 of Bühlmann and Gisler (2006) for the estimators $\hat{\tau}_1^2$, $\hat{\tau}_2^2$, $\hat{\tau}_3^2$ of structural parameters τ_1^2, τ_2^2, τ_3^2.

Credibility factors			Weighted means		
$\widehat{\alpha_g^{(3)}}$	$\widehat{\alpha_{g,h}^{(2)}}$	$\widehat{\alpha_{g,h,i}^{(1)}}$	$\widehat{B_g^{(3)}}$	$\widehat{B_{g,h}^{(2)}}$	$\widehat{B_{g,h,i}^{(1)}}$
$\frac{w_g^{(3)}}{w_g^{(3)}+\frac{\hat{\tau}_2^2}{\hat{\tau}_3^2}}$	$\frac{w_h^{(2)}}{w_h^{(2)}+\frac{\hat{\tau}_1^2}{\hat{\tau}_2^2}}$	$\frac{w_{i,\cdot}}{w_{i,\cdot}+\frac{\hat{\sigma}^2}{\hat{\tau}_1^2}}$	$\sum_g \frac{\widehat{\alpha_{g,h}^{(2)}}}{w_g^{(3)}} \widehat{B_{g,h}^{(2)}}$	$\sum_{i\in I_h} \frac{\widehat{\alpha_{g,h,i}^{(1)}}}{w_h^{(2)}} \widehat{B_{g,h,i}^{(1)}}$	$\sum_j \frac{w_{i,j}}{w_{i,\cdot}} X_{i,j}$
Other struc. param.			Weights		
$\hat{\mu}_4$			$w_{i,\cdot}$	$w_g^{(3)}$	$w_h^{(2)}$
$\sum_g \frac{\widehat{\alpha_g^{(3)}}}{w^{(4)}} \widehat{B_g^{(3)}}$			$\sum_j w_{i,j}$	$\sum_{h\in H_g} \widehat{\alpha_{g,h}^{(2)}}$	$\sum_{i\in I_h} \widehat{\alpha_{g,h,i}^{(1)}}$

Table 8: HBS parameter estimators

A.7 Generalized linear models (GLM)

GLMs, e.g. McCullagh and Nelder (1989), rely on probability distribution functions of exponential type for the response variable. The likelihood L associated to the statistical experiment generated by Y_i, $i \in I$, verifies

$$\log L(\theta \,|\, y_i) = \frac{\lambda_i(\theta) y_i - b\left(\lambda_i(\theta)\right)}{a(\phi)} + c(y_i, \phi), \quad y_i \in \mathbb{Y} \subset \mathbb{R},$$

and $-\infty$ if $y_i \notin \mathbb{Y}$, where $a : \mathbb{R} \to \mathbb{R}$, $b : \Lambda \to \mathbb{R}$ and $c : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}$ are known real-valued measurable functions and ϕ is the dispersion parameter. Table 9 gives four classic examples of probability distribution in the exponential family characterized by a, b, c and $\mathbb{Y}$. Typical application of GLMs in insurance include claim frequency modeling via the Poisson distribution, claim severity modeling via the gamma distribution, rate modeling via the normal distribution and claim fraud modeling via the Bernoulli distribution.

Distribution	$\lambda(\theta)$	ϕ	$a(x)$	$b(x)$	$c(x,\phi)$
Bernoulli $\mathcal{B}(\theta)$	$\log(\frac{\theta}{1-\theta})$	1	x	$\log(1+e^x)$	0
Gaussian $\mathcal{N}(\theta,\sigma^2)$	θ	σ^2	x	$x^2/2$	x^2/ϕ $-\frac{1}{2}\log(2\pi\phi)$
Gamma $\mathcal{G}(\nu,\theta)$	$\frac{-1}{\theta}$	$1/\nu$	x	$-\log(-x)$	$\frac{\log(x/\phi)}{\phi} - \log(x)$ $-\log(\Gamma(\frac{1}{\phi}))$
Poisson $\mathcal{P}(\theta)$	$\log(\theta)$	1	x	e^x	$-\log(x!)$

Table 9: Usual distributions in the exponential family

A.8 DL structure

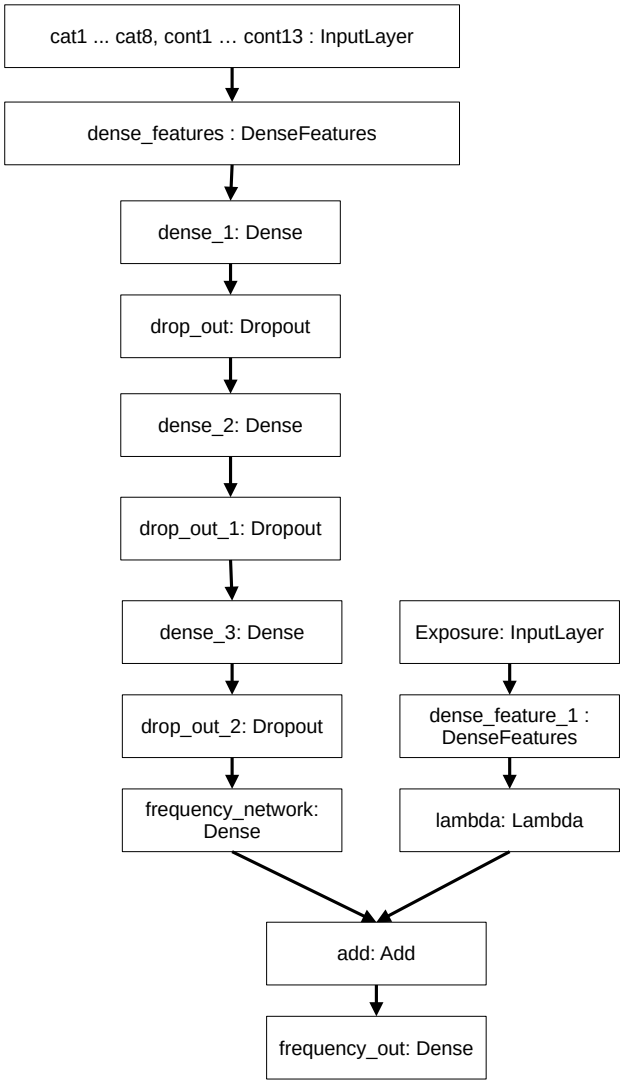


Figure 7: DL model structure

References

Ahcan, Ales, Darko Medved, Annamaria Olivieri, and Ermanno Pitacco. 2014. "Forecasting Mortality for Small Populations by Mixing Mortality Data." *Insurance: Mathematics and Economics* 54: 12–27. https://doi.org/10.1016/j.insmatheco.2013.10.013.

Antonio, Katrien, and Jan Beirlant. 2007. "Actuarial Statistics with Generalized Linear Mixed Models." *Insurance: Mathematics and Economics* 40 (1): 58–76. https://doi.org/10.1016/j.insmatheco.2006.02.013.

Bergstra, James, Daniel Yamins, and David Cox. 2013. "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures." In *International Conference on Machine Learning*, 115–23. PMLR. https://dl.acm.org/doi/10.5555/3042817.3042832.

Blier-Wong, Christopher, Jean-Thomas Baillargeon, Hélène Cossette, Luc Lamontagne, and Etienne Marceau. 2021. "Rethinking representations in P&C actuarial science with deep neural networks." *arXiv:2102.05784 [Stat]*, February. http://arxiv.org/abs/2102.05784.

Blier-Wong, Christopher, Hélène Cossette, Luc Lamontagne, and Etienne Marceau. 2021. "Machine Learning in P&C Insurance: A Review for Pricing and Reserving." *Risks* 9 (1): 4. https://doi.org/10.3390/risks9010004.

Bozikas, Apostolos, and Georgios Pitselis. 2019. "Credible regression approaches to forecast mortality for populations with Limited Data." *Risks* 7 (1): 27. https://doi.org/10.3390/risks7010027.

Breiman, Leo. 2017. *Classification and Regression Trees*. CRC Press.

Bühlmann, Hans, and Alois Gisler. 2006. *A Course in Credibility Theory and its Applications*. Springer Science & Business Media. https://doi.org/10.1007/3-540-29273-X.

Bühlmann, Hans, and Erwin Straub. 1970. "Glaubwürdigkeit für Schadensätze." *Bulletin of the Swiss Association of Actuaries* 70 (1): 111–33.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–94. https://dl.acm.org/doi/10.1145/2939672.2939785.

Cheng, Xiaojuan, Wei Luo, Guojun Gan, and Gang Li. 2019. "Fast valuation of large portfolios of variable annuities via transfer learning." In *PRICAI 2019: Trends in Artificial Intelligence*, 716–28. Springer, Cham. https://doi.org/10.1007/978-3-030-29894-4_57.

Chollet, François et al. 2018. "Keras: The Python deep learning library." *Astrophysics Source Code Library*, ascl–1806. https://ascl.net/1806.022.

Danzon, Patricia Munch. 1983. "Rating bureaus in U.S. property liability insurance markets: Anti or pro-competitive?" *The Geneva Papers on Risk and Insurance - Issues and Practice* 8 (4): 371–402. https://doi.org/10.1057/gpp.1983.42.

Delcaillau, Dimitri, Antoine Ly, Alize Papp, and Franck Vermet. 2022. "Model Transparency and Interpretability: Survey and Application to the Insurance Industry." *European Actuarial Journal* 12 (2): 443–84. https://doi.org/10.1007/s13385-022-00328-y.

Diao, Liqun, and Chengguo Weng. 2019. "Regression Tree Credibility Model." *North American Actuarial Journal* 23 (2): 169–96. https://doi.org/10.1080/10920277.2018.1554497.

Douvillé, Cécile. 2004. "Tarification Des Risques Industriels Par Le Modèle de Crédibilité : Prise En Compte de La Taille Des Risques Extension à l'assurance Des Pertes d'exploitation." *Bulletin Français d'Actuariat* 6 (12).

Dutang, Christophe, Vincent Goulet, and Mathieu Pigeon. 2008. "actuar: An R Package for Actuarial Science." *Journal of Statistical Software* 25 (7): 38. https://doi.org/10.18637/jss.v025.i07.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. "A Working Guide to Boosted Regression Trees." *Journal of Animal Ecology* 77 (4): 802–13. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Elsken, Thomas, Jan Hendrik Metzen, and Frank Hutter. 2019. "Neural architecture search: A survey." *Journal of Machine Learning Research* 20 (55): 1–21. https://jmlr.org/papers/v20/18-598.html.

Ferrario, Andrea, and Roger Hämmerli. 2019. "On Boosting: Theory and Applications." https://dx.doi.org/10.2139/ssrn.3402687.

Ferrario, Andrea, Alexander Noll, and Mario V. Wuthrich. 2020. "Insights from Inside Neural Networks." https://dx.doi.org/10.2139/ssrn.3226852.

Frees, Edward W. (Jed), Glenn Meyers, and A. David Cummings. 2014. "Insurance Ratemaking and a Gini Index." *The Journal of Risk and Insurance* 81 (2): 335–66. https://www.jstor.org/stable/24546807.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 1189–1232. https://www.jstor.org/stable/2699986.

Goldburd, Mark, Anand Khare, and Dan Tevet. 2016. *Generalized Linear Models for Insurance Rating*. 5. https://doi.org/10.2307/1270349.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT press.

Goulet, Vincent. 1998. "Principles and Application of Credibility Theory." *Journal of Actuarial Practice* 6 (18).

Goulet, Vincent, Christophe Dutang, Xavier Milhaud, and Mathieu Pigeon. 2021. "Credibility Theory Features of Actuar." Vignette of the actuar package.

Gursky, Jacob. 2020. "Boosting showdown: Scikit-Learn vs XGBoost vs LightGBM vs CatBoost in sentiment classification." https://towardsdatascience.com/.

Hachemeister, Charles A et al. 1975. "Credibility for Regression Models with Application to Trend." In *Credibility, Theory and Applications, Proceedings of the Berkeley Actuarial Research Conference on Credibility, Academic Press, New York*, 129–63.

Hanafy, Mohamed, and Ruixing Ming. 2021. "Machine Learning Approaches for Auto Insurance Big Data." *Risks* 9 (2): 42.

Henckaerts, Roel, Marie-Pier Côté, Katrien Antonio, and Roel Verbelen. 2021. "Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods." *North American Actuarial Journal* 25 (2): 255–85.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–80.

IRMI. 2022. "Rating Bureau." https://www.irmi.com/term/insurance-definitions/rating-bureau.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems* 30: 3146–54. https://dl.acm.org/doi/10.5555/3294996.3295074.

Kuo, Kevin. 2019. "DeepTriangle: A Deep Learning Approach to Loss Reserving." *Risks* 7 (3). https://doi.org/10.3390/risks7030097.

Kuo, Kevin, Bob Crompton, and Frankie Logan. 2019. "Deep Learning and Actuarial Experience Analysis." *Compact*.

Lee, Ronald D., and Lawrence R. Carter. 1992. "Modeling and forecasting U. S. mortality." *Journal of the American Statistical Association* 87 (419): 659–71. https://doi.org/10.2307/2290201.

Li, Hong, and Yang Lu. 2018. "A Bayesian Non-Parametric Model for Small Population Mortality." *Scandinavian Actuarial Journal* 2018 (7): 605–28. https://doi.org/10.1080/03461238.2017.1418420.

Lorentzen, Christian, and Michael Mayer. 2020. "Peeking into the black box: An actuarial case study for interpretable machine learning." SSRN Scholarly Paper ID 3595944. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3595944.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–74. Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3295222.3295230.

Maïssae, Haddouchi. 2020. *ForestDisc: Forest Discretization*. https://CRAN.R-project.org/package=ForestDisc.

Matthews, Spencer, and Brian Hartman. 2022. "Machine Learning in Ratemaking, an Application in Commercial Auto Insurance." *Risks* 10 (4): 80.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. Chapman; Hall.

Meel, Vidushi. 2021. "YOLOv3: Real-time object detection algorithm." https://viso.ai/deep-learning/yolov3-overview/.

Nahon, Aviv. 2019. "XGBoost, LightGBM or CatBoost — which boosting algorithm should I use?" https://medium.com/riskified-technology/xgboost-lightgbm-or-catboost-which-boosting-algorithm-should-i-use-e7fda7bb36bc.

Nigri, Andrea, Susanna Levantesi, Mario Marino, Salvatore Scognamiglio, and Francesca Perla. 2019. "A Deep Learning Integrated Lee–Carter Model." *Risks* 7 (1). https://doi.org/10.3390/risks7010033.

Noll, Alexander, Robert Salzmann, and Mario V. Wuthrich. 2020. "Case Study: French Motor Third-Party Liability Claims." https://dx.doi.org/10.2139/ssrn.3164764.

Norberg, Ragnar. 2004. "Credibility Theory." *Encyclopedia of Actuarial Science* 1: 398–406. https://doi.org/10.1002/9780470012505.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-learn: Machine learning in Python." *The Journal of Machine Learning Research* 12: 2825–30. https://dl.acm.org/doi/10.5555/1953048.2078195.

Porter, Karen, and CPCU. 2008. *Insurance Regulation*. American Institute for Chartered Property Casualty Underwriters. https://books.google.it/books?id=ob5fPgAACAAJ.

Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2017. "CatBoost: Unbiased Boosting with Categorical Features." *arXiv Preprint arXiv:1706.09516*. https://dlnext.acm.org/doi/abs/10.5555/3327757.3327770.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reback, Jeff, Wes McKinney, Joris Den Van Bossche, Tom Augspurger, Phillip Cloud, Adam Klein, Matthew Roeschke, et al. 2020. "pandas-dev/pandas: Pandas 1.0.3." https://doi.org/10.5281/zenodo.3715232.

Rentzmann, Simon, and Mario V. Wuthrich. 2019. "Unsupervised Learning: What Is a Sports Car?" https://dx.doi.org/10.2139/ssrn.3439358.

Richman, Ronald. 2021a. "AI in Actuarial Science – a Review of Recent Advances – Part 1." *Annals of Actuarial Science* 15 (2): 207–29. https://doi.org/10.1017/S1748499520000238.

———. 2021b. "AI in Actuarial Science – a Review of Recent Advances – Part 2." *Annals of Actuarial Science* 15 (2): 230–58. https://doi.org/10.1017/S174849952000024X.

Richman, Ronald, and Mario V. Wuthrich. 2019. "Lee and Carter Go Machine Learning: Recurrent Neural Networks." https://dx.doi.org/10.2139/ssrn.3441030.

Schelldorfer, Jürg, and Mario V. Wuthrich. 2019. "Nesting Classical Actuarial Models into Neural Networks." *Available at SSRN 3320525*. https://dx.doi.org/10.2139/ssrn.3320525.

Spedicato, Giorgio Alfredo, Christophe Dutang, and Leonardo Petrini. 2018. "Machine Learning Methods to Perform Pricing Optimization. A Comparison with Standard GLMs." *Variance* 12 (1): 69–89.

Tsai, Cary Chi-Liang, and T Lin. 2017. "Incorporating bühlmann Credibility Approach to Improving Mortality Forecasting." *Scandinavian Actuarial Journal* 2017: 419–40. https://doi.org/10.1080/27658449.2021.2023979.

Tsai, Cary Chi-Liang, and Adelaide Di Wu. 2020. "Incorporating Hierarchical Credibility Theory into Modelling of Multi-Country Mortality Rates." *Insurance: Mathematics and Economics*, January. https://doi.org/10.1016/j.insmatheco.2020.01.001.

Tsai, Cary Chi-Liang, and Ying Zhang. 2019. "A Multi-Dimensional Bühlmann Credibility Approach to Modeling Multi-Population Mortality Rates." *Scandinavian Actuarial Journal* 2019 (5): 406–31. https://doi.org/10.1080/03461238.2018.1563911.

Van Rossum, Guido, and Fred L Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. https://dl.acm.org/doi/book/10.5555/1593511.

Xacur, Oscar Alberto Quijano, and José Garrido. 2018. "Bayesian Credibility for GLMs." *Insurance: Mathematics and Economics* 83: 180–89. https://doi.org/10.1016/j.insmatheco.2018.05.001.

Yan, Jun, James Guszcza, Matthew Flynn, and Cheng-Sheng Peter Wu. 2009. "Applications of the Offset in Property-Casualty Predictive Modeling." In *Casualty Actuarial Society E-Forum, Winter 2009*, 366.

Zhang, Tong, and Bin Yu. 2005. "Boosting with Early Stopping: Convergence and Consistency." *The Annals of Statistics* 33 (4): 1538–79. https://doi.org/10.1214/009053605000000255.

Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. "A comprehensive survey on transfer learning." *Proceedings of the IEEE* 109 (1): 43–76. https://doi.org/10.1109/JPROC.2020.3004555.