Dutang, Christophe, Giorgio Alfredo Spedicato, and Quentin Guibert. 2025. "Adjusting Manual Rates to Own Experience: Comparing the Credibility Approach to Machine Learning." *Variance* 18 (June). https://doi.org/10.66573/001c.138733.

Ratemaking and Product Information

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

Christophe Dutang[1a], Giorgio Alfredo Spedicato[2b], Quentin Guibert[3c]

[1] Laboratory Jean Kuntzmann, [2] Leithà, [3] Université Paris-Dauphine

Keywords: Transfer learning, Hierarchical credibility theory, Bühlmann-Straub credibility theory, Boosting, Deep learning

https://doi.org/10.66573/001c.138733

Variance

Vol. 18, 2025

Credibility theory is the usual framework in actuarial science for reinforcing individual experience by transferring rates estimated from collective information. Based on the paradigm of transfer learning, this article presents the idea that a machine learning (ML) model pretrained using a rich market data portfolio improves rate predictions for individual insurance portfolios. This framework first trains several ML models on a market portfolio of insurance data. The pretrained models provide valuable information on relationships between features and predicted rates. Furthermore, features shared with the company dataset improve rate prediction compared with the same ML models trained on the insurer's dataset alone. We applied classical ML models using an anonymized dataset that included both market data and data from a European non-life insurance company; this approach is comparable with a hierarchical Bühlmann-Straub credibility model. We observed that the transfer learning strategy of combining company data with external market data significantly improved prediction accuracy compared with an ML model trained on the insurer's data alone and provided competitive results compared with those of hierarchical credibility models.

This work has been sponsored by the Casualty Actuarial Society (CAS) and the Society of Actuaries (SOA) Individual Grants Competition for 2020.

Address for Correspondence: dutangc@gmail.com

1. INTRODUCTION

Using market data to aid in setting own company rates has been a common practice in the insurance industry. External data, as provided by insurance rating bureaus (e.g., Appendix A.1), reinsurers, or advisory organizations, supplement internal company data that may be scarce or unreliable because it is not representative, has a short history, or does not yet exist (e.g., when a new business line or a new territory is entered). According to Porter and American Institute for Chartered Property Casualty Underwriters (2008), pools of insurers in the US existed as early as the second part of the 19th century. These pools supported their members in setting rates through 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 National Association of Insurance Commissioners (NAIC) laws explicitly allowed cooperation in setting rates and specified the role of rating organizations. The importance of external data has been historically recognized by regulators to support adequate rates that preserve a company's solvency,

a Christophe Dutang is an assistant professor at Laboratory Jean Kuntzmann and teaches at Ensimag, both in Grenoble. At LJK, his research focuses on actuarial science, extreme risks, and regression models. He is also a member of the French Institute of Actuaries. He is an active member of the R community, maintaining and authoring several R packages. Previously, he was assistant professor at Université Paris-Dauphine. He completed his PhD in applied mathematics at ISFA (Université Lyon 1).

b Giorgio Alfredo Spedicato is an actuarial data science manager at Leithà, an internal consulting firm within the Unipol Group in Italy. Prior to joining Unipol, he worked as a P&C pricing and reserving actuary at the Italian branches of AXA and Aviva Holdings. Giorgio holds a PhD in actuarial science and the FCAS, FSA, CPSA, and C.Stat professional credentials. He is the author of multiple peer-reviewed papers in the fields of actuarial science and medical statistics.

c Quentin Guibert is an assistant professor at the Université Paris-Dauphine and a member of the French Institute of Actuaries. His research focuses on issues encountered in actuarial science and statistical risk modeling in the insurance sector. He has also worked on topics related to risk management and insurance regulation. Previously, he served as an adjunct professor (PAST) at the Université Paris-Dauphine. He completed his PhD in management science at ISFA (Institute of Financial and Insurance Sciences, Université Lyon 1), followed by a postdoctoral fellowship at the same university funded by the ANR Lolita project.

prevent excessive competition, and ease the entrance of new players (e.g., granting a partial antitrust law exemption in the US jurisdiction; Danzon 1983).

When the insurer takes its own experience into account to enhance the credibility of its rates, it must benchmark its portfolio experience compared with that of the market. The actuarial profession traditionally used techniques based on Bayesian statistics and nonparametric credibility to optimally integrate market and insurer portfolio experiences into the technical rates. From this viewpoint, the market data contribution satisfies the two classic approaches addressed by credibility theory: the limited fluctuation credibility theory and the greatest accuracy credibility theory (e.g., Norberg 2004). The former refers to incorporating individual experience into the rate calculation to stabilize the level of individual rates. The latter corresponds to applying modern credibility theory and combines both individual and collective experiences to predict individual rates by minimizing mean square error.

Credibility theory is extensively used in non-life insurance. Early models were not based on policyholders' ratemaking variables (see, e.g., Bühlmann and Gisler (2006) for a comprehensive presentation). The actuarial literature has proposed some advanced regression credibility models, such as the Hachemeister model (Hachemeister and others 1975). Conversely, rates based on generalized linear models (GLMs), the current gold standard in personal rate pricing (Goldburd, Khare, and Tevet 2016), are based solely on the impact of ratemaking factors, giving no credit to the individual policy experience.

Nevertheless, mixed effects GLMs allow incorporating policyholders' experience within the GLM tariff structure (Xacur and Garrido 2018; Antonio and Beirlant 2007); however, they are not widely 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 in 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 from the French association of private and mutual insurers. In other countries, to the best of the authors' knowledge, public insurance bureaus exist in Italy, Germany, the UK, and Brazil. The Italian Association of Insurers aggregates data from the motor lines (but only pure premiums with few covariates) and long-term care insurance for health, while it collects extensive statistics (with many covariates) for crop insurance. The German Insurance Association provides data similar to that of the Italian association for many lines. The industry data and subscription section of the Association of British Insurers provides (at least) yearly aggregate data for many property and casualty, health, and life line of business categories. Finally, the Brazilian Insurance Regulator (SUSEP) provides aggregated losses and exposures for motor liability insurance aggregated by key rating variables.

Credibility theory is also largely used in life insurance applications for modeling mortality risks. A first attempt to stabilize mortality rates by combining the mortality data of a small population with the average mortality of the neighboring populations was proposed by Ahcan et al. (2014). Regarding this issue of limited mortality data (small population or short historical observation period), Li and Lu (2018) introduced a Bayesian nonparametric model for benchmarking a small population compared with a reference population. Bozikas and Pitselis (2019) focused on a credible regression framework to efficiently forecast populations with a short base period. To improve mortality forecasting, some recent contributions in the literature combine the usual mortality models, such as the Lee and Carter (1992) model and the Bühlmann credibility theory (Bühlmann and Gisler 2006; see also, Tsai and Lin 2017; Tsai and Zhang 2019, and Tsai and Wu 2020, among others).

The recent widespread use of machine learning (ML) has provided many more techniques for practitioner actuaries. Gradient boosting models (GBMs) and deep learning models (DLs) for motor third-party liability pricing are presented 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) showed that the random forest technique is more efficient for predicting claim occurrence (in terms of accuracy, kappa, and area under the curve values) than are logistic regression, XGBoost, decision trees, naive Bayes, and k-nearest neighbors algorithm. Matthews and Hartman (2022) compared random forest, GBM, and DL against GLM to predict claim amount and claim frequency for commercial auto insurance and demonstrated its efficiency and accuracy for future ratemaking models. Henckaerts et al. (2021) also showed that GBMs outperform classical GLMs and allow insurers to form profitable portfolios and guard against potential adverse risk selection. Furthermore, researchers have applied non-pricing approaches. For example, Spedicato, Dutang, and Petrini (2018) modeled policyholder behavior; Rentzmann and Wuthrich (2019) presented recent advances in unsupervised learning for vehicle classification as DL autoencoders; Kuo (2019) used the NAIC reserving dataset to show how neural networks with embedding may offer better prediction on tabular loss development triangles. Blier-Wong, Cossette, et al. (2021) conducted a comprehensive review of ML in property and casualty studies. On the life insurance side, applying DL to lapse modeling (Kuo, Crompton, and Logan 2019) and the DL version (Richman and Wuthrich 2019) of the classical Lee-Carter model are worth mentioning. For a more comprehensive review, see Richman (2021b, 2021a). Recently, Diao and Weng (2019) combined credibility and regression tree models. In these presentations, the ultimate goal of using ML was to improve the usual regression setup in actuarial science based on the GLM. However, techniques such as GBM and DL can also be used to transfer what the model learned on a much bigger dataset (e.g., market data) to a smaller dataset (e.g., company portfolio data). *Transfer learning* reuses knowledge learned from different data sources to improve learner performance. This ML area

has become particularly popular in recent years, especially 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 also be used in other areas, for instance, to train life insurance valuation models (Cheng et al. 2019).

Our study aim was to take advantage of ML to more easily handle complex nonlinear relationships, as compared with standard credibility-based approaches, to more accurately assess policyholder risk. Hence, our work contrasts traditional methods with those of ML for blending market data into individual portfolio experience. First, we anticipated a difference between the credibility approach and the ML used in this study. The credibility approach naturally uses a longitudinal structure to calibrate its parameters, but this is not a prerequisite for ML models, which only need to share some variables. We applied our approach on a (properly anonymized) dataset comprising both market and own portfolio experience from a European non-life insurance pool. Final comparisons focused not only on predictive performance, but also on practical applicability in terms of computational demand, ease of understanding, and interpretability of results.

The rest of this paper is organized as follows. We review the hierarchical Bühlmann-Straub (HBS) credibility model in Section 2. Section 3 presents the main ML algorithms used in this paper. Section4 compares the performance between ML and HBS models based on a market dataset and a company dataset, and Section 5 concludes this paper.

2. HIERARCHICAL CREDIBILITY MODEL

This section briefly describes the hierarchical credibility theory of Bühlmann and Gisler (2006), which we used to model 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 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 individual/policyholder clusters are repeatedly observed for a given period. Generally, the company data experience is shorter than that of the market. In addition, we assumed that market and company datasets share the same features, which allows them 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 used a five-level model structured in a hierarchical tree, as presented in Figure 1, with the usual notation. The five levels are presented from top to 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 introduced parameters related to this risk level, Ψ_g, $g = 1, \dots, G$, which are independent and identically distributed.
- Level 2: Each risk class is divided into sectors. Given Ψ_g, we denote by $\Phi_{g,h}$, $h = 1, \dots, H$ the class risk parameters, which are assumed to be conditionally independent and identically distributed.
- Level 1: Given $\Phi_{g,h}$, $\Theta_{g,h,i}$, $i = 1, \dots, I$ are the individual risk parameters, which are conditionally independent and identically distributed.
- Level 0: Given $\Theta_{g,h,i}$, data are available during the study period $[1, J_{g,h,i}]$. We denote by $\mathbf{Y}_{g,h,i} = (Y_{g,h,i,1}, \dots, Y_{g,h,i,J_{g,h,i}})$, the vector of observations over years, which are conditionally independent, identically distributed, and have a finite variance. We also introduce a vector for the relative known weights $\mathbf{w}_{g,h,i} = (w_{g,h,i,1}, \dots, w_{g,h,i,J_{g,h,i}})$ over the same observation period.

In Section 4.2, the class variable related to Level 2 results from a combination of several categorical variables. These variables comprise unobservable risk factors that allow partitioning of the data space. Seven variables 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.

To estimate credibility rates, we defined the following notations and structural parameters for $i = 1, \dots, I$ and $j = 1, \dots, J_{g,h,i}$:

- 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 $\sigma_0^2 = E[\sigma_1^2(\Theta_{g,h,i}))]$.

Similar 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 obtained 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 the credibility factor formulas $\widehat{\alpha_g^{(3)}}$, $\widehat{\alpha_{g,h}^{(2)}}$, $\widehat{\alpha_{g,h,i}}^{(1)}$ 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. The weighted means depend on structural parameters that can easily be estimated nonparametrically. Therefore,

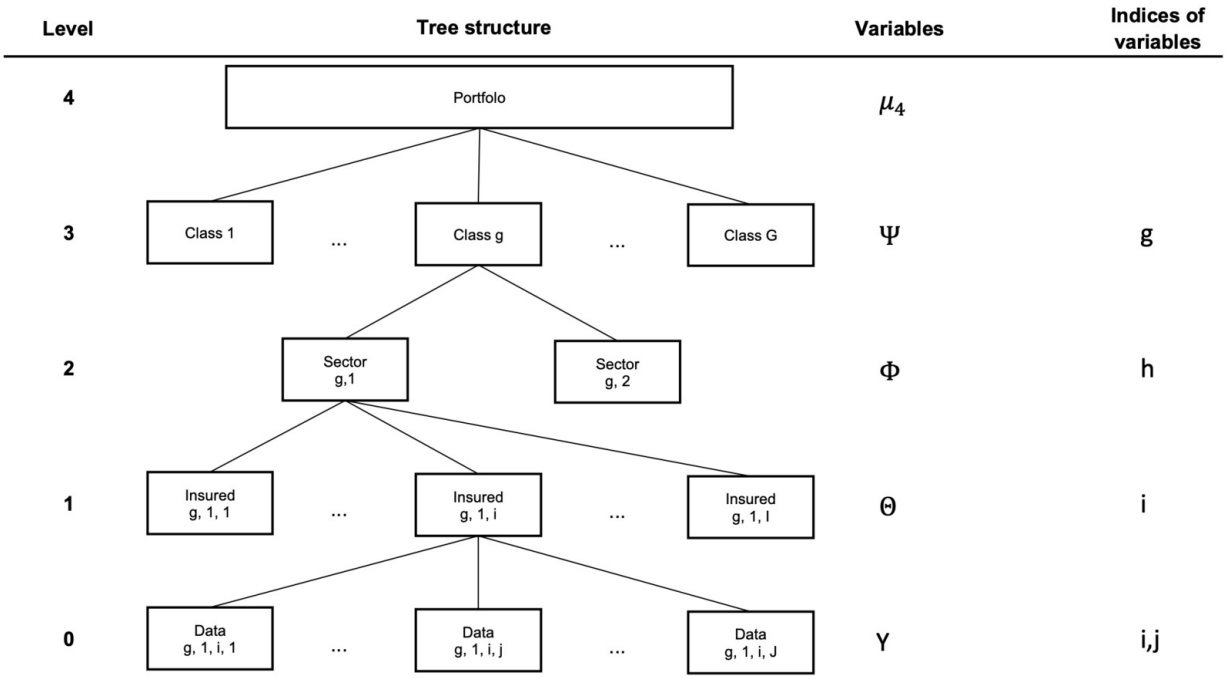


Figure 1. Representation of a five-level hierarchical tree structure.

an HBS model provides a recursive computation of weighted empirical means whose parameters minimize quadratic losses. There are no distribution assumptions when deriving estimators; thus HBS models are full nonparametric models.

3. MODELING APPROACH WITH TRANSFER LEARNING

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

3.1. TRANSFER LEARNING

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

The TRF approach relies on experience from the MKT approach and used the corresponding pretrained model as a starting point. Next, we fine-tuned the MKT model based on experience from the CPN dataset. The resulting model contained both information from the market and the company, and it should offer better predictions.

3.2. MACHINE LEARNING MODELS

Next, we focused on ML models that permit an initial estimate of losses performed on another set via transfer learning. While we explored the use of such approaches applying ML methods, traditional GLMs may be used as well (see Appendix A.7 for a brief introduction). GLMs can perform log-linear regressions to estimate both the frequency and severity of the claim. These outputs can be used as offsets in subsequent models. For instance, under a log-linear regression framework and initial log estimate of 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 for insurance pricing are strongly nonlinear and can automatically find interactions among ratemaking factors while excluding nonrelevant features. In particular, two techniques are acquiring widespread importance, boosting and deep learning. Both techniques allow the use of an initial estimate of loss or exposure to risk to train the model on previous observations.

All 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

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

where x_i is the covariates' vector and v_i is 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 $\lambda(.)$ by minimizing the Poisson loss function, typically on the test set

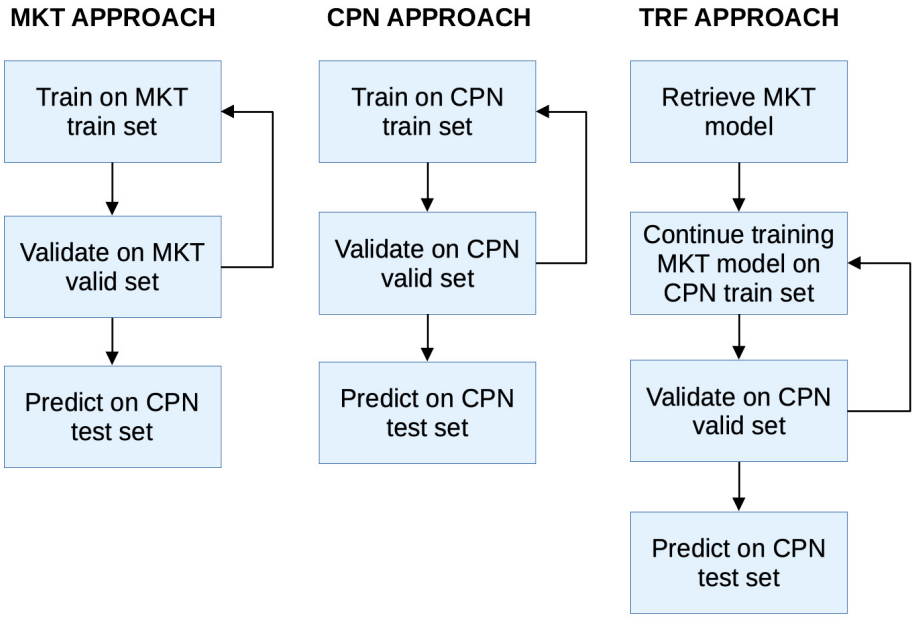


Figure 2. ML models training diagram.

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

where n_i are observed claim counts.

3.2.1. *BOOSTING TECHNIQUES*

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

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

That is, the prediction at the t-th step is given by the contribution, to the prediction of the previous step, of a weak predictor $h_t(x)$, properly weighted by a learning (shrinkage) factor η, where x is the covariate vector. The most common choice for the weak predictor $h_t(x)$ lies in the classification and regression trees (CART) family (Breiman 2017), from which the gradient boosted tree (GBT) models take the name. CARTs partition the feature space in an optimal way to receive (more) homogeneity on the resulting subsets (in terms of the modeled outcome). Such optimal partitioning 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 maximum likelihood 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 presented by Noll, Salzmann, and Wuthrich (2020), a *weak learner* is an 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 define *working weights* as $w_i = \hat{\lambda}^{(1)(x_i)} v_i$, so that N_i follows a Poisson distribution, $\mathcal{P}oi\,(\mu(x_i) \times w_i)$. 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 analog formulas 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 (shrinkage) factor and is used to make the learner even weaker, as values close to zero move the learner toward one. The estimation can be iterated M times. Since it is performed in log scale, this reduces to the formula exposed at the beginning of the paragraph.

Boosting weak predictors leads 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 introduced an efficient solution for handling categorical data.

The structural difference between XGBoost and LightGBM lies in the approach used to find tree splits. XGBoost uses a histogram based approach, where features are organized in discrete bins on which the candidate split values of the trees are determined. LightGBM focuses attention on instances characterized by large error gradients 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. Outcomes depend on the specific dataset and chosen hyperparameters (Gursky, n.d.; Nahon 2019). We chose LightGBM mainly because it is significantly faster than XGBoost, a definitive benefit when there is need to iterate the training through different combinations of hyperparameters. CatBoost was not considered because it is less mature compared with the other two algorithms.

A set of hyperparameters defines a boosted model, and even more defines a GBT model. The core hyperparameters that influence boosting are the number of models (trees), $t = 1, 2, \ldots, T$ (typically between 100 and 1,000) and the learning (shrinkage) rate η, with a typical value between 0.001 and 0.2 $h_t(x)$, and can be, when it belongs to the CART family, the maximum depth, the minimum number of observations in final leaves, or 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 approach (e.g., Bayesian optimization).

When applied to claim frequency prediction, hyperparameters 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 ($F_t(x)$) to initialize the learning process. The init score (or base margin) in the boosting approach has the same role as 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 (nonlinear) activation function to a linear combination of inputs, i.e.,

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

where $\bar{w}$ and β are 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 of one or more layers of interconnected neurons that receive a (possibly multivariate) input set and retrieve an output set (Goodfellow, Bengio, and Courville 2016). Modern deep neural networks are constructed of many (deep) layers of neurons. Deep learning has received increased interest for a decade, thanks to increased availability of massive datasets, increased computing power (in particular GPU computing), and newer approaches to reduce overfitting that prevented widespread adoption of such techniques in previous decades. Different architectures have reached state-of-the-art performance in many fields. For example, convolutional 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 neural networks) provide excellent results for 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 (2021b, 2021a) for deep neural networks.

Simpler structures are needed for claim frequency regression. Multilayer perception architecture basically consists of stacked simple neuron layers, from the input layer to the single output layer. This structure handles the relationship between ratemaking factors and 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 $\lambda(x_i) = \beta_0 + \sum_{j=1}^{Q} \beta_j \phi(z_i)$, where Q is 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, and then this exposure is added in a specific layer just before the final layer (which has a dimension of one).

Training a DL model involves providing batches of data to the network, evaluating the loss performance, and updating the weights in the direction that minimizes training (back-propagation). The whole dataset is provided to the fitting algorithms many times (epochs) split into batches. A common practice to avoid overfitting 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

This section compares prediction performance between our ML and credibility models. We conducted the analysis on two real and anonymized datasets, CPN and MKT, preprocessed and split into training, validation, and test sets as discussed in Section 3.1. As mentioned previously, the predictive performance of the fitted models is 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 training set and predictive performance is assessed on the test set. The validation set is used in the DL and LightGBM models to avoid overfitting. Finally, the 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 NGI has become quite popular in actuarial academia and among practitioners for comparing competing risk models.

Let y_i be the actual number of claims ranked by their modeled score $v_i \times \hat{\lambda}(x_i)$. 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 the NGI, we also compute the mean absolute error (MAE) and the root mean square error (RMSE), which are also popular metrics for comparing predictive performance (see Appendix A.2).

4.1. DATASET STRUCTURE

Two anonymized datasets were provided, one for the market (`mkt_anonymized_data.csv`) and one for the company (`cpn_anonymized_data.csv`), henceforth referred to as MKT and CPN datasets, respectively (see Appendix A.4). These datasets share the same structure, as each company provides its data to the pool in the same format. The pool aggregates individual filings into a market-wide file that is provided back to the companies. 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 categorical variables. Losses are the number of damaged units, while exposures 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 are considered a synonym for claim numbers. Our aggregated dataset contains the variables listed below:

- `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 training, validation, and test sets
- `cat1`: categorical variable 1 available in the original file (aggregating variable) (This can be considered a risk classification and, possibly, the most important predictor. The number of exposures insured strongly depends on this variable. Also, the `cat1` distribution can 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). Furthermore, the last available year (2008) is used as the test set, while data from previous years are randomly split between training and validation sets on an 80/20 basis (see Table 1). Market data are available for 11 years and company data for the last five years. Also, the number of exposures is widely dependent on the `cat1` variable.

Tables 2 and 3 compare explanatory variables by domain. The frequency distribution is reported for categorical variables, while summary statistics are computed for continuous variables (mean, standard deviation, minimum, and maximum). The `zone_id` and `cat1` statistics are shown in Appendix A.5 for the sake of synthesis. Note also that the variable `year` is not taken as an explanatory variable for the ML or credibility models. We implicitly assumed that the claims process was stationary.

4.2. IMPLEMENTATION DETAILS

This section presents the operations performed on the data and the implementation of the models. Dataset preprocessing was performed in a `Python` 3.8 environment, using the *Pandas* and *Scikit-Learn* libraries (Reback et al. 2020; Pedregosa et al. 2011) for the extraction, transformation, and loading stages. `R` Software (R Core Team 2022) and `Python` programming language were used for the analyses.

4.2.1. BOOSTING APPROACH

The LightGBM (LGB) model was used to apply boosted trees on the provided datasets, minimizing the Poisson deviance. As for most modern ML methods, an 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 close formula to identify the best combination for the given data.

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

4.2.2. DEEP LEARNING

Several approaches may be considered for building a DL architecture. Since the hyperparameter space of a DL architecture is vast, designing the best search strategy and network architecture (number of layers, number of neurons within, search, etc.) is challenging. Many practitioners start with a known working architecture for a similar task and perform moderate changes. It is also worth mentioning that more sophisticated approaches to DL architecture optimization are being developed (e.g., the neural architec-

Table 1. Dataset sizes.

	Test	Training	Validation	Total
CPN	19,124	50,430	12,519	82,073
MKT	89,805	527,388	130,995	748,188

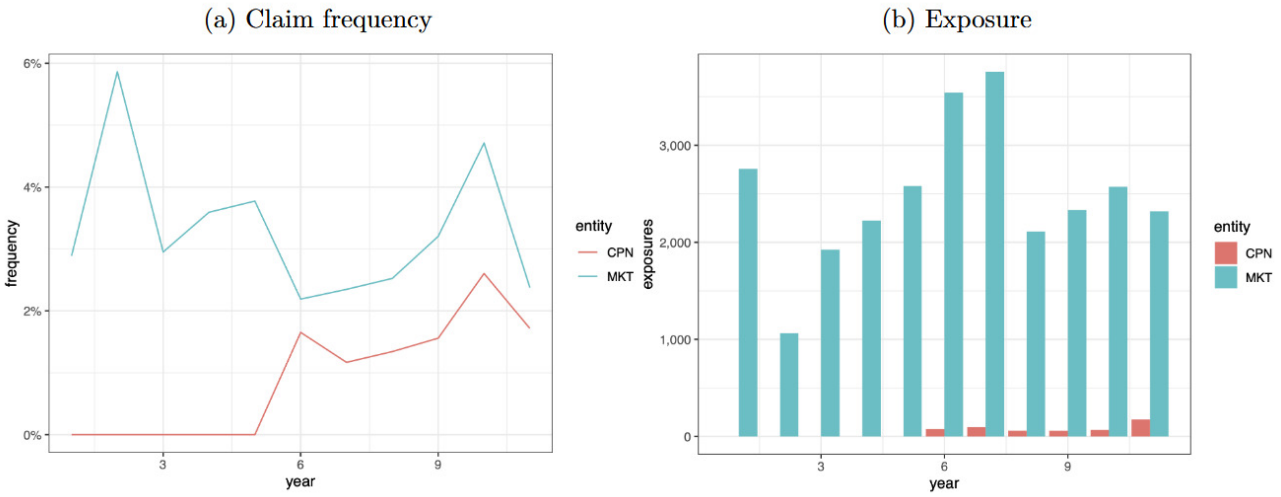


Figure 3. Claim frequences and exposures.

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

ture search; Elsken, Metzen, and Hutter 2019), but presenting such techniques is beyond the scope of this paper.

We set the chosen DL architecture by several trials based on previous experiments found in the literature for tabular data analysis (Schelldorfer and Wuthrich 2019; Kuo, Crompton, and Logan 2019). Our approach introduced a dense layer to collect the inputs and handled categorical variables using embedding. Three hidden layers performed feature engineering and knowledge extraction from the input. Dropout layers were added to increase the robustness of the process. As anticipated in the methodological section, the exposure part was handled separately in another branch and then merged in the final layer. The same model structure was used for the CPN, MKT, and TRF models. The TRF model was built first using the pretrained weights calculated on the MKT data and continued 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 was 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 was not improved for more than 20 epochs. DL models were

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		

trained in *Keras Tensorflow* 2.4 (Chollet 2018), taking on average 40s per epoch.

4.2.3. *CREDIBILITY MODELS*

For the credibility approach, the original datasets that were in longitudinal format were processed into a wide format (also called unbalanced) needed by the *actuar* `R` package (Dutang, Goulet, and Pigeon 2008). Furthermore, as required by the hierarchical Bühlmann-Straub model, continuous variables were discretized using the entire dataset (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 discretized continuous variables into three or four 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 was performed by the `cm` function of the R package *actuar,* which allows fitting various forms of credibility models (see, e.g., Goulet et al. (2021)). The response variable used for credibility models was the claim frequency (and not the number of claims). Therefore, predicted claim frequencies were multiplied by exposure to obtain the number of predicted claims.

Several credibility models were compared in terms of performance. We first conducted 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` was created to distinguish the company and market data. This base Bühlmann-Straub model is referred to as `BSbase` in the following.

Then, we selected the most appropriate HBS model by the most appropriate permutation of categorical explanatory variables `cat1–cat8`, since there was no particular order among them, except `cat1–cat3`. More precisely, we considered 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). There were 4!=24 possible HBS models. The best categorical HBS model that minimized the mean squared error when fitting models is referred to as `HBScateg` in the following.

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

Given the high number of HBS models fitted and used for prediction on the validation dataset, we used parallel computation using the `R` (core) package *parallel*, while model comparisons were performed in an `R` environment. The running times are summarized in Table 4 and show that MKT and TRF approaches took particularly long to validate. Indeed, fitting time contained only a call to `cm()` for every HBS model, while validation time made a prediction for every policy of the validation set (see Table 1). The prediction computation was particularly long, but it requires for each policy the exact location in the credibility tree structure starting from the top.

As explained previously, the fitting of HBS models was conducted on the training dataset, the best model (in terms of RMSE) was selected on the validation dataset, and the overall comparison was done on the test dataset.

4.2.4. *PERFORMANCE ASSESSMENT*

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 portfolio risks heterogeneity composition. In this regard, the performance assessment considered not only the discrepancy between actual and predicted losses, but also the ability of the model to rank risks, thereby 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 obtaining an acceptable estimate of the pure premium is more challenging (e.g., due to a systematic unmodeled social or environmental trend in either frequency or in severity). The ability of ML to identify nonlinear patterns and interactions is useful both to model the pure premium and to rank risks.

To compare the credibility and ML models within or between model classes, we used the NGI, the ratio between the sum of observed claims and the sum of expected claims (denoted by *actual to predicted ratio*), the MAE, and the RMSE. The NGI is a discriminant metric that ranks models according to their ability to predict, while the *actual to predicted ratio* is used to check whether 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 model's 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 considered; indeed, it may become an essential selection criterion in some contexts where the ability to explain a model is essential for regulatory or marketing purposes.

4.3. INTERPRETATION AND PREDICTIVE PERFORMANCE OF MODEL RESULTS

This section focuses on interpreting the ML and credibility models. We examined variable importance for ML models and analyzed credibility factor densities related to the best HBS models. In a second step, we assessed performance of both approaches.

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

Variables used	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

4.3.1. MODEL INTERPRETATION FOR ML

ML models have long been considered black boxes, but methods have been developed to provide explanations of model structure and provide outputs, even in actuarial science (see, e.g., Lorentzen and Mayer (2020)). In our application, we simply focused on the variable importance analysis internally calculated by the LightGBM model. That measure of variable importance broadly reflects the gain of using that feature in LightGBM trees to reduce training losses. Variable importance analysis in DL models is not automatically calculated during the training stage and requires the use of a separate algorithm (e.g., one with Shapley values (SHAP); Lundberg and Lee 2017), which is outside the scope of this paper. Another possible approach would be to use permutation importance, but there are no readily available routines for Tensorflow datasets. However, it is reasonable to assume that variable ranking is similar between the two ML approaches.

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

1. The `cat1` and `cat2` variables are consistently ranked as the most important predictors both for MKT and CPN approaches, as shown in Subfigures 4a and 4b.
2. The TRF plot, Subfigure 4c, is more difficult to interpret. It indicates which variables most likely correct the difference between the MKT and CPN models. While `cat1` keeps first place, the relative importance of other variables is higher than in the MKT and CPN plots.

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 limited ourselves to an illustration of the overall interpretability of the variables, which allows an actuary to understand the overall effect of variables on the tariff. Depending on the audience involved in the interpretability analysis (Delcaillau et al. 2022) (e.g., an actuary or a policyholder interested in its tariff), it may be necessary to discuss in depth the local interpretability and variable interaction issues.

4.3.2. MODEL INTERPRETATION FOR CREDIBILITY

The approach we used to select the best HBS model was based on permutations, which implicitly accounts for the importance of variables when building the hierarchical tree structure. Therefore, the structure of the 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 one, 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 were fitted even for `cat2` and `cat3`.

These analyses 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 model structure. In addition, its hierarchical structure and the value of the credibility coefficients allow us 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.

However, this approach to interpreting the HBS model is constrained by the choice of a credibility-based approach, which depends on the claims history of the policyholder. From this viewpoint, the model predictions do not necessarily depend only on a variable's importance, but also depend on the experience accumulated in the claims history. In some situations, the seniority of the claim is not important; recent information may better represent the current nature of the risk. Further research is needed to develop indicators that would distinguish the relationship between

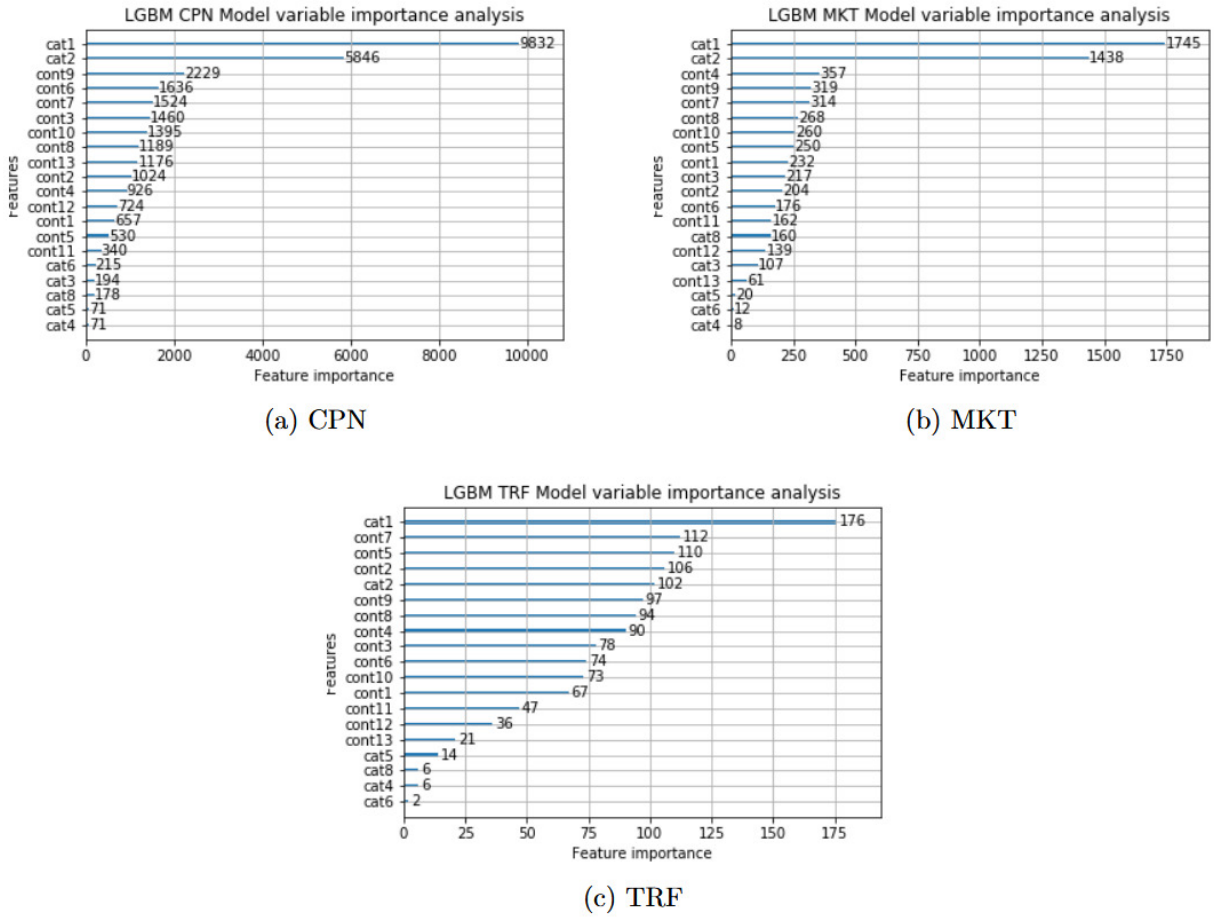


Figure 4. LightBM variable importance for CPN, MKT, and TRF approaches.

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 DL, light gradient boosting, and credibility models, whereas Figure 6 displays the normalized NGI against other metrics for each model point. The column *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 that the predictive performance analyses of the different approaches were conducted on the (same) test company dataset to ensure comparability.

First, the actual/predicted ratio was between 0.9 and 1.1 for all models, but as expected, CPN's was the worst. This result was indeed expected since the MKT dataset includes the company's data. We anticipated that since the test set considers a different year than the training and validation sets, the predictions may be structurally biased because the insured risk strongly depends on the year's context, and frequency trending is not considered in the modeling framework at all. Nevertheless, the results show that the MKT data, in this case, provided a superior experience compared with the CPN only data.

The LightGBM model results showed the best performance with the TRF and MKT approaches when measured by the NGI and MAE. We also noted that the LightGBM with the TRF approach was the best model in terms of RMSE. The HBS model built with categorical explanatory variables performed well with the MKT approach and was the second or third best model, depending on the metric considered. Given its nonparametric nature, this model is very flexible for adjusting to different feature effects. We generally observed that combining company data with external market data had a significant advantage in predictive performance for both the ML and credibility models. However, only the LightGBM model seemed able to exploit the TRF approach in an appropriate way. Indeed, it seems that the TRF approach penalized the credibility methods, which can be explained by a more important weight given to the information related to the company.

Regarding the predictive accuracy, 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 account for the effect of different covariates

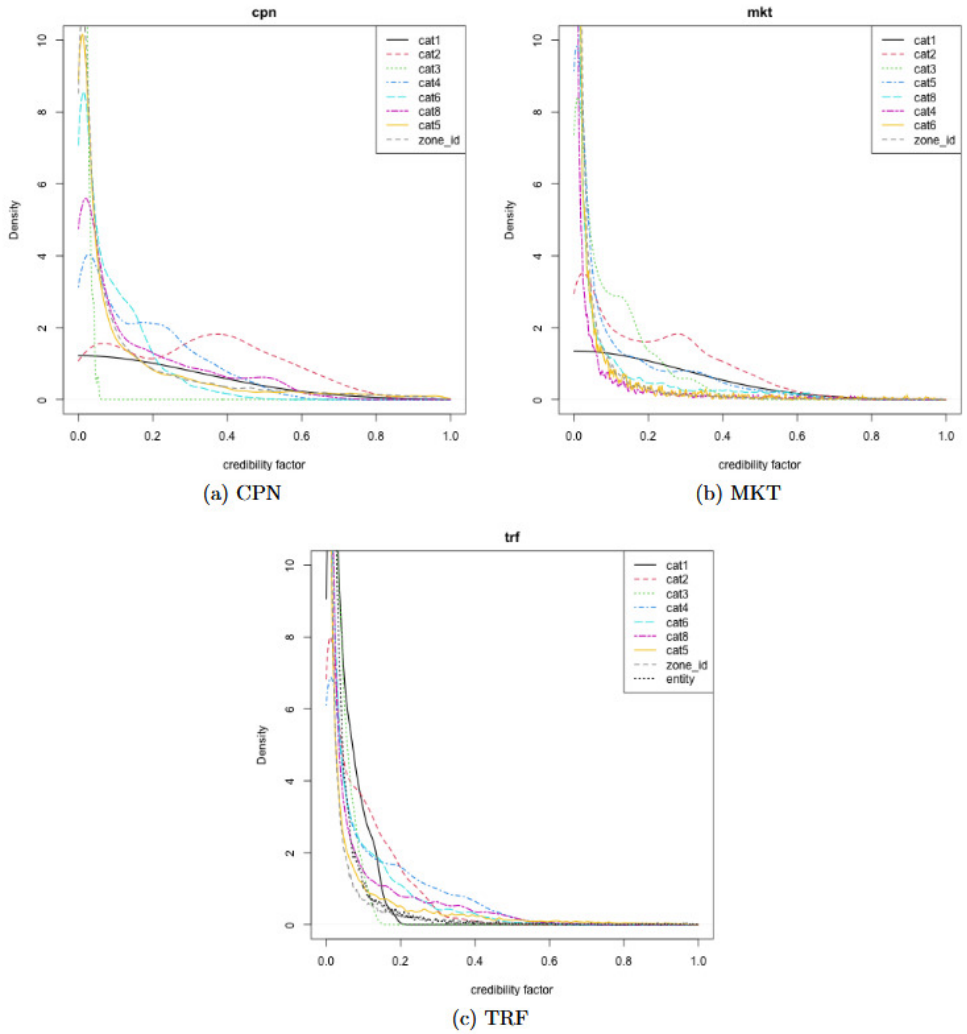


Figure 5. Empirical densities of credibility factors of the best HBS models.

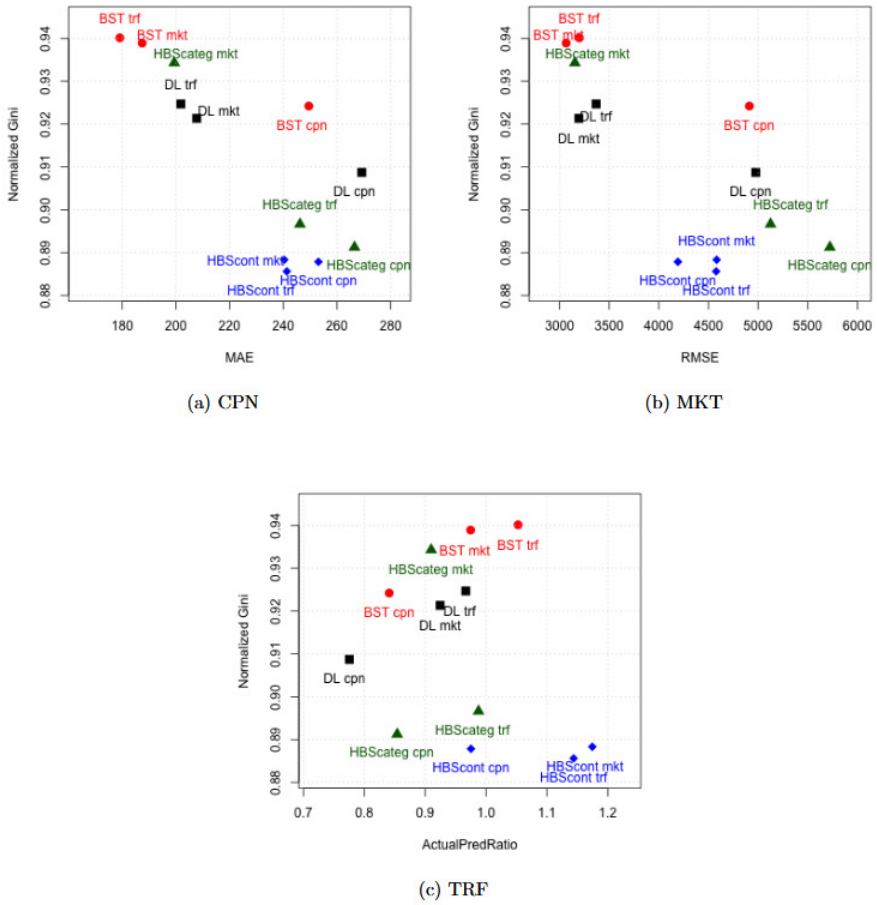
on the premium by splitting the portfolio into different levels. Hierarchical models are easily interpretable and provide actuaries with a clear picture of the pricing process by classifying policyholders according to their risk and claim history. However, they are not very flexible and make it difficult to capture nonlinearities or interaction effects between variables.

In this paper, we present an application of ML methods, the LightGBM model and a deep learning model, which can be compared with 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 were examined: the first directly applied an ML model pretrained on market data, while the second relied on transfer learning logic, where the pretrained model was fitted on the insurer's data. We performed our empirical analysis transferring loss experience from an external insurance bureau to a specific company portfolio. We focused 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 approach allowed us to significantly improve the prediction performance of an ML model compared with a model only trained on the insurer's data. Our results show the advantage and efficiency of pretraining an ML model on a reference dataset. We also observed that HBS models performed well on market data or company data alone in our application, indicating that the transfer does not improve the prediction power compared with the MKT or the CPN approaches, depending on the chosen metrics (MAE or RMSE). Finally, ML approaches obtained more competitive results compared with credibility models with this dataset. However, it is reasonable to expect that 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 handle other types of data or business in insurance applications when reference data are available. The only disadvantage is the training validation test computation time, which might be too long for big datasets. However, applying MKT or TRF approaches should be transposed to a specific context by replacing a market/company situation with, for example, a holding group/entity or company/line of business situation, because, in practice, the loss experience of competitors remains unknown. ML models also have a practical advantage in that their implemen-

Table 5. Model comparisons 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



(a) CPN (b) MKT

(c) TRF

Figure 6. Normalized Gini (NGI) against other metrics.

tation is relatively automated, while HBS model implementation may require a manual and expensive selection phase to derive the best feature combinations. Moreover, the code to train an ML model, as shown in this or similar studies, is

readily available (see, e.g., Appendix A.3) and can be replicated on an adequately resourced PC without too much effort.

This technique can be used to rate many insurance products. Although our exercise applied it to agricultural insurance, in theory it can be applied to any insurance industry context where the set of ratemaking variables shared between two distinct portfolios is nonempty, holding the common ratemaking variables in 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 lines of business, are created. Initial loss estimates may be performed on the initial product and then applied as initial scores to 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 to blend 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 being priced, in order to properly extrapolate the results, as previously anticipated.

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

These connections between credibility theory and ML techniques open pathways for future research. We used an empirical approach to build the hierarchical tree structure of the credibility model. One way to improve this is to define the tree structure through different partitioning tree models, similar 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 would be interesting to measure the interest of an approach based on transfer learning on this type of model.

6. ACKNOWLEDGMENTS

The authors wish to give a special thanks to CAS research and publications 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.

Submitted: March 31, 2022 EDT. Accepted: April 10, 2023 EDT. Published: June 27, 2025 EDT.

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.

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., 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.

Hachemeister, Charles A. and others. 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.

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): 97. 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 and Hall.

Meel, Vidushi. 2021. "YOLOv3: Real-Time Object Detection Algorithm." February 25, 2021. https://viso.ai/deep-learning/yolov3-overview/.

Nahon, Aviv. 2019. "XGBoost, LightGBM or CatBoost — Which Boosting Algorithm Should I Use?" December 30, 2019. https://medium.com/riskified-technology/xgboost-lightgbm-or-catboost-which-boosting-algorithm-should-i-use-e7fda7bb36bc.

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.

APPENDIX

A.1. RATING BUREAUS

According to IRMI (n.d.), 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 rating bureaus has become progressively less compulsory in recent decades, and their activities have become increasingly consultative; single carriers may purchase their data collection service and decide whether and how to use them. In the US market, the most relevant rating bureaus are the NCCI (for workers' compensation), the Insurance Services Office that serves most personal and commercial lines, the Surety Association of America that operates in the surety and crime insurance areas, and the American Association of Insurance Services that specializes in many commercial lines different from workers' compensation.

A.2. USUAL METRICS

Consider a set of observations y_i and their corresponding predictions $\hat{y}_i$ for $i = 1, \dots, n$. 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 and squared norms of residual vectors.

A.3. CODE

The modeling was 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 ANONYMIZING

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

Then, the available data were split threefold; the last available year was designated the test set, while the remaining years were split into training and validation sets, using an 80/20 ratio. Therefore, we had available three datasets for the market data and another three for the company data.

A.5. OTHER DESCRIPTIVE STATISTICS

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

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, respectively) 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 model presented in Section 2 are given in Table 8 (see Theorem 6.4 of Bühlmann and Gisler (2006) for details). We referred 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.

A.7. GENERALIZED LINEAR MODELS

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 distributions in the exponential family characterized by a, b, c, and $\mathbb{Y}$. Typical applications 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.

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.

Nb. levels	Market most frequent levels						**Nb. levels**	Company Market most frequent levels					
	158	245	167	119	154	273		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

Table 8. HBS parameter estimators.

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 9. Usual distributions in the exponential family.

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	$-\log(x!)$

A.8. DL STRUCTURE

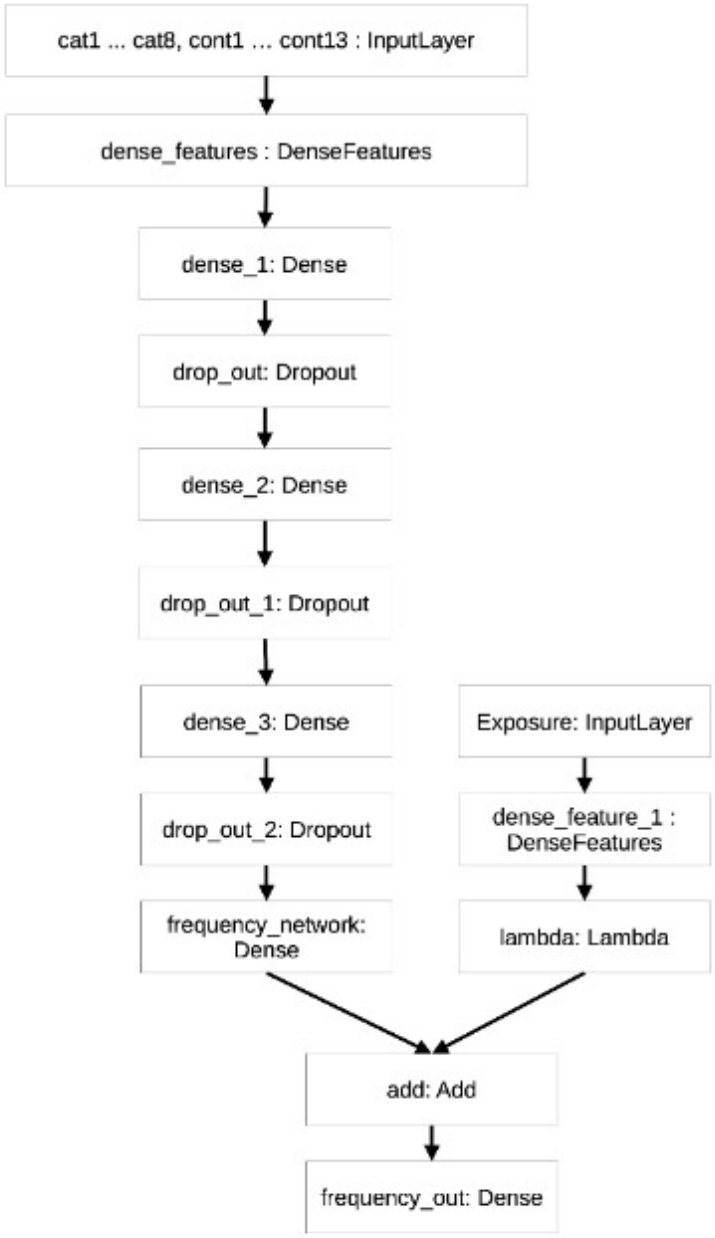


Figure 7. DL model structure.