Model-independent prediction intervals: a widely applicable and interpretable uncertainty measure

Clemens Schmid[1*], Johannes Schupp[2] and Hans-Joachim Zwiesler[1]

[1]Institute of Insurance Science, Ulm University, Ulm, 89069, Germany.
[2]Institute for Finance and Actuarial Sciences (ifa), Lise-Meitner-Str. 14, Ulm, 89081, Germany.

*Corresponding author(s). E-mail(s): clemens-1.schmid@uni-ulm.de;
Contributing authors: j.schupp@ifa-ulm.de;
hans-joachim.zwiesler@uni-ulm.de;

Abstract

Prediction intervals can be used to better understand uncertainty of forecasts, especially when using modern data science methods. We present and extend a model-independent and hence widely applicable approach for the estimation of prediction intervals. It is particularly suitable for the extension of established and productive models. Without much effort, prediction intervals can be estimated to better evaluate the quality of their forecasts. We apply the method to a real health insurance data set. Here, we use a random forest for the estimation of the target variable and subsequently also for the estimation of the prediction intervals. In addition, we compare the results to an alternative but random forest specific method and observe highly competitive results.

Keywords: Prediction intervals, health insurance, random forest, uncertainty measure

1 Introduction

Modern data science approaches are used for many actuarial tasks in insurance business today. Compared to well established actuarial methods, the application is often motivated by a higher accuracy, which is measured by simple point estimators, see e.g. [1]. An unbiased and accurate estimator is beyond any doubt important. However, uncertainty is also a crucial quantity in actuarial considerations. Essentially, the

uncertainty is insured by the insurance company. Therefore, an appropriate assessment of this quantity is required, e.g. for risk management or the determination of adequate safety margins in pricing.

In this letter, we discuss interpretable prediction intervals taking both noise and estimation uncertainty into account and illustrate the application for a realistic scenario. Thereby, we further improve an existing bootstrap approach based on [12]. In the latter, neural networks and prediction intervals are used to assess the quality of mortality forecasts, where only point estimates have been considered so far, see e.g. [9, 10]. In this letter instead, we use random forests offering advantages like less tuning parameters and easy applicability to different scales [2]. For a comparison and validation of our modification, we use a random forest specific method with quantile regression forests [8]. The independence of the underlying model is a main advantage of the bootstrap approach as established and productive applications with a focus on point estimates can be straightforwardly extended to achieve a better understanding of the inherent uncertainty. Extensive reviews of techniques regarding prediction intervals can be found in [7] for neural networks and in [13] or [11] for random forests.

Throughout the letter, $X := (X_1, \ldots, X_n)$, Y and ϵ denote real random vectors or variables on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ with $n \in \mathbb{N}$ and (X, Y) following a distribution $\mathcal{D}$. The general training data set is defined as

$$T := \{(X_i, Y_i) := (X_{i1}, \ldots, X_{in}, Y_i) \overset{\text{i.i.d.}}{\sim} \mathcal{D} | i \in \{1, \ldots, m\}\}$$

with $m \in \mathbb{N}$, and for convenience, we use $Y(x) := Y|X = x$ as the conditional response with $x \in \mathbb{R}^n$.

The remainder is structured as follows: In Section 2, we first give a definition of prediction intervals and then introduce the new adapted bootstrap approach as an extension of [12]. Also, the quantile regression forest method and metrics for comparing different techniques are summarized. Section 3 describes the data set and results of the different procedures are presented. Finally, Section 4 concludes.

2 Prediction intervals

2.1 Definition

The following definition is based on the so-called type III interval in [13].

Definition 1. *We define a prediction interval as*

$$I_\alpha(X, T) := \left[I_\alpha^-(X, T), I_\alpha^+(X, T)\right]$$

with $\mathrm{P}\left(Y \in I_\alpha(X, T) | X = x\right) \geq \alpha$ *for* $x \in \mathbb{R}^n$ *and* $\alpha \in (0, 1)$.

The prediction interval contains the response with a probability of α, which leads to an intuitive assessment of the prediction uncertainty of a specific model. This concept is also better interpretable and more informative than a mere conditional variance estimation, which neglects probabilistic statements and more heavily depends on the underlying scale, in particular on extreme, but rare outcomes of the target variable.

2.2 Adapted bootstrap approach

The underlying approach for estimating prediction intervals was first introduced in [6] and has been applied in various fields since then, including insurance related data sets like in [12]. Due to parts of its structure, the procedure is often referred to as a bootstrap approach. We pick up this technique based on the named publications but slightly modify it in order to improve its consistency and unlike most implementations use random forests as basis to show model independency as well as to take advantage of their favorable properties. Hence, we call this adapted bootstrap approach (AB).

The approach requires rather common assumptions like $Y = f(X) + \epsilon$ for a function $f : \mathbb{R}^n \to \mathbb{R}$ (model), $\mathrm{E}(\epsilon | X = x) = 0$ (unbiased errors), $f(x) - \hat{f}(x) \perp \epsilon(x)$ (independence) as well as $\epsilon(x) \sim N(\mu_\epsilon(x), \sigma_\epsilon^2(x))$ (normal errors) for a fixed $x \in \mathbb{R}^n$ and an estimator $\hat{f}(x)$ derived from T (e.g. a random forest). Further assumptions are $\mathrm{E}(\hat{f}(x) - f(x)) = 0$ (unbiased estimator) and $f(x) - \hat{f}(x) \sim N(\mu(x), \sigma^2(x))$ (normal estimator).

The unbiasedness is critical, though following [6] bias is assumed to be neglectable in comparison to variance. The assumption with regard to the distribution is based on the application in [12], but note that different versions are possible, see [6].

Based on these assumptions, a straightforward calculation leads to the formula

$$I_\alpha^\pm(x) = \hat{f}(x) \pm \Phi^{-1}\left(\frac{1+\alpha}{2}\right) \cdot \sqrt{\sigma^2(x) + \sigma_\epsilon^2(x)} \tag{1}$$

for symmetric prediction intervals, where Φ^{-1} denotes the quantile function of the standard normal distribution and $x \in \mathbb{R}^n$ corresponds to a realization of the explanatory variables for which the prediction interval of the response should be retrieved. Therefore, it remains to estimate the model variance $\sigma^2(x)$ and noise variance $\sigma_\epsilon^2(x)$, which are in general unknown.

The model variance $\sigma^2(x) = \mathrm{Var}(\hat{f}(x))$ is approximated using $N \in \mathbb{N}$ bootstrap samples derived from T and the corresponding random forests on these sets. Specifically, each of the N resamples is of the same size as the training set T and drawn according to the standard approach introduced in [5], in particular with replacement. The final estimator $\hat{\sigma}^2(x)$ equates to the empirical variance calculated with the predictions for x from the models each trained on one of the N bootstrap samples. Hence, we approximate the variance of the base random forest $\hat{f}(x)$ (ensemble of decision trees) with N additional random forests (ensemble of random forests). This approach differs slightly from [12] or [6] as in these papers the variance of a single member of the ensemble (of neural networks) is estimated.[1] This does not correspond to the actually desired value. Theoretically, our modification should result in a coherent theory and in a higher accuracy by avoiding unnecessarily wide intervals. In practice, the differences depend on the variance reduction achieved by using an ensemble instead of a single member. In [12], there is apparently only a slight deviation from the desired values, but this can only be observed ex post and therefore not be assumed in general. A similar modification of the method is also motivated in [4].

[1] This is similar to estimating the variance of a random forest based on the variance of a single decision tree.

For the noise variance $\sigma_\epsilon^2(x) = \mathrm{Var}(\epsilon(x))$, we implicitly assume that a noisy functional dependency

$$(Y(X) - \hat{f}(X))^2 - \sigma^2(X) = g(X) + \delta(X)$$

with $\mathrm{E}(\delta|X = x) = 0$ exists, where g can be estimated consistently. Hence, $\sigma_\epsilon^2(x) = g(x)$ holds, and we approximate g using a separate random forest estimator $\hat{g}$ on the modified learning data set

$$\tilde{T} := \{(X_i, R_i)|R_i := ((Y_i - \hat{f}(X_i))^2 - \hat{\sigma}^2(X_i))^+,\ (X_i, Y_i) \in T,\ i \in \{1, \ldots, m\}\}.$$

Here, we follow [12]. Moreover, $\hat{f}(X_i)$ in R_i refers to the out of bag prediction [3] in order to gain a better estimate. Plugging in both approximations of the variances in Formula 1, we obtain an estimator for a symmetric prediction interval that should approximately fulfil the condition of Definition 1 for a large training data set. The precise steps of the presented approach can be found concisely in Procedure 1.

Procedure 1 Adapted bootstrap approach

Required: training data set $T := \{(X_i, Y_i) := (X_{i1}, \ldots, X_{in}, Y_i) \overset{\text{i.i.d.}}{\sim} \mathcal{D}|i \in \{1, \ldots, m\}\}$, realization $x \in \mathbb{R}^n$ of X, coverage probability $\alpha \in (0, 1)$.

1: Train an estimator $\hat{f}$ on T (e.g. a random forest).
2: Generate $N \in \mathbb{N}$ bootstrap samples of size m from T and train an estimator $\hat{f}_i$ ($i \in \{1, \ldots, N\}$) on each of them.
3: Set $\hat{\sigma}^2(x) := \frac{1}{N-1} \sum_{i=1}^N \left(\hat{f}_i(x) - \overline{\hat{f}(x)}\right)^2$ with $\overline{\hat{f}(x)} := \frac{1}{N} \sum_{i=1}^N \hat{f}_i(x)$.
4: Set $\tilde{T} := \{(X_i, R_i)|R_i := ((Y_i - \hat{f}(X_i))^2 - \hat{\sigma}^2(X_i))^+,\ (X_i, Y_i) \in T,\ i \in \{1, \ldots, m\}\}$ and train an estimator $\hat{g}$ on this set.
5: Set $\hat{\sigma}_\epsilon(x) := \hat{g}(x)$.
6: Set $\hat{I}_\alpha^\pm(x, T) := \hat{f}(x) \pm \Phi^{-1}\left(\frac{1+\alpha}{2}\right) \cdot \sqrt{\hat{\sigma}^2(x) + \hat{\sigma}_\epsilon^2(x)}$.

Result: estimated prediction interval $\hat{I}_\alpha(x, T) := [\hat{I}_\alpha^-(x, T), \hat{I}_\alpha^+(x, T)]$ at point x.

2.3 Quantile regression forest

As a comparison to a model-specific procedure, we briefly summarize a method based on quantile regression forests (QRF) [8]. Here, a prediction interval is defined as

$$I_\alpha(x) := \left[Q_{0.5-\frac{\alpha}{2}}(x), Q_{0.5+\frac{\alpha}{2}}(x)\right]$$

with $Q_{0.5\pm\frac{\alpha}{2}}(x) := \inf\{y \in \mathbb{R}|\ \mathrm{F}(y|X = x) \geq 0.5 \pm \frac{\alpha}{2}\}$ as the quantiles of the conditional distribution of Y. To obtain an estimator, $\mathrm{F}(y|X = x)$ is first approximated with quantile regression forests and then used to gain empirical quantiles $\hat{Q}_{0.5\pm\frac{\alpha}{2}}(x)$. For an in-depth introduction, we refer to [8].

This method relies on the explicit structure of a random forest and therefore is not model-independent. Also, the intervals are possibly not symmetric around a prediction as there is not a point prediction in this context, which also means that the variance of the point estimate is not taken into account. On the other hand, the procedure only needs one instance of a random forest and does not have any distributional assumptions since the distribution itself is approximated. Hence, this approach can be seen as non-parametric, whereas the method presented in the preceding section is parametric.

2.4 Performance measures

In order to assess the quality of prediction intervals, we use two measures that reflect the coverage probability and the informativeness. Both are also used in [12] and described in [7]. As usual, these measures are evaluated on a separate testing data set V that is equally structured as the learning data set T and consists of $m \in \mathbb{N}$ elements in our case.

The first measure is called prediction interval coverage probability (PICP) and defined as

$$\text{PICP} := \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{Y_i \in [\hat{I}_\alpha^-(X_i,T), \hat{I}_\alpha^+(X_i,T)]\}}.$$

It can be seen as an approximation of $\mathrm{P}(Y(X) \in [\hat{I}_\alpha^-(X,T), \hat{I}_\alpha^+(X,T)])$ and hence should be close to the desired α.

The second measure is the so-called mean prediction interval width (MPIW) and described as

$$\text{MPIW} := \frac{1}{m} \sum_{i=1}^{m} (\hat{I}_\alpha^+(X_i,T) - \hat{I}_\alpha^-(X_i,T)).$$

Therefore, it estimates $\mathrm{E}(\hat{I}_\alpha^+(X,T) - \hat{I}_\alpha^-(X,T))$ and should be preferably small under the condition that the PICP is still close to α.

In their original form, both measures do not reflect the conditionality on X. This inaccuracy can be reduced by applying the measures to subsets of the testing data set, which is done hereafter.

3 Application to insurance data

3.1 Data description

In this section, we analyse and compare the approaches exemplary for a health insurance portfolio. The data set contains 13,203 substitutive private health insurance claims for insured persons over one year. The data mainly contains information about the insured person, contract characteristics, premiums and refunds over historic years. In total, there are 19 explanatory variables and one response, the sum of claims in one year, similar to a severity model in P & C insurance. As the claims are right-tailed, we choose to model the log response variable. However, the results hereafter are on the original scale.

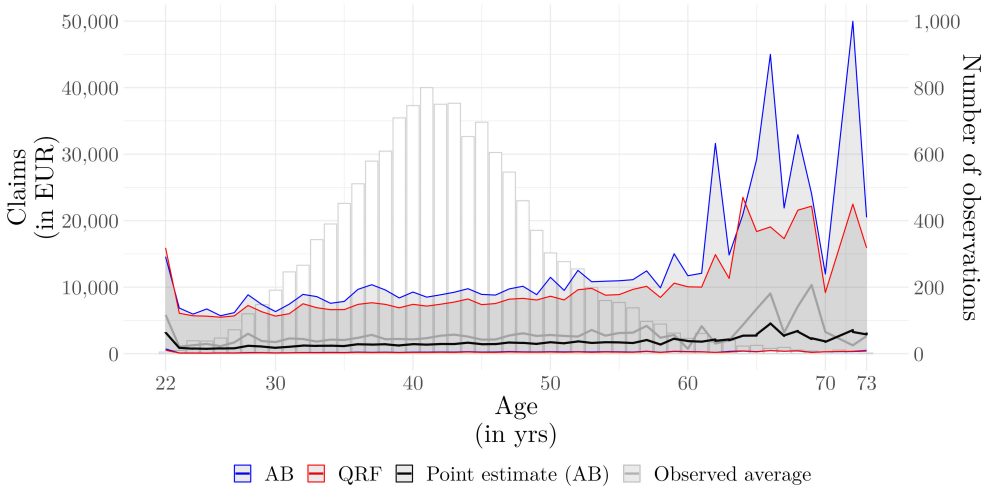

Fig. 1 Average predicted claims (boundaries of prediction interval, point estimate and observed average) by age for the adapted bootstrap approach and the quantile regression forest method with additional information about the number of observations

The data is split into a training and testing set at the ratio of 3:1 under stratification of the response. The random forest models of the corresponding method are estimated using 500 decision trees in one ensemble. Moreover, the minimum number of observations in a single leaf is 5, and the value for the randomly selected variables at each split is sustained through a tuning procedure. Specifically for the adapted bootstrap approach, we choose 100 bootstrap samples for estimating the model variance. In general, we select $\alpha = 90$ % for the prediction intervals and evaluate them on the testing data.

3.2 Results

We compare the approaches with respect to the metrics PICP and MPIW. First, we investigate the absolute values of the prediction intervals, where we split the testing data with respect to age. Especially in areas with only few observations, the uncertainty is high, see Figure 1. In particular, above 60 the upper bounds of the prediction intervals take values with up to 50,000 EUR. Also, the mean forecasts deviate from the observed averages for ages with only a few observations. Notably, both approaches show comparable prediction intervals, which are slightly larger for the adapted bootstrap method, which could be a result of incorporating the deviation of the point prediction into the estimation of the intervals. The MPIW corresponds to 9,235 EUR for the adapted bootstrap approach and 7,542 EUR for the quantile regression forest method. Both techniques result in presumably large intervals. However, this only reflects the high uncertainty in prediction, which is not surprising for an experienced actuary. Therefore, a limited view on the mean prediction could be dangerous and potentially results in the conclusion that uncertainty is only huge for highest ages.

Both methods meet the desired value of 90 % for their PICP, which implies validity of the intervals on average (0.9001 for the adapted bootstrap approach and 0.9013 for the quantile regression forest technique). We also analyzed the PICP for a similar age partition as in Figure 1. The PICP for ages with a large number of observations was close to 0.9 in any case. For highest ages, the uncertainty is higher due to small sample size, and hence, the PICP values sometimes deviate.

4 Conclusion

The adapted bootstrap method is a consistent modification of the bootstrap approach for estimating prediction intervals. It is especially useful for an extension of established and productive applications due to its model independence. Here, prediction intervals can give a better understanding of the quality of a model, and almost solely, more computational resources are required. The results are competitive to a tailored model-specific procedure like the quantile regression forest method in our example. The prediction intervals can be intuitively used for assessing the uncertainty of a target variable and are more informative than single point estimates. Therefore, the concept of prediction intervals should be of particular interest for actuaries that want to better understand predictions of applied data science methods.

References

[1] Ausschuss Actuarial Data Science (2020) Anwendung von Künstlicher Intelligenz in der Versicherungswirtschaft. URL https://aktuar.de/unsere-themen/fachgrundsaetze-oeffentlich/2020-02-14_Ergebnisbericht_Anwendungen_KI_Versicherungswirtschaft.pdf, accessed 18 October 2022

[2] Biau G, Scornet E (2016) A random forest guided tour. TEST 25(2):197–227

[3] Breiman L (2001) Random Forests. Machine Learning 45(1):5–32

[4] Carney J, Cunningham P, Bhagwan U (1999) Confidence and prediction intervals for neural network ensembles. In: The 1999 International Joint Conference on Neural Networks Proceedings, vol 2. IEEE, New York, pp 1215–1218

[5] Efron B (1979) Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics 7(1):1–26

[6] Heskes T (1997) Practical confidence and prediction intervals. In: Advances in Neural Information Processing Systems, vol 9. MIT Press, Cambridge, pp 176–182

[7] Khosravi A, Nahavandi S, Creighton D, et al (2011) Comprehensive Review of Neural Network-Based Prediction Intervals and New Advances. IEEE Transactions on Neural Networks 22(9):1341–1356

[8] Meinshausen N (2006) Quantile Regression Forests. Journal of Machine Learning Research 7:983–999

[9] Nigri A, Levantesi S, Marino M, et al (2019) A Deep Learning Integrated Lee-Carter Model. Risks 7(1):33

[10] Richman R, Wüthrich MV (2019) Lee and Carter go Machine Learning: Recurrent Neural Networks. Tutorial, SSRN

[11] Roy MH, Larocque D (2020) Prediction intervals with random forests. Statistical Methods in Medical Research 29(1):205–229

[12] Schnürch S, Korn R (2022) Point and interval forecasts of death rates using neural networks. ASTIN Bulletin 52(1):333–360

[13] Zhang H, Zimmerman J, Nettleton D, et al (2020) Random Forest Prediction Intervals. The American Statistician 74(4):392–406