Inhaltsverzeichnis

[Introduction 1](#_Toc164941952)

[Literature review 1](#_Toc164941953)

[Methods 4](#_Toc164941954)

[Derive benchmarking prediction intervals 4](#_Toc164941955)

[Method 1 4](#_Toc164941956)

[Method 2 5](#_Toc164941957)

[Conformal prediction 5](#_Toc164941958)

[Measures to assess reliability and sharpness 5](#_Toc164941959)

[(True) coverage / PICP 5](#_Toc164941960)

[ACE 5](#_Toc164941961)

[Upper coverage 5](#_Toc164941962)

[Lower coverage 5](#_Toc164941963)

[MIS 6](#_Toc164941964)

[MSIW (Mean Scaled Interval Width) 6](#_Toc164941965)

[MSIWW (Mean Scaled Interval Width Weighted) 6](#_Toc164941966)

[SWR (Sharpness Width Ratio) 6](#_Toc164941967)

[Bias 6](#_Toc164941968)

[Results 7](#_Toc164941969)

[Managerial / research discussion 7](#_Toc164941970)

[Conclusion 7](#_Toc164941971)

[References 8](#_Toc164941972)

[Appendix 8](#_Toc164941973)

# Introduction

…

# Literature review

## Methodological literature overview

The underlying CLVTools package make use of two models to predict customer lifetime value, the pnbd model and the gamma-gamma model. As models in general are incomplete representations of the reality, they are never achieving a correct result and hence there is some amount of uncertainty in the forecasts. For users of models, it is hence vital to understand and quantify this uncertainty. To achieve this for the underlying model and its context, first must be determined where the uncertainty in the prediction of customer lifetime value comes from. To achieve this, it follows a non-exhaustive literature overview about the sources of uncertainty. It is split into two parts where the first covers uncertainty sources related to “outside of the model”, i.e. customer behavior and the second will focus on uncertainty coming from the modeling part.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 14PI | 20 | 24 | 16 | 9 | 12 | 23 | 27 | 1 | 28 | 29 | 30 | 31 | 32 |
| Customer behavior / model extern |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Campaigns of competitors |  | x |  |  |  |  |  |  |  |  |  |  |  |  |
| Marketing contacts |  |  |  | x |  |  |  |  |  |  |  |  |  |  |
| State of the economy |  |  |  | x |  |  |  |  |  |  |  |  |  |  |
| Customer retention or churn (not observable) |  | x | x |  |  |  |  |  |  |  |  | x[[1]](#footnote-1) | x1 | X1 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Model (related) errors (named as such) / model intern | x |  |  |  | x | x |  |  |  |  |  |  |  |  |
| Parameter estimation errors | x |  |  |  | x |  | x | x |  | x |  |  |  |  |
| Wrong form of the point  forecasting model | x |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Random variations in data  generating process | x |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data uncertainty |  |  |  |  | x | x |  |  |  |  | x |  |  |  |
| Epistemic and aleatory uncertainty |  |  |  |  |  | x |  |  |  |  |  |  |  |  |

In the following, the most important points are discussed in more detail, starting with customer behavior. As the whole CLV calculation attempts to model customer behavior, it is self-explaining that uncertainty is based here. Influences found in the literature that increase the uncertainty of customer behavior and hence the CLV do so because they are not considered in the model. Examples are campaigns of competitors, marketing contacts in the past, presence and future and state of the economy in a sense that people change their consumption behavior between recession and boom times. What is also a dominant issue that creates uncertainty is the possibility of a customer leaving the company forever either to switch to a competitor or stop consuming. The problem here is that the probability of being “alive” is included in the model but still, most customers won’t notify the company when they churn, so it stays a mere probability. Considering a customer still alive who churned quickly at the very beginning may obviously lead to overestimation of the CLV and vice versa, hence uncertainty. The second part considers uncertainty that comes from inside the model. It is to note that the papers quoted here are not necessarily concerned with CLV estimation but treat forecasting models in general or in other contexts, often time series or wind forecasting. Nevertheless, since the model concerned here suffers from similar issues as other forecasting models, these aspects are relevant here as well. Especially often addressed is the fact that a model based on parameters needs its parameters to be estimated first which is connected to some amount of uncertainty. This issue will be addressed in this work as well in the following part. Also often mentioned is data uncertainty that might not be a problem here as the process how these data are obtained and treated before the model is employed is not too complex. Nevertheless, problems as mentioned in (29), i.e. inconsistencies in data integration from various sources (e.g. different branches), any data corruption in the process or the employment privacy policies. The latter can become an issue when mistakes are made when anonymizing the records of single customers.

With these points raised, it is evident that mere point forecasts are an insufficient statement about the likely future values as they do not provide information about uncertainty (2PI,6PI,9PI,12PI). Hence, point estimates are often accompanied or even replaced by intervals, so-called prediction intervals in the context of forecasts or confidence intervals in the context of e.g. parameter estimation. (5PI) The advantages shall be briefly discussed in the following. An interval (forecast) is offering a range of possible values of future outcomes with a specific level of confidence. (11PI) This means that the true value of the prediction will fall into this declared interval with a specific percentage value, e.g. 95%. (4PI) point out 4 main points why PIs are of such importance: 1. They assess future uncertainty and 2. hence enable the user to plan “different strategies for the range of possible outcomes”. This should mean that one can prepare a strategy in case a high value inside the interval is realized and one for a low value but in the context of CLV, this could also mean to realize that a customer has a high variability in their CLV and therefore target them. There are 2 rationales behind this approach: First, (16L) state that there is often a right tail distribution for CLV, and it hence makes sense to target a customer with high variability to realize that potential. Second, (33L) state one should focus on those as it offers the opportunity to learn and reduce uncertainty. The 3. point (4PI) makes about the importance of PIs is that they enable to assess different methods of forecasting more thoroughly and 4. PIs allow to compare forecasts made under different assumptions more carefully. Another point, made by (32PI) “forecasts cannot be expected to be perfect, and intervals emphasize this” which underlines maybe the most important characteristic of PIs, namely pointing out to the user of forecasts that they are most probably wrong and hence treat them appropriately. Thinking one step beyond PIs, a more sophisticated option are density predictions, which are comparable with PIs but they assign probabilities to each area in the interval and provide so more information about uncertainty. (6PI)

As the importance of PIs is justified now, it shall be revisited how such intervals may be obtained. In different contexts, methods are applicable that are not in other contexts. To begin with the context classical regression, (33PI) summarizes 4 methods where 3 of them might not be applicable in the specific CLV framework, Bayes methods (where “training” observations are taken into account step by step and with each step the parameters and with each step the parameters and predictions recalculated), direct interval estimation methods (which are especially made for interval estimations, like quantile regression) and conformal prediction. An applicable method from this paper might be the ensemble method what would include calculating the CLV for each customer with several methods/models and taking the predicted values as observations from which to calculate mean and variance and from this setting, assuming a normal distribution, calculate a classic prediction interval. In the environment of time-series, prediction intervals for forecast are usually computed in a different way, using the observed forecasting errors observed in the calibration period. The procedure that is inter alia suggested in (2PI) includes to first observe the errors between true value and prediction of the employed model and the “bootstrap” (randomly choose) from these observations 1 error for the first prediction. From this first prediction, do the second prediction and so on, incorporating all the previous predictions. In this way, a whole future is being constructed. Repeating this process e.g. 1000 times will create 1000 possible futures from which one can get for each time step the 95% central values which can then be treated as prediction interval. Instead of randomly draw from the distribution of errors, one can also assume that the errors are normally distributed and draw from a normal distribution. Unfortunately, in the context of the pnbd model, no such error observations from a training period are available which makes it impossible to apply this approach. Nevertheless, the CLVTools still applies a bootstrap method in terms of randomly choosing which customers to include in a model parameter calculation. Deriving a significant number of models (and getting the CLV for each customer for each of the models) allows to retrieve a prediction interval. Approaches to derive prediction intervals for this CLV model will be introduced later in this work.

To eventually come to the purpose of this work, the derived PIs shall be evaluated. The following table gives an overview about possible measures found in the literature and measures that will be deployed in this work.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2L | 15 | 16 | 20 | 31 | 35 | 34 | 37 | 38 | 39 | This work |
| MIS | (x) |  |  |  | (x) |  |  |  |  | (x) | x |
| (True) coverage / PICP / ECP | x | x | x | x |  | x |  |  | x |  | x |
| Upper coverage | x |  |  |  |  |  |  |  |  |  | x |
| Lower coverage |  |  |  |  |  |  |  |  |  |  | x |
| ACE |  |  | x | x | x |  |  |  | x |  | x |
| MSIW |  |  |  |  |  | (x) | (x) | (x) | (x) |  | x |
| MSIWW |  |  |  |  |  |  |  |  |  |  | x |
| SWR |  |  |  |  |  |  |  |  |  |  | x |
| Bias | (x) |  |  |  |  |  |  |  |  |  | x |

This table is not exhaustive in a sense that it does not contain all papers that dealt with evaluating their PIs but contains a collection of recent studies that employed inter alia methods that are applicable for the CLV context as well. The measures listed above are also employed in this work. Additional measures like MSIS (2L, 31, 39), NMPIL (34,35), PINAW (37) or Bias (2L) were not directly translatable to the CLV context[[2]](#footnote-2), but their idea is used in a slightly changed measure, hence they are noted in brackets. All the measures used will be explained in the section below. The bias has a different position as it does not evaluate PIs but the point forecasts.

## Role of uncertainty in marketing

…

# Methods

In this section, methods to derive alternative PIs to benchmark the by CLVTools suggest intervals will be presented. Also, all measures that aim to assess reliability and sharpness of PIs and will be used in this work will be introduced.

## Derive benchmarking prediction intervals

### Method 1

To predict the CLV, CLVTools predicts DERT (Discounted Expected Residual Transactions) and multiplies it predicted mean spending. To derive both sizes, two parametric models are used, the pnbd- and gg-model, respectively. For the pnbd-model, there are 4 parameters estimated, r and α as shape and scale parameter of the gamma distribution (purchase rate) and s and β for the attrition rate of individual customers (*CLVTools documentation*). Estimating both models results in a direct estimate for each parameter and a covariance matrix of all parameters (1 per model) which represents the uncertainty connected with the estimates. Assuming a normal distribution with the just named characteristics, a set of 1000 observations for the parameters for both models is estimated. This results in 1000 different models with 1000 estimated CLVs from which one takes the central 90% observations to receive the corresponding prediction interval.

### Method 2

This method uses the already predicted 90% PI borders that were estimated for DERT and predicted mean spending. Again, the assumption is made that the possible values for DERT and predicted mean spending are normally distributed. Therefore, for each customer, there exists theoretically a normal distribution. From this, one takes 1000 values for DERT and predicted mean spending and calculates the CLV manually. From these values, one can again take the 90% prediction interval for the CLV.

### Conformal prediction (will be reformulated to algorithm form)

In this approach, a version of conformal prediction for regression / continuous data will be implemented. The big advantage of CP is that it can be used to derive prediction intervals model-independently and without having a closed form for e.g. the variance as it is the case for the CLV context. The implementation in this work shall be introduced in the following. The historic data of purchases is separated into 3 parts, a training, calibration, and testing part where each customer (and all their purchases) cannot be split over several parts. After splitting, the pnbd and gg models are trained for all 3 parts, so that 6 models are resulting. This step is necessary to get access to all parameters of the pnbd and gg models. The parameters of the training model are now taken to replace the estimated parameters of the calibration and testing part models so that all models of a kind (pnbd or gg) have the same parameters but different data. With this setting, it is possible to make predictions for the calibration and test partition while doing so for the training partition is not required any further. The difference of predictions and true values of the calibration part are taken to derive an absolute error distribution. From this distribution, one can take the desired confidence interval, e.g. take the error for which 90% of the errors are smaller. Note that by now, the error is taken as difference between true value and prediction and scaled by the prediction. This is necessary to avoid issues that come with heterogeneity of CLVs which range from $1 to $500 and would make an unscaled measure meaningless[[3]](#footnote-3). As a final step, the quantile that was previously derived is subtracted/added from/to the predictions of the test partition to get the lower/upper quantiles, now in absolute values. This results in symmetric PIs. From there, one can derive the true coverage on the test data. The described algorithm is applied several times (as often as it takes that every customer is sampled at least 1 time), each time sampling new training, calibration and test sets and finally averaging over lower and upper bounds of the PIs. An issue that must be noted is that the performance of the PIs on the test set is reasonable while applying it to the “normal” predictions outside of CP results in far too high coverage rates. Probably that is due to the way the data were simulated in relation to those predictions and/or the model performance loses precision when the training set is too small, resulting in higher errors, resulting in larger prediction intervals which deliver a too good coverage rate.

#### Validity

The measure validity assesses how exact the prediction intervals perform. They must cover on average the true value in (1-α)\*100% of the cases to be called valid. The current approach delivers in this regard (and with the simulated test data) satisfying results when repeated often enough (what is typically not the case when only as many runs are conducted as are necessary to sample each customer at least once as described in the approach above).

#### Exchangeability

…

## Measures to assess reliability and sharpness

Besides the measure in table … above, some additional ones are created or taken from the literature and adopted to this very specific CLV context to ensure a reasonable analysis.

### (True) coverage / PICP

This measure holds the percentage of cases when the true value lays indeed in the PI:

With n being the number of customers, yi the true observation for a customer’s CLV and PIi the respective prediction interval.

### ACE

This measure indicates how much on average the intended nominal prediction coverage, PINC, e.g. 90% and the true coverage, PICP differ

### Upper coverage

This measure indicates the percentage of times the upper prediction limit was not exceeded by the true value

With ULi being the upper limit of the prediction interval.

### Lower coverage

This measure indicates the percentage of times the lower prediction limit was not exceeded by the true value

With ULi being the upper limit of the prediction interval.

### MIS

The Mean Interval Score is based on the interval score, proposed by (35) and averaged over all customers instead of timesteps as it originates from time series forecasting. Due to the scale difference between small and large values for CLV, the differences in the initial formula were scaled by the estimated value (scaling by the true value would result in division by 0 if the customer would not buy again). Also, it is not to be confounded with MSIS which would in addition be scaled by mean absolute seasonal difference. (31) This measure assesses reliability and sharpness at the same time because it penalized interval width and true values outside of the CI at the same time.

### MSIW (Mean Scaled Interval Width)

This measure assesses the interval width with respect to the level of the true value. This takes account for the special CLV situation where an interval having a width of $10 is of different value for a CLV of $20 or $200 what is a realistic case for different customers. The actual MSIW is the mean over all customers.

### MSIWW (Mean Scaled Interval Width Weighted)

Lower CLVs tend to have a higher (worse) MSIW for all methods (see appendix). At the same time, firms may want to put more emphasize on higher CLV customers. This measure will therefore weigh customers with higher CLV more, i.e. the weight for a customer will be its CLV relative to the sum of all CLVs.

### SWR (Sharpness Width Ratio)

This measure evaluates reliability per width achieved by the intervals. A higher value is better than a lower.

### Bias

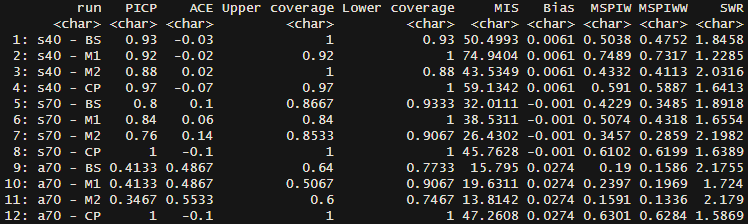
The bias is not calculated for the intervals but for the point predictions. It measures if the predictions are systematically wrong in one or the other direction, compared to the actual value. Therefore, it will not be different for different methods to derive prediction intervals.

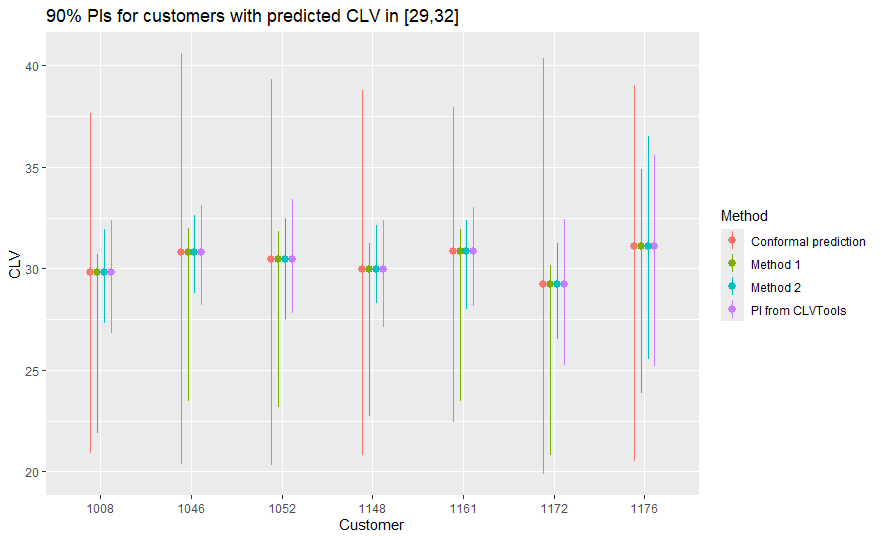
With yEsti being the estimate. The formula was taken from (2L) and was slightly adapted.

# Results

To assess the CLVTools bootstrap intervals, they will be benchmarked with the above outlined measures against alternative method(s) to derive intervals. A further distinction will be made in the differentiation between customers with higher and lower CLV because intervals tend to be wider relative to their actual CLV value, for low CLV customers, see appendix. In addition, this methodology will be applied to several datasets and the results averaged over them.

See in the following the resulting table[[4]](#footnote-4).



Also, see a visualization of the PIs for all methods in a sample for customers with an estimated CLV between 29 and 32.

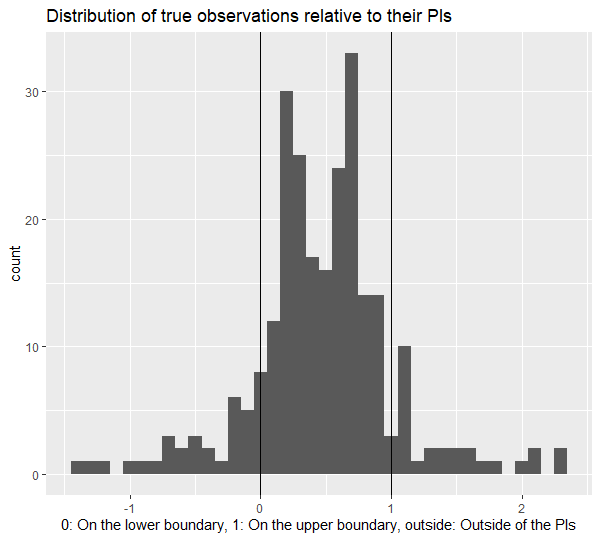
\**Interpret the results*\*

* Differences between datasets
* In which cases is the performance of intervals good and when not? -> further analyses

As the results suggest, the characteristics of the different methods can also be observed graphically, e.g. very low interval boundaries for the first method and therefore very good Lower coverage. …

**Additional analyses**

After assessing the bias of point forecasts, at least a visual bias analysis shall be conducted for the intervals, too. To account for different scales of (predicted) CLVs for different customers, the true observations are scaled to their respective interval boundaries where 0 means the true value lays on the lower boundary and 1 it lays on the upper boundary. Everything outside [0,1] is outside of the prediction interval. It should be roughly 5% for 95% intervals but it is in this case x%. The summary graphic … shows the distribution of observations with respect to their interval boundaries over all customers.



# Managerial / research discussion

…

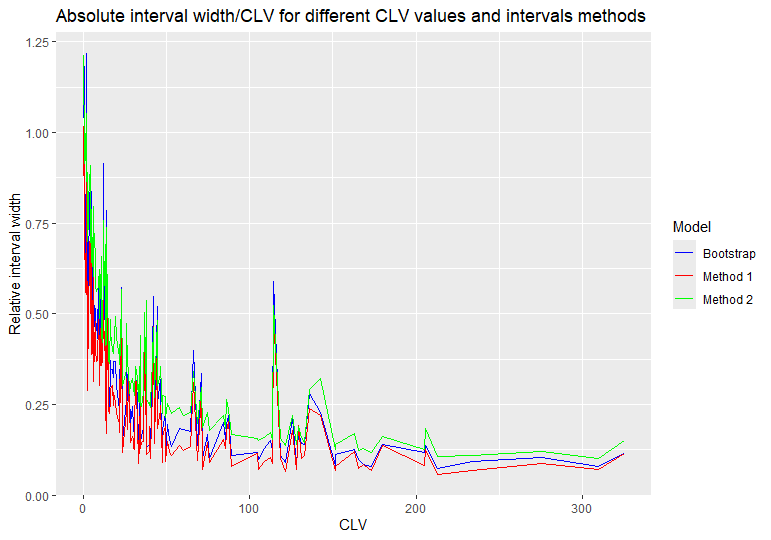
# Conclusion

…

# References

…

# Appendix



1. Even though those papers do not state explicitly that churn contributes to uncertainty in the CLV estimation, they do make the connection from churn to CLV estimation (20, p. 2) [↑](#footnote-ref-1)
2. Since the literature on benchmarking PIs in the CLV context is very limited, it was necessary to conduct research in other fields where it is more common to calculate and assess PIs. Typical examples are time series forecast competitions (31) or the prediction of wind (16, 20, 37) or electricity-related forecasts (35, 38) [↑](#footnote-ref-2)
3. Also, this approach comes with issues regarding scale as lower CLVs tend to vary in more proportionally to their absolute magnitude. This issue must be addressed but stating how much more they vary is not trivial. [↑](#footnote-ref-3)
4. Notation for „run“-column: „cluster of customers – PI method”. For cluster of customers: E.g. “s33” means customers with 33% lowest CLV estimated [↑](#footnote-ref-4)