Application of Conformal Prediction (CP) and Quantile Regression (QR) for the PNBD model

Supplement to my Master's thesis
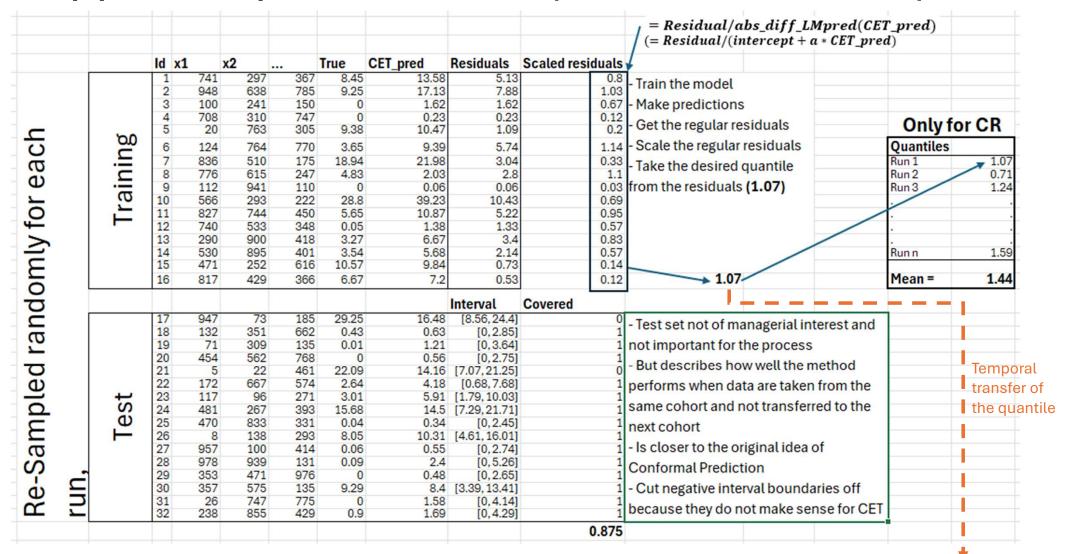
„Uncertainty in time-series modeling: An application to the prediction of individual customer lifetime values"

Conformal Prediction (CP)

General Procedure (no pnbd-context)

		Id	x1	x2	...	True	Prediction		Description
Training	**Proper Training Set**	1	741	297	367	8.45	-		- train the model (parameters)
		2	948	638	785	9.25	-		**1.**
		3	100	241	150	0	-		
		4	708	310	747	0	-		
		5	20	763	305	9.38	-		
		6	124	764	770	3.65	-		
		7	836	510	175	18.94	-		
		8	776	615	247	4.83	-		
		9	112	941	110	0	-		
		10	566	293	222	28.8	-		
		11	827	744	450	5.65	-		
		12	740	533	348	0.05	-		
		13	290	900	418	3.27	-		
								Residual	
	Validation Set	14	530	895	401	3.54	2.47	1.07	- apply the model, make predictions
		15	471	252	616	10.57	9	1.57	- take residuals from validation set to avoid
		16	817	429	366	6.67	30.25	23.58	overfitting/too small residuals because
		17	947	73	185	29.25	25	4.25	model was trained on these data already
		18	132	351	662	0.43	0	0.43	- calculate the quantile of the residuals
		19	71	309	135	0.01	3.96	3.95	-> here, 15.029
		20	454	562	768	0	4	4	
		21	5	22	461	22.09	22	0.09	**2.**
		22	172	667	574	2.64	12.62	9.98	
		23	117	96	271	3.01	0.62	2.39	
		24	481	267	393	15.68	0.09	15.59	
		25	470	833	331	0.04	0	0.04	**15.029**

		Id	x1	x2	...	True	Prediction	Interval	Covered	Description
Test Set		26	8	138	293	8.05	0.28	[-14.749, 15.309]	1	- apply the trained model, make predictions
		27	957	100	414	0.06	16.65	[1.621, 31.679]	0	- calculate the prediction intervals by adding
		28	978	939	131	0.09	1.38	[-13.649, 16.409]	1	and subtracting the quantile
		29	353	471	976	0	10.8	[-4.229, 25.829]	1	**3.**
		30	357	575	135	9.29	5.22	[-9.809, 20.249]	1	- 90% should be covered
		31	26	747	775	0	28.5	[13.471, 43.529]	0	- 71% < 90% -> undercoverage in this case
		32	238	855	429	0.9	0	[-15.029, 15.029]	1	

Coverage: 0.71

Applied in pnbd-context (Part 1, old cohort)

$$= Residual/abs\_diff\_LMpred(CET\_pred)$$
$$(= Residual/(intercept + a * CET\_pred)$$

	Id	x1	x2	...	True	CET_pred	Residuals	Scaled residuals	
Training	1	741	297	367	8.45	13.58	5.13	0.8	- Train the model
	2	948	638	785	9.25	17.13	7.88	1.03	- Make predictions
	3	100	241	150	0	1.62	1.62	0.67	
	4	708	310	747	0	0.23	0.23	0.12	- Get the regular residuals
	5	20	763	305	9.38	10.47	1.09	0.2	- Scale the regular residuals
	6	124	764	770	3.65	9.39	5.74	1.14	- Take the desired quantile
	7	836	510	175	18.94	21.98	3.04	0.33	from the residuals (1.07)
	8	776	615	247	4.83	2.03	2.8	1.1	
	9	112	941	110	0	0.06	0.06	0.03	
	10	566	293	222	28.8	39.23	10.43	0.69	
	11	827	744	450	5.65	10.87	5.22	0.95	
	12	740	533	348	0.05	1.38	1.33	0.57	
	13	290	900	418	3.27	6.67	3.4	0.83	
	14	530	895	401	3.54	5.68	2.14	0.57	
	15	471	252	616	10.57	9.84	0.73	0.14	
	16	817	429	366	6.67	7.2	0.53	0.12	

Only for CR

Quantiles	
Run 1	1.07
Run 2	0.71
Run 3	1.24
.	
.	
.	
Run n	1.59
Mean =	**1.44**

1.07

	Id	x1	x2	...	True	CET_pred	Interval	Covered	
Test	17	947	73	185	29.25	16.48	[8.56, 24.4]	0	- Test set not of managerial interest and
	18	132	351	662	0.43	0.63	[0, 2.85]	1	not important for the process
	19	71	309	135	0.01	1.21	[0, 3.64]	1	- But describes how well the method
	20	454	562	768	0	0.56	[0, 2.75]	1	performs when data are taken from the
	21	5	22	461	22.09	14.16	[7.07, 21.25]	0	same cohort and not transferred to the
	22	172	667	574	2.64	4.18	[0.68, 7.68]	1	next cohort
	23	117	96	271	3.01	5.91	[1.79, 10.03]	1	- Is closer to the original idea of
	24	481	267	393	15.68	14.5	[7.29, 21.71]	1	Conformal Prediction
	25	470	833	331	0.04	0.34	[0, 2.45]	1	- Cut negative interval boundaries off
	26	8	138	293	8.05	10.31	[4.61, 16.01]	1	because they do not make sense for CET
	27	957	100	414	0.06	0.55	[0, 2.74]	1	
	28	978	939	131	0.09	2.4	[0, 5.26]	1	
	29	353	471	976	0	0.48	[0, 2.65]	1	
	30	357	575	135	9.29	8.4	[3.39, 13.41]	1	
	31	26	747	775	0	1.58	[0, 4.14]	1	
	32	238	855	429	0.9	1.69	[0, 4.29]	1	

0.875

Re-Sampled randomly for each run,

Temporal transfer of the quantile

Applied in pnbd-context (Part 2, new cohort)

$$= abs\_diff\_LMpred(CET\_pred) * quantile$$
$$= CET\_pred \pm scaled\ quantile$$

Id	x1	x2	...	True	CET_pred	Scaled quantile	Interval	Covered	Coverage
5000	741	297	367	1.86	1.03	3.18	[0, 4.21]	1	0.9375
5001	948	638	785	11.42	12.83	8.89	[3.94, 21.72]	1	
5002	100	241	150	6.11	4.8	5.01	[0, 9.81]	1	- apply the PNBD model to the new
5003	708	310	747	0	0.1	2.73	[0, 2.83]	1	cohort and make CET predictions
5004	20	763	305	0.55	2.23	3.76	[0, 5.99]	1	- get the individal interval lengths by
5005	124	764	770	3.88	1.33	3.33	[0, 4.66]	1	scaling the quantile with the expected
5006	836	510	175	0	1.38	3.35	[0, 4.73]	1	abs. difference of the specific customer
5007	776	615	247	7.2	7.92	6.52	[1.4, 14.44]	1	- add and subtract the result form the
5008	112	941	110	5.87	4.72	4.97	[0, 9.69]	1	respective point prediction
5009	566	293	222	0.18	0.74	3.04	[0, 3.78]	1	- check if the resulting interval covers the
5010	827	744	450	14.25	18.53	11.65	[6.88, 30.18]	1	true value
5011	740	533	348	12.77	7.23	6.18	[1.05, 13.41]	1	- calculate the overall coverage
5012	290	900	418	0.7	1.94	3.62	[0, 5.56]	1	
5013	530	895	401	5.13	0.08	2.72	[0, 2.8]	0	
5014	471	252	616	1.96	0.81	3.07	[0, 3.88]	1	
5015	817	429	366	18	22.95	13.79	[9.16, 36.74]	1	
5016	947	73	185	9.8	17.82	11.31	[6.51, 29.13]	1	
5017	132	351	662	9.86	1.72	3.51	[0, 5.23]	0	
5018	71	309	135	0.06	1.93	3.62	[0, 5.55]	1	
5019	454	562	768	14.28	22.8	13.72	[9.08, 36.52]	1	
5020	5	22	461	4.48	1.49	3.40	[0, 4.89]	1	
5021	172	667	574	0.02	1.75	3.53	[0, 5.28]	1	
5022	117	96	271	1.55	5.35	5.27	[0.08, 10.62]	1	
5023	481	267	393	3.44	1.25	3.29	[0, 4.54]	1	
5024	470	833	331	4.13	5.09	5.15	[0, 10.24]	1	
5025	8	138	293	0.01	0.5	2.92	[0, 3.42]	1	
5026	957	100	414	2.5	0.14	2.75	[0, 2.89]	1	
5027	978	939	131	2.66	3.27	4.26	[0, 7.53]	1	
5028	353	471	976	1.21	0.87	3.10	[0, 3.97]	1	
5029	357	575	135	7.11	5.36	5.28	[0.08, 10.64]	1	
5030	26	747	775	0.66	0.64	2.99	[0, 3.63]	1	
5031	238	855	429	6.73	12.6	8.78	[3.82, 21.38]	1	

Quantile Regression (QR)

Applied in pnbd-context (Part 1, old cohort)

Id	x1	x2	...	True
1	741	297	367	8.45
2	948	638	785	9.25
3	100	241	150	0
4	708	310	747	0
5	20	763	305	9.38
6	124	764	770	3.65
7	836	510	175	18.94
8	776	615	247	4.83
9	112	941	110	0
10	566	293	222	28.8
11	827	744	456	5.65
12	740	533	348	0.05
13	290	900	418	3.27
14	530	895	401	3.54
15	471	252	616	10.57
16	817	429	366	6.67

1. Create a grid of n possible parameter combinations
2. Go through the grid and provide every run the whole cohort and 1 parameter combination to the model
3. Make predictions with this combination and get the upper and lower coverage
4. Assign the results to their respective parameter combination
5. Find the minimum deviation of the results to the desired quantile, e.g. 5% or 95%
6. Select this parameter combination to produce e.g. the upper prediction interval boundaries

pnbd model

run	r	alpha	s	beta	CET_Upper_Diff	CET_Lower_Diff
1	1	1	1	0.5	0.0750	0.8250
2	2	1	1	0.5	0.1375	0.7625
3	3	2	1	0.5	0.0001	0.7309
4	4	2	1	0.5	0.6400	0.0421
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n	.	.	.	.	.	.
			Min	0.0001	0.0421	

Parameter combination for upper boundary 3, 2, 1, 0.5
Parameter combination for lower boundary 4, 2, 1, 0.5

Id	TRUE	Pred. Run 1	Above Run 1	Below Run 1	Pred. Run 2	Above Run 2	Below Run 2	Pred. Run 3	Above Run 3	Below Run 3
1	8.45	13.58	0	1	12.40	0	1	...	...	...
2	9.25	17.13	0	1	15.23	0	1			
3	0	1.62	0	1	0.28	0	1			
4	0	0.001	0	1	0.39	0	1			
5	9.38	10.47	0	1	9.91	0	1			
6	3.65	9.39	0	1	8.50	0	1			
7	18.94	21.98	0	1	5.20	1	0			
8	4.83	2.03	1	0	0.25	1	0			
9	0	0.06	0	1	1.25	0	1			
10	28.8	39.23	0	1	37.81	0	1			
11	5.65	10.87	0	1	9.00	0	1			
12	0.05	1.38	0	1	1.07	0	1			
13	3.27	6.67	0	1	5.70	0	1			
14	3.54	5.68	0	1	4.27	0	1			
15	10.57	9.84	1	0	8.42	1	0			
16	6.67	7.2	0	1	7.20	0	1			
	Average		0.125	0.875		0.1875	0.8125		...	...
	Average - alpha/2		0.075	0.825		0.1375	0.7625			

Note
- With a "confidence level" of 10%, 5% should be above the upper interval boundary and 5% below
  Therefore, deduct 5% (alpha/2) from the achieved value
- The model can not predict exactly 0, so it makes sense to introduce a small tolerance
  for the lower boundary to avoid misleading results (not shown here)

Applied in pnbd-context (Part 2, new cohort)

Parameter combination for upper boundary: **3, 2, 1, 0.5**
Parameter combination for lower boundary: **4, 2, 1, 0.5**

Interval boundaries
(predictions from the model)

Id	x1	x2	...
5000	741	297	367
5001	948	638	785
5002	100	241	150
5003	708	310	747
5004	20	763	305
5005	124	764	770
5006	836	510	175
5007	776	615	247
5008	112	941	110
5009	566	293	222
5010	827	744	450
5011	740	533	348
5012	290	900	418
5013	530	895	401
5014	471	252	616
5015	817	429	366

1.

pnbd model

2.

Id	TRUE	CET Lower	Below	CET Upper	Above
5000	13.58	8	1	14	1
5001	17.13	10	1	18	1
5002	1.62	0.01	1	2	1
5003	0	0.01	0	3	1
5004	10.47	10	1	15	1
5005	9.39	9	1	20	1
5006	21.98	14	1	22	1
5007	2.03	0.3	1	3	1
5008	0	0.1	0	5	1
5009	39.23	30	1	25	0
5010	10.87	10	1	15	1
5011	0	0.026	0	19	1
5012	6.67	3	1	7	1
5013	0	0.009	0	5	1
5014	9.84	4	1	23	1
5015	7.2	9	0	8	1

3. Coverage **0.6875** **0.9375**
(95% are desired here for both boundaries)

1. Provide the model with the needed parameter combination to predict the lower/upper interval boundary
2. Make predictions for the whole cohort
3. Check the coverage

Note Again, introducing a small tolerance for the lower boundary makes sense