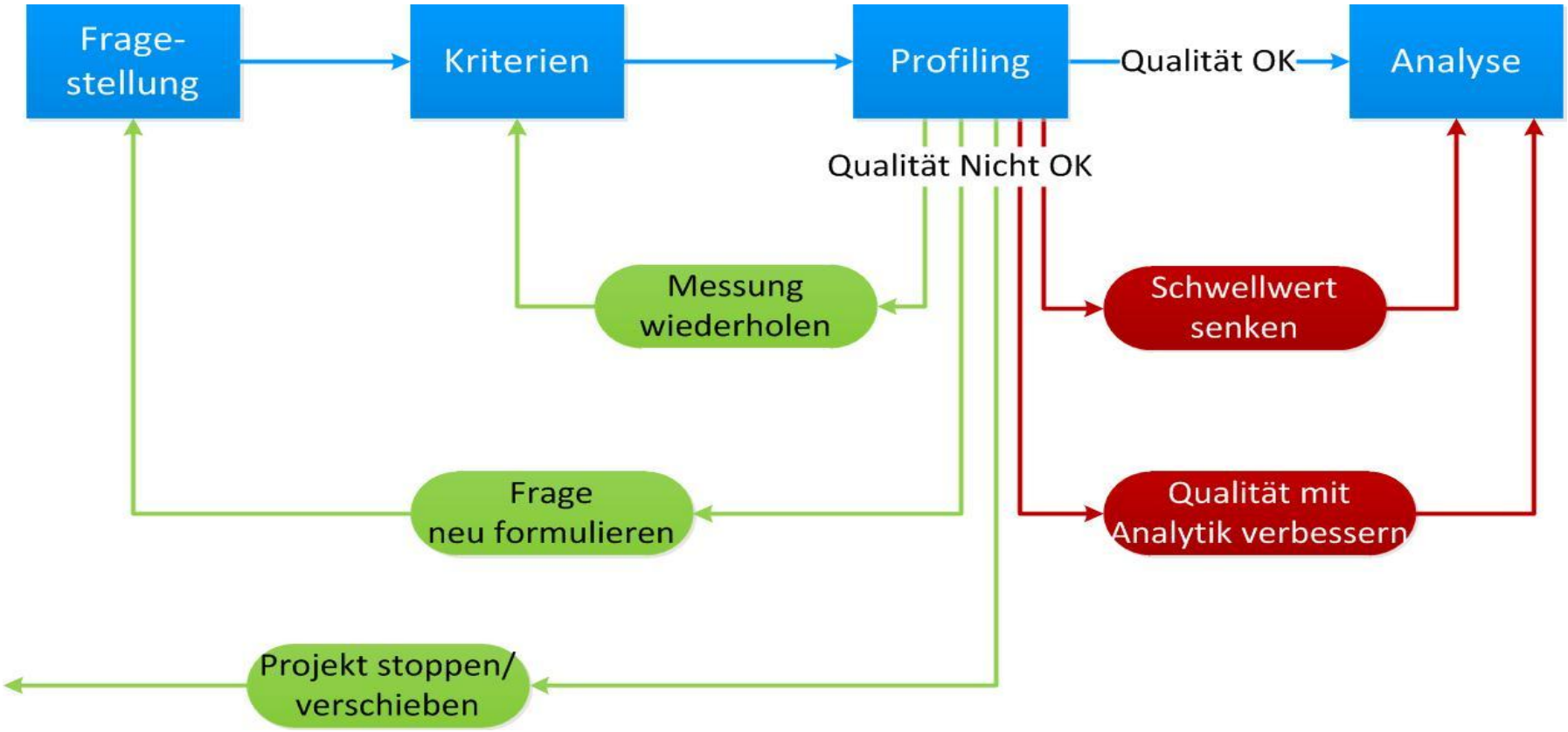
Ihre Datenqualität ist nicht so gut, wie sie dachten.

Was sind die Konsequenzen für analytische Auswertungen?

Dr. Gerhard Svolba

THE POWER TO KNOW.

Konsequenzen schlechter Datenqualität



Kosten	**Vertrauen**
Zeit	**Risiko falscher**
Verzögerungen	**Entscheidungen**
Keine Resultate	**Insignifikanz**

Die Konsequenzen der folgenden Effekte wurden untersucht:

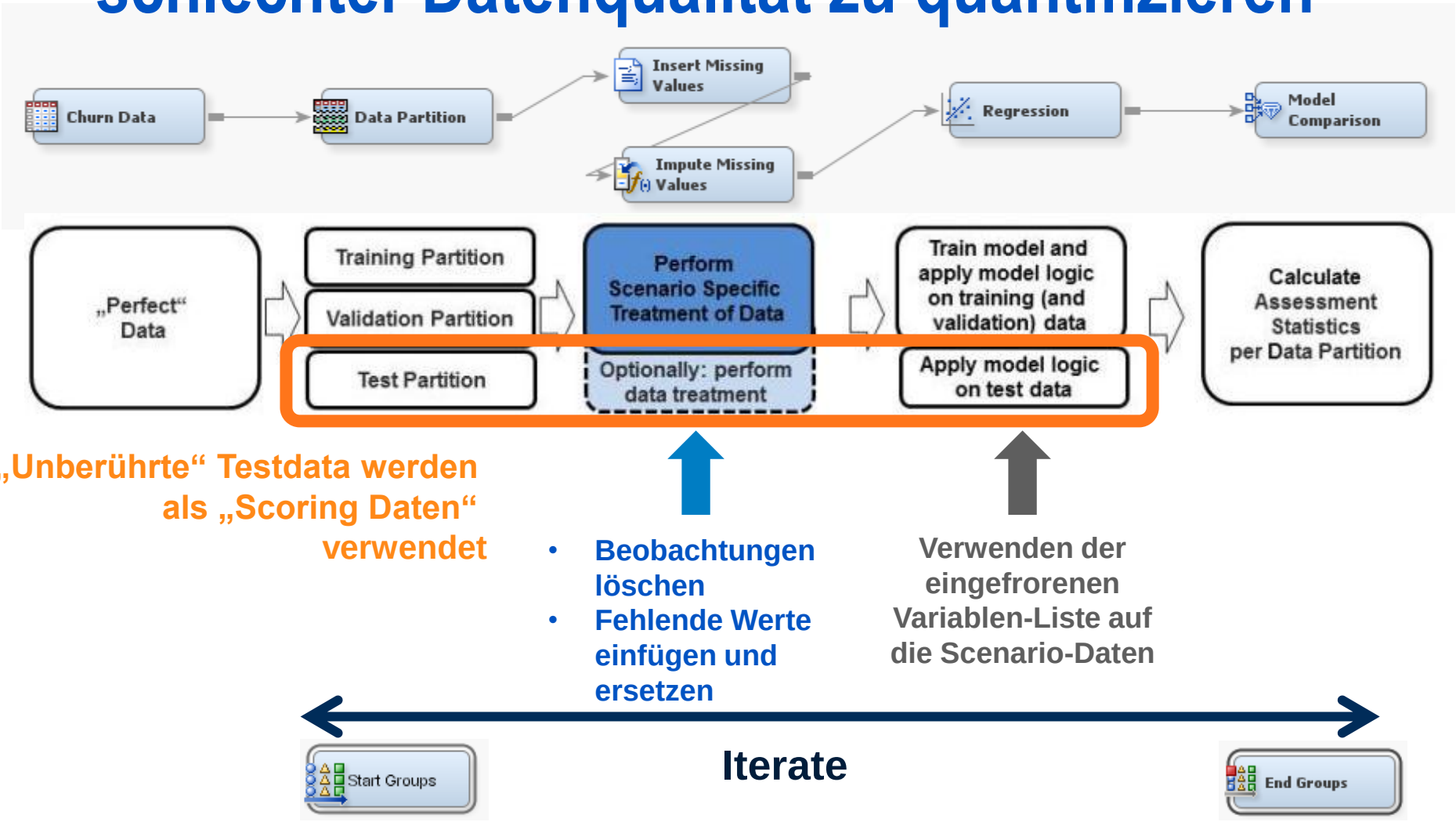
- Wie beeinflussen fehlende Werte die Vorhersagekraft? (SIM_1)
- Variation der Anzahl der verfügbaren Beobachtungen und Ereignisse (SIM_2)
- Schrittweise Erhöhung der verfügbaren Länge der Datenhistorie (SIM_3)
- Andere Fragen / Simulationen
  - Entfernen der wichtigsten Variablen aus dem Input-Set
  - Einführung zufälliger und systematischer Verzerrungen in der Input- und Zielgröße im Predictive Modeling
  - Auswirkung von zufälligen und systematischen fehlenden Werte und Fehlern in der Zeitreihen Prognose

Echtdaten wurden für die Simulationsstudien verwendet

- 4 Echt-Datensätze aus unterschiedlichen Branchen mit einem binären Target wurden verwendet
- Variablen mit > 5 % fehlenden Werten wurden entfernt
- Bei Variablen mit ≤ 5 % fehlenden Werten wurden Beobachtungen mit Missings gefilert
- Durchlauf multipler Modell-Zyklen um ein stabiles Modell mit guter Vorhersagekraft zu bekommen

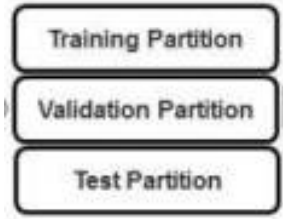
ID	PLCYDAT	TRAVTIM	CAR_USE	POLICYN	BLUEBOO	INITDAT	RETAINE	NPOLICY	CAR_TYP	RED_CA	AGE	HOMEKID	YOJ	INCOME
13276104	29Jul1998	22	Commerci	24836	$14.940	29Jul1998	1	1	Sedan	yes	43	0	11	$91.449
3577610	06May1994	30	Commerci	93553	$5.900	########		1	SUV	no	44	2	12	$43.486
8568865	18May1996	48	Commerci	145326	$18.510	########	7	1	Van	no	50	0	7	$106.952
3764824	18Sep1995	25	Commerci	157851	$17.600	02Oct1988	7		Van	yes	55	0	11	$59.162
69165559	30Jan1995	7	Commerci	328102	$25.660	30Jan1995		1	Panel Truck	no	46	0	14	$59.162
29481629	14Aug1997	5	Commerci	427058	$27.200	########	4	2		no	36	2	12	$130.540
2693415	19Jan1997	12	Commerci	479082	$25.810	21Jan1991	6	1	Panel Truck	no	41	0	7	$92.842
3969722	12Apr1997	38	Commerci	492179	$16.640	12May1993		3	Van	yes	39	0	10	$35.081
17373789	05Sep1995	38	Commerci	536131	$29.450	27Sep1989	6	2	Van	no	43	2	17	$145.353
70821537	02Jan1994	48	Commerci	568498	$9.280	02Jan1994	1		SUV	no	52	0	0	$0
25895102	24Nov1998	22	Commerci	574750	$6.450	24Nov1998	1	1	Pickup	no	35	0	7	$14.508
25561360	22Nov1997	5	Commerci	589112	$29.270	23Nov1993		1	Panel Truck	yes	40	0	11	$57.474
75303717	08Nov1995	62	Commerci	595977	$4.600	09Nov1989		1	Sedan	yes	31	1	12	$26.520
88374800	20May1998	35	Commerci	609510	$19.450	21May1992	6	1	SUV	no	47	0	8	$18.444
21305754	22Sep1994	24	Commerci	680397	$23.450	22Sep1994	1	1	Sedan	no	42	0	8	$52.988
40120026	07Apr1999	14	Commerci	746346	$28.560	12May1986	13	2	Panel Truck	yes	48	0	8	$52.988
9350798	01Jul1994	31	Commerci	809783	$33.710	02Jul1991	3		Panel Truck	no	49	0	9	$52.988
2990245	21Feb1994	22	Commerci	825520	$33.320	22Feb1990	4		Panel Truck	yes	46	0	14	$52.988
3939254	19May1996	30	Commerci	838306	$10.910	########	13	2	Pickup	yes	45	0	11	$61.931
53660483	22Jul1998	40	Commerci	891607	$23.230	06Jun1992	6		Panel Truck	yes	48	0	9	$61.509
80332520	04Jul1995	35	Commerci	1026106	$22.050	06Jul1985	10	1	Sports Car	no	44	2	15	$139.330
46606719	12Jun1997	41	Commerci	1049545	$17.470	15Jun1987	10	1	Van	no	38	0	0	$0
59798463	30Aug1995	21	Commerci	1103124	$5.760	########		1		yes	55	0	0	$84.964
59358283	04Jul1993	18	Commerci	1253235	$23.390	07Jul1980		1	Panel Truck	yes	42	0	9	$76.226
17048324	28May1997	36	Commerci	1331376	$15.120	05Jun1997	1	2	Sedan	no	52	0	9	$68.992
42131636	01Jul1996	8	Commerci	1344988	$30.150	########	16		Panel Truck	no	43	0	11	$132.561
37074846	27Jul1994	21	Commerci	1350845	$7.500	29Jul1985	9	1	Pickup	yes	60	0	9	$125.893
71829426	21Oct1995	14	Commerci	1440036	$17.550	21Oct1995		1	Van	no	37	0	12	$123.520
53887573	17May1997	33	Commerci	1441776	$29.210	02Jun1991		3	Panel Truck	yes	47	0	13	$45.257
52095933	11Dec1998	53	Commerci	1488986	$13.050	12Dec1993	5	1	Sports Car	no	40	3	9	$75.516
56847373	28Dec1994	40	Commerci	1597160	$36.120	16Dec1990	4	2	Panel Truck	yes	47	0	12	$104.271
45386733	30Apr1998	35	Commerci	1668247	$28.180	########	1	3	Van	no	33	1	12	$111.427
1307371	17Oct1998	5	Commerci	1699144	$19.800	19Sep1994		2	Pickup	no	44	0	14	$44.790
58208610	20Apr1998	11	Commerci	1787809	$17.300	########		2	Van	yes	50	0	14	$44.790
68042599	19Mar1995	32	Commerci	1816497	$11.620	30Mar1995	1	2	Pickup	no	51	0	9	$50.166
93412915	06Dec1994	50	Commerci	1875494	$14.530	09Dec1982	12	1	Sports Car	no	43	3	14	$48.184
46157391	18Jun1994	24	Commerci	2030290	$21.990	21Jun1983	11		Pickup	no	49	0	8	$22.059
22521555	13Apr1995	35	Commerci	2035565	$12.180	########	7		Pickup	yes	34	1	10	$23.571
68337846	28Oct1993	45	Commerci	2070960	$27.890	05Nov1989	4	2	Panel Truck	no	55	0	14	$55.409
50390643	02Sep1997	48	Commerci	2137830	$8.460	05Sep1987	10	1	Pickup	yes	48	0	10	$39.613
19707619	06Jun1995	9	Commerci	2219845	$34.510	07Jun1988	7	1	Van	no	45	3	15	$23.773
34577136	01Jun1996	17	Commerci	2235985	$31.390	25Apr1996	1	2	Panel Truck	yes	41	0	8	$55.364
83085566	29Aug1993	44	Commerci	2321817	$23.320	01Sep1983	10	1	Panel Truck	no	55	0	16	$163.158
64298739	18Jan1996	59	Commerci	2364349	$10.350	21Jan1982	14	1	Sedan	yes	52	0	12	$24.590
93092921	03Apr1999	45	Commerci	2365874	$27.020	04Apr1992	7	1	Pickup	no	54	0	15	$107.808
39176001	20Dec1994	42	Commerci	2392645	$7.470	22Dec1987	7	1	Pickup	no	39	2	12	$59.685
84194086	25Jun1995	51	Commerci	2465546	$6.720	26Jun1991	4	1	Pickup	no	44	0	4	$146.267
60189132	20Apr1996	10	Commerci	2546478	$6.120	23Apr1986	10	1	SUV	no	45	3	4	$1.158
79219605	07Oct1997	32	Commerci	2577125	$13.160	23Oct1984	13	2	Sedan	no	43	0	11	$1.158

Simulationsstudien helfen die Konsequenzen schlechter Datenqualität zu quantifizieren



Prozess für die Szenarios mit fehlenden Werten (SIM_1)

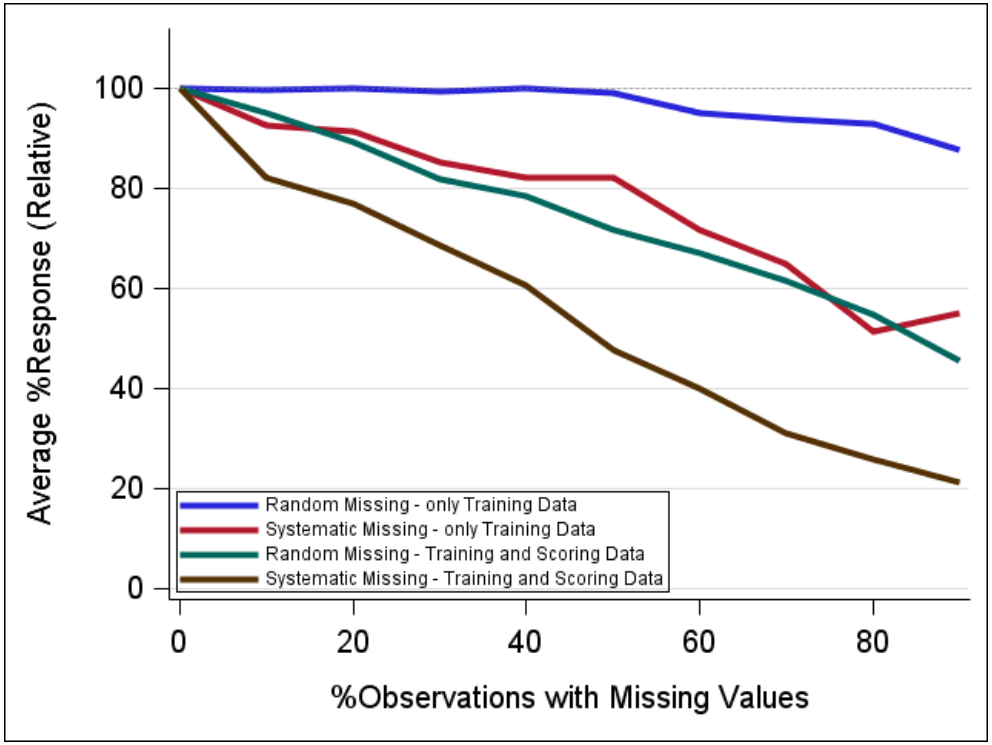
- Für einen bestimmten Anteil von Beobachtungen (10%, ...)
  - Setzen von Intervall-skalierten und nominalen Input-Variablen auf fehlend
    - » Zufällige Auswahl
    - » Systematische Auswahl auf Basis von Segmenten
  - Ersetzen der fehlende Werte durch den Impute Node im SAS Enterprise Miner
- Trainieren des Modells mit dem eingefrorenen Set an Variablen
- Optional: Anwenden der "Behandlung" auch für Scoring-Daten
- Bewerten der Modellqualität



ID	TRAVTIM	BLUEBOO	INITDATE	RED_C	AGE	INCOME
13276104	22	$14.940	29Jul1998	yes	43	$91.449
8568865	48	$18.510	########	no	40	$106.952
3764824	25	$17.600	02Oct1988	yes	55	$59.162
6916555	7	$17.230	30Jan1995	no	46	$59.162
2693415	12	$25.810	21Jan1991	no	40	$59.162
3969225	38	$16.640	12May1993	yes	39	$35.081
1737378	38	$29.450	27Sep1989	no	43	$145.353
7082153	48	$9.280	02Jan1994	no	52	$0
2589510	22	$6.450	24Nov1998	no	35	$14.508
2556136	23	$29.270	23Nov1993	yes	40	$57.474
7530371	62	$4.600	09Nov1989	yes	31	$26.520
2130575	24	$23.450	22Sep1994	no	42	$52.988
4012002	14	$28.560	12May1986	yes	48	$52.988
9350798	31	$33.710	02Jul1991	no	49	$52.988
2990245	23	$33.320	22Feb1990	yes	40	$52.988
3939254	30	$10.910	########	yes	45	$61.931
5366048	40	$23.230	06Jun1992	yes	48	$61.509
8033252	35	$22.050		no	44	$139.330
4660671	41	$17.470	15Jun1987	no	38	$0
5935828	18	$23.390	07Jul1980	yes	42	$76.226
1704832	36	$17.230	05Jun1997	no	52	$59.162
42131636	8	$30.150	########	no	43	$132.561
3707484	21	$7.500	29Jul1985	yes	40	$125.893
7182942	14	$17.550	21Oct1995	no	40	$123.520
5388757	33	$29.210	02Jun1991	yes	47	$45.257
5209593	53	$13.050	12Dec1993	no	40	$75.516
5684737	40	$36.120	16Dec1990	yes	47	$104.271
4538673	35	$28.180	########	no	33	$111.427
5820861	11	$17.300	########	yes	50	$111.427
6804259	23	$11.620	30Mar1995	no	51	$50.166
93412915	50	$14.530	09Dec1982	no	43	$48.184
46157391	24	$21.990	21Jun1983	no	40	$22.059
2252155	35	$12.180	########	yes	40	$23.571
6833784	45	$27.890	05Nov1989	no	55	$55.409
5039064	48	$8.460	05Sep1987	yes	48	$39.613
19707619	9	$17.230	07Jun1988	no	45	$23.773
3457713	17	$31.390	25Apr1996	yes	40	$55.364
8308556	44	$23.320	23Jan1900	no	55	$163.158
6429873	23	$10.350	21Jan1982	yes	52	$59.162
9309292	45	$27.020	04Apr1992	no	54	$107.808
39176001	42	$7.470	22Dec1987	no	39	$59.685
8419408	51	$6.720	26Jun1991	no	44	$146.267
60189132	10	$6.120	23Apr1986	no	45	$1.158
7921960	32	$13.160	23Oct1984	no	43	$1.158

Ergebnisse aus den Missing Value Szenarios

- Zufällig fehlende Werte nur in den Trainingsdaten haben nur beschränkte Auswirkungen.
- Fehlende Werte in den Scoringdaten beeinflussen das Ergebnis viel stärker.
- Systematische fehlende Werte haben einen viel stärkeren Effekt
- Punkte die wichtig sind:
  - Nicht nur der Anteil fehlender Werte, sondern insbesondere der Typ
  - Fehlende Werte in den Scoring Daten



Quantifizieren der Ergebnisse der Missing Value Szenarios

- Durchführen eines allgemeinen linearen Modells:

  Response = f(%missing, Systematic_YN, ScoringData_YN)

Parameter	Value	Interpretation
Intercept	19.29	Response ohne fehlende Werte
%missing	- 0.1	10 % fehlend ~ 1% weniger Response
Systematic_YN	- 3.6	Systematische Fehler resultieren in 3.6 % weniger Response
Scoring_YN	- 4.23	Fehlende Werte in den Scoring Daten resultieren in 4.23 % weniger Response

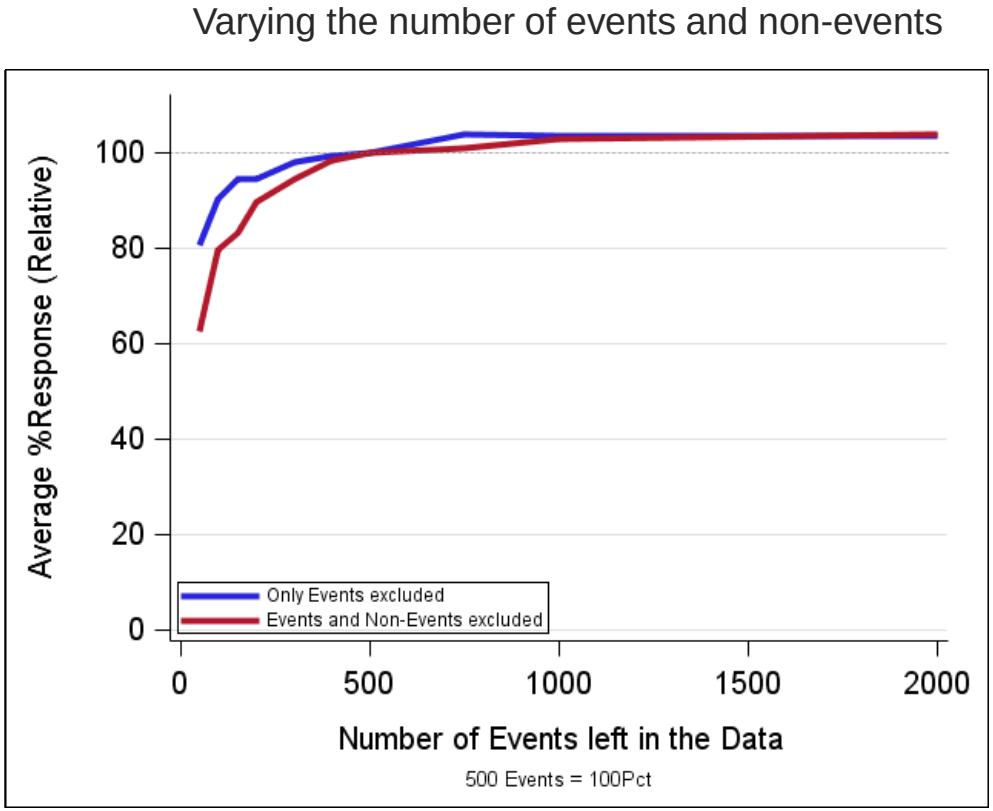
Analyse des Effekts der Datenquantität in der Ereignisvorhersage (SIM_2)

- Zufällig ausgewählte Beobachtungen wurden aus den Daten entfernt
- Zusätzliche Beobachtungen und Ereignisse resultieren in einer höheren %Correct Response Rate
- Aber:
  - Linearer oder nicht-linearer Effekt
  - Wie kann dieser Effekt quantifiziert werden?
  - Tragen auch Non-Events zu einer Steigerung bei?
  - Zahlt es sich aus, auf weitere Events zu warten?

ID	TRAVTIM	BLUEBOO	INITDATT	RED_C/	AGE	INCOME
13276104	22	$14.940	29Jul1998	yes	43	$91.449
8568865	48	$18.510	########	no	50	$106.952
3764824	25	$17.600	02Oct1988	yes	55	$59.162
26934155	12	$25.810	21Jan1991	no	41	$92.842
3969722	38	$16.640	12May1993	yes	39	$35.081
17373785	38	$29.450	27Sep1989	no	43	$145.353
7082153	48	$9.280	02Jan1994	no	52	$0
25561360	5	$29.270	23Nov1993	yes	40	$57.474
7530371	62	$4.600	09Nov1989	yes	31	$26.520
21305754	24	$23.450	22Sep1994	no	42	$52.988
4012002	14	$28.560	12May1986	yes	48	$52.988
9350798	31	$33.710	02Jul1991	no	49	$52.988
2990245	22	$33.320	22Feb1990	yes	46	$52.988
3939254	30	$10.910	########	yes	45	$61.931
5935828	18	$23.390	07Jul1980	yes	42	$76.226
17048324	36	$15.120	05Jun1997	no	52	$68.992
42131636	8	$30.150	########	no	43	$132.561
3707484	21	$7.500	29Jul1985	yes	60	$125.893
7182942	14	$17.550	21Oct1995	no	37	$123.520
53887578	33	$29.210	02Jun1991	yes	47	$45.257
5209593	53	$13.050	12Dec1993	no	40	$75.516
5684737	40	$36.120	16Dec1990	yes	47	$104.271
58208610	11	$17.300	########	yes	50	$111.427
6804259	32	$11.620	30Mar1995	no	51	$50.166
93412915	50	$14.530	09Dec1982	no	43	$48.184
46157391	24	$21.990	21Jun1983	no	49	$22.059
22521555	35	$12.180	########	yes	34	$23.571
68337845	45	$27.890	05Nov1989	no	55	$55.409
5039064	48	$8.460	05Sep1987	yes	48	$39.613
19707619	9	$34.510	07Jun1988	no	45	$23.773
8308556	44	$23.320	01Sep1983	no	55	$163.158
64298735	59	$10.350	21Jan1982	yes	52	$24.590
93092928	45	$27.020	04Apr1992	no	54	$107.808
39176001	42	$7.470	22Dec1987	no	39	$59.685
60189132	10	$6.120	23Apr1986	no	45	$1.158
79219605	32	$13.160	23Oct1984	no	43	$1.158
34577130	17	$31.390	25Apr1996	yes	41	$55.364
8308556	44	$23.320	01Sep1983	no	55	$163.158
64298735	59	$10.350	21Jan1982	yes	52	$24.590
93092928	45	$27.020	04Apr1992	no	54	$107.808
39176001	42	$7.470	22Dec1987	no	39	$59.685
84194080	51	$6.720	26Jun1991	no	44	$146.267
60189132	10	$6.120	23Apr1986	no	45	$1.158
79219605	32	$13.160	23Oct1984	no	43	$1.158

Ergebnisse aus den Datenquantitäts Scenarios

- Grenznutzen wird im Bereich von 500 bis 1000 Ereignissen flach
- Auch Nicht-Ereignisse bieten zusätzliche Informationen, insbesondere im Bereich von bis zu 500 Ereignissen

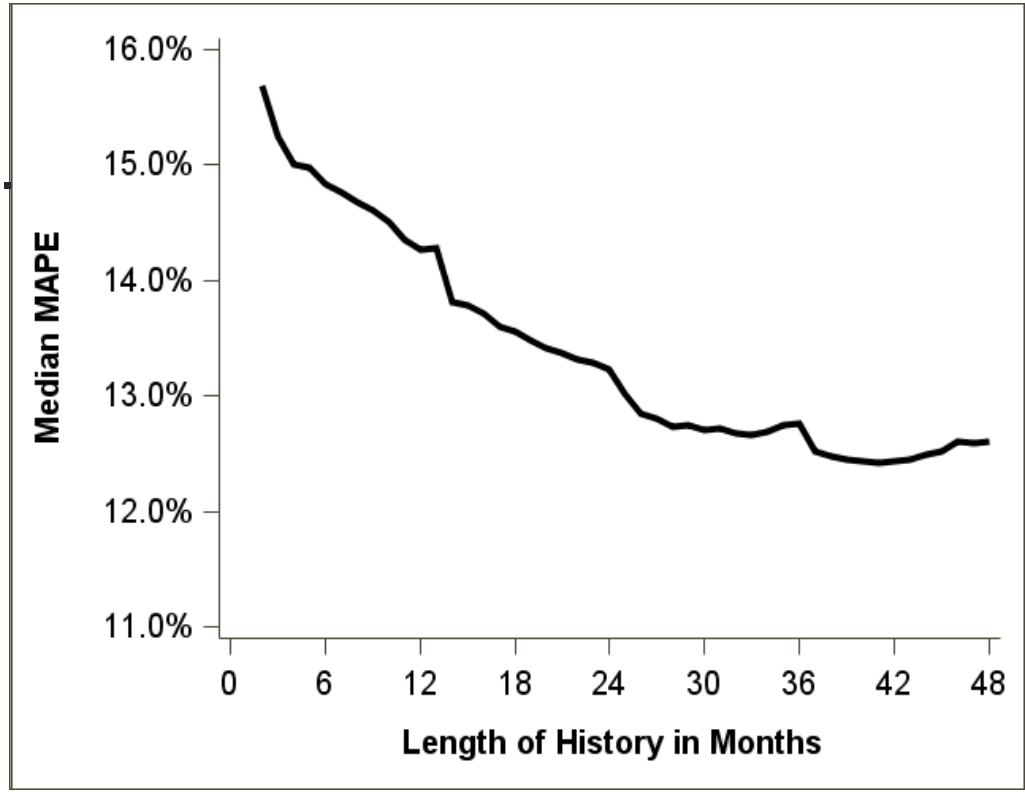


Kontinuierliches Erweitern der verfügbaren Länge der Historie im Zeitreihen-Forecasting

- Fachliche Fragen
  - Ist es möglich, Zeitreihen-Analyse zu starten, wenn nur 18 Monate Historie verfügbar sind?
  - Soll auf weitere Historien-Monate gewartet werden?
  - Was ist der Vorteil zusätzlichen Daten-Management-Aufwands?
  - Was ist die beste Länge der Datenhistorie für Zeitreihen Prognose?
- Methoden
  - Simulationsumgebung mit SAS High Performance Forecasting
  - Auf Basis von 788 Zeitreihen auf monatlichen Daten aus verschiedenen Branchen
  - Minimum Historie für jede Zeitreihe: 48 Monate
  - Eingeschränkt auf Prognoseverfahren "exponentielle Glättung
  - Validierung auf MAPE auf 12 Zukunftsmonate
  - Iteration durch Verschieben der „Zero-Time“

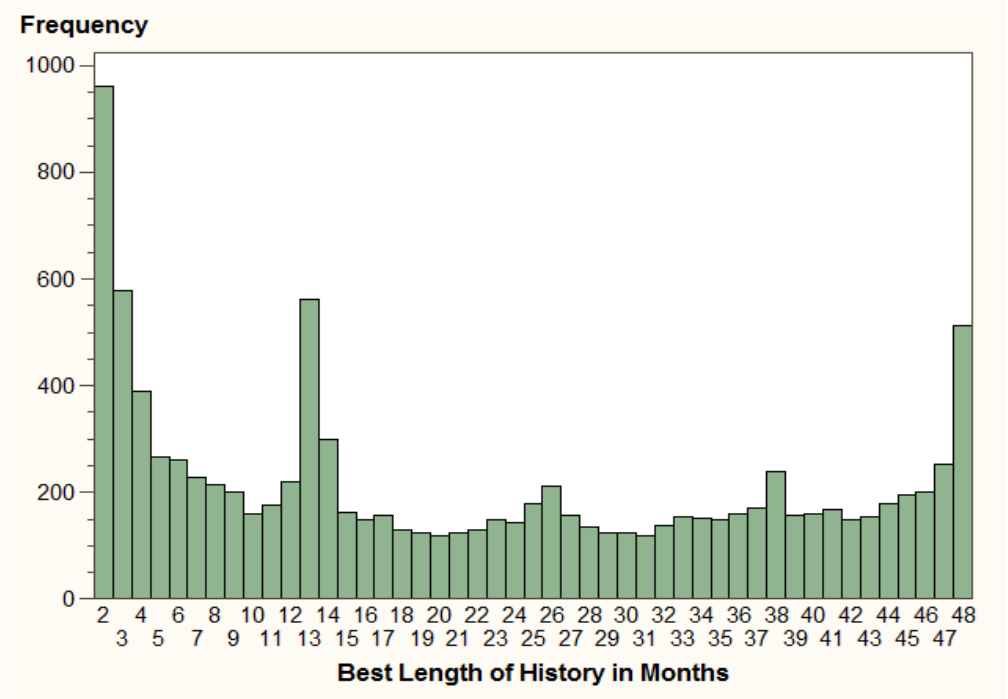
Wie weit sollten wir uns „zurückerinnern“?

- Die erwartete Abnahme des MAPE mit zunehmender Historienlänge ist erkennbar.
- Es gibt eine exponentielle Abnahme der zusätzlichen Werts weiterer Monate.
- Stärkere Sprünge nach 12, 24 und 36 Monaten sind sichtbar.



Was ist die beste Länge der Datenhistorie für Zeitreihen-Forecasting?

- Methode: für jede Zeitreihe wurde ermittelt, welche Anzahl von Historien-Monaten den kleinsten Fehler für die künftigen 12 Monate ergibt.
- Ergebnis: nicht in allen Fällen ist es vorteilhaft die lange Zeithistorie zu verwenden.



Abschließende Kommentare

- Datenqualität für Analytik ist mehr
  - Mehr Anforderungen
  - Mehr Möglichkeiten
- Gehen Sie ins Detail!
  - Zufällige oder systematische Fehler?
  - Permanenten oder historische/temporäre Qualitätsprobleme?
- Datenmenge macht einen Unterschied!
  - Aber balancieren Sie Aufwand und Nutzen!
- SAS unterstützt beim
  - Profiling, Verbessern, Bewerten, Simulieren der Datenqualität.