Prävention von Abonnementkündigungen mit Hilfe von Predictive Analytics

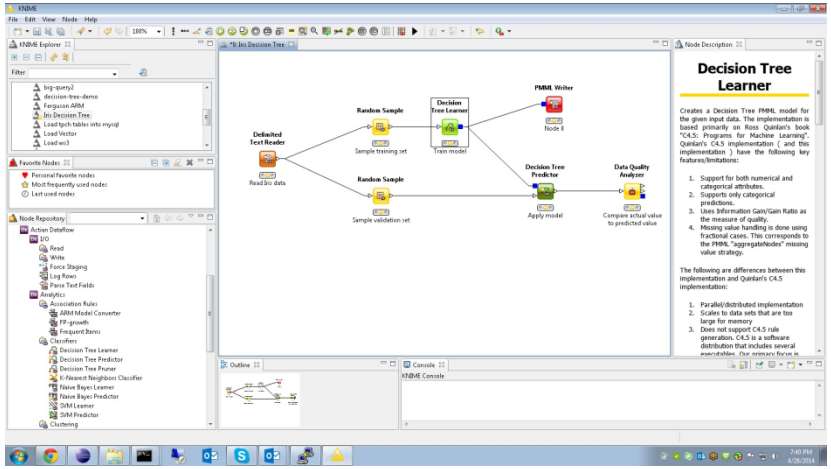
Ziel

- Entwicklung eines Modells zur Vorhersage von Kundenabgängen (Churn Prediction)
  → Kündigt Kunde 23447283 in einer definierten künftigen Periode?

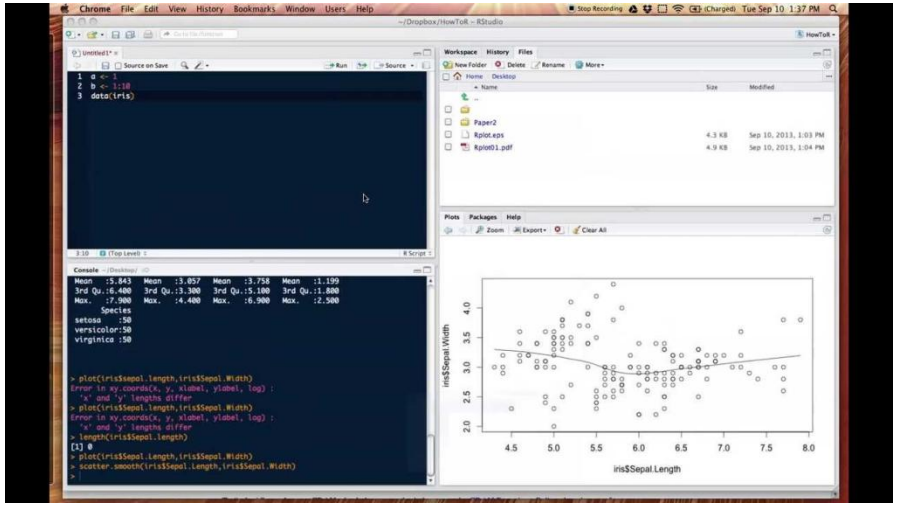
- Ableitung der wichtigsten Prädiktoren (Indikatoren) für einen Kundenabgang
  → Bsp.: Kunden mit Zahlungsart „Rechnung“ kündigen häufiger
  → Bsp.: Kunden, die ihre Kündigung androhten, kündigen danach häufiger

Tools

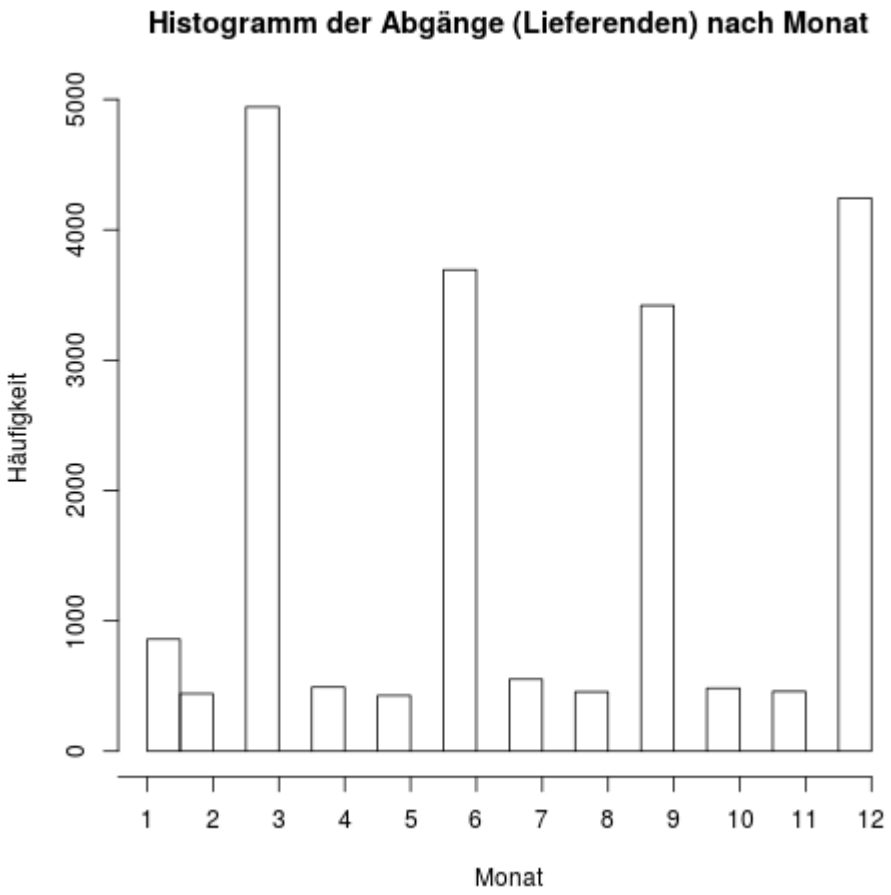
- KNIME Analytics Plattform



- R (Studio)



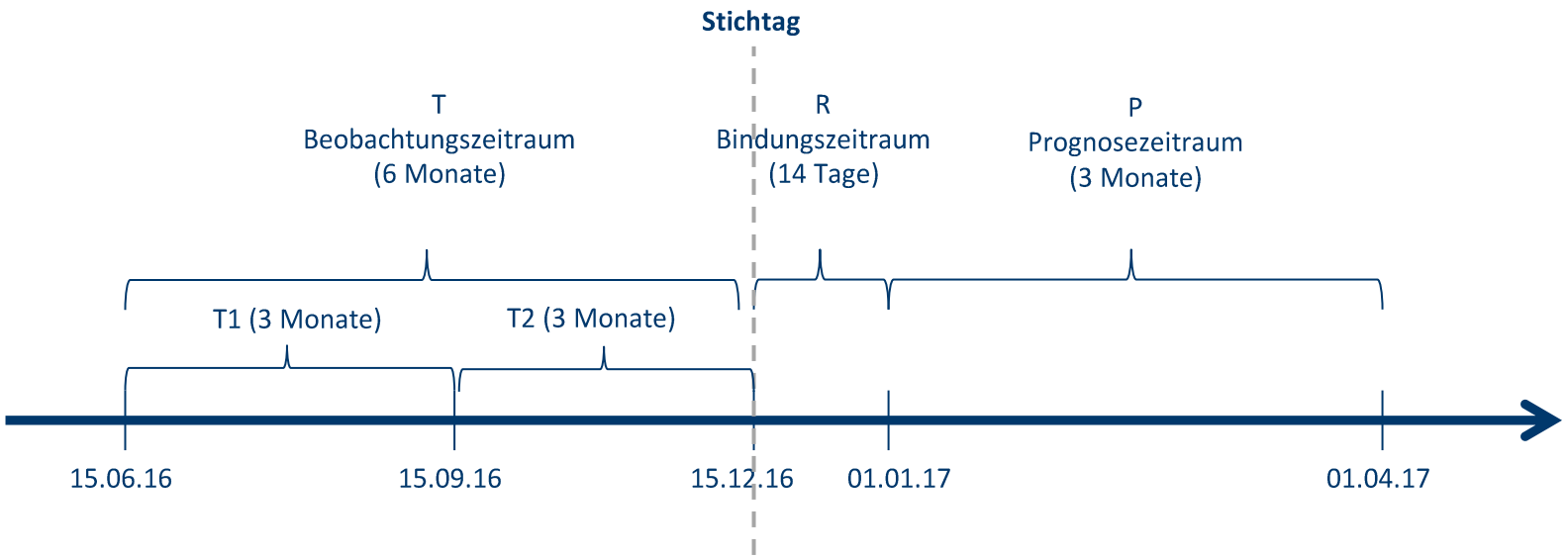
Kundenabgänge der letzten 5 Jahre



Methode

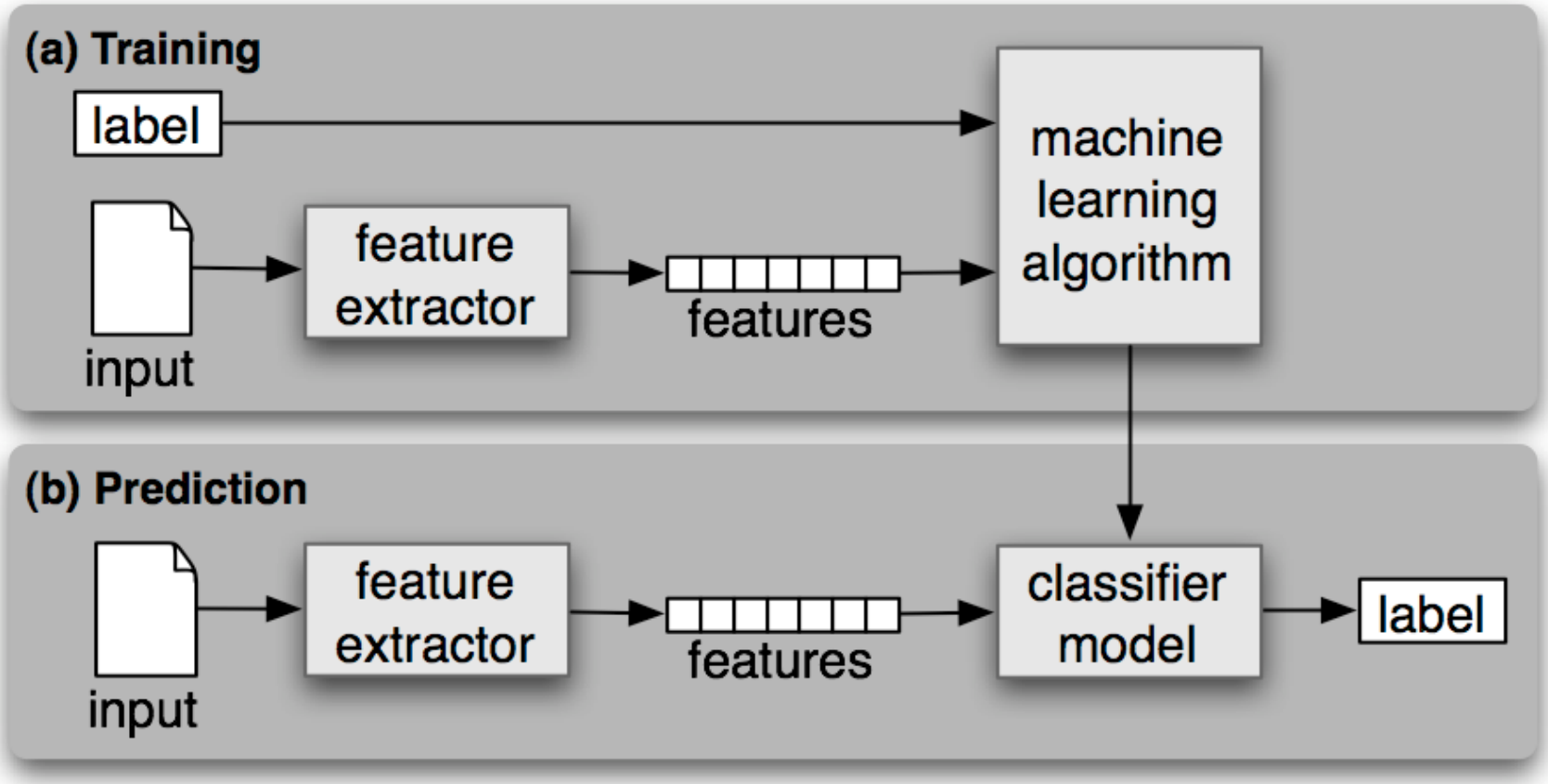
- Binäre Klassifikation
  - 1: kündigt im nächsten Quartal
  - 0: kündigt nicht im nächsten Quartal
- Algorithmen:
  - Tree-Modelle
    - Decision Tree (Entscheidungsbaum)
    - Random Forest (kombinierte Entscheidungsbäume)
    - Cost-sensitive Random Forest
  - Naive Bayes
  - Logistische Regression

Beobachtungszeitraum

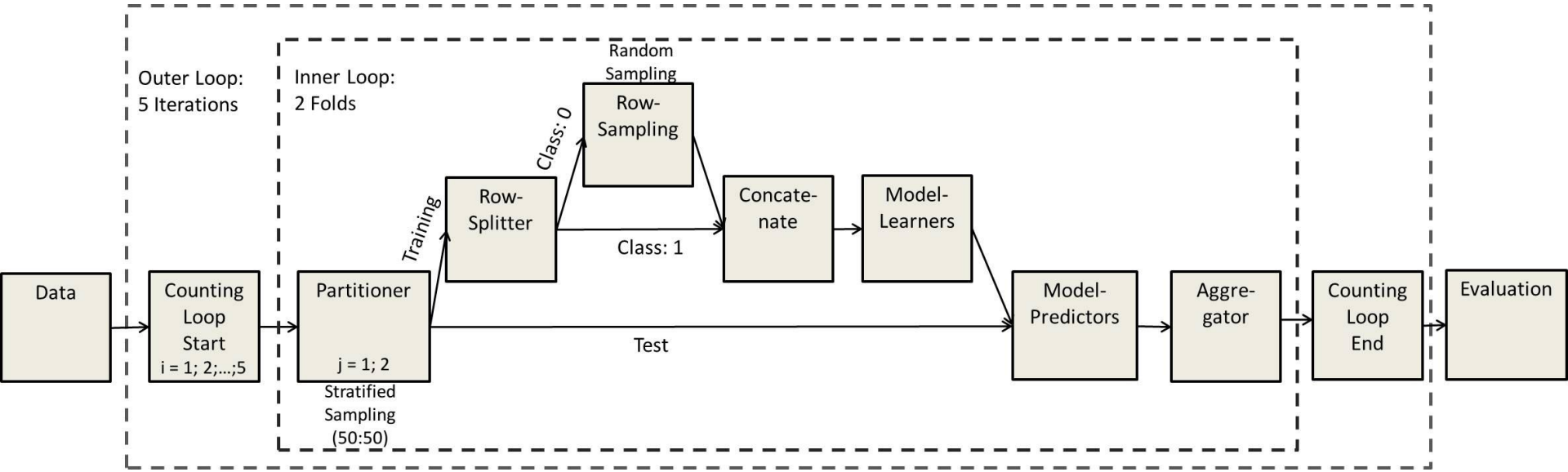


→ 8 Prognose-Quartale (Q3 2014 – Q2 2016) zu einem Trainingsdatenset kombiniert

Modellbildung und Evaluierung



Parameter-Tuning mittels 5x2 Cross-Validierung



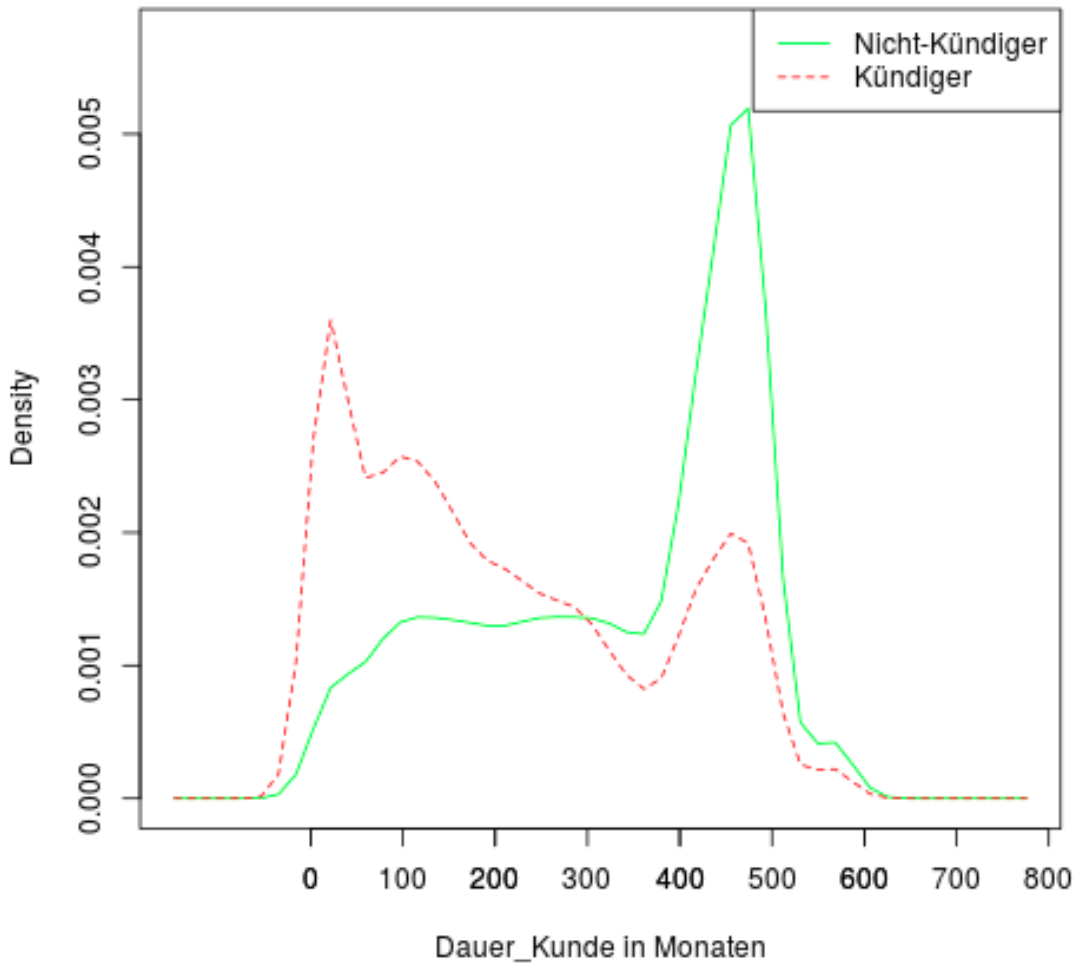
Datenkategorien

- Abo- und Abonnentendaten
  (u.a. Abotyp, Zahlungsart, Bestellkanal etc.)
- Soziodemographische Daten auf Gebäudeebene
  (u.a. Sinus Milieu, Status, Anzahl Haushalte/Gewerbe im Haus)
- Kontaktdaten
  (u.a. Anzahl Kontakte mit Unternehmen, Reklamationen, angedrohte Kündigungen)

Vorgehen

1. Data Integration
2. Pre-processing (Fehlende Werte, Binning)
3. Feature Elimination
4. Parameter-Tuning
5. Modellvergleich

Visualisierung – Beispiel: Dauer_Kunde



Problematik „Class Imbalance“

- Ausgangssituation: nur 0,5% Kündiger in den Trainingsdaten

Lösungsansätze

- Angemessene Evaluationskriterien
  - Sensitivity (Recall), Specificity & Precision
  - ROC/AUC
  - Lift/Gain Charts
- (random) Under-sampling
  - 5% Churner, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%
- Cost Sensitive Learning
  - Cost Sensitive Random Forest

Sensitivity, Specifity, Precision

- Sensitivity (Recall):
  - Wie viel Prozent der Kündiger wurden auch als Kündiger vorhergesagt/erfasst?

Tatsächlich \ Vorhersage	Nicht-Kündiger	Kündiger
Nicht-Kündiger	80.000 (TN)	5.000 (FP)
Kündiger	700 (FN)	200 (TP)

- Specificity
  - Wie viel Prozent der Nicht-Kündiger wurden auch als Nicht-Kündiger vorhergesagt?

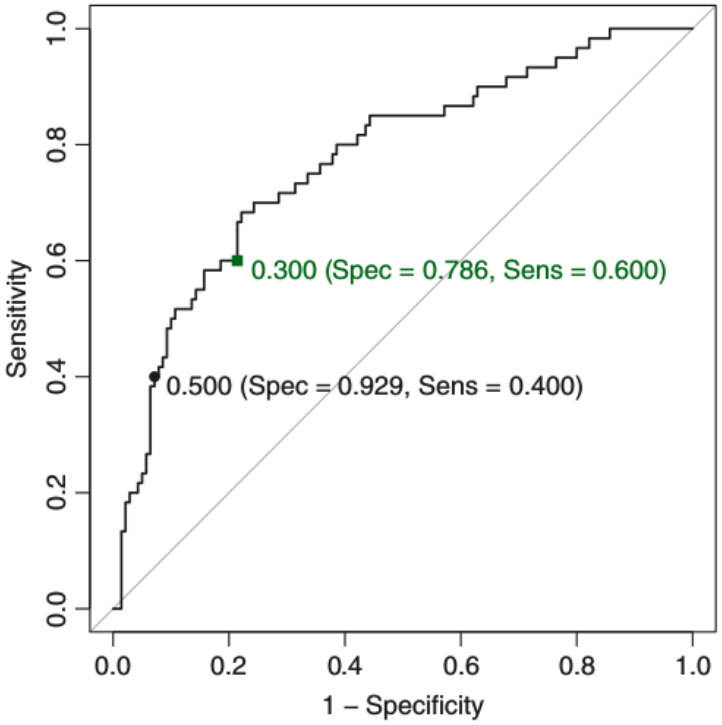
Tatsächlich \ Vorhersage	Nicht-Kündiger	Kündiger
Nicht-Kündiger	80.000 (TN)	5.000 (FP)
Kündiger	700 (FN)	200 (TP)

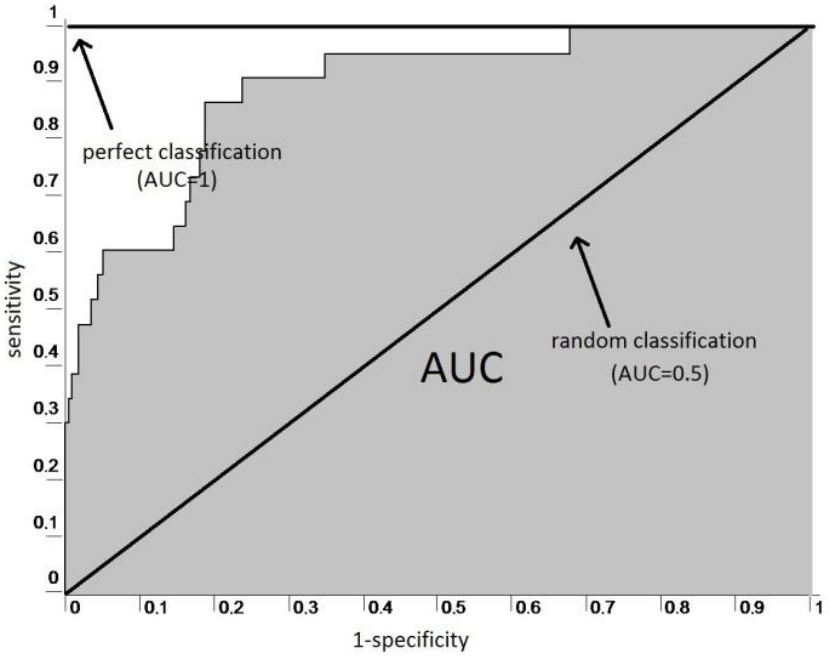
- Precision
  - Wie viel Prozent der vorhergesagten Kündiger sind auch tatsächlich Kündiger

Tatsächlich \ Vorhersage	Nicht-Kündiger	Kündiger
Nicht-Kündiger	80.000 (TN)	5.000 (FP)
Kündiger	700 (FN)	200 (TP)

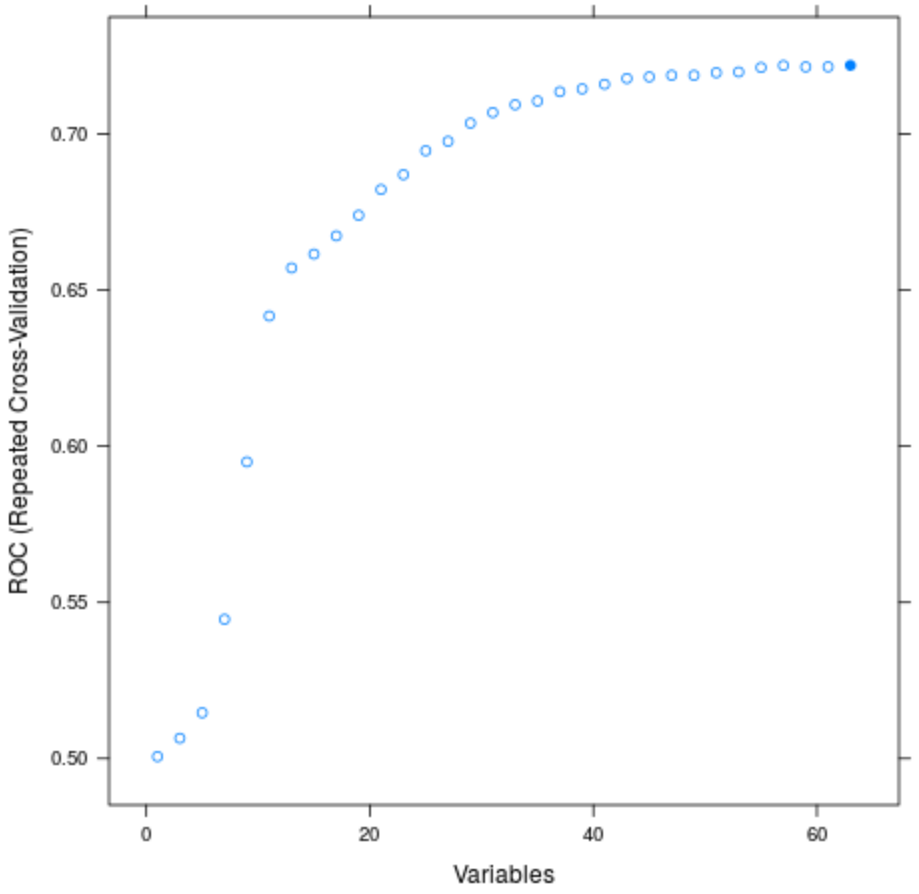
ROC und AUC

- ROC → Receiver Operating Characteristics





Die wichtigsten Prädiktoren (RFE mit R Package caret)



Cost-sensitive Learning

- Grundannahme:

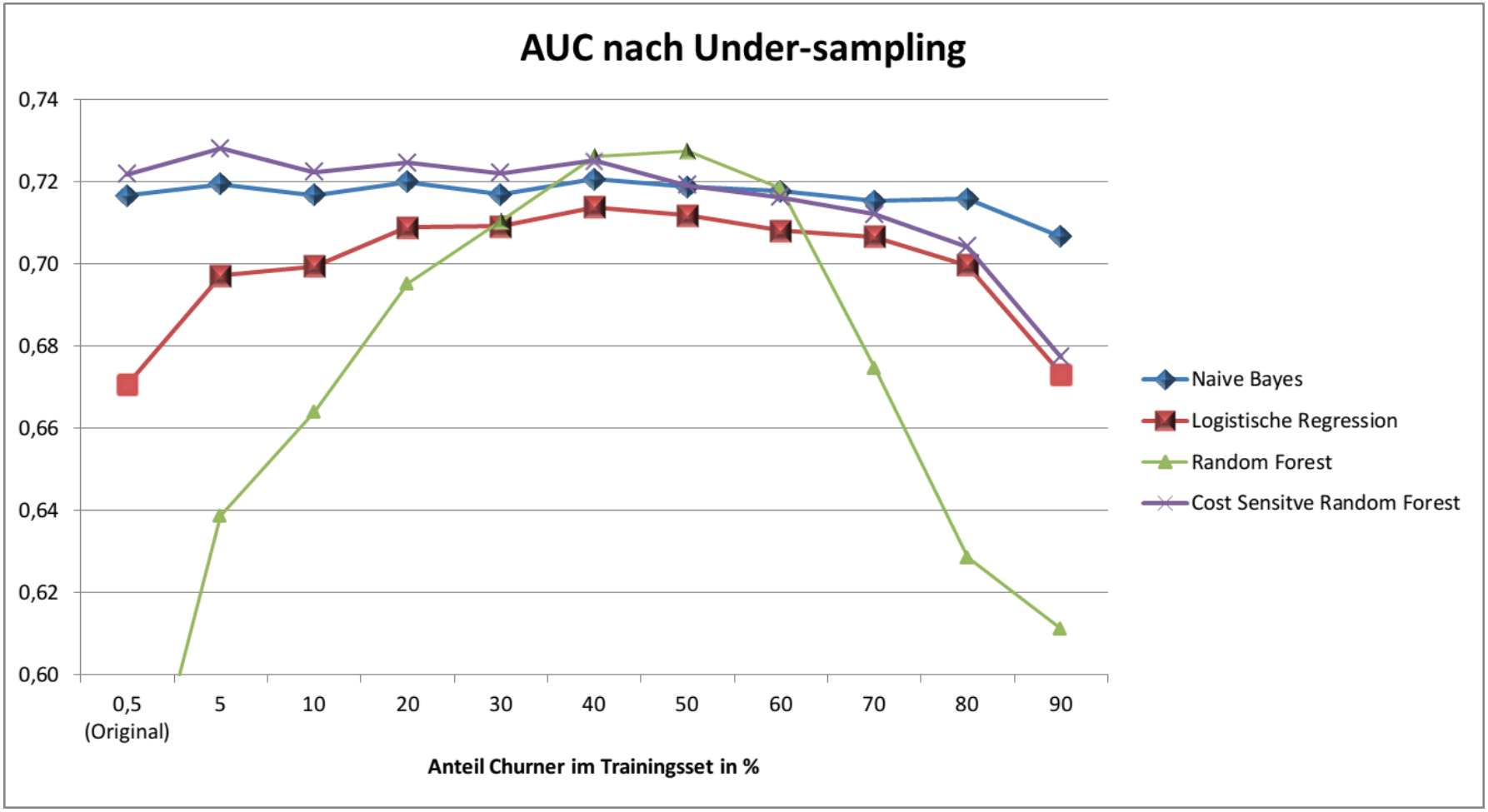
Es ist für das Unternehmen ungünstiger (teurer) einen tatsächlichen Kündiger nicht zu erkennen, als einen loyalen Kunden irrtümlich als Kündiger vorherzusagen

- Cost-Matrix

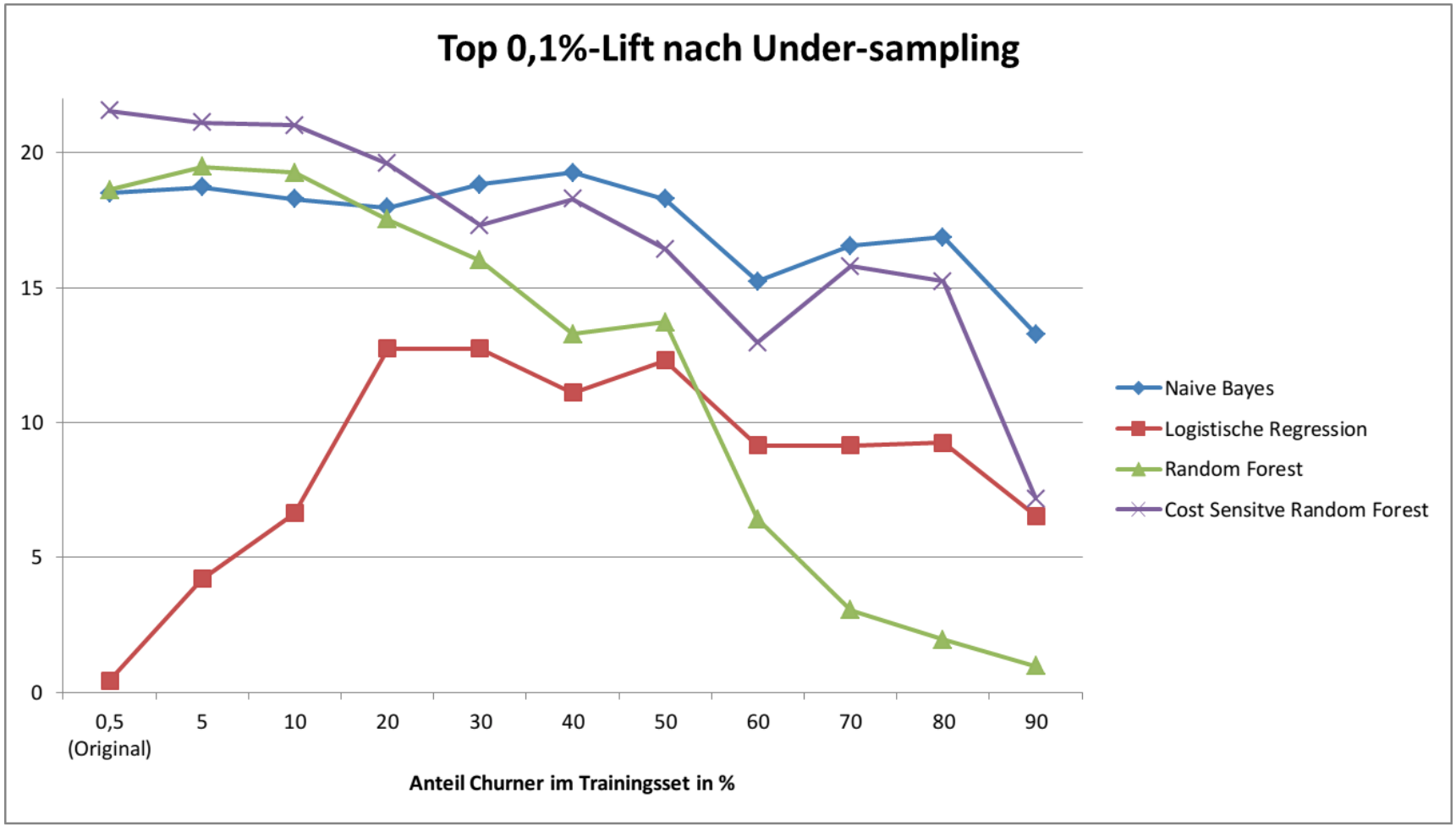
	Actual negative	Actual positive
Predict negative	C(0,0), or TN	C(0,1), or FN
Predict positive	C(1,0), or FP	C(1,1), or TP

	Actual negative	Actual positive
Predict negative	0	10
Predict positive	1	0

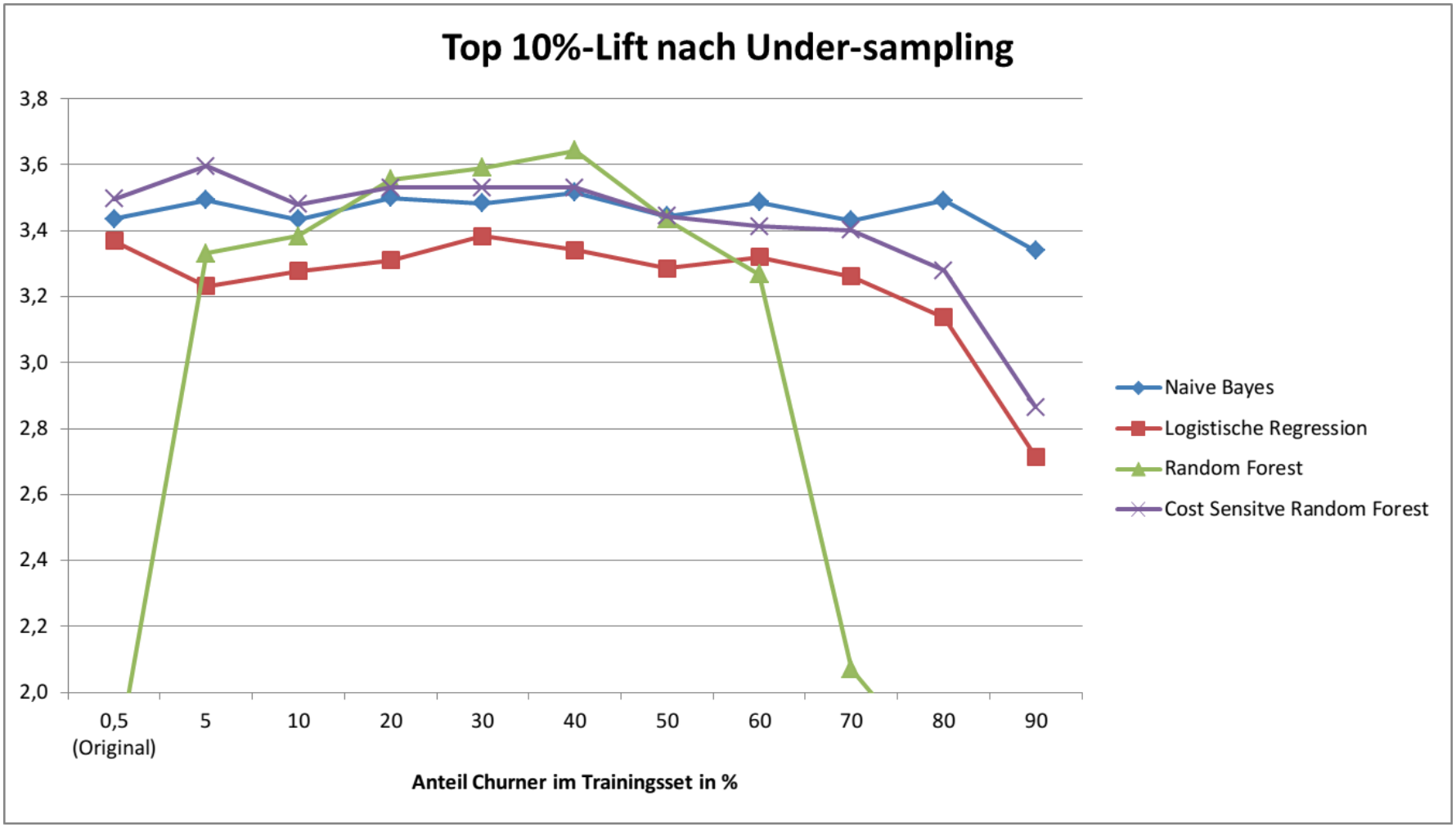
Ergebnisse – Under-sampling (AUC)



Ergebnisse – Under-sampling (0,1%-Lift)



Ergebnisse – Under-sampling (10%-Lift)



Ergebnisse mit neuen Daten
Confusion Matrix – Random Forest

Schwelle für Vorhersage (1) : 0.5

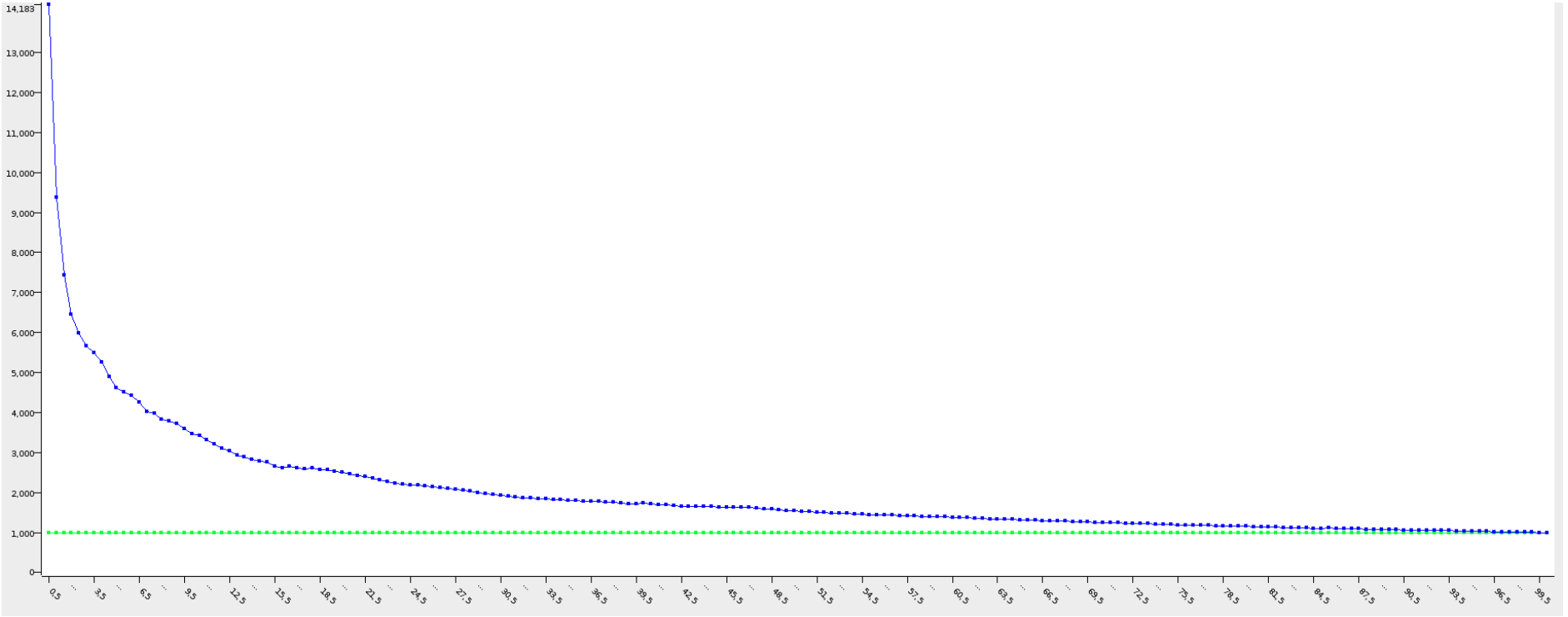
Tatsächlich \ Vorhersage	Nicht-Kündiger	Kündiger
Nicht-Kündiger	74.886 (TN)	7.549 (FP)
Kündiger	266 (FN)	128 (TP)

Sensitivity: 32,4% der Kündiger erfasst

Specificity: 90,8% der Nicht-Kündiger richtig vorhergesagt

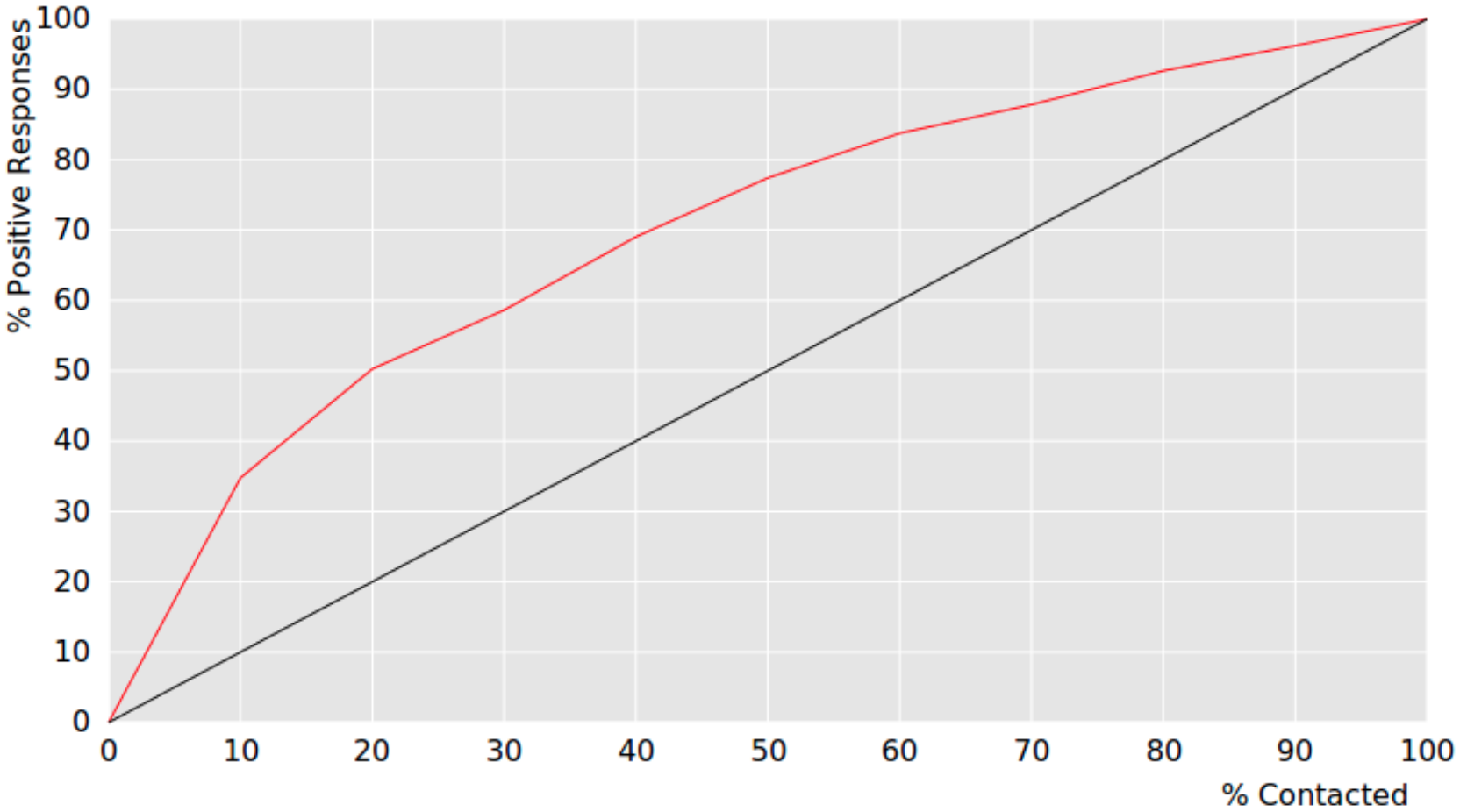
Precision: 1,7% der vorhergesagten Kündiger kündigten tatsächlich

Ergebnisse mit neuen Daten
Lift Chart – Cost Sensitive Random Forest



→ In den Top-Segmenten ist die Dichte der tatsächlichen Kündiger deutlich höher als bei einer zufälligen Auswahl von Kunden (Top-0,5%: ca. 14 mal höher)

Ergebnisse mit neuen Daten
Cumulative Gain Chart – Cost Sensitive Random Forest



→ Wenn die Top-20% der Kunden angesprochen werden, sind darunter 50% aller Kündiger

Ergebnisse mit neuen Daten
Algorithmen im Vergleich

	Algorithm.	Evaluationskrit.	Wert	TP	FP	TP:FP-Ratio
1	NB	AUC	0,690			
2	NB	Sensitivity	69,0%			
3	NB	Precision	0,8%			
4	NB	0,1%-Lift	17,5	7	77	1:11,0
5	NB	1%-Lift	6,8	27	802	1:29,7
6	NB	10%-Lift	3,1	123	8,161	1:66,3
7	DT	AUC	0,683			
8	DT	Sensitivity	41,6%			
9	DT	Precision	1,0%			
10	DT	0,1%-Lift	5,0	2	82	1:41,0
11	DT	1%-Lift	4,6	18	811	1:45,1
12	DT	10%-Lift	2,9	113	8,171	1:72,3
13	RF	AUC	0,702			
14	RF	Sensitivity	32,5%			
15	RF	Precision	1,7%			
16	RF	0,1%-Lift	10,0	4	80	1:20,0
17	RF	1%-Lift	8,6	34	795	1:23,4
18	RF	10%-Lift	3,4	134	8,622	1:60,8
19	CSRF	AUC	0,705			
20	CSRF	Sensitivity	0,0%			
21	CSRF	Precision	0,0%			
22	CSRF	0,1%-Lift	22,5	9	75	1:8,3
23	CSRF	1%-Lift	9,4	37	792	1:21,4
24	CSRF	10%-Lift	3,5	137	8,147	1:59,5

Fazit

- Klassifikationsergebnisse signifikant besser als eine zufällige Klassifikation
  - → Kundendaten enthalten Informationen, die eine Vorhersage zumindest ansatzweise ermöglichen
- Allerdings: hohe False-Positive Rate
  - → viele loyale Kunden werden als Kündiger vorhergesagt
- Naive Bayes Algorithmus und Cost-sensitive Random Forest liefern die besten Ergebnisse
- Insbesondere die Abonnement-Daten dienen als Indikatoren für einen kurzfristigen Kundenabgang
- Kontaktdaten geben bedingt Aufschluss über künftiges Verhalten
  - → insbesondere eine angedrohte Kündigung scheint tatsächlich ein guter Indikator zu sein
- Auch die Sinus Milieus scheinen Informationsgehalt zu besitzen

Mögliche nächste Schritte

- Testen weiterer Parameter-Werte (z.B. andere Cost-Matrix)
- Break-Even-Analyse für Kundenbindungsmaßnahmen
- Weitergehende Tests mit den Top-Features
- Erneute Feature-Eliminierung mit Indikator-Variablen
- Testen anderer Prognosezeiträume (z.B. 6M, 12M)
- Testen einer Survival Regression