WHITE PAPER

Abwanderung erkennen, bevor sie passiert

Warum ein gutes Modell allein kein Retention-System ist – heterogene Ensembles, kalibrierte Wahrscheinlichkeiten und die ökonomisch richtige Entscheidungsschwelle

findic

Inhalt

Über dieses White Paper

In vielen Churn-Projekten kommt früher oder später dieselbe Entscheidung auf: Ab welcher prognostizierten Wahrscheinlichkeit soll ein Kunde aktiv kontaktiert werden? Häufig wird dafür die Standardschwelle von 0,5 verwendet. Für ein betriebswirtschaftliches Retention-Szenario ist diese Schwelle jedoch meist nicht geeignet, weil sie nicht aus den tatsächlichen Kosten der Fehlentscheidungen abgeleitet ist.

Ein einfaches Beispiel zeigt das Problem: Eine unnötige Retentionsaktion – etwa ein Gespräch, ein Rabatt oder ein Pause-Angebot – verursacht im Mittel etwa 50 €. Ein übersehener Abwanderer kann je nach Vertragswert dagegen ein Vielfaches davon kosten. Wenn der falsch-negative Fehler deutlich teurer ist als der falsch-positive, ist eine pauschale Schwelle von 0,5 nicht sinnvoll. In unserem Test lag die kostenoptimierte Schwelle bei 0,27. Gegenüber der Standardschwelle reduzierten sich die realisierte Fehlklassifikationskosten dadurch um 40,5 Prozent bei gleichzeitig 73 zusätzlich erkannten Abwanderern, also fast halb so vielen übersehenen Fällen.

Dieses Whitepaper beschreibt eine Churn-Pipeline, die diese Schwellenentscheidung systematisch berücksichtigt. Zusätzlich geht es um zwei Punkte, die im Betrieb häufig unterschätzt werden: die Kalibrierung der vorhergesagten Wahrscheinlichkeiten und die Frage, ob ein einzelnes Modell für stabile Entscheidungen ausreicht. Ziel ist kein vollständiges Implementierungsrezept, sondern eine nachvollziehbare Darstellung der getroffenen Modellierungs- und Bewertungsentscheidungen.[1]

1 · Warum Churn-Modelle im Betrieb scheitern

Wenn Churn-Projekte nach dem Go-live enttäuschen, liegt das selten am Algorithmus. Es liegt an drei Stellen, die in der Entwicklung harmlos aussehen und im Betrieb teuer werden.

Die Mehrheitsklasse gewinnt immer

Bei einer Churn-Quote von 23 Prozent erreicht ein Modell, das schlicht niemanden als Abwanderer markiert, 77 Prozent Accuracy – und null Nutzen. Genau in diese Richtung driftet jedes Standardverfahren ohne Gegensteuer: Die Mehrheitsklasse dominiert die Verlustfunktion, und die Fälle, um die es eigentlich geht, werden statistisches Beiwerk. Hinzu kommt: Die gern zitierte ROC-AUC schmeichelt bei unausgewogenen Daten systematisch.[2]

Unsere Antwort ist zweistufig. Erstens trainieren alle Basismodelle mit Klassengewichten, die aus dem realen Verhältnis der Trainingsdaten abgeleitet werden – bei uns rund 3,3 zu 1 zugunsten der Verbleiber, nichts davon wird pauschal angenommen. Zweitens wird die Hyperparametersuche konsequent auf Average Precision optimiert, eine Metrik, die die Minderheitsklasse tatsächlich in den Blick nimmt.

---

[1] Die Pipeline wurde auf einem B2B-SaaS-Datensatz mit 8.214 Kunden und einer Churn-Quote von 23,4 % entwickelt und evaluiert (stratifizierter Split: 60 % Training, 20 % Kalibrierung, 20 % Test)

[2] Saito, T. & Rehmsmeier, M. (2015): The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE 10(3)

Eine 0,7 ist selten eine 0,7

Gradient-Boosting-Modelle können Kunden häufig sehr gut nach Risiko sortieren. Ob eine prognostizierte Wahrscheinlichkeit von 0,7 tatsächlich einer beobachteten Churn-Rate von ungefähr 70 Prozent entspricht, ist damit aber noch nicht gesagt. Gerade baumbasierte Ensembles liefern oft gut geordnete Scores, deren absolute Wahrscheinlichkeiten ohne Kalibrierung verzerrt sein können.[3]

Für den operativen Einsatz ist dieser Unterschied entscheidend. Ein Ranking hilft bei der Priorisierung; eine belastbare Wahrscheinlichkeit wird benötigt, wenn Budget, Reaktionszeiten oder Eskalationsstufen daran geknüpft werden. Abschnitt 5 zeigt die Abweichung der Rohscores (Expected Calibration Error 0,081) und das Ergebnis nach der Kalibrierung (0,027).

Der Schwellenwert ist eine betriebswirtschaftliche Frage

Klassifizieren heißt entscheiden, und jede Entscheidung hat einen Preis. Eine Kostenmatrix macht das explizit: der falsch-negative Fehler (übersehener Abwanderer) kostet 500 €, der falsch-positive (unnötige Aktion) 50 €. Für kalibrierte Wahrscheinlichkeiten lässt sich die optimale Schwelle direkt ausrechnen: t* = C_FP / (C_FP + C_FN) – bei unseren Annahmen also theoretisch 0,09.[4]

Dass unsere gemessene Schwelle mit 0,27 deutlich höher liegt als dieser theoretische Wert, hat einen handfesten Grund, den Abschnitt 6 erklärt. Festhalten lässt sich schon hier: Wer die Schwelle bei 0,5 belässt, hat implizit entschieden, dass beide Fehlerarten gleich teuer sind. Im Retention-Geschäft stimmt das praktisch nie.

2 · Die Architektur in vier Schichten

Die Pipeline besteht aus vier Schichten. Jede Schicht übernimmt eine klar abgegrenzte Aufgabe und baut auf den Ergebnissen der vorherigen Schicht auf.

Schicht 1 – Merkmale. Aus elf Rohfeldern (Vertragslaufzeit, Support-Tickets, Login-Verhalten, NPS, Zufriedenheitswerte) entstehen 57 Merkmale: Raten statt Absolutwerte, RFM-inspirierte Proxys, Interaktionsterme und ein regelbasierter Risikoscore, der bekannte Gefahrenmuster explizit bündelt. Jede Statistik, Median, Quantile, Maxima, wird ausschließlich auf den Trainingsdaten geschätzt und anschließend eingefroren. Dadurch wird vermieden, dass Informationen aus Validierungs- oder Testdaten in die Merkmalsbildung einfließen.

Schicht 2 – Modelle. Drei Gradient-Boosting-Verfahren, CatBoost, LightGBM, XGBoost, werden getrennt trainiert und per Out-of-Fold-Stacking kombiniert. Jeder Kunde erhält seine Metamerkmale von Modellen, die ihn im Training nicht gesehen haben. Eine logistische Regression lernt daraus die finale Kombination. Die Hyperparameter werden über eine bayessche Suche mit frühzeitigem Abbruch

[3] Niculescu-Mizil, A. & Caruana, R. (2005): Predicting Good Probabilities with Supervised Learning. Proceedings of ICML 2005

[4] Elkan, C. (2001): The Foundations of Cost-Sensitive Learning. Proceedings of IJCAI 2001

schlechter Versuche bestimmt; die Baumanzahl wird per Early Stopping auf einem eigenen Validierungssplit begrenzt.[5]

Schicht 3 – Kalibrierung. Die Kalibrierung erfolgt zweistufig: zuerst für die Ausgaben der Basismodelle, anschließend für die Ausgaben des Meta-Lerners, jeweils auf einer eigenen Kalibrierungsmenge. Verwendet wird isotonische Regression, ein nichtparametrisches und monotones Verfahren ohne vorgegebene Kurvenform. Da dieses Verfahren ausreichend viele Kalibrierungsfälle benötigt, enthält die Pipeline eine Absicherung: Unter 1.000 Kalibrierungsfällen wird automatisch auf das robustere Platt Scaling umgestellt.

Schicht 4 – Entscheidung. Die kalibrierte Wahrscheinlichkeit wird mit einer Kostenmatrix verknüpft. Die Schwelle wird so gewählt, dass die erwarteten Euro-Kosten auf der Kalibrierungsmenge minimiert werden. Darauf aufbauend werden drei Risikostufen mit festen Reaktionszeiten definiert. Das Ergebnis ist keine isolierte Modellkennzahl, sondern eine priorisierte Arbeitsliste für den operativen Einsatz.

3 · Warum drei Boosting-Maschinen besser sind als eine

CatBoost, LightGBM und XGBoost gehören zur selben Modellfamilie. Dabei schlägt auf tabellarischen Daten diese Familie nach wie vor deutlich aufwendigere Deep-Learning-Ansätze.[6]

Die drei Modelle unterscheiden sich jedoch in ihrer Fehlerstruktur. CatBoost kodiert kategoriale Merkmale leckagefrei anhand der Zielvariable, bildet dabei aber seltene Kategorien ungenauer ab. LightGBM wächst Bäume blattweise und kann dadurch tiefere Strukturen erfassen, ist aber anfälliger für Overfitting. XGBoost setzt stärkere Regularisierung ein. Dadurch entstehen teilweise unterschiedliche Fehlklassifikationen. Ein Ensemble kann diese Unterschiede nutzen, sofern die Fehler der Einzelmodelle nicht vollständig korreliert sind.

Die Koeffizienten des Meta-Lerners zeigen eine stärkere Gewichtung von CatBoost (2,31) und LightGBM (2,04) gegenüber XGBoost (1,18). Das deutet darauf hin, dass die Ausgaben der ersten beiden Modelle in der finalen Kombination einen größeren Beitrag leisten. Gemessen wird ein PR-AUC von 0,817 gegenüber 0,772 beim besten Einzelmodell und 0,793 bei Soft Voting.

Der Zugewinn durch Stacking ist messbar, aber begrenzt. PR-AUC steigt gegenüber dem besten Einzelmodell um 0,045 Punkte und gegenüber Soft Voting um 0,024 Punkte. Der praktische Nutzen liegt vor allem darin, dass die nachgelagerten Schritte (Kalibrierung und Schwellenlogik) auf einer stabileren Modellbasis aufsetzen.

[5] CatBoost: Prokhorenkova et al. (2018), NeurIPS 31. LightGBM: Ke et al. (2017), NeurIPS 30. XGBoost: Chen & Guestrin (2016), KDD. Stacking: Wolpert (1992), Neural Networks 5(2). Hyperparametersuche: Optuna, Akiba et al. (2019), KDD

[6] Grinsztajn, L., Oyallon, E. & Varoquaux, G. (2022): Why do tree-based models still outperform deep learning on typical tabular data? NeurIPS 35

4 · Kalibrierung: Aus Scores werden Wahrscheinlichkeiten

Der unkalibrierte Meta-Lerner unserer Pipeline lag auf dem Testset bei einem Expected Calibration Error von 0,081.[7] Nach der zweistufigen isotonischen Kalibrierung lag der Wert bei 0,027. In den entsprechenden Wahrscheinlichkeitsbereichen nähert sich die beobachtete Churn-Rate damit stärker den prognostizierten Wahrscheinlichkeiten an.

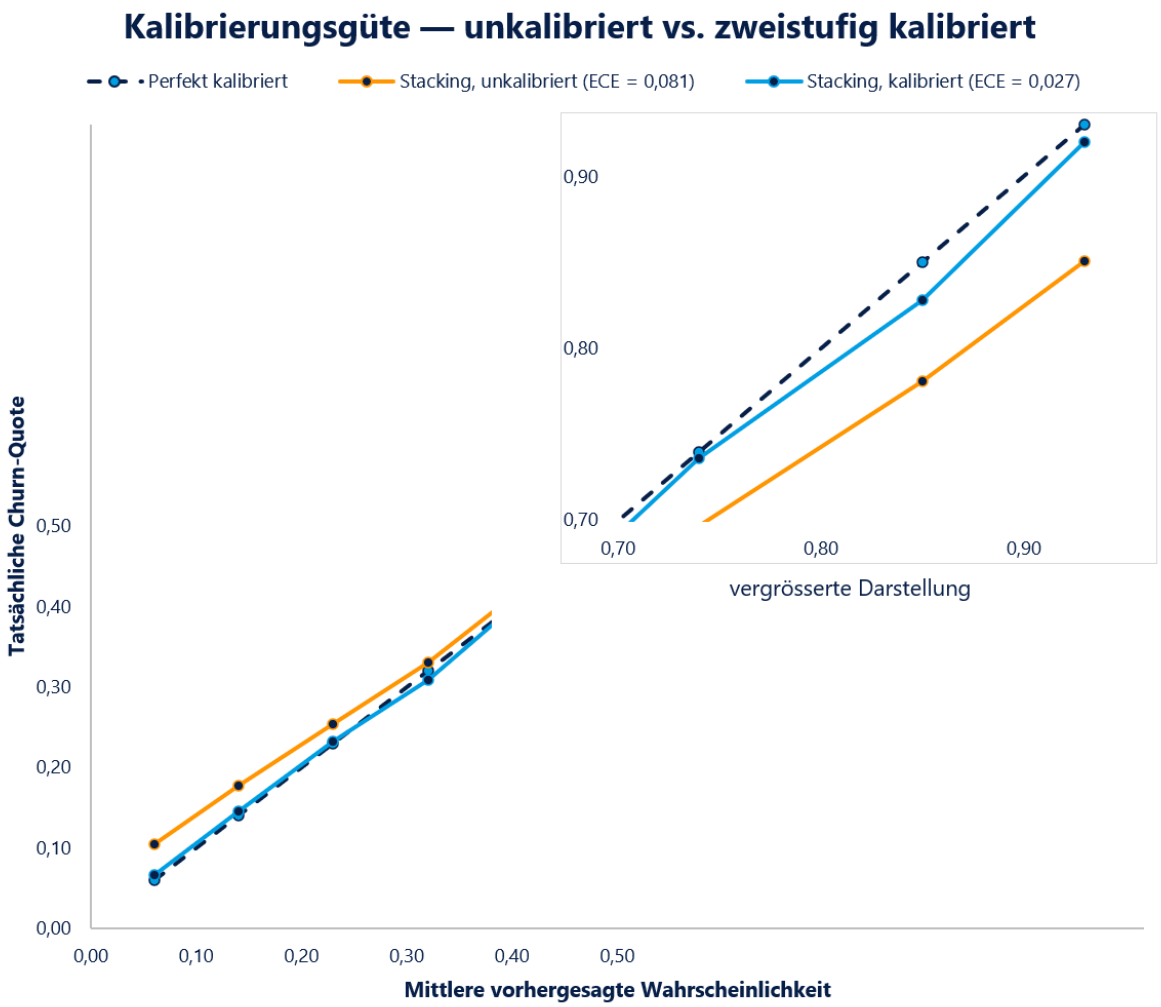

Abbildung 1: Reliability-Diagramm auf dem Testdatensatz. Die rohe Ensemble-Ausgabe (orange) weicht systematisch von der Diagonalen ab; nach der Kalibrierung (türkis) liegen die Punkte auf der Ideallinie.

Die zweistufige Kalibrierung adressiert unterschiedliche Verzerrungsquellen. Die erste Stufe korrigiert die Ausgaben der Basismodelle vor der Kombination. Die zweite Stufe korrigiert mögliche Verzerrungen des Meta-Lerners, etwa durch Klassengewichtung. Um Überanpassung zu begrenzen, wird die

[7] Naeini, M. P., Cooper, G. F. & Hauskrecht, M. (2015): Obtaining Well Calibrated Probabilities Using Bayesian Binning. Proceedings of AAAI 2015

Kalibrierung über getrennte Datenbereiche abgesichert und der ECE auf dem unabhängigen Testdatensatz ausgewiesen.

Ohne Kalibrierung wären die Risikostufen aus Abschnitt 6 nur eingeschränkt belastbar. Durch die Kalibrierung werden die Modellwerte besser als Wahrscheinlichkeiten interpretierbar und können für Budget- und Priorisierungsentscheidungen genutzt werden.

5 · Die Schwelle als ökonomische Entscheidung

Elkans Formel t* = C_FP / (C_FP + C_FN) ergibt bei 50 € und 500 € den theoretischen Wert 0,09. Die empirisch bestimmte Schwelle liegt in unserem Fall bei 0,27. Diese Abweichung ist wichtig: Die theoretische Regel setzt unter anderem sehr gut kalibrierte Wahrscheinlichkeiten und eine direkt anwendbare Kostenannahme voraus. In der Praxis können Kalibrierungsfehler, Stichprobenrauschen und die tatsächliche Wirkung von Retentionsmaßnahmen dazu führen, dass das empirische Kostenminimum höher liegt. Deshalb wird die Schwelle auf der Kalibrierungsmenge anhand der realisierten Kostenkurve bestimmt:

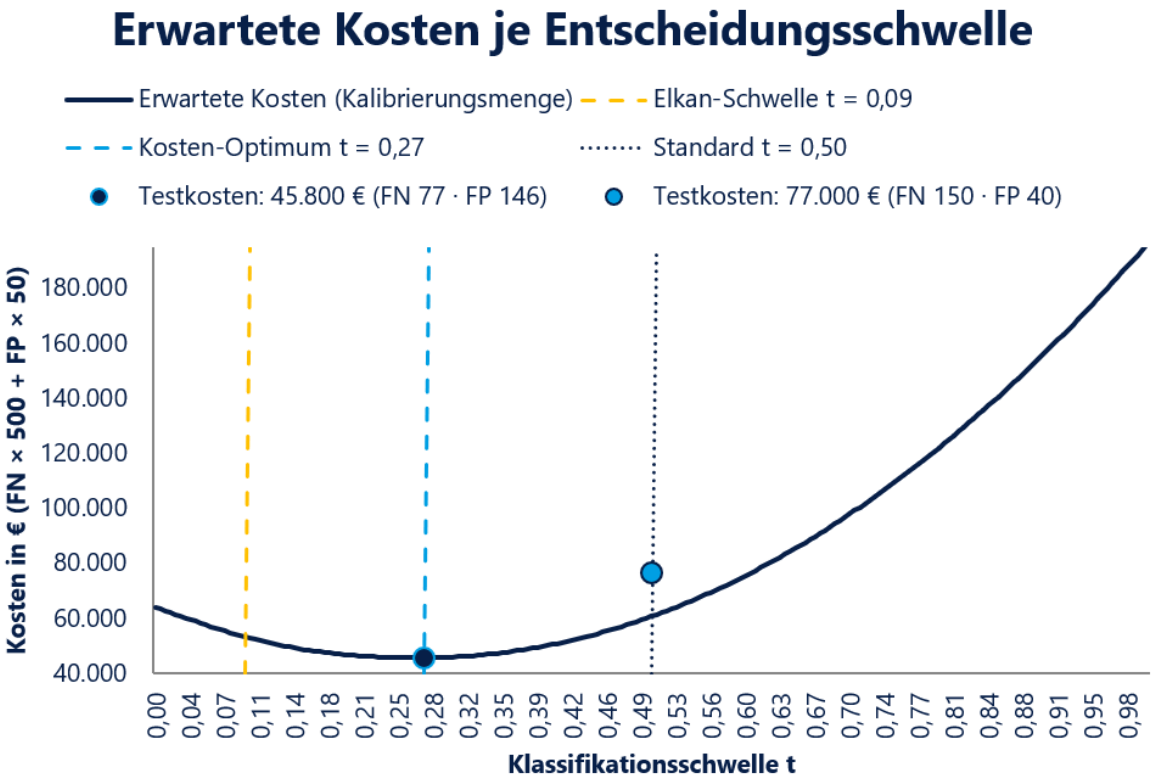

Abbildung 2: Erwartete Kosten über der Klassifikationsschwelle (Kalibrierungsmenge). Das empirische Minimum liegt bei t = 0,27 – deutlich über der theoretischen Elkan-Schwelle, weit unter dem Standard 0,5.

Der Vergleich der gängigen Strategien auf dem Testset:

Strategie	Schwelle	Übersehene Abwanderer (FN)	Fehlalarme (FP)	Gesamtkosten
Standard (Default)	0,50	150	40	77.000 €
Youden-optimiert	0,34	86	118	48.900 €
F2-optimiert	0,31	82	128	47.400 €
Kosten-optimiert	0,27	77	146	45.800 €

Tabelle 1: Schwellenstrategien im Vergleich (Testdatensatz; Kostenannahmen FN = 500 €, FP = 50 €).

Die letzte Zeile zeigt den zentralen Effekt: Die kostenoptimierte Schwelle erzeugt die meisten Fehlalarme aller Strategien (146), überlässt aber auch die wenigsten Abwanderer sich selbst: Nur 77 bleiben unentdeckt, fünf weniger als bei der F2-Variante und 73 weniger als bei der Standardschwelle. Das ist bei den gewählten Kostenannahmen konsequent: Ein übersehener Abwanderer wiegt mit 500 € genau zehn unnötige Maßnahmen à 50 € auf. Gegenüber der Standardschwelle sinken die Kosten um 40,5 Prozent. Der Effekt entsteht dabei nicht durch ein anderes Modell, sondern durch eine an den Kosten ausgerichtete Entscheidungsschicht.

Die operative Umsetzung übersetzt die kalibrierten Wahrscheinlichkeiten in drei Stufen mit festen Reaktionszeiten:

Risikostufe	Wahrscheinlichkeit	Reaktionszeit	Maßnahme
Low Risk	unter 0,27	Asynchron	In-App-Hinweis, Informationsmail
Medium Risk	0,27 bis 0,70	3 Tage	Kontakt durch Customer Success, Nutzungscoaching
High Risk	über 0,70	24 Stunden	persönlicher Kontakt, individuelles Angebot

Tabelle 2: Risikostratifizierung und Maßnahmenzuordnung.

6 · Gemessene Ergebnisse

Alle Werte stammen vom unabhängigen Testdatensatz (20 % der 8.214 Kunden, stratifiziert), der weder Training noch Kalibrierung noch Schwellenwahl je gesehen hat. Die Intervalle in Klammern sind 95-%-Konfidenzintervalle aus 1.000 Bootstrap-Ziehungen.

Metrik	Stacking (kalibriert)	Soft Voting	Bestes Einzelmodell (CatBoost)
ROC-AUC	0,913 [0,891–0,932]	0,905	0,894
PR-AUC	0,817 [0,774–0,856]	0,793	0,772
Brier-Score	0,104 [0,093–0,116]	0,121	0,138
ECE	0,027	0,052	0,096
MCC	0,647	0,601	0,562

Tabelle 3: Testergebnisse. Konfidenzintervalle dort, wo Bootstrap sinnvoll skaliert.

Aus den Ergebnissen lassen sich folgende Punkte ableiten. Erstens erzielt das Stacking in allen betrachteten Metriken die besten Werte, der Abstand zum Soft Voting bleibt jedoch moderat. Zweitens ist die Kalibrierung ein wesentlicher Qualitätsfaktor: Der ECE sinkt gegenüber dem besten Einzelmodell von 0,096 auf 0,027, und der Brier-Score verbessert sich deutlich. Drittens zeigt sich der größte praktische Nutzen in der kostenbasierten Entscheidungsschicht: Die Fehlklassifikationskosten fallen um 40,5 Prozent, weil die Schwelle an den wirtschaftlichen Fehlerkosten ausgerichtet wird.

7 · Was das System nicht kann

Drei Einschränkungen sind für die Einordnung wichtig.

Es ist prädiktiv, nicht kausal. Eine hohe Churn-Wahrscheinlichkeit sagt nichts darüber, ob eine Maßnahme diesen Kunden hält. Vielleicht ist die Kündigung längst beschlossen; vielleicht wäre der Kunde auch ohne Anruf geblieben. Wer die Wirkung von Interventionen optimieren will, braucht Uplift-Modelle oder kontrollierte Experimente. Unsere Kostenlogik gilt streng genommen nur, wenn die Maßnahmen bei den kontaktierten Kunden überhaupt wirken.

Es ist an seine Daten gebunden. RFM-Proxys, Vertragsflags und Zufriedenheitsindizes setzen voraus, dass solche Variablen erhoben werden. Eine Übertragung auf andere Branchen erfordert angepasste Merkmalsdefinitionen, damit fachliche Arbeit, kein Neutraining des Prinzips.

Es altert. Kundenverhalten verändert sich: Preise, Produkte und Marktbedingungen bleiben nicht konstant. Deshalb müssen Monitoring und geplantes Retraining Teil des Betriebs sein. Jede Modellversion dokumentiert Kennzahlen und Entscheidungsparameter in einer maschinenlesbaren Model Card, sodass Veränderungen zwischen Versionen überprüfbar werden.

8 · Der nächste Schritt

Für einen Pilotbetrieb reichen in der Regel ein Export der relevanten Kundendaten – etwa Verträge, Nutzung und Support-Interaktionen –, ein Zeitraum von sechs bis acht Wochen sowie eine fachliche Ansprechperson im Customer Success. Am Ende steht eine priorisierte Liste von Kunden, bei denen eine Kontaktaufnahme wirtschaftlich sinnvoll erscheint, einschließlich Risikostufe und Begründung.

Entscheidend ist nicht nur, wie hoch die aktuelle Abwanderungsquote ist, sondern wie früh gefährdete Kunden zuverlässig erkannt und sinnvoll priorisiert werden können.

9 · Quellen / weiterführende Literatur

Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631. Framework für Bayessche Hyperparametersuche mit automatischem Abbruch aussichtsloser Versuche.

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. Skalierbares Gradient Boosting mit starker Regularisierung.

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI), 973–978. Grundlegende Herleitung kostensensitiver Entscheidungsregeln – die optimale Klassifikationsschwelle folgt direkt aus dem Kostenverhältnis der Fehlerarten.

Fridrich, M. (2018). Cost-Benefit Metrics in Customer Churn Prediction: A Review. MMK 2018 – International Masaryk Conference for Ph.D. Students and Young Researchers, 178–185. Überblick über geschäftsorientierte Bewertungsmetriken und die Grenzen rein technischer Klassifikationskennzahlen im Retention-Kontext.

Grinsztajn, L., Oyallon, E. & Varoquaux, G. (2022). Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data? Advances in Neural Information Processing Systems 35 (NeurIPS). Umfangreicher Benchmark, der zeigt, dass baumbasierte Verfahren auf typischen Tabellendaten auch deutlich aufwendigeren Deep-Learning-Ansätzen überlegen bleiben.

Imani, M., Joudaki, M., Beikmohammadi, A. & Arabnia, H. R. (2025). Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning. Machine Learning and Knowledge Extraction, 7(3), 105. Systematischer Überblick über aktuelle Verfahren, Herausforderungen und Einsatzfelder der Churn-Vorhersage.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (NeurIPS). Histogrammbasiertes, blattweises Gradient Boosting mit hoher Trainingsgeschwindigkeit.

Naeini, M. P., Cooper, G. F. & Hauskrecht, M. (2015). Obtaining Well Calibrated Probabilities Using Bayesian Binning. Proceedings of the AAAI Conference on Artificial Intelligence, 29(1), 2901–2907. Führt den Expected Calibration Error (ECE) als Maß für die Abweichung zwischen prognostizierter und beobachteter Ereignisquote ein.

Niculescu-Mizil, A. & Caruana, R. (2005). Predicting Good Probabilities with Supervised Learning. Proceedings of the 22nd International Conference on Machine Learning (ICML), 625–632. Zeigt, dass Boosted Trees exzellent ranken, ihre Rohwahrscheinlichkeiten aber systematisch verzerrt sind – und dass Platt Scaling und isotonische Regression daraus belastbare Wahrscheinlichkeiten machen.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. (2018). CatBoost: Unbiased Boosting with Categorical Features. Advances in Neural Information Processing Systems 31 (NeurIPS). Gradient Boosting mit unverzerrter Kodierung kategorialer Merkmale ohne Zielwissen-Leckage.

Saito, T. & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), e0118432. Belegt, dass ROC-Kurven bei unausgewogenen Daten ein zu optimistisches Bild liefern und Precision-Recall-Kurven die realistischere Bewertung erlauben.

Wolpert, D. H. (1992). Stacked Generalization. Neural Networks, 5(2), 241–259. Grundlagentext des Stacking: Kombination mehrerer Modelle über einen Meta-Lerner auf Out-of-Fold-Prognosen.

10 · Kontakt

Udo Jacobasch
Partner

Udo.Jacobasch@findic.de
Office Berlin

Dr. Josef Dirnberger
Manager

Josef.Dirnberger@findic.de
Office München

Hamburg

Kurze Mühren 20,

D-20095 Hamburg

Phone +49 40/303740-0

E-Mail info@findic.de

Zürich

Gutenbergstrasse 1,

CH-8002 Zürich

Phone +41 44/56098-00

E-Mail info@findic.ch