Survival Analyse zur Optimierung von Marketingkampagnen

Maike Küffer
Hochschule Darmstadt
Fachbereich Mathematik & Naturwissenschaften und Fachbereich Informatik

Zusammenfassung

Ziel der vorliegenden Arbeit ist es, Anwendungsmöglichkeiten von Verfahren der Survival Analyse im Bereich des Marketings zur Optimierung von Werbemaßnahmen aufzuzeigen. Zielgröße der Survival Analyse ist in diesem Kontext die Zeit, die ein Kunde benötigt um einen Folgekauf zu tätigen. Werbemaßnahmen können inhaltlich auf Kundengruppen abgestimmt werden. Doch auch die Frequenz, mit der Werbemittel, wie Newsletter oder Werbepost verschickt werden, können an das individuelle Kaufintervall eines Kunden angepasst werden. Kauffreudige Kunden erhalten häufiger Werbepost als Gelegenheitskäufer. Ist das Kaufintervall eines Kunden bekannt können auch Gutscheine und Rabatte entsprechend platziert werden. Tätigt ein Kunde keinen Folgekauf in der prognostizierten Zeit, so können Reaktivierungsmaßnahmen erfolgen um die Abwanderung des Kunden zu verhindern.

Als statistische Verfahren der Survival Analyse werden der Kaplan-Meier-Schätzer und -Kurven, der Log-Rank-Test und das Cox-Regressionsmodell vorgestellt. Es wird ein Analyseverfahren entwickelt, um auf Basis der Vorhersage des Cox-Regressionsmodells Kunden in Scoreklassen einzugruppieren. Diese Scoreklassen unterscheiden sich hinsichtlich ihrer Kaufintervalle und ermöglichen eine individuelle Anpassung von Marketingmaßnahmen an die Scoreklassen.

Einleitung

Im Laufe der letzten Jahre nahmen die Umsätze, die im Onlinehandel erzielt werden konnten, stetig zu. Betrug der Umsatz 2010 in Deutschland noch 20,1 Milliarden Euro, so steigerte sich dieser Betrag in den letzten 10 Jahren kontinuierlich auf 57,8 Milliarden Euro (Prognose für 2019, nach [1]).

Im Gegensatz zum Einzelhandel ergibt sich im Onlinehandel oder auch allgemeiner im Distanzhandel die Möglichkeit, die Kauffrequenz eines Kunden zu analysieren.

Die Untersuchung, welche Faktoren die Zeit vom Kauf bis zum Folgekauf eines Kunden beeinflussen, ist im Rahmen der Survival Analyse möglich. Sie eignet sich, um die Kauffrequenz von Kunden besser zu verstehen und Fragen wie

„Tätigen Frauen schneller einen Folgekauf als Männer?“
„Kaufen junge Menschen in kürzeren Abständen als ältere?“
„Wenn ein Kunde in der Vergangenheit ein Produkt einer bestimmten Warengruppe bestellt hat, mit welcher Wahrscheinlichkeit bestellt er innerhalb der nächsten 30 Tage wieder?“

zu beantworten.

Die Survival Analyse hat ihren Ursprung in der Medizin und betrachtet als Zielgröße meist die Überlebenszeit eines Patienten. So kann die Überlebenszeit bei unterschiedlichen Therapieansätzen, Diagnosen etc. verglichen werden.

Als Methoden der Survival Analyse werden der Kaplan-Meier-Schätzer, der Log-Rank-Test und das Cox-Regressionsmodell mit seinen Voraussetzungen, Parameterschätzern und den Cox-Snell Residuen zur Beurteilung der Güte des Modells eingeführt.

Ziel der vorliegenden Arbeit ist es, mit der Survival Analyse Möglichkeiten der Anwendung im Marketing aufzuzeigen, um Merkmale zu identifizieren, die die Zeit bis zum Folgekauf beeinflussen und Scoreklassen mithilfe eines Scorings abzuleiten. Dank dieser gebildeten Scoreklassen können Kunden mit geringerer Kauffrequenz von Kunden mit höherer Kauffrequenz unterschieden werden.

Survival Analyse

Die Survival Analyse untersucht die Zeit bis ein bestimmtes Ereignis eingetreten ist. Besonders beachtet werden zensierte Daten (das interessierende Ereignis tritt nicht für alle Fälle während des Beobachtungszeitraums auf).

Die Zeit sei nun beschrieben durch die stetige Zufallsvariable T mit Ausprägung t. Eine wichtige Größe in der Survival Analyse ist die Überlebensfunktion $S(t)$, die Sterbefunktion $F(t)$ und die Hazard-Funktion $h(t)$

$$S(t) = P(T > t) = \int_t^\infty f(x)dx \tag{1}$$

$$F(t) = P(T \leq t) = \int_0^t f(x)dx \tag{2}$$

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{3}$$

Die Überlebensfunktion beschreibt die Wahrscheinlichkeit, den Zeitpunkt t zu überleben.
Mithilfe des **Kaplan-Meier Schätzers** kann die Überlebensfunktion $\hat{S}(t)$ geschätzt werden. Es handelt sich um eine nicht-parametrische Schätzung.
Mit dem **Log-Rank Test** kann getestet werden ob sich die Überlebensfunktionen $S_0(t)$ und $S_1(t)$ zweier Gruppen 0 und 1 statistisch signifikant unterscheiden. Die **Cox-Regression** ist ein semiparametrisches multivariates Verfahren, das den Einfluss von Merkmalen auf die Zeit bis zu einem Ereignis untersucht. Der Modellansatz nach Cox für m Einflussgrößen lautet:

$$h(t, x) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m). \tag{4}$$

Dabei bezeichnet $h_0(t)$ die Baseline-Hazard-Rate, auch Basisrisiko genannt. Diese Baseline-Hazard-Rate wird nicht weiter spezifiziert. Die Parameterschätzung erfolgt mit partiellen Likelihood Funktionen.

Survival Analyse im Marketing

Die folgende Liste zeigt einen Prozessablauf der Analyse zur Entwicklung von Scoreklassen mit unterschiedlicher Kauffrequenz.

1. **Datenaufbereitung** und Feature Engineering
2. **Bivariate Analyse mit Kaplan-Meier Schätzer**
3. **Variablenauswahl** für multivariate Analyse
   AIC Kriterium und Erkenntnisse der bivariaten Analyse
4. **Multivariate Analyse mit Cox-Regressionsmodell**
5. **Entwicklung von Scoreklassen**
   auf Basis der Vorhersage des Modells
   Ziel: Kunden verschiedener Scoreklassen unterscheiden sich hinsichtlich ihrer Zeit bis zum Folgekauf
6. **Betrachtung der Kauffunktionen**
   „Welche Scoreklasse kauft im Mittel wann?“
7. **Validierung des Modells**
   mittels Cox-Snell Residuen und
   Vergleich der Kauffunktionen von Test- und Trainingsdaten
8. Angepasste **Marketingmaßnahmen** für verschiedene Scoreklassen

Survival Analyse von Kaufdaten mit R

Der praktische Teil dieser Arbeit beinhaltet eine Analyse von Verkaufsdaten eines Jahres eines Versandhändlers in R. Dabei werden zunächst in einer bivariaten Analyse mithilfe des Kaplan-Meier Schätzers mögliche Einflussgrößen identifiziert. Zur **Entwicklung von Scoreklassen** werden Cox-Regressionsmodelle gebildet, um anschließend die unterschiedlichen Kauffunktionen der Scoreklassen zu vergleichen. Für die Scoreklassen lassen sich Prognosen über die Zeit bis zum Folgekauf aufstellen. Anschließend werden Möglichkeiten aufgezeigt, den Erfolg der **optimierten Marketingmaßnahmen** mithilfe von Key Performance Indikatoren zu messen.

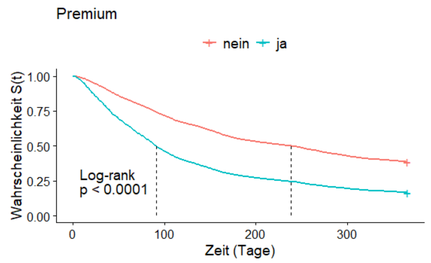

Abbildung 2: Kaplan-Meier Kurve für das Merkmal Premium

Im Datensatz enthalten sind 2,27 Mio. Einkäufe von 1,2 Mio. Kunden. Abbildung 2 zeigt exemplarisch den Einfluss des Merkmals Premium auf die Zeit bis zum Folgekauf. Premiumkunden kaufen im Schnitt 147 Tage früher als Nicht-Premium-Kunden. Die Ergebnisse der bivariaten Analyse liefern zudem erste Erkenntnisse für vorbereitende Schritte der Modelle.
Bei der multivariaten Analyse mittels Cox-Regression zeigte sich u.a., dass die Anzahl der Bestellungen im Vorjahr und das Alter großen Einfluss auf die Zeit bis zum Folgekauf eines Kunden haben.
Mithilfe der Vorhersage des linearen Terms der Vorhersage der Cox-Regression werden auf Basis der Dezile 10 Scoreklassen gebildet:

$$h(t) = h_0(t)\exp(\underbrace{\beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m}_{\text{lineare Vorhersage}})$$

Für diese 10 Scoreklassen werden wiederum die 10 Kauffunktionen (mathematisch: Sterbefunktionen) betrachtet. Die folgende Tabelle zeigt die Quantile der Kauffunktionen für die unterschiedlichen Scoreklassen. Die farbliche Markierung beschreibt die Dringlichkeit von Marketingmaßnahmen.

		Tage							
		$Q_{0,1}$	$Q_{0,2}$	$Q_{0,3}$	$Q_{0,4}$	$Q_{0,5}$	$Q_{0,6}$	$Q_{0,7}$	$Q_{0,8}$
Scoreklasse	1	74	174	327	-	-	-	-	-
	2	58	112	178	280.5	-	-	-	-
	3	53	101	157	226	304	-	-	-
	4	40	77	112	159	234	316	-	-
	5	36	64	95	129	168	251	341	-
	6	32	55	82	106	141	186	273	-
	7	28	46	68	90	114	150	212	307
	8	24	41	59	78	97	123	162	257
	9	21	33.5	45	61	77	95	121	167
	10	13	19	27	35	44	58	74	95

Je nach Dringlichkeit können der Einsatz von Gutscheinen/ Rabatten, die Newsletter-Frequenz oder die Gutscheinbedingungen (z. B. Mindestbestellwert) variiert werden.
Die Modelle wurden mit 80% der Daten trainiert. Abbildung 3 zeigt die Gegenüberstellung der Trainings-und Testdaten.

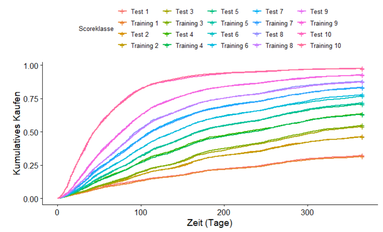

Abbildung 3: Gegenüberstellung der Kauffunktionen von Trainings- und Testdaten

Scoreklasse 10 tätigt am schnellsten eine Folgekauf. Scoreklasse 1 ist die schlechteste Scoreklasse.
Die Modelle können zusätzlich mithilfe von **KPIs** überprüft werden. Mögliche KPIs sind die **Klick-Rate** (*Prozentualer Anteil an Kunden die einen E-Mail Newsletter geöffnet haben und einen Link der E-Mail angeklickt haben*), die **Conversion-Rate** (*Prozentualer Anteil an Kunden, die einen Newsletter geöffnet, angeklickt und infolgedessen einen Kauf getätigt haben*) oder die **Abmelderate** (*Der Anteil an Kunden, die sich aus dem Newsletterverteiler abmelden*).

Fazit und Ausblick

Die verwendeten Verfahren der Survival Analyse eignen sich trotz ihres Ursprungs in der Medizin auch im Marketingbereich, um die Zeit bis zu einem Ereignis zu untersuchen. Die Modelle lieferten eine gute Vorhersage der Wahrscheinlichkeit, dass Kunden bis zu einem gewissen Zeitpunkt kaufen. Zudem wurden Möglichkeiten aufgezeigt, die Kenntnis der Kaufintervalle der Scoreklassen zu nutzen, um die Frequenz und den Inhalt von Werbemaßnahmen zu optimieren.
Die Survival Analyse bietet noch weitere Verfahren, wie das exponentielle Regressionsmodell, das Weibull Regressionsmodell und AFT-Modelle (Accelerated-Failure-Time), die im Marketing verwendet werden könnten.
Eine weitere Möglichkeit der Anwendung der Survival Analyse im Marketing sind baumbasierte Ansätze (Survival Trees).

Literatur

[1] HDE. Umsatz durch e-commerce (b2c) in deutschland in den jahren 1999 bis 2018 sowie eine prognose für 2019 (in milliarden euro), 20. Mai, 2019. Gesehen am 20. Januar 2020.

[2] J. Hedderich and L. Sachs. *Angewandte Statistik: Methodensammlung mit R.* Springer Berlin Heidelberg, 2018.

[3] J.P. Klein and M.L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer New York, 2006.