
Hochschule Darmstadt

Fachbereiche Mathematik
und Naturwissenschaften
& Informatik

# Survival Analyse zur Optimierung von Marketing Kampagnen

**zur Erlangung des akademischen Grades**
**Master of Science (M. Sc.)**
**im Studiengang Data Science**

**vorgelegt von**
**Maike Küffer**

| | |
|---|---|
| **Referent(in):** | Prof. Dr. Jutta Groos |
| **Korreferent(in)** | Prof. Dr. Inge Schestag |
| **Betreuer** | Dr. Johannes Gladitz |
| | |
| **Ausgabedatum:** | 01.08.2019 |
| **Abgabedatum:** | 23.01.2020 |

# Selbstständigkeitserklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht. Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen. Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

*Darmstadt, den*

# 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 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.

Im praktischen Teil dieser Arbeit wird mit R das entwickelte Verfahren für 2,27 Millionen Käufe von 1,2 Millionen Kunden angewandt. Für die Scoreklassen ergeben sich Kauffunktionen, die Aussagen über die Kaufintervalle der Scoreklassen ermöglichen. Aus den Quantilen der Kauffunktionen lassen sich Handlungsbedarf und Dringlichkeit von Marketingmaßnahmen ableiten. Mögliche Marketingmaßnahmen und KPIs zur Überprüfung des Erfolgs der Maßnahmen werden vorgestellt. Um die Modelle zu testen, werden sie zunächst mit 80% der Daten trainiert und im Anschluss mit den übrigen 20% der Daten getestet. Der Abgleich der prognostizierten Kauffunktionen für die Testdaten zeigt gute Ergebnisse. Die Kaufwahrscheinlichkeiten der Scoreklassen der Testdaten werden durch die Modelle gut beschrieben.

# Abstract

The objective of this thesis is to show possible applications of methods of the Survival Analysis in the area of marketing with the aim of optimizing advertising activities. In this context the target variable of the Survival Analysis is the time a customer needs to make a followup purchase.

The content of advertising activities can be tailored to customer groups. However, the frequency with which marketing materials, such as newsletters or advertising mail are sent can also be adapted to a customer's individual buying interval. Customers who are keen to buy, receive advertising mail more often than casual buyers. If a customer's purchase interval is known, vouchers and discounts can also be placed accordingly. If a customer does not make a follow-up purchase in the forecast time, reactivation activities can be taken to prevent the customer from churn.

The Kaplan-Meier estimator and curves, the log-rank test and the Cox regression model are presented as statistical methods of the Survival Analysis. An analysis method is developed to group customers into score classes based on the prediction of the Cox regression model. These score classes differ in terms of their purchase intervals and allow marketing activities to be individually adapted to the score classes.

In the practical part of this work, the developed method is used in R for 2.27 million purchases made by 1.2 million customers. For the score classes, there are purchase functions that enable statements about the purchase intervals of the score classes. The need for action and the urgency of marketing measures can be derived from the quantiles of the purchase functions. Possible marketing measures and KPIs to check the success of the marketing activities are presented.

To validate the models, the models are first trained with 80% of the data and then tested with the remaining 20% of the data. The comparison of the forecast purchase functions for the test data shows good results. The purchase probabilities of the score classes of the test data are well described by the models.

# Inhaltsverzeichnis

# 1 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 [HDE19]).

Im Gegensatz zum Einzelhandel ergibt sich im Onlinehandel oder auch allgemeiner im Distanzhandel die Möglichkeit, die Kauffrequenz eines Kunden zu analysieren. Im Einzelhandel erfolgt der Einkauf anonym. Doch auch hier ist die Erhebung der Kauffrequenz des Kunden dank Kundenkarten und Bonussystemen möglich.

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. Eine Besonderheit der Survival Analyse ist die Berücksichtigung sogenannter zensierter Daten. Um zensierte Daten handelt es sich, wenn zum Studienende noch nicht für alle Patienten der Tod eingetreten ist. Diese Patienten werden bei der Survival Analyse nicht aus der Analyse ausgeschlossen, sondern besonders beachtet.

Analog zur Survival Analyse in der Medizin kann im Marketing statt der Überlebenszeit die Zeit bis zum Folgekauf eines Kunden betrachtet werden. Liefert die Survival Analyse Erkenntnisse über die individuelle Kauffrequenz eines Kunden, so können Marketingmaßnahmen entsprechend abgestimmt werden.

Ziel der vorliegenden Arbeit ist es, mit der Survival Analyse Möglichkeiten 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.

In einer Umfrage von 2017 des Online Portals Statista gaben 50% der Befragten an, einen Newsletter abzubestellen, wenn sie zu viele Newsletter von dem Absender bekommen [Sta17]. Ist die Kauffrequenz eines Kunden bekannt, kann die Frequenz des

Newsletterversands (oder Werbepost) entsprechend gestaltet werden, um die Abmelderate zu verringern: Kunden einer guten Scoreklasse erhalten häufiger einen Newsletter als Kunden mit einer schlechteren Scoreklasse. Ein Kunde erhält folglich einen Newsletter zu dem Zeitpunkt, zu dem er wieder bereit ist zu kaufen.

Nicht nur die Frequenz der Marketingmaßnahmen kann an die Kauffrequenz angepasst werden. Auch der Inhalt von Newslettern kann auf die Kauffrequenz abgestimmt werden. Wird die Abwanderung eines Kunden erkannt („Eigentlich hätte der Kunde schon kaufen müssen“), kann mit einer gezielten, personalisierten Marketingmaßnahme die Möglichkeit der Reaktivierung des Kunden geschaffen werden.

Zudem können Gutscheine und Rabatte an die Kauffrequenz eines Kunden angepasst werden, denn ein zu früh verschickter Gutschein bedeutet möglicherweise einen geringeren Umsatz, während ein Gutschein, der zum richtigen Zeitpunkt kommt, die Kundenbindung stärken kann.

In Kapitel 2 dieser Arbeit werden die mathematischen Grundbegriffe der Survival Analyse erläutert. Es werden die Begriffe Überlebensfunktion, Sterbefunktion und Hazardfunktion definiert sowie die verschiedenen Typen von Zensierungen vorgestellt.

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.

In Kapitel 3 werden Anwendungsmöglichkeiten der Survival Analyse im Marketing genannt sowie der Prozessablauf einer Analyse zur Entwicklung und Validierung von Scoreklassen beschrieben.

Der praktische Teil dieser Arbeit, Kapitel 4, 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.

# 2 Survival Analyse

Dieses Kapitel behandelt gängige Verfahren der Survival Analyse, die auch im praktischen Teil dieser Arbeit verwendet werden. In Abschnitt 2.1 werden Grundbegriffe, wie die Überlebensfunktion und die Hazard Funktion, definiert sowie die Besonderheiten der Daten bei Überlebensdaten beschrieben. Im folgenden Abschnitt 2.2 wird der Kaplan-Meier-Schätzer zur Schätzung der Überlebensfunktion vorgestellt. Zum Vergleich von Überlebenszeiten wird in Abschnitt 2.3 der Log-Rank-Test erläutert. Die Cox-Regression wird in Abschnitt 2.4 vorgestellt. Sie dient der Beschreibung des Einflusses von Parametern auf die Überlebenszeit.
Die folgenden Ausführungen sind angelehnt an [KM06] sowie an [HS18]. Wurden weitere Quellen verwendet, wird im Text darauf verwiesen.

## 2.1 Grundlagen

Die Survival Analyse untersucht die Zeit bis ein bestimmtes Ereignis eingetreten ist. In der Literatur wird dieses Ereignis dem Namen Survival Analyse entsprechend meist als „Tod“ definiert. Je nach Anwendungsgebiet der Survival Analyse, ergeben sich andere Begrifflichkeiten für das Ereignis. Die Analyse von Einflüssen auf die Zeit bis zum Eintreten eines Ereignisses spielt in verschiedenen wissenschaftlichen Bereichen, wie beispielsweise der Medizin, Soziologie, den Ingenieurwissenschaften und der Betriebswirtschaftslehre eine Rolle. In diesem Kapitel wird das Ereignis auch als „Sterben“ bzw. das Gegenereignis als „Überleben“ bezeichnet. Bei der Betrachtung von Überlebenszeiten ergeben sich Schwierigkeiten, weshalb beispielsweise die klassische lineare Regression nicht als Analysemethode geeignet ist.

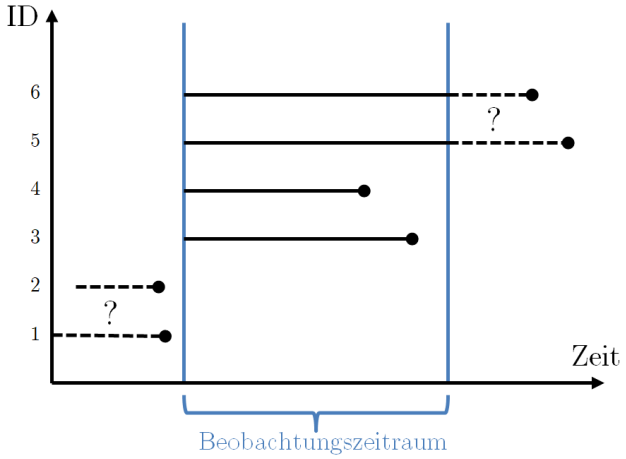



Abbildung 1: Exemplarische Darstellung links- und rechtszensierter Daten

Abbildung 1 veranschaulicht das Problem der Zensierung. Um eine Zensierung handelt es sich, wenn das untersuchte Ereignis nicht während des Beobachtungszeitraums

eingetreten ist. Dabei wird zwischen rechts-, links- und intervallzensierten Daten unterschieden. Bei rechtszensierten Daten ist das Ereignis im Beobachtungszeitraum noch nicht eingetreten, während bei linkszensierten Daten das Ereignis vor dem Beobachtungszeitraum unbemerkt bereits eingetreten ist und somit keine Daten über den Zeitpunkt vorliegen. Um intervallzensierte Daten handelt es sich, wenn nicht der Zeitpunkt sondern nur ein Zeitraum bestimmt werden kann, wann ein Ereignis eingetreten ist. Bei Personen mit ID 5 und 6 aus Abbildung 1 handelt es sich um rechtszensierte Daten, demgegenüber sind die Daten von Person 1 und 2 linkszensiert. Die Information, ob ein Ereignis zensiert oder nicht-zensiert ist, wird im Folgenden durch eine sogenannte Statusvariable beschrieben.

Würde man beispielsweise rechtszensierte Fälle bei einer Analyse ausschließen, statt sie als zensiert zu betrachten, ergäben sich Fehler. Analyseverfahren wie die lineare Regression würden die Überlebenszeit unterschätzen, da ausgeschlossene Personen genau die Personen sind, die über den Beobachtungszeitraum hinaus überleben und somit länger leben.

Die Zeit sei nun beschrieben durch die stetige Zufallsvariable $T$ mit Ausprägung $t$. Eine wichtige Größe in der Survival Analyse ist die Überlebensfunktion $S(t)$. Sie ist definiert als

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

und ist das Integral über die Dichtefunktion $f(x)$. Sie beschreibt die Wahrscheinlichkeit, den Zeitpunkt $t$ zu überleben. Es gelten folgende Eigenschaften:

- $S(t)$ ist eine streng monoton fallende Funktion, d.h. für $t_1 < t_2$ gilt $S(t_1) > S(t_2)$
- $t$ hat einen Wertebereich von $[0; \infty]$
- $S(0) = 1$, d.h. die Wahrscheinlichkeit, den Zeitpunkt 0 zu überleben, liegt bei 1
- $\lim\limits_{t\to\infty} S(t) = 0$

Die zugehörige Sterbefunktion $F(t)$ als Gegenstück zur Überlebensfunktion mit $S(t) = 1 - F(t)$ wird definiert mit

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

und beschreibt entsprechend die Wahrscheinlichkeit, bis zum Zeitpunkt $t$ zu sterben. Eine weitere wichtige Größe in der Survival Analyse ist die Hazardfunktion. Die Hazardfunktion definiert die Rate, dass ein Ereignis zum Zeitpunkt $t$ eintritt unter

der Bedingung, dass das Ereignis bis zum Zeitpunkt $t$ noch nicht eingetreten ist. Die Hazardfunktion lautet

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

Die Hazardfunktion nimmt bei der Cox-Regression eine zentrale Rolle ein (siehe Abschnitt 2.4). Mit $H(t)$ bezeichnet man die kumulative Hazardfunktion.

$$H(t) = \int_0^t h(u)du = -\ln(S(t)) \tag{4}$$

Die sich wie unter 4 aus der Überlebensfunktion berechnen lässt. Die kumulative Hazardfunktion kann als „Anhäufung von Hazards“ beschrieben werden. Abbildung 2 zeigt den typischen Verlauf von Überlebens-, Sterbe- und Hazardfunktion.

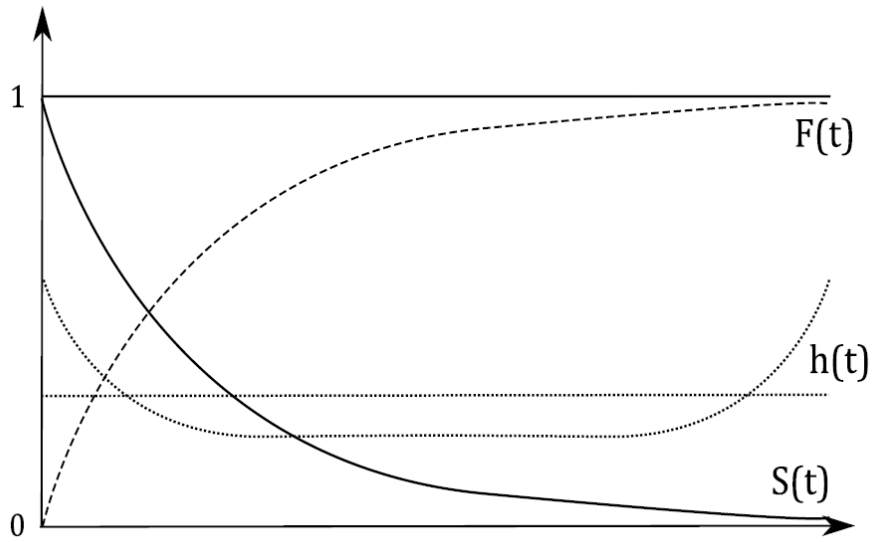



Abbildung 2: Typischer Verlauf von Überlebens-, Sterbe- und Hazardfunktion (zeitabhängig und zeitunabhängig); Darstellung aus [HS18, S.883]

Bisher wurde die Zeit $T$ als metrische Zufallsvariable definiert. Betrachtet man die Zeit als diskrete Größe mit $n$ Ereignissen zu den Zeitpunkten $t_1, t_2, ..., t_n$ so folgt

$$h_i = P(T = t_i | T \geq t_i) = \frac{p_i}{\sum_{j=1}^{n} p_j} \text{ mit } P(T = t_i) = p_i, i = 1, ..., n \tag{5}$$

$$H(t) = \sum_{i:t_i \leq t} h_i \tag{6}$$

$$S(t) = \prod_{i:t_i \leq t} (1 - h_i) \tag{7}$$

Im Gegensatz zu den stetigen Funktionen in Abbildung 2 ergeben sich im diskreten

Fall Treppenfunktionen für die Hazard-, Sterbe- und Überlebensfunktion.

## 2.2 Kaplan-Meier-Schätzer

Ein weit verbreitetes Verfahren zur Schätzung der Überlebensfunktion ist der Kaplan-Meier-Schätzer. Er wurde 1958 von Edward L. Kaplan und Paul Meier veröffentlicht [KM58] und liefert eine nicht-parametrische Schätzung $\hat{S}$ für rechtszensierte Daten.

$$\hat{S}(t) = \prod_{i:t_{(i)}<t} \frac{n_i - d_i}{n_i} \tag{8}$$

mit $t_{(i)}$ sind geordnete Zeitpunkte und $n_i$ als die Anzahl der Ereignisse unter Risiko („at risk“), d. h. die Ereignisse, die bis zum Zeitpunkt $t_i$ noch nicht eingetreten sind und $d_i$ die Anzahl an bereits eingetretenen Ereignissen. Die Wahrscheinlichkeit zu überleben, ist also definiert als das Produkt aus den beiden vorherigen Wahrscheinlichkeiten. Da nur die Ereignisse unter Risiko betrachtet werden, fließen zensierte Fälle zum Zeitpunkt ihrer Zensierung nicht mehr in die Berechnung ein.

Der Standardfehler kann mithilfe der Greenwood-Formel für die Schätzung der Varianz berechnet werden:

$$SE(\hat{S}(t)) = \sqrt{\hat{S}^2(t) \sum_{i:t_{(i)}<t} \frac{d_i}{n_i(n_i - d_i)}} \tag{9}$$

Das $1 - \alpha$-Konfidenzintervall lautet somit:

$$[\hat{S}(t) - z_{1-\frac{\alpha}{2}} \cdot SE(\hat{S}(t)); \hat{S}(t) + z_{1-\frac{\alpha}{2}} \cdot SE(\hat{S}(t))] \tag{10}$$

Alternativ kann das Konfidenzintervall nach Kalbfleisch und Prentice [KP80] geschätzt werden. Diese Berechnung wird im praktischen Teil dieser Arbeit verwendet, da sie die Schätzung der Überlebenswahrscheinlichkeit auf den Wertebereich
von 0 bis 1 beschränkt:

$$[\exp(\ln \hat{S}(t) - z_{1-\frac{\alpha}{2}} \cdot SE(\hat{H}(t))); \exp(\ln \hat{S}(t) + z_{1-\frac{\alpha}{2}} \cdot SE(\hat{H}(t)))]. \tag{11}$$

Abbildung 3 zeigt eine Kaplan-Meier-Kurve. Es handelt sich hier um eine Treppenfunktion, bei der jedes Ereignis durch eine Stufe dargestellt ist. Zensierte Fälle werden mit einem Kreuz markiert, beeinflussen jedoch nicht den Verlauf der Kurve. Die Treppenfunktion beginnt bei $S(0) = 1$, muss jedoch nicht zwangsläufig bei $S(t_{max}) = 0$ enden. Sie endet genau dann nicht bei 0, wenn es sich bei der letzten Beobachtung $t_{max}$ um einen zensierten Fall handelt. Dies ist in Abbildung 3 nicht der Fall: Hier ist

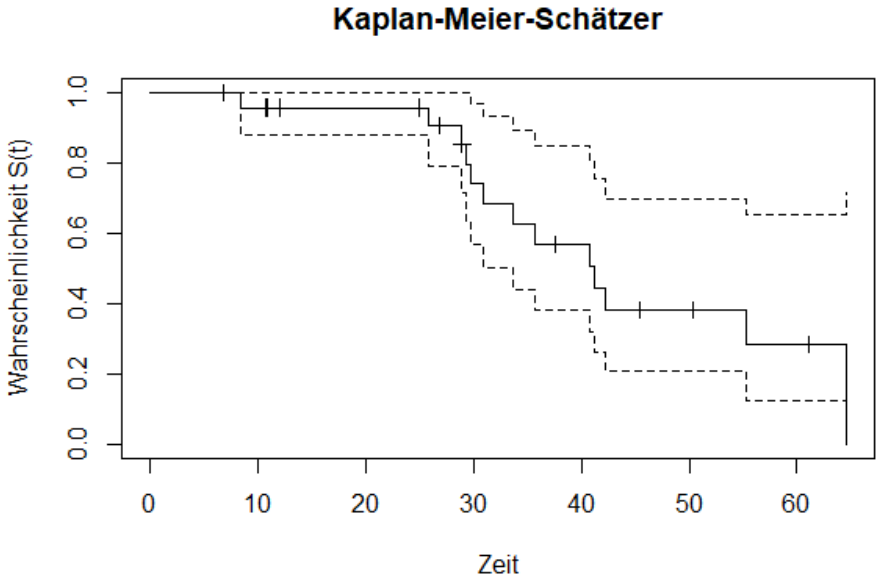



Abbildung 3: Beispiel für Kaplan-Meier-Kurve

der letzte Fall unzensiert. Die gestrichelten Linien in der Abbildung stellen das Konfidenzintervall nach 11 dar. Zu erkennen ist, dass das Konfidenzintervall im Verlauf der Kurve, also zu einem späteren Zeitpunkt, breiter wird. Die Anzahl der Patienten unter Risiko reduziert sich und somit auch die Genauigkeit der Schätzung.

Die Quantile $t_q$ der Überlebenszeit können berechnet werden mit

$$\hat{t_q} = \min\{t_i : \hat{S}(t_i) \leq 1 - q\}. \tag{12}$$

Für das $t_{0,5}$-Quantil ergibt sich somit die Aussage, dass zum Zeitpunkt $t$ bereits 50% der Ereignisse eingetreten sind. Es wird auch als das „mediane Überleben“ bezeichnet. Ein Quantil kann nicht berechnet werden, wenn

$$\hat{S}(t_i) > q \text{ für alle t.} \tag{13}$$

In diesem Fall ist während der Beobachtungszeit für weniger als für $q$% der Patienten ein Ereignis eingetreten.

Wahlweise kann auch das „mittlere Überleben“ über die Berechnung der Fläche unter der Kurve von $S(t)$ bestimmt werden, mit $\mu = \int_0^\infty S(t)dt$. Da es sich bei der Schätzung der Überlebensfunktion nach Kaplan-Meier um eine Treppenfunktion handelt, kann die Fläche durch die Summe der Rechteckflächen bestimmt werden. Die letzte Beobachtung sollte unzensiert sein, da die Fläche ansonsten nicht begrenzt ist. Alternativ kann auch der RMST (Restricted Mean Survival) berechnet werden, der nur die Fläche in dem Intervall $[0, \tau]$, mit $\tau$ ist letztes unzensiertes Ereignis, betrachtet. Im praktischen Teil dieser Arbeit wird das mediane Überleben verwendet.

## 2.3 Log-Rank-Test

Der Log-Rank-Test untersucht die Überlebenszeiten für zwei unabhängige Stichproben (definiert durch eine Gruppenvariable) auf statistisch signifikante Unterschiede.

$$H_0 : S_0(t) = S_1(t) \text{ gegen } H_1 : S_0(t) \neq S_1(t) \tag{14}$$

Die folgenden Ausführungen sind angelehnt an [HL99, Kapitel 2.4]. Der Test basiert auf einer Kontingenztabelle mit der Gruppe, dem Status (zensiert/ nicht zensiert) für jeden beobachteten Zeitpunkt $t_i$. Eine solche Kontingenztabelle zeigt Tabelle 1.

Tabelle 1: Kontingenztabelle des Log-Rank-Tests zum Vergleich von Überlebensfunktionen zur Zeit $t_{(i)}$; Darstellung nach [HL99, S.59]

| | | Gruppe 1 | Gruppe 0 | Gesamt |
|---|---|---|---|---|
| Ereignis | ja | $d_{1i}$ | $d_{0i}$ | $d_i$ |
| | nein | $n_{1i} - d_{1i}$ | $n_{0i} - d_{0i}$ | $n_i - d_i$ |
| | Unter Risiko | $n_{1i}$ | $n_{0i}$ | $n_i$ |

mit

- $n_{0i}, n_{1i}$: Anzahl der Fälle unter Risiko zum Zeitpunkt $t_{(i)}$ in Gruppe 0 bzw. Gruppe 1
- $d_{0i}, d_{1i}$: Anzahl der eingetretenen Ereignisse zum Zeitpunkt $t_{(i)}$ in Gruppe 0 bzw. Gruppe 1
- $d_i$: Gesamtanzahl an Ereignissen zum Zeitpunkt $t_{(i)}$
- $n_i$: Gesamtanzahl der Fälle unter Risiko zum Zeitpunkt $t_{(i)}$.

Unter der Annahme, dass die Überlebensfunktionen in den beiden Gruppen gleich sind, wird nun die erwartete Anzahl an eingetretenen Ereignissen $\hat{e}_{0i}$ in Gruppe 0 (alternativ in Gruppe 1) mit

$$\hat{e}_{0i} = \frac{n_{0i} d_i}{n_i} \tag{15}$$

geschätzt. Die Varianz von $d_{0i}$ kann über die hypergeometrische Verteilung geschätzt werden.

$$\hat{v}_{0i} = \frac{n_{0i} n_{1i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)} \tag{16}$$

Die Teststatistik kann über die gewichtete Summe der beobachteten Überlebenszeiten berechnet werden:

$$Q = \frac{[\sum_{i=1}^{m} w_i(d_{0i} - \hat{e}_{0i})]^2}{\sum_{i=1}^{m} w_i^2 \hat{v}_{0i}} \tag{17}$$

Der Beitrag zur Teststatistik wird zu jedem Zeitpunkt durch Berechnung der erwarteten Anzahl von Todesfällen in Gruppe 1 oder 0 unter der Annahme erhalten, dass die Überlebensfunktion in jeder der beiden Gruppen gleich ist.

Die Teststatistik ist unter der Annahme, dass die Zensierungen in beiden Gruppen dem gleichen Muster folgen und die Stichprobengröße ausreichend groß ist, asymptotisch $\chi^2$-verteilt mit einem Freiheitsgrad.

Mit $w_i = 1$ handelt es sich hier um die Teststatistik nach Mantel und Haenszel, die auch im R Paket `survival` für die Berechnung des Log-Rank-Tests mit der Funktion `survdiff()` verwendet wird. Setzt man die Gewichtung $w_i = n_i$, beschrieben u.a. von Breslow (1970) [Bre70], so nennt man den Test auch generalisierten Wilcoxon-Test. Diese Gewichtung bevorzugt frühere Ereignisse gegenüber späteren Ereignissen. Zudem kann der Test verallgemeinert werden, um mehr als zwei unabhängige Stichproben zu vergleichen. Die paarweisen Vergleiche können in R mit der Funktion `pairwise_survdiff()` des Pakets `survminer` erfolgen. Voraussetzung, dass der Log-Rank-Test berechnet werden darf, ist, dass sich die Kurven der Überlebensfunktionen der beiden unabhängigen Stichproben nicht schneiden (vgl. [HS18, S.890]).

## 2.4 Cox-Regression

Die Cox-Regression wurde 1972 von David Cox [Cox72] veröffentlicht und ermöglicht es als multivariates Verfahren, aus der Survival Analyse den Zusammenhang von Einflussgrößen auf die Überlebenszeit $T$ zu untersuchen. Als weit verbreitetes Verfahren ist es in allen gängigen Statistik Programmen enthalten.

Der Modellansatz nach Cox für m Einflussgrößen lautet:

$$h(t, x) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m). \tag{18}$$

Dabei bezeichnet $h_0(t)$ die Baseline-Hazard-Rate, auch Basisrisiko genannt. Diese Baseline-Hazard-Rate wird nicht weiter spezifiziert. Aus diesem Grund wird das Cox-Regressionsmodell auch den semi-parametrischen Verfahren zugeordnet. Nur für die Effekte der Einflussgrößen wird eine Verteilung angenommen. Nach Definition 18 ist zu erkennen, dass die Zeit nur einen Einfluss auf das Basisrisiko hat. Der Vektor der Einflussgrößen wirkt multiplikativ auf die Baseline-Hazard-Rate [BHM86, S.138]. Die Baseline-Hazard-Rate kann durch die Werte der Einflussgrößen erhöht oder verringert

werden.

Die Cox-Regression wird auch Proportionales Hazard Modell genannt. Dies liegt daran, dass das Hazard Ratio für zwei Individuen $x$ und $x^*$ von der Zeit unabhängig und konstant ist, wie in 19 zu sehen.

$$\frac{h(t,x)}{h(t,x^*)} = \frac{h_0(t)\exp(\sum_{k=1}^{m}\beta_k x_k)}{h_0(t)\exp(\sum_{k=1}^{m}\beta_k x_k^*)} = \frac{\exp(\sum_{k=1}^{m}\beta_k x_k)}{\exp(\sum_{k=1}^{m}\beta_k x_k^*)} = \exp[\sum_{k=1}^{m}\beta_k(x_k - x_k^*)] \tag{19}$$

### 2.4.1 Parameterschätzung

Die Parameterschätzung erfolgt mithilfe einer partiellen Likelihood Funktion $S$. Es handelt sich hier um ein abgewandeltes Maximum Likelihood Verfahren von Cox [Cox72], das bedingte Wahrscheinlichkeiten verwendet.

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(x_i'\beta)}{\sum_{t_{(j)}\geq t_i}\exp(x_i'\beta)} \tag{20}$$

Hier wird die Vektorschreibweise verwendet: $x_i'$ definiert den Vektor der m Einflussgrößen $x_1, ..., x_m$. $t_{(i)}$ beschreibt die geordneten Ereigniszeiten.

Es wird das Ereignis eines Individuums zum Zeitpunkt $t_{(i)}$ ins Verhältnis zu allen Individuen, die zu diesem Zeitpunkt noch unter Risiko stehen, gesetzt. Die Maximierung über $\beta$ erfolgt numerisch. Hier wird nochmals deutlich, dass die Baseline-Hazard-Rate $h_0(t)$ nicht für die Schätzung der Parameter erforderlich ist. Zensierte Fälle erhalten einen Einfluss auf die Schätzung im Nenner des Quotienten, da sie Bestandteil der Fälle unter Risiko zum jeweiligen Zeitpunkt sind (Abschnitt nach [HS18, Kapitel 8.7.4.3]).

Der hier beschriebene Ansatz behandelt die Parameterschätzung ohne das Auftreten von Bindungen. Um Bindungen in Daten handelt es sich, wenn mehr als ein Ereignis zum Zeitpunkt $t_{(i)}$ vorliegt.

Liegen Bindungen vor, so kann die Likelihood nach Breslow, Likelihood nach Efron oder der Diskrete Likelihood berechnet werden. Eine ausführliche Gegenüberstellung der drei Verfahren findet sich in [KM06, Kapitel 8].

Die Interpretation der Parameter erfolgt über das Hazard Ratio (kurz: HR, siehe auch 18) und gleicht der Interpretation der Odds Ratios bei der logistischen Regression [HS18, S. 904].

Betrachtet wird das Hazard Ratio von zwei Gruppen A und B, also HR $= \frac{h_B(t)}{h_A(t)}$. Dabei ist Gruppe A die Referenzgruppe. Dann lässt sich das Hazard Ratio folgendermaßen interpretieren [ZBH11]:

- HR $\approx 1$: Das Risiko ist in Gruppe A und B ungefähr gleich groß.
- HR $> 1$: Das Risiko ist in Gruppe B größer als in Gruppe A.

- HR $< 1$: Das Risiko ist in Gruppe B kleiner als in Gruppe A.

### 2.4.2 Prüfung der Voraussetzungen

Die Voraussetzung der proportionalen Hazards kann grafisch mithilfe der Schoenfeld Residuen, benannt nach David Schoenfeld [Sch82], überprüft werden. Die Darstellung erfolgt nach [HS18, S. 862f.].

$$r_{S_{ji}} = x_{ji} - \frac{\sum_{t_{(l)} \geq t_{(i)}} x_{jl} \exp(x'\hat{\beta})}{\sum_{t_{(l)} \geq t_{(i)}} \exp(x'\hat{\beta})} \tag{21}$$

Schoenfeld Residuen werden für jeden nicht zensierten Fall $i$ und jede Einflussgröße $j$ berechnet. Es wird die Differenz zwischen dem beobachteten Wert $x_{ji}$ und dem erwarteten Wert gebildet.

Die Schoenfeld Residuen sollen sich, bei Erfüllung der Voraussetzung, im Zeitverlauf nicht ändern. Aus diesem Grund werden für jede Einflussgröße die Schoenfeld Residuen gegen die Zeit geplottet. Entsprechende Plots finden sich im praktischen Teil dieser Arbeit in Abschnitt 4.4.4.

Zusätzlich zu der grafischen Überprüfung der Schoenfeld Residuen können Tests gerechnet werden. Entsprechende Tests auf proportinale Hazards sind beschrieben bei [GT94].

Sind die Voraussetzungen zur Modellbildung verletzt, können weitere Schritte unternommen werden (siehe: Abschnitt 2.4.4).

### 2.4.3 Cox–Snell Residuen zur Beurteilung der Modellanpassung

Die Beurteilung des Modells kann mithilfe der sogenannten Cox-Snell Residuen erfolgen. Die Cox-Snell Residuen zum Zeitpunkt $t_i$ für das $i$-te Individuum sind definiert als

$$r_{C_i} = \hat{H}_0 \exp(x'\hat{\beta}) = \hat{H}(t_i) = -\ln(\hat{S}(t_i)). \tag{22}$$

Es lässt sich zeigen, dass die $r_{C_i}$ einer Exp(1)-Verteilung folgen. Um zu überprüfen, ob die $r_{C_i} \approx \text{Exp}(1)$ verteilt sind, wird ein Plot entwickelt:

Mithilfe des Nelson-Aalen-Schätzer[1] wird die Kumulative Hazard Rate von $r_{C_i}$ geschätzt. Wird dies gegen $r_{C_i}$ geplottet, so sollte sich eine durch den Ursprung gehende Gerade mit Steigung 1 ergeben.

In R können die Cox-Snell-Residuen über die Martingale-Residuen, die in R mit der Funktion `residuals(..., type = "martingale")` des Paktets `survival` berechnet

[1] oder mit der Korrektur nach Fleming-Harrington bei Vorliegen von Bindungen

werden, abgeleitet werden. Zwischen den Martingale Residuen $r_{M_i}$ und den Cox-Snell-Residuen $r_{C_i}$ besteht die folgende Beziehung:

$$r_{C_i} = \delta_i - r_{M_i}. \tag{23}$$

Dabei bezeichnet $\delta_i$ die Statusvariable mit Codierung: 0 = zensiert, 1 = nicht zensiert eines Individuums zum Zeitpunkt $t_i$. Mehr über Martingale-Residuen ist nachzulesen bei [KM06, Kapitel 11].

Im praktischen Teil dieser Arbeit, in Kapitel 4.4.7 ist ein Cox-Snell-Plot abgebildet.

### 2.4.4 Erweiterungsmöglichkeiten

Das „einfache“ Cox-Regressionsmodell bietet diverse Erweiterungsmöglichkeiten.

Thematisches Vorwissen, dass die Effekte der Einflussgrößen nicht proportional sind, müssen zu einer Anpassung des Modells führen. So kann beispielsweise im Marketingbereich im Vorfeld bekannt sein, dass Kundengruppe 1 innerhalb kürzester Zeit nach Erscheinen einer Werbekampagne ein Produkt kauft. Nach Ablauf von ein paar Tagen kauft aus dieser Gruppe kaum noch ein Kunde. Kundengruppe 2 hat hingegen eine längere Vorlaufzeit und beginnt erst im späteren Zeitverlauf Käufe zu tätigen. Die Hazards der beiden Kundengruppen sind somit nicht proportional.

Eine Möglichkeit, dieses Problem zu lösen, wäre, das Modell nach dem Faktor Kundengruppe zu stratifzieren. Die folgende Liste zeigt diese und weitere Erweiterungsmöglichkeiten des Cox-Regressionsmodells:

- *Stratifizierung*: Für jedes Stratum werden unterschiedliche Baseline Hazard Raten angenommen.
- *Zeitabhängige Kovariaten*: Einflussparameter, die sich über die Zeit ändern (z. B. Wohnortwechsel eines Individuums) können in das Modell aufgenommen werden.
- *Linkszensierte Daten*: Das Modell kann entsprechend angepasst werden, so dass auch linkszensierte Daten verarbeitet werden können.
- *Interaktionseffekte*: Interaktionseffekte können analog zur linearen oder logistischen Regression über einen Interaktionsterm ($x_1 \times x_2$) in das Modell aufgenommen werden und untersucht werden.

Die Erweiterungsmöglichkeiten sind ausführlich nachzulesen bei [KM06, Kapitel 9].

# 3 Survival Analyse im Markeing

Bei Anwendung der Verfahren der Survival Analyse im Marketingbereich wird statt der Überlebenszeit eine andere Zielgröße betrachtet. Dabei kann es sich um die Zeit bis zum Folgekauf, die Dauer zwischen der Teilnahme an Werbeaktionen, die Zeit bis zur Abmeldung von einem Newsletter uvm. handeln. In der vorliegenden Arbeit wird die Zeit vom Kauf eines Kunden bis zu seinem Folgekauf, genannt Kaufintervall, betrachtet. In diesem Zusammenhang wird zudem nach Formel 12, statt vom medianen Überleben vom medianen Kaufen für das $t_{0.5}$-Quantil der Kauffunktion berichtet. Die Kauffunktion steht für die Sterbefunktion $F(t)$ nach Formel 2.

Wie in der Medizin ergibt sich das Problem der zensierten Daten, die eine Bewertung der Kauffrequenz auf Basis des Durchschnitts der vergangenen Kaufintervalle verhindern. Kunden, die nur einmal gekauft haben, müssten aus der Analyse ausgeschlossen werden, da ansonsten das Ergebnis verzerrt würde. In die Survival Analyse können demnach auch im Marketing Merkmale einbezogen werden, welche die Zeit bis zum Folgekauf beeinflussen.

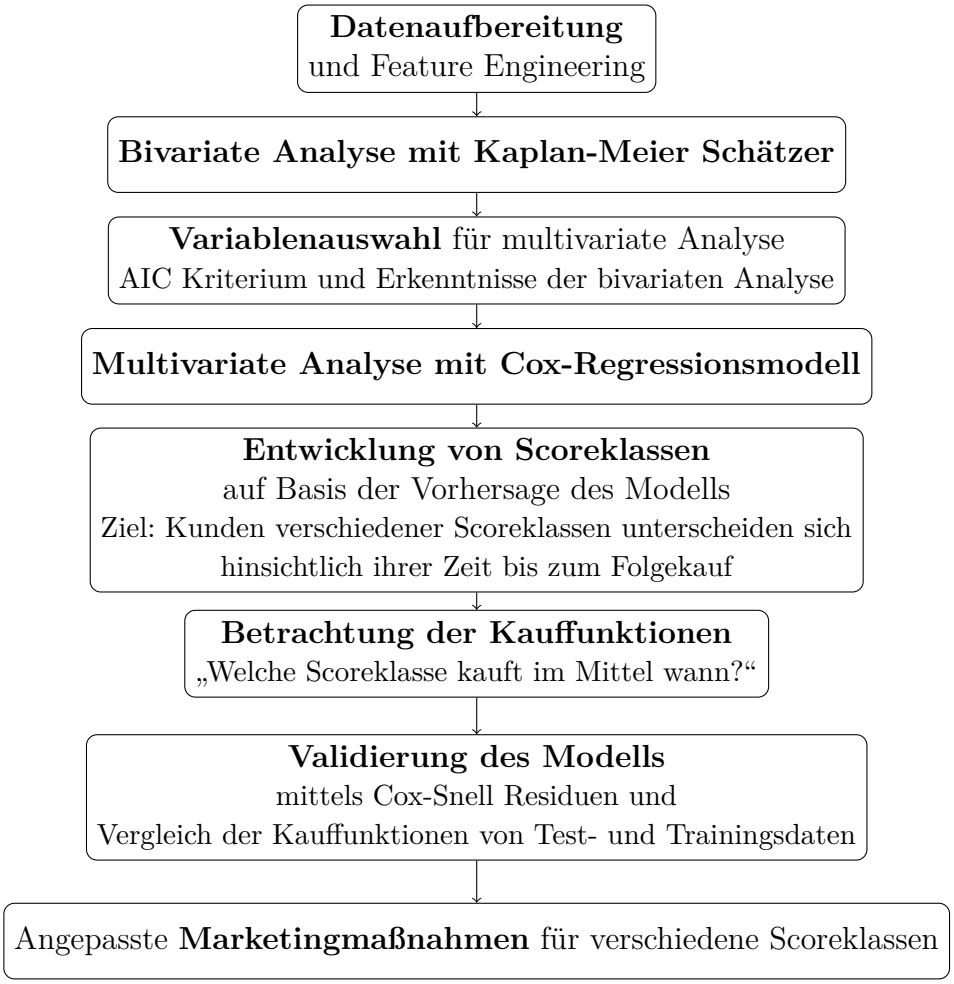



Abbildung 4: Diagramm zum Ablauf der Analyse zur Entwicklung von Scoreklassen mit unterschiedlicher Kauffrequenz

Abbildung 4 zeigt den möglichen Ablauf einer Survival Analyse zur Entwicklung von Scoreklassen, um Marketingmaßnahmen zu optimieren. Nachfolgend werden die Schritte näher erläutert.

1. Datenaufbereitung und Feature Engineering
   Die Daten müssen in geeigneter Form aufbereitet werden, mit Zeit bis Folgekauf, Status (zensiert/ nicht zensiert) sowie weiteren interessierenden Einflussgrößen.
   Feature Engineering meint das Ableiten von neuen Merkmalen, die für die Survival Analyse von Interesse sind. Dass Kunden mehrmals kaufen, also mehrere Kaufintervalle betrachtet werden, ist bei der Survival Analyse unter Verwendung von Kaplan-Meier-Schätzern und Cox-Regressionsmodellen nicht möglich. Diese Messwiederholung kann aber durch geeignetes Feature Engineering indirekt in die Modelle einfließen.
   Mögliche Merkmale können sein:
   - Gesamtanzahl der Bestellungen eines Kunden in der Vergangenheit
   - Umsatz eines Kunden in der Vergangenheit

2. Bivariate Analyse mit Kaplan-Meier Schätzer
   Mithilfe der bivariaten Analyse können erste Einflüsse auf die Zeit bis zum Folgekauf eines Kunden aufgedeckt werden. Zudem können in diesem Schritt Kategorien nominaler Merkmale für die spätere Scorebildung zusammengefasst werden, wenn sie keine relevanten signifikanten Unterschiede hinsichtlich der Zeit bis zum Folgekauf zeigen. Hinweise zur geeigneten Kodierung von Einflussgrößen in der Survival Analyse sind nachzulesen bei [KM06, Kapitel 8.2]. Zudem kann an dieser Stelle der Log-Rank-Test gerechnet werden.

3. Variablenauswahl für multivariate Analyse
   Für das Cox-Regressionsmodell kann im nächten Schritt die schrittweise Variablenauswahl auf Basis des AIC Kriteriums sowie auf Basis von Erkenntnissen der bivariaten Analyse erfolgen.

4. Multivariate Analyse mit Cox-Regressionsmodell
   Die Cox-Regression wird mit den Einflussgrößen der Variablenauswahl des vorherigen Schrittes für die Trainingsdaten modelliert. Die Voraussetzung der proportionalen Hazards wird mithilfe der Schoenfeld Residuen untersucht.

5. Entwicklung von Scoreklasse
   Mit der Vorhersage des linearen Terms des Cox-Regressionsmodells wird ein Score für alle Kunden berechnet. Dieser Score wird auf Basis der Dezile klassiert, um 10 gleich große Kundengruppen zu erhalten. Jeder Kunde kann nun einer Scoreklasse zugeordnet werden. Die Scoreklassen unterscheiden sich hinsichtlich ihrer

Kaufintervalle. Zusätzlich können die Scoreklassen deskriptiv betrachtet werden. Mögliche Fragestellungen:

- *In welcher Scoreklasse ist der Männeranteil am größten?*
- *In welche Scoreklasse sind die jungen Kunden eingeordnet?*
- *Kaufen Kunden einer bestimmten Scoreklasse vermehrt online?*

6. Betrachtung der Kauffunktionen
   Für die 10 Scoreklassen können nun die Kauffunktionen betrachtet werden. Diese Kauffunktionen werden wiederum mit dem Kaplan-Meier Verfahren geschätzt. Nun werden die Quantile der Kauffunktionen der Scoreklassen betrachtet. Das $t_{0.5}$-Quantil beschreibt, zu welchem Zeitpunkt 50% der Kunden einer Scoreklasse bereits einen Folgekauf getätigt haben.

7. Validierung des Modells
   Die Güte der Modellanpassung des Modells wird mithilfe der Cox-Snell Residuen überprüft. Zudem werden, um Abweichungen sichtbar zu machen, die ermittelten Quantile und Kauffunktionen des Modells den Testdaten gegenübergestellt.

8. Angepasste Marketingmaßnahmen
   Nun können Marketingmaßnahmen auf Basis der Kauffunktionen der Scoreklassen optimiert werden. Eine „gute“ Scoreklasse bekommt früher Werbepost, als eine „schlechte“ Scoreklasse. Mögliche Marketingmaßnahmen sind in Kapitel 4.4.6 beschrieben.

Das beschriebene Vorgehen wird im praktischen Teil dieser Arbeit, Kapitel 4 umgesetzt.

Ein Problem bei der Survival Analyse im Marketing ist die Bewertung von statistisch signifikanten Ergebnissen. Einer medizinischen Studie geht (im besten Fall) eine Fallzahlplanung voraus, um die Power und das Signifikanzniveau zu bestimmen. Durch die optimale Stichprobengröße werden klinisch relevante Effekte mithilfe statistischer Tests erkannt. Je größer jedoch der Stichprobenumfang, desto eher ergibt sich ein statistisch signifikantes Ergebnis, obwohl das Ergebnis inhaltlich nicht relevant ist [BS11, Kapitel 7.7].

Tabelle 2 zeigt Ergebnisse des Log-Rank-Tests bei unterschiedlicher Stichprobengröße. Dazu wurden zwei unabhängige Stichproben mit exponentialverteilter[2] Zeit bis zum Folgekauf mit $\lambda = \frac{1}{180}$ simuliert. Mit dem Erwartungswert $E(T) = \frac{1}{\lambda}$ ergibt sich eine durchschnittliche Zeit bis zum Folgekauf von 180 Tagen [HS18, S. 893]. In Anhang D findet sich das zugehörige Listing zur Berechnung von Tabelle 2.

[2]Die Exponentialverteilung und auch die Weibullverteilung werden häufig verwendet, um die Zeit bis zu einem Ereignis, Ausfallzeiten etc., parametrisch zu beschreiben.

Tabelle 2: Ergebnisse des Log-Rank-Tests bei unterschiedlicher Stichprobengröße

| Stichprobengröße | $\chi^2$ | p-Wert |
|---|---|---|
| 1.000 | 0,1 | 0,9 |
| 10.000 | 0,2 | 0,8 |
| 100.000 | 1,6 | 0,5 |
| 500.000 | 7,8 | 0,005 |
| 1.000.000 | 15,4 | 0.05 |
| 2.500.000 | 15,4 | 0.002 |
| 5.000.000 | 15,4 | $< 0.001$ |

Für Gruppe 1 wurde die Zeit nun um einen halben Tag verschoben. Gruppe 1 kauft also 12 Stunden später als Gruppe 2. Nun wurden die beiden Stichproben auf statistisch signifikante Unterschiede hinsichtlich ihrer Zeit bis zum Folgekauf untersucht. Bei unterschiedlichen Stichprobengrößen zeigen sich andere Ergebnisse: Für eine Stichprobe von 1000 Kunden ist kein statistisch signifikanter Unterschied nachweisbar. Bei 2,5 Millionen und mehr Kunden ist das Ergebnis des Log-Rank-Tests statistisch signifikant. Der Unterschied der beiden Kauffunktionen ist in Wahrheit jedoch unbedeutend. Bei einem Kaufintervall von 180 Tagen macht es für eventuelle Marketingmaßnahmen keinen Unterschied, ob ein Kunde einen halben Tag später oder früher kauft. Betrachtet man Kaufdaten aus Kundendatenbanken, ist ein Umfang mit über einer Million Kunden nicht ungewöhnlich. Im praktischen Teil dieser Arbeit werden 2,37 Mio. Einkäufe von 1,27 Millionen betrachtet.

Entscheidungen im Marketing sollten demnach, neben der Betrachtung des p-Werts, auch auf inhaltlich relevanten, medianen Unterschieden im Kauf sowie auf grafischen Methoden wie Kaplan-Meier-Kurven basieren.

# 4 Survival Analyse von Kaufdaten mit R

Das folgende Kapitel beinhaltet eine Analyse in R mit dem Ziel der Optimierung von Marketingkampagnen eines deutschen Versandhändlers. Die Analyse erfolgt im Rahmen meiner Tätigkeit bei Statistik-Service Dr. Gladitz, Zionskirchstraße 27, D-10119 Berlin. Die bereitgestellten Daten für die vorliegende Arbeit sind anonymisiert. Zudem wird der Name des Versandhändlers nicht genannt.

Für die Optimierung der Marketingkampagnen stehen Verkaufsdaten von einem Geschäftsjahr zu Verfügung und sollen im Folgenden mithilfe der Survival Analyse untersucht werden. Dieses Kapitel gliedert sich in folgende Abschnitte: In Kapitel 4.1 werden zunächst die Aufgabenstellung und die zugehörigen Rahmenbedingungen erläutert. In 4.2 werden die Verkaufsdaten deskriptiv beschrieben.

Mithilfe von Kaplan-Meier-Schätzern, dem Log-Rank Test, sowie der Cox Regression wird in den folgenden Kapitel die Dauer von Kauf bis Folgekauf der Kunden analysiert. Zudem werden mithilfe der Cox-Regression Scoreklassen abgeleitet, die Kundengruppen auf Basis Ihrer Kauffrequenz bewerten.

## 4.1 Aufgabenstellung

Aus dem Kaufdatum und dem Datum des Folgekaufs lässt sich die Anzahl der Tage ermitteln, die ein Kunde bis zu einem Folgekauf benötigt. Zu verstehen, was diese Zeitdauer von Kauf bis Folgekauf beeinflusst und wie dieses Wissen verwendet werden kann, ist das Ziel dieser Analyse. Dabei werden Methoden der Survival Analyse genutzt, um den Einfluss verschiedener Merkmale wie z. B. das Geschlecht, Alter, generierter Umsatz bei vorherigen Käufen auf die Zeit bis zum Folgekauf zu analysieren.

Auf Basis der Survival Analyse der Kaufdaten sollen zudem Handlungsempfehlungen für verschiedene Marketingmaßnahmen entwickelt werden. Die Kunden werden mithilfe eines Scorings in verschiedene Käufergruppen einsortiert, wobei die verschiedenen Käufergruppen unterschiedlich schnell Folgekäufe tätigen.

Die Marketingmaßnahmen sollen so an die individuelle Kauffrequenz der Kunden angepasst werden können, indem definiert wird, zu *welchem Zeitpunkt* für *welche Käufergruppe welche Marketingmaßnahmen* zu empfehlen sind.

Die Analyse erfolgt in R. Es werden R Markdown-Dokumente angefertigt, die es erlauben auch neue Daten in den Folgejahren zu analysieren. Die entwickelten Modelle können so in den Folgejahren ohne großen Aufwand aktualisiert werden. Die kompilierten R Markdown-Dokumente sind dem elektronischen Anhang dieser Arbeit beigefügt.

## 4.2 Beschreibung des Datensatzes und deskriptive Statistik

Für die Analyse steht ein Datensatz mit Verkaufsdaten eines deutschen Versandhändlers mit Niederlassungen in mehreren europäischen Ländern zur Verfügung. Die Schwerpunkte im Verkauf des Händlers liegen in den Bereichen Mode und Wohnen. Die Waren werden ausschließlich über Produktkataloge und per Internet angeboten - stationäre Geschäfte sind nicht vorhanden. Der zu analysierende Datensatz beinhaltet die Verkaufsdaten deutscher Kunden eines Geschäftsjahres.

Der Datensatz für das Geschäftsjahr wurde erzeugt, indem 12 Monatsdateien zusammengefügt wurden. Diese 12 Monatsdateien enthalten jeweils den letzten Kauf eines Kunden in dem entsprechenden Monat (auch wenn er mehrmals in diesem Monat gekauft hat). Ein Kunde kann also bis zu 12 mal im Datensatz enthalten sein. Für das Feature Engineering des Datensatzes wurden drei Geschäftsjahre verwendet. Merkmale, wie der generierte Umsatz beziehen sich auf die 12 Monate vor dem zu analysierenden Jahr. Der beobachtete Folgekauf erfolgt innerhalb der 12 Monate nach dem zu analysierenden Jahr.

Eine Übersicht der Merkmale des Datensatzes mit zugehöriger Beschreibung und Kodierung findet sich im Anhang A.

Es wurden 2,37 Mio. Einkäufe von 1,27 Millionen Kunden getätigt. Davon erzielten 87.533 (3,7 %) der Einkäufe keinen Umsatz. Diese Einkäufe mit $\text{Umsatz} = 0$ werden aus der Analyse ausgeschlossen. Hier handelt es sich um Kunden, die aufgrund von Stornierungen keine Umsätze generiert haben.
Nach Ausschluss dieser Fälle verbleiben 1,2 Mio. Kunden mit 2,27 Mio. Einkäufen im Datensatz. Die Kunden tätigen im Mittel 1,9 ($\pm$ 1,4) Einkäufe. Der Großteil der Kunden ist weiblich: den 1,14 Mio. (95,0%) Frauen stehen 59.885 (5,4%) Männer gegenüber. Die Kunden sind durchschnittlich 54,5 ($\pm$ 11,4) Jahre alt. Die Angabe des Alters fehlt bei 3.887 (0,3%) der Kunden. Um diese Kunden nicht aus der Analyse ausschließen zu müssen, wurden die fehlenden Werte des Alters durch den Mittelwert ersetzt. Für die übrigen Kunden wurde das Alter zum Stichtag 31.12. gebildet.

Abbildung 5 zeigt die Anzahl der Einkäufe pro Monat und dabei ein unterschiedliches Kaufverhalten je nach Monat. Die Einführung der Frühjahr/Sommer und Herbst/Winter Kollektionen führt zu den umsatzstarken Monaten April und Oktober. Zum Ende der jeweiligen Saison verringern sich die Umsätze gegenüber der Vormonate. Im Dezember, Februar und August wird am wenigsten gekauft. Aufgrund der Saisonalität der Sortimente und der Datenstruktur erfolgt die Analyse des Kaufverhaltens auf monatlicher Basis. Die Survival Analyse untersucht die Zeit bis zu einem bestimmten Ereignis. Da ein Kunde mehrmals im Datensatz vorkommen kann, also mehrere Ereignisse zu einem Kunden vorliegen können, wäre die Datenstruktur bei nicht monatlicher Betrachtung nicht für die Anwendung des Kaplan-Meier-Schätzers sowie der Cox-Regression geeignet.

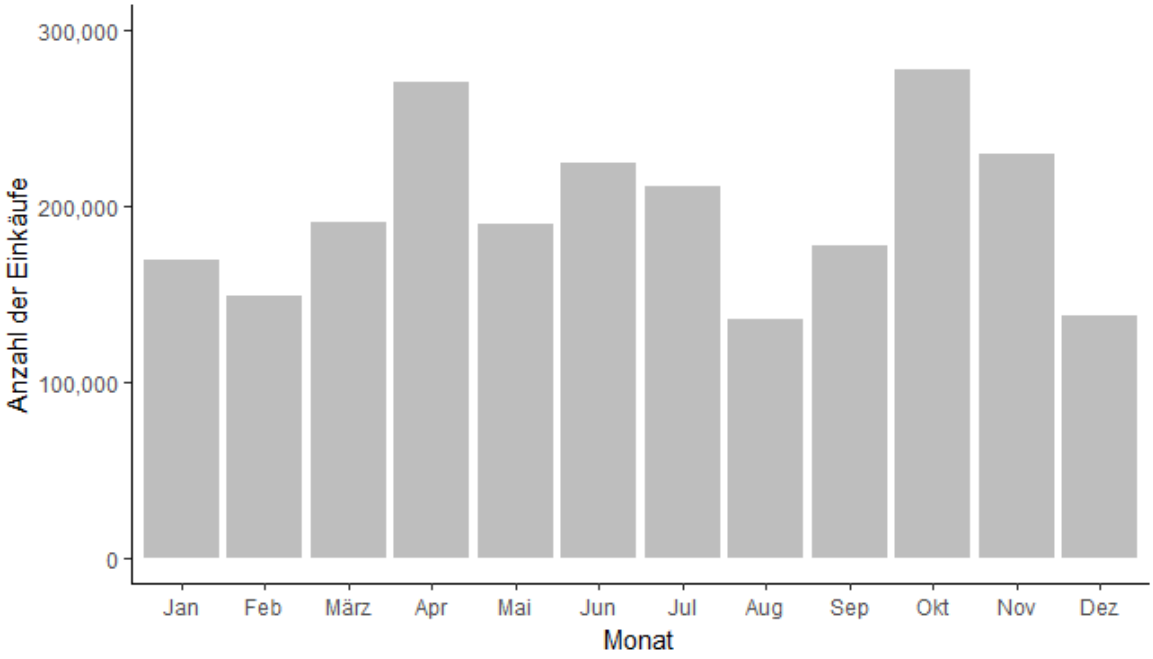



Abbildung 5: Balkendiagramm mit Anzahl an Einkäufen pro Monat

Aus den Merkmalen Kaufdatum und Folgebestellung wurde das Merkmal Tage abgeleitet. Innerhalb eines Jahres erfolgt bei 1,6 Mio. (70,4 %) Bestellungen eine Folgebestellung. In der Analyse betrachtet werden soll das Kaufverhalten der Kunden innerhalb eines Jahres. Einkäufe, die nicht innerhalb eines Jahres erfolgten (Tage > 365) werden deshalb als zensierte Daten (29,6 %) betrachtet. Entsprechend wird das Merkmal Status mit den Ausprägungen Folgekauf und Zensiert gebildet.

Tabelle 3: Exemplarische Datensätze zum Verständnis der Merkmale Tage und Status

| ID | Bestellung | Folgebestellung | Tage | Status |
|---|---|---|---|---|
| 101 | 23.10.2018 | 06.11.2018 | 14 | Folgebestellung |
| 102 | 03.03.2018 | 07.04.2018 | 35 | Folgebestellung |
| 103 | 16.09.2018 | 20.12.2019 | 366 | Zensiert |
| 104 | 13.02.2018 | – | 366 | Zensiert |
| ... | ... | ... | ... | ... |

Das Merkmal Umsatz enthält den Umsatz der vergangenen 12 Monate eines Kunden. Im Schnitt beträgt der Umsatz 711,5 Euro (± 880.2, Median: 441 Euro, Max: 28.486 Euro). Um den Umsatz als Gruppierungsvariable für den Kaplan-Meier Schätzer verwenden zu können und zur besseren Interpretierbarkeit als Einflussparameter für die Cox-Regression, wird der Umsatz, auf Basis seiner Quantile klassiert um fünf nahezu gleich große Gruppen zu erhalten. Abbildung 6 zeigt die rechtsschiefe Verteilung des Umsatzes sowie die Klassierung.

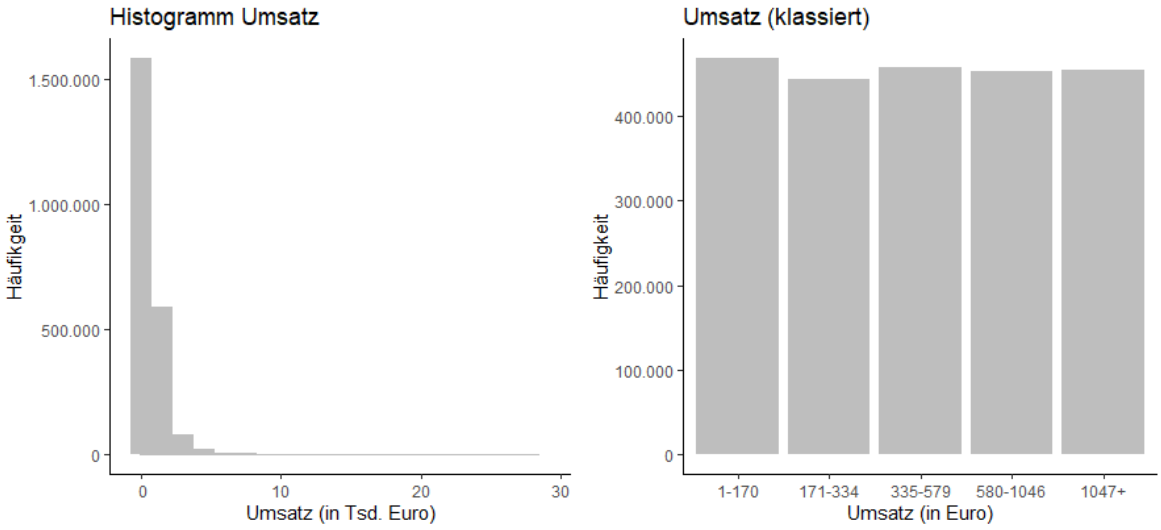



Abbildung 6: Histogramm und Klassierung von Umsatz

Die metrischen Merkmale Alter sowie AnzahlBestellungen werden für die weitere Analye ebenfalls klassiert. Die Klassierung sowie absolute und relative Häufigkeiten der beiden Merkmale finden sich in Tabelle 4.

Tabelle 4: Absolute und relative Häufigkeiten von Alter (klassiert) und Anzahl der Bestellungen in den letzten 12 Monaten (klassiert)

| | | n | % |
|---|---|---|---|
| Alter<br>Basis: Käufer | [18,35] | 66.103 | 5,5 |
| | (35,45] | 159.908 | 13,3 |
| | (45,55] | 431.178 | 35,8 |
| | (55,65] | 354.905 | 29,5 |
| | (65,75] | 142.855 | 11,9 |
| | 76 + | 48.275 | 4,0 |
| | Gesamt | 1.203.224 | 100,0 |
| Anzahl Bestellungen<br>Basis: Käufe | 1 | 640.762 | 28,1 |
| | 2 | 469.421 | 20,6 |
| | 3 | 338.277 | 14,8 |
| | 4-5 | 397.260 | 17,4 |
| | 6-8 | 245.248 | 10,8 |
| | 9 + | 187.049 | 8,2 |
| | Gesamt | 2.278.017 | 100,0 |

Das Merkmal Premium beschreibt, ob ein Kauf mit Premiumstatus des Kunden getätigt wurde oder nicht. Dabei kann sich der Premiumstatus eines Kunden im Laufe des Jahres ändern. Der Status ändert sich bei 15,5 % der Kunden, 84,5 % der Kunden behalten ihren Status im Laufe des Jahres. Da die Analyse monatsweise erfolgt und ein Kunde nur einmal pro Monat im Datensatz enthalten ist, ergeben sich durch den Statuswechsel keine Plausibilitätsprobleme. 933.319 (41,0 %) der Käufe erfolgen mit

Premiumstatus, 1.344.698 (59,0 %) ohne.

Der Kanal der ersten Bestellung bzw. des Erstkontakts mit dem Kunden ist durch das Merkmal Anlauf definiert. Bei 565.910 (47.0%) Kunden erfolgte der Anlauf über Katalog, bei 333.500 (27,7 %) online und bei 303.814 (25,2%) über sonstige Kanäle.

Grafik 7 zeigt statistisch signifikante Unterschiede (ANOVA, p-Wert $< 0.001$) im Alter hinsichtlich des Anlaufkanals Online mit Katalog bzw. Sonstige. Kunden mit dem Erstkontakt Online sind mit 48,9 ($\pm$ 11,8) Jahren jünger als Kunden mit Anlaufkanal Katalog (56,8 $\pm$ 9,6) oder Anlaufkanal Sonstige (56,3 $\pm$ 12,0).

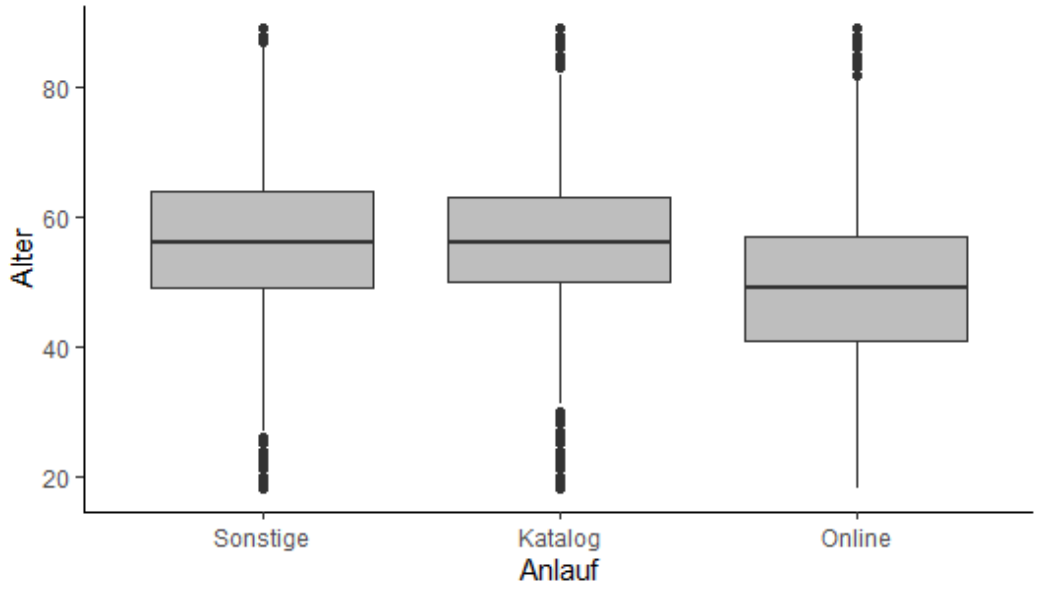



Abbildung 7: Boxplots von Alter für verschiedene Anlaufkanäle (Basis: Kunde)

Der Newsletter Status (NLStatus) beschreibt welche Art von Newsletter ein Kunde zum Zeitpunkt des Datenexports abonniert hat. Das Verhalten des Kunden bei Erhalt eines Newsletters in den letzten 12 Monaten vor einem Kauf beschreibt das Merkmal nl_verhalten. Da ein Kunde in diesem Zeitraum mehrmals einen Newsletter erhält, ist hier das häufigste Verhalten bei Erhalt gespeichert. Tabelle 5 beschreibt den Status und das Verhalten bei dem Empfang von Newslettern.

Die Merkmale Warengruppe 1 bis Warengruppe 6 beschreiben ob ein Kunde in den vergangenen 12 Monaten einen Kauf in der entsprechenden Warengruppe getätigt hat. Grafik 8 zeigt die Häufigkeiten der Einkäufe in den verschiedenen Warengruppen. Die beliebtesten Warengruppen sind Warengruppe 1 und 2. Am wenigsten wird in Warengruppe 3 gekauft.

Ein Kunde kann online, telefonisch oder über das Bestellformular des Katalogs eine Bestellung aufgeben, sowie Produkte kaufen, die online oder im Katalog angeboten werden. Die Merkmale KanalOnline, KanalPrint, SortimentOnline und SortimentPrint beschreiben, ob ein Kunde mindestens einmalig in den vergangenen 12 Monaten über den entsprechenden Kanal bzw. Artikel des entsprechenden Sortiments gekauft hat.

Tabelle 5: Absolute und relative Häufigkeiten von Newsletter Status und Verhalten

| | | n | % |
|---|---|---|---|
| Newsletter Status<br>Basis: Käufer | NL & DL | 338.586 | 28,1 |
| | NL/kein NL | 339.934 | 28,3 |
| | DL | 524.704 | 43,6 |
| | Gesamt | 1.203.224 | 100,0 |
| Newsletter Verhalten<br>Basis: Käufe | angeklickt | 680.527 | 29,9 |
| | kein NL erhalten | 1.331.978 | 58,5 |
| | nicht geöffnet | 155.507 | 6,8 |
| | geöffnet | 110.005 | 4,8 |
| | Gesamt | 2.278.017 | 100,0 |

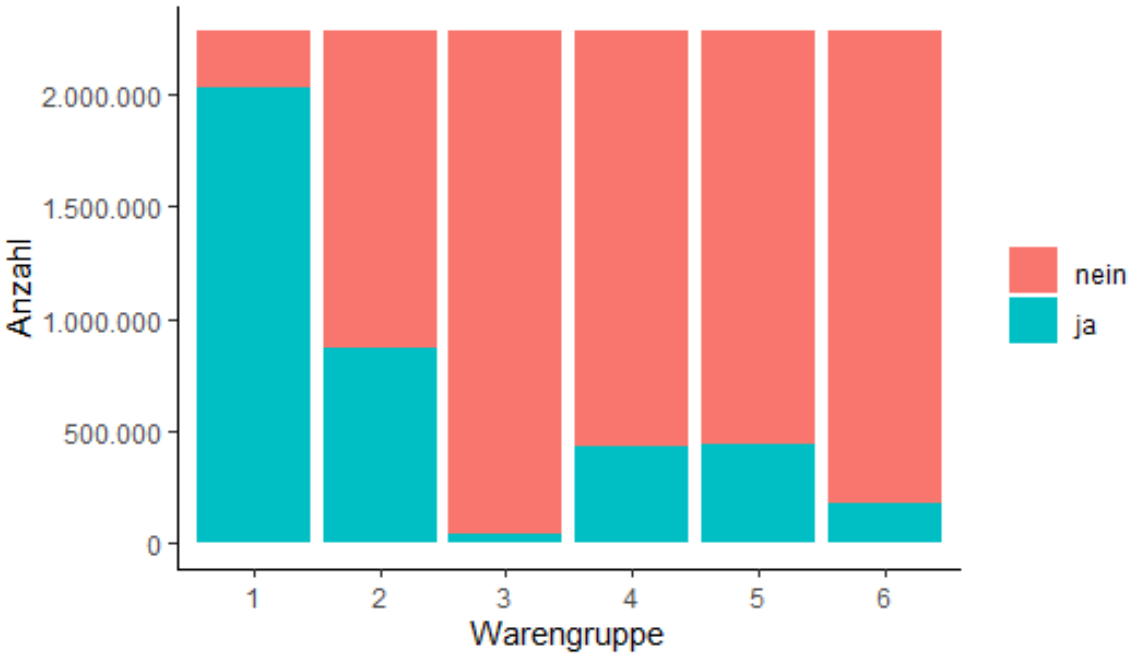



Abbildung 8: Einkäufe in den letzten 12 Monaten in den Warengruppen 1 bis 6 (Basis: Kauf)

Zusammenfassend werden die Annahmen für die weitere Analyse getroffen:

- Einkäufe mit $\texttt{Umsatz} = 0$ werden aus der Analyse ausgeschlossen.
- Augrund der Saisonalität und der Datenstruktur werden für die Monate getrennte Modelle gerechnet.
- Ein Geschäftsjahr wird betrachtet. Die Anzahl der Tage bis zum Folgekauf, wenn nicht innerhalb des Geschäftsjahres gekauft wird, wird auf 366 Tage gesetzt und der Fall als zensiert betrachtet.
- Die Kovariaten $\texttt{Alter}$, $\texttt{Umsatz}$, $\texttt{AnzahlBestellungen}$ für die Cox-Regression werden klassiert. Dies erleichtert die Interpretierbarkeit und die Überprüfung der Voraussetzungen. Zudem können Kaplan-Meier-Schätzer verwendet werden.

## 4.3 Kaplan-Meier-Schätzer und Log-Rank-Tests

Um zu verstehen, welche Merkmale die Dauer bis zum Folgekauf beeinflussen, werden im Folgenden Kaplan-Meier-Schätzer betrachtet, sowie Log-Rank-Tests gerechnet. Die Berechnung und Darstellung der Kurven erfolgt mithilfe der R-Pakete `survival` und `survminer`. Listing 2 zeigt den Quellcode für die Berechnung der Kaplan-Meier-Kurven aus Abbildung 9 und die Berechnung des medianen Kaufens aus Tabelle 6 für 12 Monate

```
library(survival)
library(survminer)

# Kaplan-Meier-Kurve berechnen
fit <- survfit(Surv(tage365, status) ~ monat, data = df)

# Plot Kaplan-Meier Kurve getrennt nach Monaten
ggsurvplot(fit,
           data = df,
           title = "Kaplan-Meier-Schaetzer",
           ylab = "Wahrscheinlichkeit S(t)",
           xlab = "Zeit (Tage)",
           xlim = c(0, 366),
           surv.median.line = "v",   # vertikale Linien medianes Kaufen
           size = 0.1)

# Tabelle medianes Kaufen
print(fit)
```

Listing 1: Quellcode Kaplan-Meier-Kurve und medianes Kaufen

Abbildung 9 zeigt die Kaplan-Meier-Schätzer für die Monate Januar bis Dezember.

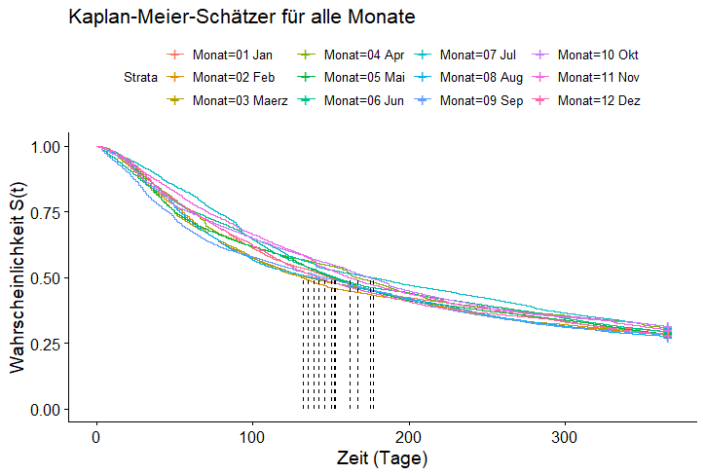



Abbildung 9: Kaplan-Meier-Schätzer für alle Monate

Die einzelnen Kurven sind kaum zu unterscheiden und kreuzen sich mehrfach. Die gestrichelten, vertikalen Linien beschreiben das mediane Kaufen, d. h. nach welcher

Anzahl an Tagen 50 % der Kunden einen Folgekauf getätigt haben. Zu erkennen ist, dass sich das mediane Kaufen der Monate unterscheidet.

Tabelle 6: Medianes Kaufen für 12 Monate

| Monat | n | Events | Median | 95% KI |
|---|---|---|---|---|
| Jan | 162.440 | 114.638 | 150 | [148, 151] |
| Feb | 143.451 | 102.934 | 132 | [131, 133] |
| März | 184.701 | 132.060 | 135 | [133, 137] |
| April | 264.357 | 182.900 | 167 | [165, 168] |
| Mai | 184.423 | 131.809 | 152 | [151, 153] |
| Jun | 211.090 | 149.147 | 153 | [152, 154] |
| Jul | 205.932 | 141.914 | 177 | [175, 179] |
| Aug | 131.096 | 94.841 | 139 | [137, 141] |
| Sep | 165.929 | 119.031 | 146 | [144, 147] |
| Okt | 266.068 | 182.300 | 175 | [174, 176] |
| Nov | 225.099 | 157.478 | 162 | [161, 163] |
| Dez | 133.431 | 95.433 | 142 | [141, 143] |

Das mediane Kaufen ist mit Angabe der 95 % Konfidenzintervalle in Tabelle 6 angegeben. Es zeigt sich, dass im Juli die Dauer, bis 50 % der Kunden gekauft haben, mit knapp 6 Monaten bzw. 177 Tagen am längsten ist. Hingegen beträgt das mediane Kaufen im Februar nur 132 Tage. Die Berechnung des Log-Rank-Tests zum Vergleich der Monate ist nicht möglich, dass die Stichproben nicht unabhängig voneinander sind (ein Kunde kann in mehreren Monaten kaufen).

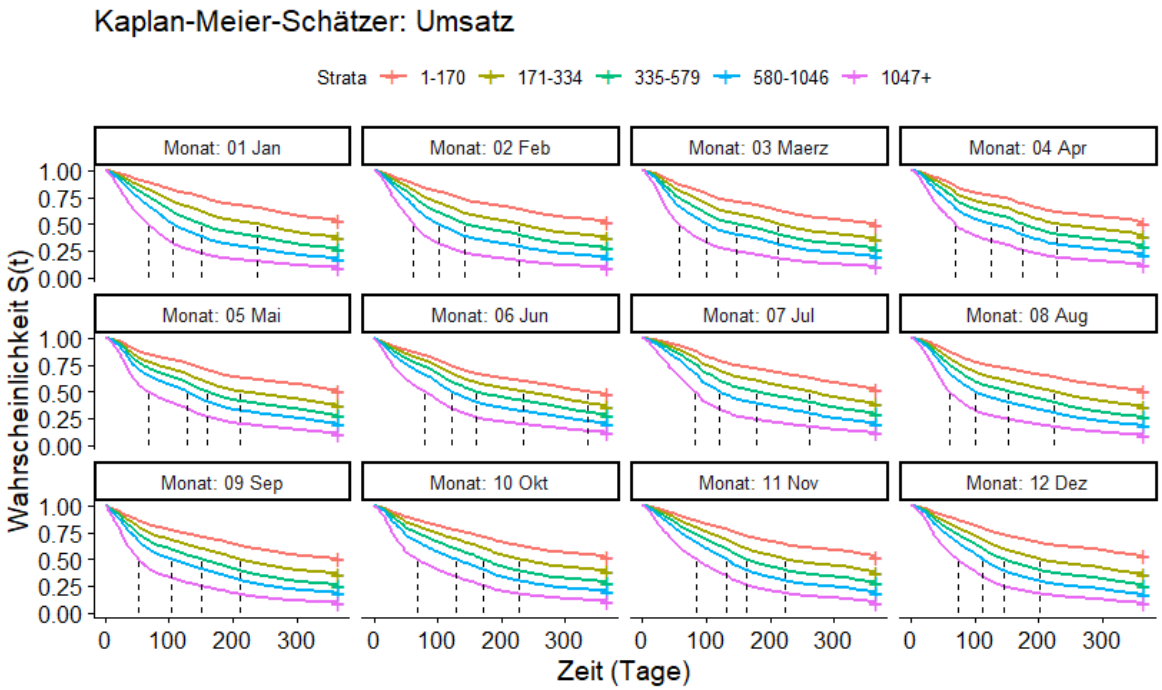



Abbildung 10: Kaplan-Meier-Schätzer getrennt nach Umsatz für 12 Monate

In Abbildung 10 werden die Zeiten bis zum Folgekauf getrennt für die erzielten Umsätze in den letzten 12 Monaten verglichen. Dabei zeigen sich für alle Monate die gleichen Effekte: Je mehr Umsatz in den vorherigen 12 Monaten erzielt wurde, desto schneller erfolgt der Folgekauf. Die Kunden mit Umsätzen von 1047+ Euro tätigen am schnellsten einen Folgekauf. Kunden mit Umsatz von 1 bis 170 Euro benötigen am meisten Zeit bis zum Folgekauf.

Da sich auch für die weiteren Merkmale des Datensatzes bei Betrachtung der Kaplan-Meier-Kurven die gleichen Effekte pro Monat zeigen, werden aus Gründen der Übersichtlichkeit im Folgenden nur die Ergebnisse des Januars diskutiert.

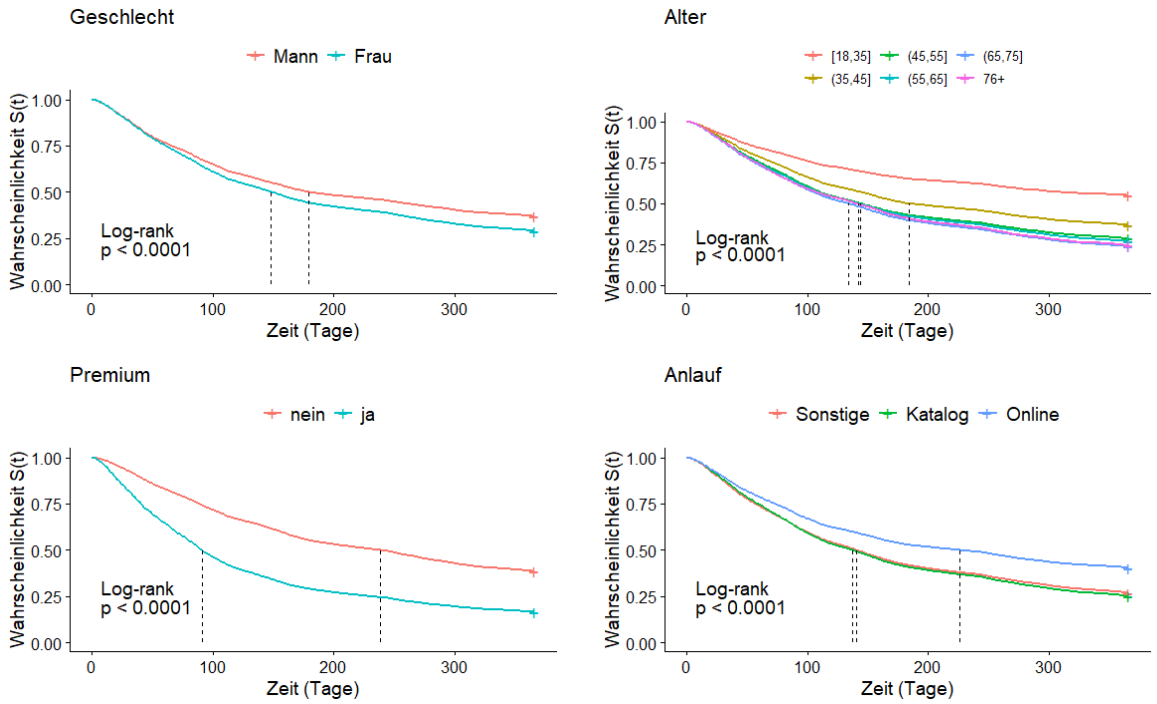



Abbildung 11: Kaplan-Meier-Schätzer getrennt nach Geschlecht, Alter, Premium und Anlauf für Januar

Für Männer und Frauen ist der Zeitraum bis zum Kauf unterschiedlich lang. Das mediane Kaufen liegt bei Männern bei 179 (95% KI: [172,191]) Tagen, bei Frauen hingegen bei 148 (95% KI: [147, 150]) Tagen. Die zugehörigen Kaplan-Meier-Kurven sind in Abbildung 11 dargestellt.

Bei Premiumkunden und Nicht-Premium-Kunden zeigt sich ein großer Unterschied im medianen Kaufen. Premiumkunden kaufen im Schnitt 147 Tage früher als Nicht-Premium-Kunden.

Bei verschiedenen Altersgruppen zeigt sich ein unterschiedliches Kaufverhalten: Für Kunden unter 35 Jahren liegt der Kaplan-Meier-Schätzer in der gesamten Beobachtungszeit von einem Jahr bei über 50%. Das mediane Kaufen ist nicht zu bestimmen, da weniger als die Hälfte der unter 35 Jährigen einen Folgekauf getätigt haben. Hingegen hat nach 184 (95% KI: [179 ,194]) Tagen die Hälfte der 36 bis 45-Jährigen einen Fol-

gekauf getätigt. Bei den über 45-Jährigen lassen sich die Kaplan-Meier-Kurven kaum mehr unterscheiden (siehe Abbildung 11). Fasst man diese Altergruppen zusammen zeigt sich, dass nach 141 (95% KI: [140 , 142]) Tagen über die Hälfte von ihnen einen Folgekauf getätigt haben.

Tabelle 7: Medianes Kaufen für Anlauf für Januar

| | n | Events | Median | 95% KI |
|---|---|---|---|---|
| Sonstige | 42.246 | 30.848 | 141 | [139, 143] |
| Katalog | 80.280 | 60.011 | 137 | [135, 139] |
| Online | 39.914 | 23.779 | 226 | [218, 233] |

Für den Anlauf der Kunden zeigen sich Unterschiede zwischen den Kaplan-Meier-Kurven von Katalog und Online gegenüber der Kunden mit Erstkontakt Online. Sowohl der globale Log-Rank-Test (p-Wert $< 0.001$) als auch die paarweisen Vergleiche, siehe Tabelle 8 mit Bonferroni-Holm, liefern statistsch signfikante Unterschiede der Zeit bis zum Folgekauf zwischen den 3 Kurven.

Tabelle 8: Paarweise Vergleiche: Log-Rank-Tests für Anlauf mit Bonferroni-Holm-Korrektur

| | p-Wert | |
|---|---|---|
| | Sonstige | Katalog |
| Katalog | $< 0.001$ | |
| Online | $< 0.001$ | $< 0.001$ |

```
fit <- survival::survfit(Surv(tage365, status) ~ anlauf,
          data = df_januar)

ggsurvplot(fit, ...,
          pval = TRUE, pval.method = TRUE) # Log-Rank Test

# Paarweise Tests
pairwise_survdiff(Surv(tage365, status) ~ anlauf,
          data = df_januar,
          p.adjust.method = "BH") # Bonferroni-Holm-Korrektur
```

Listing 2: Quellcode paarweise Vergleiche mit Log-Rank-Tests und Bonferroni-Holm-Korrektur

Ein praktisch relevanter Unterschied ergibt sich für die Zeit bis zum Kauf zwischen Katalog und Sonstige nicht. Das mediane Kaufen unterscheidet sich hier nur um 5 Tage (siehe Tabelle 7).

Abbildung 12 zeigt den Einfluss der Anzahl an Bestellungen in den vergangenen 12 Monaten auf die Zeit bis zu Folgekauf. Je mehr Bestellungen ein Kunde in der Vergangenheit getätigt hat, desto schneller kauft er wieder. Nach 365 Tagen haben nur 44% der Kunden, die in der Vergangenheit nur eine Bestellung getätigt haben, einen Folgekauf gemacht. Bei Kunden mit 9+ Bestellungen in der Vergangenheit liegt die Folgekauf-Wahrscheinlichkeit nach 365 Tagen hingegen bei 97%.

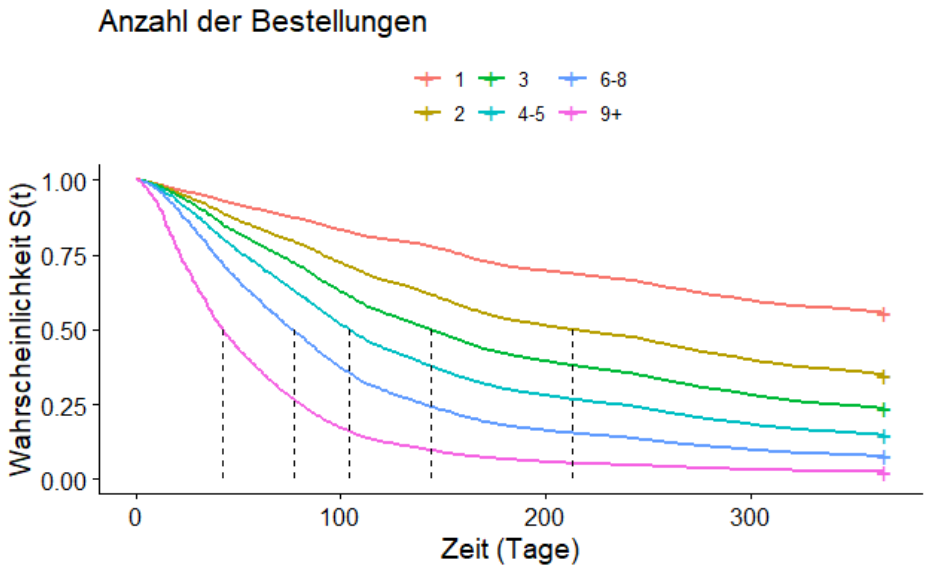



Abbildung 12: Kaplan-Meier-Schätzer getrennt nach der Anzahl an Bestellungen in den letzten 12 Monaten für Januar

Nach 110 (95% KI: [109, 111]) Tagen hat die Hälfte der Kunden mit Newsletter-Status NL & DL einen Folgekauf getätigt. Bei den Kunden mit NL/ kein NL bzw. DL liegt diese Zahl bei 168 (95% KI: [165, 170]) bzw. 166 (95% KI: [164, 168]) Tagen.

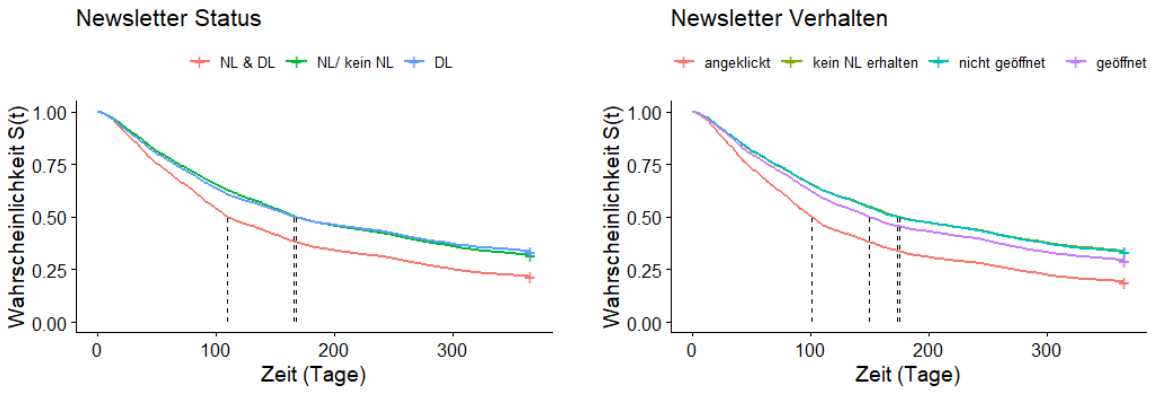



Abbildung 13: Kaplan-Meier-Schätzer getrennt nach Newsletter Verhalten und Newsletter Status für Januar

Die Kaplan-Meier-Kurven liegen aufeinander, die Kunden zeigen ein ähnliches Kaufverhalten. Die Kaplan-Meier-Kurven zum Newsletter Status und zum Newsletter Verhalten sind in Abbildung 13 dargestellt. Kunden, die in der Vergangenheit Newsletter

angeklickt haben, kaufen schneller erneut als Kunden, die einen Newsletter nur geöffnet haben. Ob ein Kunde einen Newsletter nicht geöffnet oder gar nicht erst erhalten hat, spielt hingegen keine Rolle bezüglich der Zeit bis zum Folgekauf.

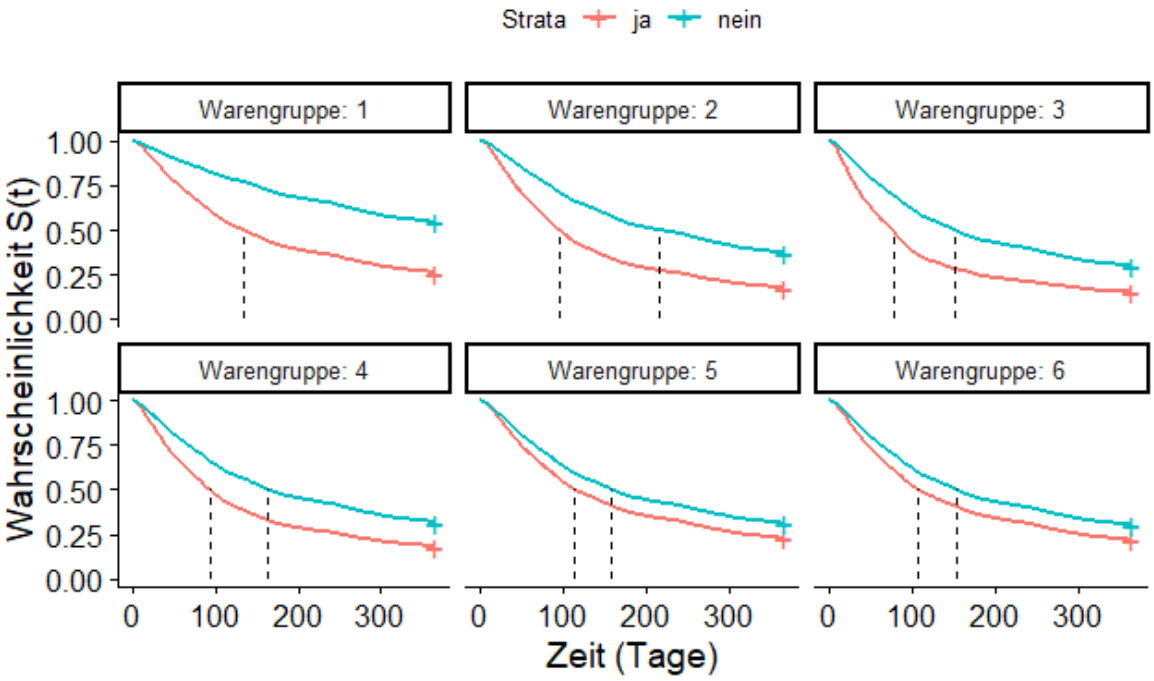



Abbildung 14: Kaplan-Meier-Schätzer für die Warengruppen 1 bis 6 für Januar

Kauften Kunden in der Vergangenheit Waren der Warengruppen 1 bis 6, so unterscheiden sich ihre Kaplan-Meier Kurven. Die Kurven sind in Abbildung 14, das mediane Kaufen in Tabelle 9.

Tabelle 9: Medianes Kaufen in den Warengruppen 1 bis 6 für Januar

| Warengruppe | | n | Events | Median | 95% KI | p-Wert |
|---|---|---|---|---|---|---|
| 1 | nein | 19.370 | 8.836 | - | - | $< 0.001$ |
| | ja | 143.070 | 105.802 | 134 | [133, 135] | |
| 2 | nein | 98.981 | 62.316 | 216 | [212, 219] | $< 0.001$ |
| | ja | 63.459 | 52.322 | 95 | [94, 96] | |
| 3 | nein | 159.731 | 112.344 | 152 | [150, 153] | $< 0.001$ |
| | ja | 2.709 | 2.294 | 77 | [73, 80] | |
| 4 | nein | 134.784 | 91.989 | 163 | [162, 165] | $< 0.001$ |
| | ja | 27.656 | 22.649 | 93 | [91, 94] | |
| 5 | nein | 130.124 | 89.646 | 158 | [157, 160] | $< 0.001$ |
| | ja | 32.316 | 24.992 | 113 | [111, 115] | |
| 6 | nein | 148.708 | 10.3875 | 154 | [153, 155] | $< 0.001$ |
| | ja | 13.732 | 10.763 | 107 | [105, 110] | |

Die Kaplan-Meier-Kurven, ob ein Kunde in der Vergangenheit im Online Sortiment oder über online bestellt hat, unterscheiden sich nicht, wie in Abbildung 15 zu sehen.

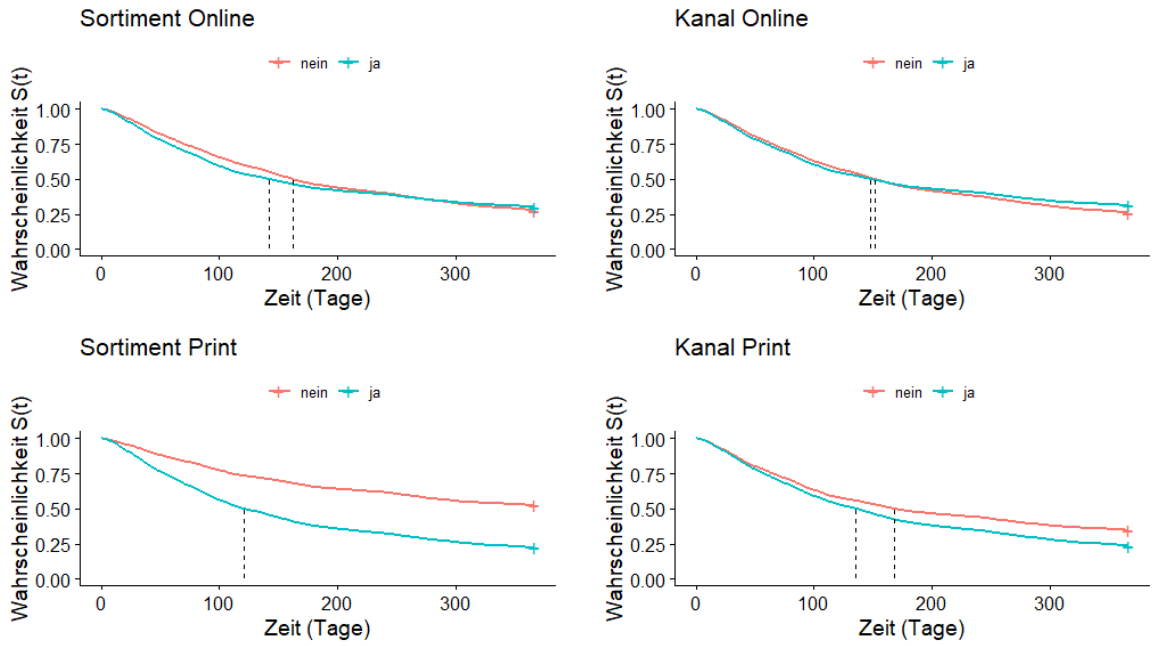



Abbildung 15: Kaplan-Meier-Schätzer getrennt nach Sortiment Online, Kanal Online, Sortiment Print und Kanal Print für Januar

Hat ein Kunde ein Produkt des Print Sortiments gekauft, unterscheidet er sich im Kaufverhalten von einem Kunden ohne entsprechenden Kauf. Nach 121 Tagen haben über die Hälfte der Kunden mit Kauf eines Print-Sortiments-Produkts einen Folgekauf getätigt. Nach 365 Tagen haben nur 48% der Kunden ohne Kauf eines Print-Sortiments-Produkts einen Folgekauf getätigt. Das mediane Kaufen ist für diese Kundengruppe also nicht zu bestimmen. Tabelle 10 zeigt das mediane Kaufen und die Konfidenzintervalle. Ebenfalls in Abbildung 15 zu sehen ist, dass sich die Kaplan-Meier-Kurven unterscheiden je nachdem, ob ein Kunde in der Vergangenheit über den Kanal Print bestellt hat oder nicht.

Tabelle 10: Medianes Kaufen für Sortiment Online, Kanal Online, Sortiment Print und Kanal Print für Januar

| | | n | Events | Median | 95% KI |
|---|---|---|---|---|---|
| Sortiment Print | nein | 38761 | 18463 | - | - |
| | ja | 123679 | 96175 | 121 | [120, 122] |
| Kanal Print | nein | 84187 | 54921 | 168 | [165, 170] |
| | ja | 78253 | 59717 | 135 | [134, 137] |
| Sortiment Online | nein | 50871 | 36688 | 162 | [160,163] |
| | ja | 143.070 | 105.802 | 142 | [140, 143] |
| Kanal Online | nein | 56987 | 42099 | 152 | [150, 153] |
| | ja | 105453 | 72539 | 148 | [147, 150] |

## 4.4 Cox-Regression und Scoring

### 4.4.1 Vorbereitende Schritte

Die gewonnenen Erkenntnisse der bivariaten Analyse in Kapitel 4.3 werden für Datenaufbereitungsschritte für die Cox-Regression genutzt. Bei Betrachtung der Kaplan-Meier-Kurven zeigte sich, dass einzelne Merkmale bzw. Ausprägungen nicht geeignet sind, aktivere von weniger aktiven Kunden zu unterscheiden.

Als vorbereitende Schritte werden folgende Variablen umcodiert, indem Ausprägungen zusammengefasst werden:

- Alter: [18, 35], (35,45], 46+
- Anlauf: Online, Katalog/Sonstige
- Newsletter Status: NL & DL, NL/kein NL, DL
- Newsletter Verhalten: angeklickt, geöffnet, kein NL

Die folgenden Merkmale werden aus der Analyse ausgeschlossen, da sich die Kaplan-Meier-Kurven schneiden und somit einen Hinweis liefern, dass die Proportional Hazard Annahme verletzt ist. In Anhang E sind die Kaplan-Meier-Kurven für alle 12 Monate abgebildet. Es zeigt sich, dass die folgenden Merkmale in keinem der Monate für die weitere Analyse geeignet sind:

- Sortiment Online
- Katalog Online.

Als weiterer Schritt werden die Daten pro Monat in 80% Trainingsdaten und 20 % Testdaten aufgeteilt.

### 4.4.2 Schrittweise Variablenselektion

Mithilfe der R Funktion `stepAIC()` aus dem Paket MASS werden die Merkmale schrittweise für Modelle der 12 Monate ausgewählt. Als Informationskriterium wird das Akaike-Informationskriterium für die schrittweise Modellbildung gewählt. [KM06] beschreibt das AIC zur Modell-Bildung im Rahmen der Survival-Analyse mit

$$AIC = -2 \text{ Log } L + 2p, \tag{24}$$

mit $p$ ist die Anzahl der Parameter im Modell und $L$ ist die Likelihood Funktion. Das AIC Kriterium betrachtet also die Anpassungsgüte mithilfe der Likelihood Funktion und enthält zusätzlich einen Strafterm um zu komplexe Modelle zu bestrafen. Ein Model mit dem kleineren AIC Wert ist zu bevorzugen.

Tabelle 11 zeigt Ergebnisse der 12 Modelle für 12 Monate der schrittweisen Variablenselektion nach dem AIC-Kriterium. Kein Merkmal wird in allen 12 Monaten aus den Modellen entfernt. Geschlecht ist im Juni nicht in dem Model enthalten. Unregelmäßigkeiten lassen sich ebenfalls für die Warengruppen 5 und 6, sowie Kanal Print feststellen. Alle weiteren Merkmale werden für alle Monate als relevant für die Modelle betrachtet.
Listing 3 zeigt den Quellcode für die schrittweise Variablenselektion.

```
library(dplyr)

# Aufteilen von Trainings-und Testdaten
train <- as.data.frame(df %>% group_by(Monat) %>% sample_frac(0.8))
test  <- as.data.frame(anti_join(df, train, by = c('id', 'Monat')))

# Funktion zur schrittweisen Variablenselektion der Cox-Regression
# return: Formel mit Parametern des Endmodels
step_cox <- function(mymonat){

          # Auswahl des Monats
          train_monat <- train[train$Monat == mymonat, ]

          # Modell mit allen Variablen
          full_model <- coxph(Surv(tage365, status) ~ geschlecht +
                    alter_kl +
                    premium +
                    umsatz_kl +
                    anz_bestellungen_kl +
                    anlauf +
                    nl_status + nl_verhalten +
                    wg1 + wg2 + wg3 + wg4 + wg5 + wg6 +
                    kanal_print +
                    sortiment_print,
                    data = train_monat)
          # Schrittweise
          step_model <- stepAIC(full_model,
                    direction = "both",
                    trace = FALSE)
          return(toString(step_model$formula))
          }

monate <- c('01 Jan', '02 Feb','03 Maerz', '04 Apr' , '05 Mai', '06 Jun',
          '07 Jul', '08 Aug', '09 Sep', '10 Okt', '11 Nov', '12 Dez')
lapply(monate, step_cox) # Variablenselektion pro Monat
```

Listing 3: Definition von Test- und Trainingsdaten und schrittweise Variablenselektion getrennt für 12 Monate für Cox-Regressionsmodelle

Aus pragmatischen Gründen[3] werden für alle Modelle die gleichen Einflussparameter selektiert. Da die Cox-Regression zur Berechnung von Scores und der Entwicklung von Kundenklassen und nicht zur inhaltlichen Interpretation genutzt wird, verbleiben die Merkmale für alle Monate im Modell. Durch die geschätzten $\beta$ Parameter nahe 0 in diesen Fällen verändert sich der Score nicht bzw. nur minimal. Die Parameterschätzer für Januar finden sich in Kapitel 4.4.5.

Tabelle 11: Modelle der 12 Monate nach schrittweiser Variablenselektion anhand des AIC-Kriteriums

| | Monat | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Geschlecht | • | • | • | • | • | | • | • | • | • | • | • |
| Alter | • | • | • | • | • | • | • | • | • | • | • | • |
| Premium | • | • | • | • | • | • | • | • | • | • | • | • |
| Anlauf | • | • | • | • | • | • | • | • | • | • | • | • |
| Umsatz | • | • | • | • | • | • | • | • | • | • | • | • |
| Anzahl Bestellungen | • | • | • | • | • | • | • | • | • | • | • | • |
| NL Status | • | • | • | • | • | • | • | • | • | • | • | • |
| NL Verhalten | • | • | • | • | • | • | • | • | • | • | • | • |
| WG 1 | • | • | • | • | • | • | • | • | • | • | • | • |
| WG 2 | • | • | • | • | • | • | • | • | • | • | • | • |
| WG 3 | • | • | • | • | • | • | • | • | • | • | • | • |
| WG 4 | • | • | • | • | • | • | • | • | • | • | • | • |
| WG 5 | | • | • | • | | • | • | • | • | • | | |
| WG 6 | | • | | | • | | • | • | | | | • |
| Sortiment Print | • | • | • | • | • | • | • | • | • | • | • | • |
| Kanal Print | • | • | • | • | | • | • | • | | • | | • |

[3]Die Modelle werden künftig in einem Produktivsystem mit SQL untergebracht. Es bietet sich an, für alle 12 Modelle eines Jahres die gleichen Merkmale auszuwählen und nur die $\beta$-Gewichte anzupassen.

### 4.4.3 Ergebnisse der Cox-Regression

Die Darstellung der Ergebnisse des Cox-Regressionsmodells erfolgt mittels Wald-Diagramm der Funktion `ggforest()` des Paktets `survminer`. In Abbildung 16 sind die Ergebnisse für das Cox-Regressionsmodell für Januar dargestellt und werden im Folgenden erläutert. Die Ergebnisse der Modelle Februar bis Dezember sind in Anhang B enthalten. Die Berechnung für Januar zeigt das Listing 4.

```
# Cox-Regression mit Rueckgabe des Wald-Diagramms der
    Parameterschaetzungen
mycox <- function(mymonat){

          # Auswahl des Monats
          train_monat <- train[train$Monat == mymonat, ]
          test_monat <- test[test$Monat == mymonat, ]

          # Cox Modell rechnen
          fit <- coxph(Surv(tage365, status) ~ geschlecht +
                    alter_kl +
                    premium +
                    umsatz_kl +
                    anz_bestellungen_kl +
                    anlauf +
                    nl_status + nl_verhalten +
                    wg1 + wg2 + wg3 + wg4 + wg5 + wg6 +
                    kanal_print +
                    sortiment_print,
                    data = train_monat)

          forest <- ggforest(fit,
                    data = train_monat,
                    fontsize = 1.7) # Wald-Diagramm

          return(forest)
}

mycox("01 Jan") # Funktionsaufruf Cox-Regression fuer Januar
```

Listing 4: Berechnung des Cox-Regressionsmodells für Januar und Darstellung der Ergebnisse mit Wald-Diagramm der Parameterschätzungen

Die Ergebnisse der bivariaten Analyse aus Kapitel 4.3 lassen sich im Grundsatz bestätigen. Im Januar zeigen die Merkmale mit Ausnahme von Warengruppe 5, Warengruppe 6 und Kanal Print einen statistisch signifikanten Einfluss auf das Kaufen.

Kunden mit 9+ Bestellungen in den letzten 12 Monaten (`anz_bestellungen_kl`) haben ein 5,19-fach so hohes Risiko für einen Folgekauf im Vergleich zu Kunden mit nur

einer Bestellung (Referenzkategorie). Bei Kunden mit 6-8 Bestellungen liegt das Hazard Ratio bei 3,07. Es zeigt sich: Je mehr Bestellungen ein Kunde in der Vergangenheit getätigt hat desto höher das Hazard Ratio.

Das Risiko für einen Folgekauf für Frauen ist 1,09 mal so hoch wie für Männer. Kunden mit einem Alter von 46 und mehr haben ein 40 % höheres Risiko für einen Folgekauf als Kunden der Alterklasse 18 bis 35. Bei Premiumkunden ist das Risiko 9% höher als bei Nicht-Premiumkunden.

Auch für den Umsatz gelten die bereits in Kapitel 4.3 gezeigten Effekte: Je höher der Umsatz, desto höher das Hazard Ratio. Im Anlauf der Kunden zeigt sich, wenn der Erstkontakt online war, das Risiko für einen Folgekauf gegenüber den Kunden mit Anlauf Katalog/ Sonstige um 0,87 vermindert war .

Bezieht ein Kunde keinen Newsletter oder hat den Newsletter nur geöffnet, nicht aber angeklickt, so ist das Risiko einen Folgekauf zu tätigen um 0,95 bzw. 0,92-fach kleiner als bei einem Kunden der den Newsletter auch angeklickt hat. Ein Kunde mit NL/kein NL/DL hat gegenüber einem Kunden mit Newsletter Status NL&DL ein um 0,9 geringeres Risiko.

Die Hazard Ratios der Warengruppen 1 bis 4 zeigen, dass Kunden mit einem Kauf in der Warengruppe ein erhöhtes Risiko für einen Folgekauf gegenüber den Kunden ohne einen entsprechenden Kauf haben. Zudem gilt: Hat ein Kunde einen Kauf aus dem Print Sortiment getätigt, so ist sein Risko 32% höher als bei Kunden ohne entsprechenden Kauf. Für Kunden mit Kauf über den Kanal Print lässt sich keine Aussage gegenüber den Kunden ohne Bestellung über diesen Kanal treffen. Die 1 ist im Konfidenzintervall des Hazard Ratios enthalten, es konnte also kein statistsch signifkanter Einfluss nachgewiesen werden (p-Wert = 0,141).

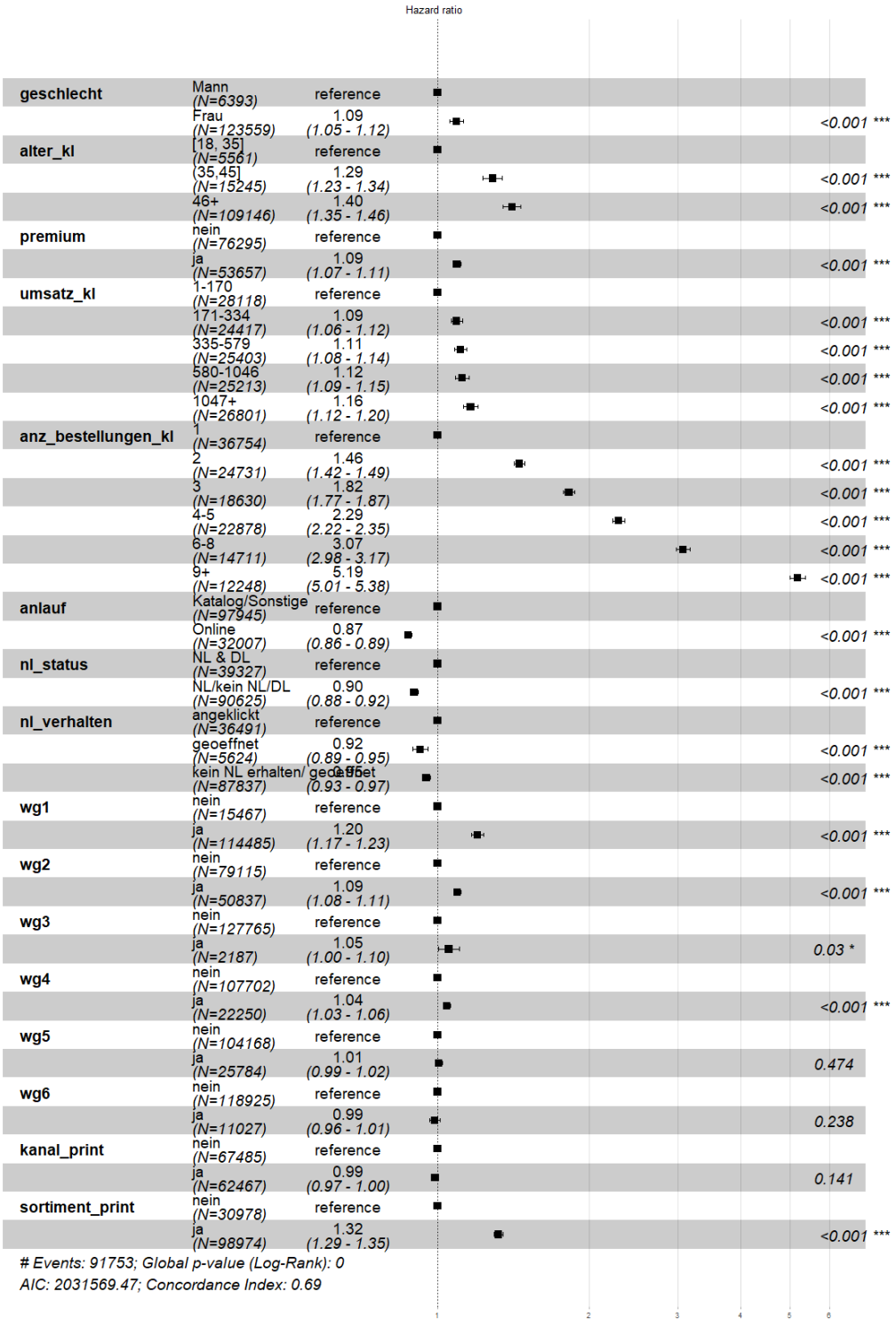



Abbildung 16: Wald-Diagramm mit Ergebnissen des Cox-Regressionsmodells für Januar

### 4.4.4 Überprüfung der Voraussetzung

Die Überprüfung der zentralen Annahme der proportionalen Hazards, also, dass das Hazard Ratio über die Zeit konstant bleibt, erfolgt mithilfe der Schoenfeld Residuen. Bereits die Kaplan-Meier-Kurven aus Kapitel 4.3 liefern einen ersten Hinweis, dass die Proportinale-Hazard-Annahme nicht verletzt ist. Die Vorgehensweise in R für Januar zur Berechnung und Erzeugung der Abbildungen der Schoenfeld Residuen zeigt Listing 5.

```
1 # Testen der proportional Hazards Annahme fuer ein Cox-Regression Modell-
      fit
2 zph <- cox.zph(fit)
3
4 plot(zph, resid= TRUE, col = "red", lwd = 3) # Plot Schoenfeld Residuen
```

Listing 5: Überprüfung der Proportional-Hazard-Annahme

Die Abbildungen sollen, falls die proportinale-Hazard-Annahme erfüllt ist, eine horizontale Linie bei 0 zeigen, da die Schoenfeld Residuen zeitunabhängig sind. Die statistschen Tests zur Überprüfung der Annahme nach [GT94] werden aufgrund der großen Stichprobengröße nicht betrachtet. Die Abbildungen 17, 18 und 19 zeigen die Graphen der skalierten Schoenfeld Residuen mit einer Spline Interpolation 4ten Grades (rote Linie).

Zu erkennen ist, dass alle Plots eine nahezu horizontale Linie um die 0 zeigen. Die proportinale Hazard-Annahme Januar Modells ist erfüllt. Da sich für die Monate Februar bis Dezember gleichartige Ergebnisse zeigen, wird auf die Darstellung von diesen verzichtet.

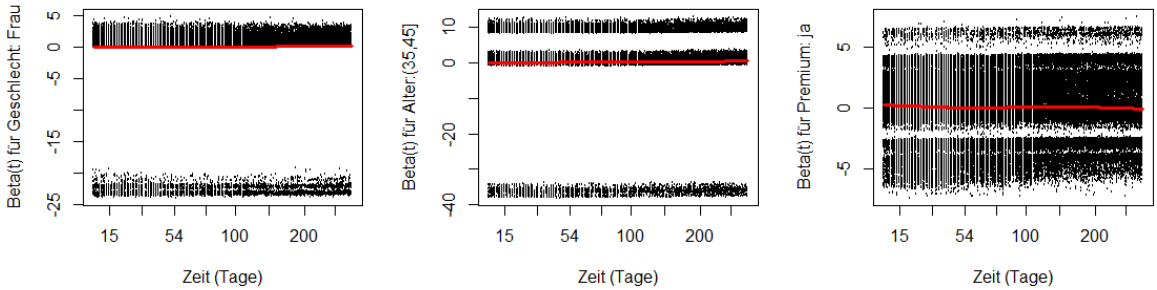



Abbildung 17: Schoenfeld-Residuen zur Überprüfung der Proportional-Hazard-Annahme (I)

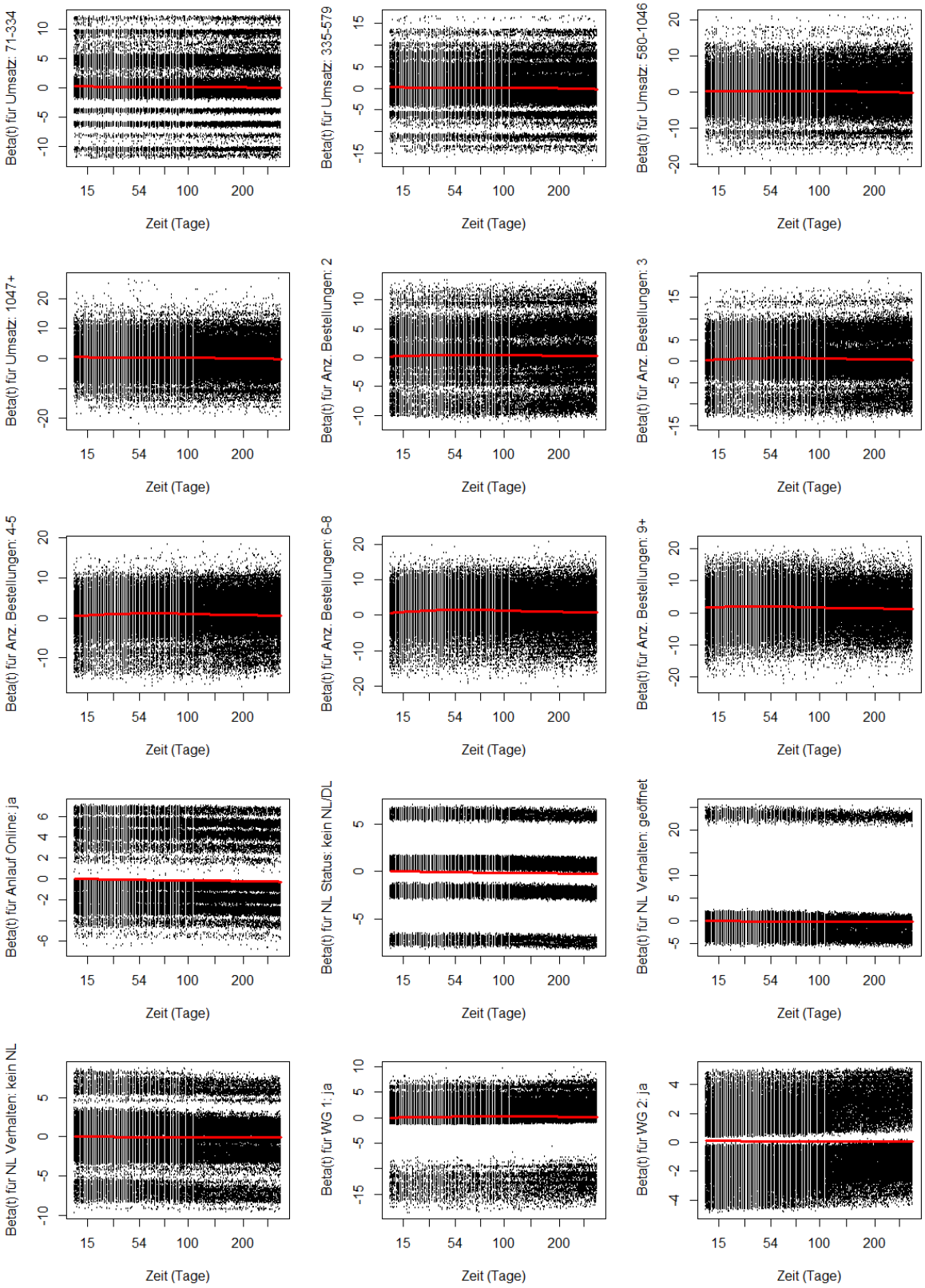



Abbildung 18: Schoenfeld-Residuen zur Überprüfung der Proportional-Hazard-Annahme (II)

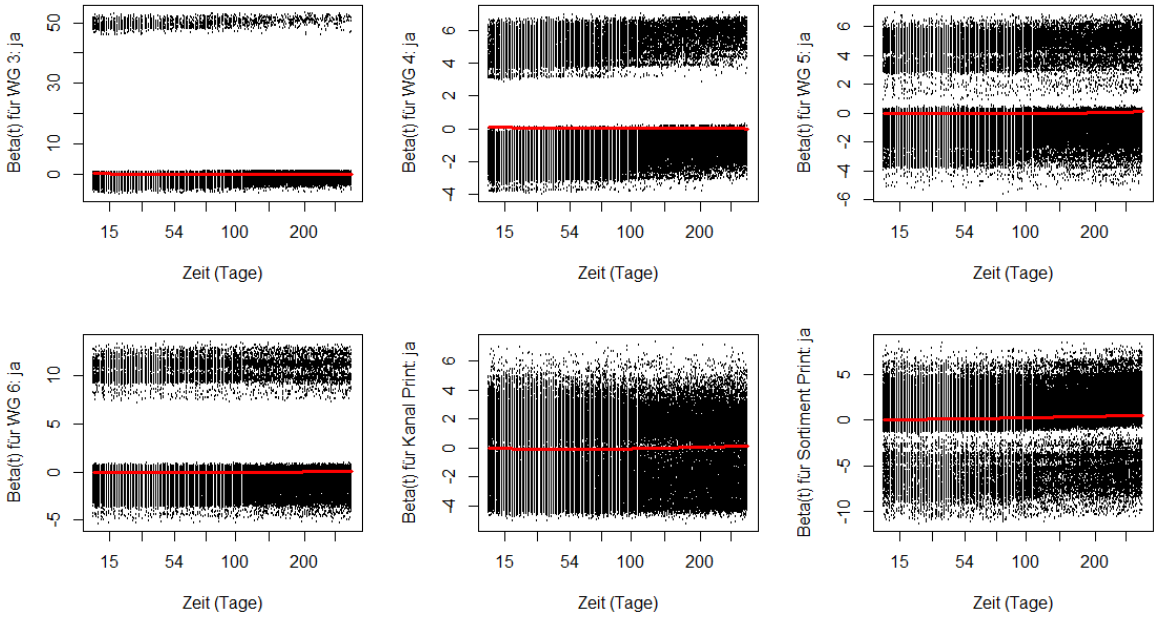



Abbildung 19: Schoenfeld-Residuen zur Überprüfung der Proportional-Hazard-Annahme (III)

### 4.4.5 Berechnung der Scoreklassen

Nun sollen auf Basis des Cox-Regressionsmodells 10 Scoreklassen entwickelt werden, die den Aktivitätsgrad der Kunden widerspiegeln. Dazu wird für jeden Kunden die lineare Vorhersage mithilfe von `predict(fit, type="lp")` (siehe: Listing 6) berechnet.

Im Cox-Regressionmodell entspricht dies dem folgenden Term:

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

mit m Einflussparameter, t Zeit, $\beta$ die zu schätzenden Regressionskoeffizienten.

```
# Auswahl des Monats
train_monat <- train[train$Monat == "01 Jan", ]

# Scoreklassen und Grenzen bilden
train_monat$score <- predict(fit, train_monat, type="lp") # lp = Lineare
    Vorhersage fuer coxph Objekt von Januar

# 10 Klassen bilden auf Basis der Quantile der Vorhersage
train_monat$kl <- ntile(train_monat$score, 10)

# Klassengrenzen von Score bestimmen
scores <- train_monat %>% group_by(kl) %>%
summarise(max = max(score), mw = mean(score))
klassengrenzen <- c(-Inf, scores$max[-10], +Inf)
klassengrenzen

fit_scores <- survfit(Surv(tage365, status) ~ kl,
data = train_monat)

ggsurvplot(fit_scores,
data = train_monat,
fun = "event") # Plot fuer Kaufwahrscheinlichkeit der Klassen

# Quantile der Klassen
quantile(fit_scores, probs = c(0.1 ,0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8),
conf.int = FALSE)
```

Listing 6: Berechnung der Scoreklassen auf Basis der Cox-Regression (Januar)

Zur inhaltlichen Interpretation wurden in Kapitel 4.4.3 die Hazard Ratios besprochen. Tabelle 12 zeigt hingegen die Parameterschätzer für $\beta$. Zwischen den Hazard Ratios und den Parameterschätzungen besteht die Beziehung $HR = \exp(\beta)$.

Für die 10 Scoreklassen werden nun die Vorhersagewerte auf Basis der Dezile Ihrer Verteilung zusammengefasst. Abbildung 20 zeigt die Verteilung des Scores mit der

Tabelle 12: Parameterschätzer des Cox-Regressionsmodells für Januar

| Merkmal | $\beta$ |
|---|---|
| geschlechtFrau | 0.08531 |
| alter_kl(35,45] | 0.25194 |
| alter_kl46+ | 0.33992 |
| premiumja | 0.08826 |
| umsatz_kl171-334 | 0.08768 |
| umsatz_kl335-579 | 0.10423 |
| umsatz_kl580-1046 | 0.11283 |
| umsatz_kl1047+ | 0.15057 |
| anz_bestellungen_kl2 | 0.37506 |
| anz_bestellungen_kl3 | 0.6009 |
| anz_bestellungen_kl4-5 | 0.82694 |
| anz_bestellungen_kl6-8 | 1.12218 |
| anz_bestellungen_kl9+ | 1.64684 |
| anlaufOnline | -0.13577 |
| nl_statusNL/keinNL/DL | -0.1071 |
| nl_verhaltengeoeffnet | -0.07944 |
| nl_verhaltenkeinNL | -0.05156 |
| wg1ja | 0.1816 |
| wg2ja | 0.09012 |
| wg3ja | 0.05164 |
| wg4ja | 0.04337 |
| wg5ja | 0.00608 |
| wg6ja | -0.01394 |
| kanal_printja | -0.01123 |
| sortiment_printja | 0.27664 |

Klasseneinteilung auf Basis der Dezile. Die Klassengrenzen der Vorhersage sind in Tabelle 13 dargestellt. Dabei bezeichnet Klasse 10 die „beste“ Kundengruppe und Klasse 1 die „schlechteste“ hinsichtlich der Zeitdauer bis zum Folgekauf.

Tabelle 13: Tabelle der Scoreklassen mit Klassengrenzen und Mittelwert

| Scoreklasse | Klassengrenzen | Mittelwert |
|---|---|---|
| 1 | (-Inf, -0.921] | -1.124 |
| 2 | (-0.921, -0.664] | -0.808 |
| 3 | (-0.664, -0.473] | -0.59 |
| 4 | (-0.473, -0.195] | -0.317 |
| 5 | (-0.195, 0.001] | -0.102 |
| 6 | (0.001, 0.192] | 0.093 |
| 7 | (0.192, 0.384] | 0.296 |
| 8 | (0.384, 0.607] | 0.486 |
| 9 | (0.607, 0.945] | 0.758 |
| 10 | (0.945, Inf] | 1.308 |

Anschließend soll nun für die 10 Kundenklassen geschätzt werden, wann ein Folgekauf erfolgt. Eine Vorhersage der Überlebenszeit analog zur linearen oder logistschen Funktion mit `predict(..., type = "response")` ist bei der Cox-Regression nicht möglich. Ein Vorteil der Cox-Regression als semi-parametrisches Verfahren ist, dass die Baseline Hazard Rate $h_0(t)$[4] eben nicht geschätzt werden muss. Die Funktion `basehaz()` aus dem R Paket `survival` berechnet nicht eine geschätzte Baseline Hazard Funktion, sondern die Überlebenskurve für die Mittelwerte der Kovariaten (bzw. der Referenzkategorien bei kategorialen Merkmalen).

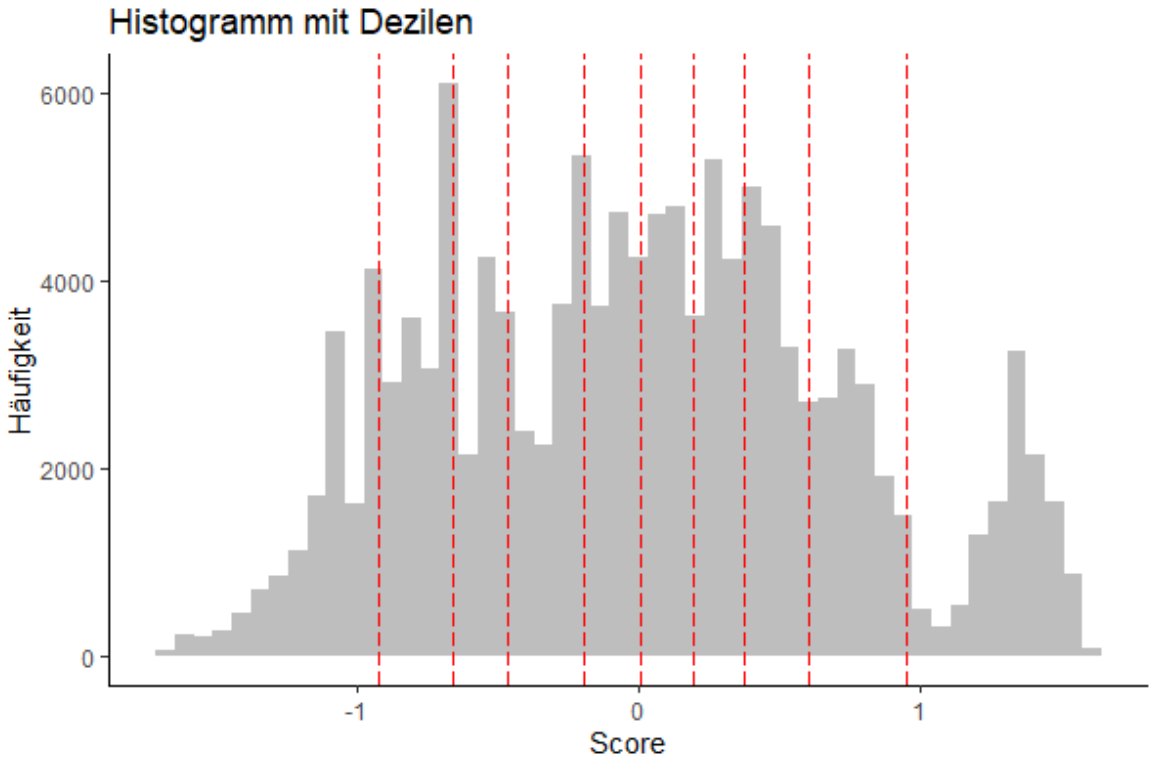



Abbildung 20: Histogramm des Scores mit Angabe der Dezile

Für die Schätzung der Kauffunktionen der 10 Scoreklassen werden stattdessen Kaplan-Meier-Schätzer verwendet. Dafür wird für jede der 10 Scoreklassen der Mittelwert des Scores bestimmt und für diesen die Kauffunktion berechnet.

Abbildung 21 zeigt die Kauffunktion für die 10 Scoreklassen. Die Kurven schneiden sich nicht und sind eindeutig zu unterscheiden. Für Scoreklasse 1 ergibt sich ein flacher Verlauf, während die Kaufwahrscheinlichkeit für Klasse 10 schnell ansteigt.

In Scoreklasse 10 haben bereits nach 44 Tagen die Hälfte der Kunden einen Folgekauf getätigt. Nach 168 Tagen haben in Klasse 5 die Hälfte der Kunden wieder gekauft, während in Klasse 1 das mediane Kaufen nicht berechnet werden kann, da weniger als 50 % der Kunden im Beobachtungszeitraum von einem Jahr gekauft haben.

Bei der Betrachtung der Scoreklassen zeigt sich Folgendes:

[4]Die Baseline Hazard Rate kann jedoch mit dem Breslow Schätzer geschätzt werden (siehe [KM06, Kapitel 8.8])

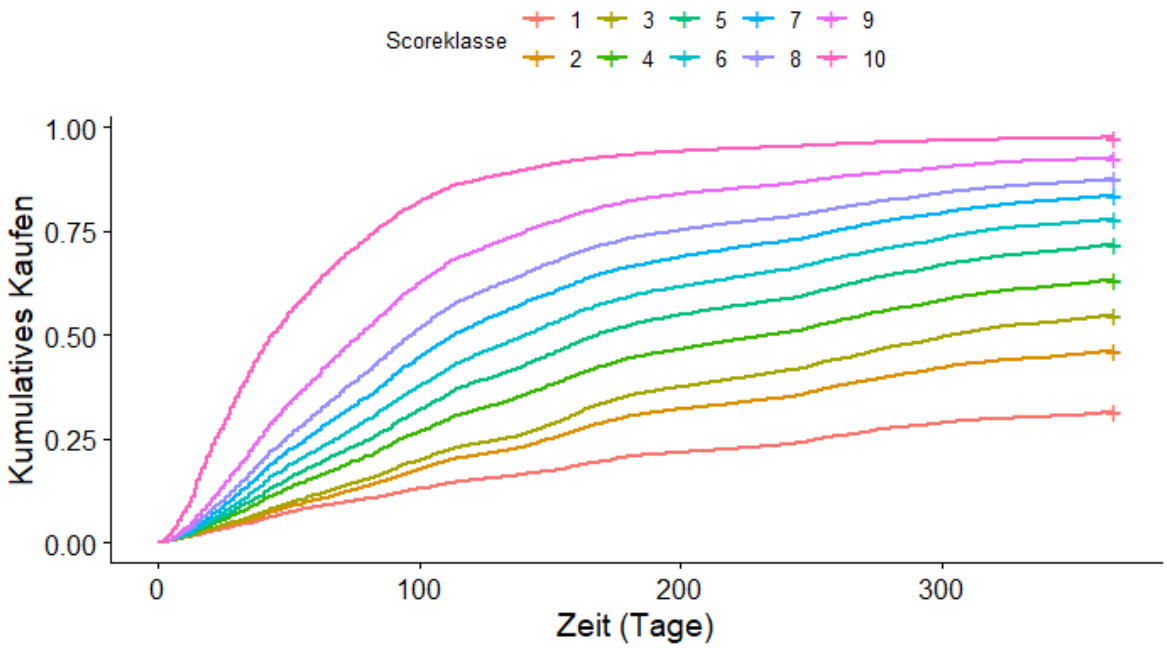



Abbildung 21: Darstellung der Kauffunktionen der 10 Scoreklassen

In Scoreklasse 1 sind 10,9% Männer, während in Scoreklasse 10 nur 4,2% der Kunden männlich sind.

Der Anteil der Kunden mit Premiumstatus liegt bei Scoreklasse 1 bei 5,7%. Es zeigt sich, dass mit steigender Scoreklasse dieser Anteil ansteigt: Scoreklasse 2 mit 12,1%, Scoreklasse 3 mit 13,1%, Scoreklasse 4 mit 22,5% bis Scoreklasse 10 mit 90,5% Premiumkunden.

Der größte Anteil an jungen Kunden von 18 bis 35 Jahren ist in Scoreklasse 1 zu finden. Hingegen sind in Scoreklasse 10 90,5% der Kunden 46 Jahre und älter.

Auch bei der Anzahl der Bestellungen wird ein Unterschied der Scoreklassen ersichtlich (siehe Tabelle 14).

Tabelle 14: Kreuztabelle mit relativen Häufigkeiten von Scoreklasse und Anzahl an Bestellungen

| | | Scoreklasse | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Anzahl Bestellungen (in %) | 1 | 98,1 | 92,1 | 79,6 | 14,8 | | | | | | |
| | 2 | 1,9 | 7,4 | 19,0 | 73,0 | 72,4 | 15,3 | 0,4 | | | |
| | 3 | | 0,5 | 1,3 | 11,3 | 23,7 | 70,2 | 33,0 | 2,9 | | |
| | 4-5 | | | 0,1 | 1,0 | 3,8 | 14,1 | 64,2 | 82,4 | 10,1 | |
| | 6-8 | | | | | 0,1 | 0,4 | 2,4 | 14,7 | 88,3 | 7,4 |
| | 9+ | | | | | | | | | 1,5 | 92,6 |
| | Gesamt | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

94,4% der Kunden in Scoreklasse haben keinen Artikel aus dem Sortiment Print bestellt, wohingegen in Scoreklasse 10 98,1% einen solchen Artikel bestellt haben.

Mithilfe der Cox-Regression und dem Scoring konnten Kundengruppen identifiziert werden, die sich hinsichtlich der Zeitdauer bis zum Folgekauf, demographischer Faktoren wie Geschlecht und Alter, sowie in ihrem Kaufverhalten unterscheiden. Eine ausführliche deskriptive Beschreibung der Scoreklassen in tabellarischer Form findet sich in Anhang C.

### 4.4.6 Ableitung von individuellen Marketingmaßnahmen

Die Kauffunktionen der entwickelten Scoreklassen zeigen eine unterschiedliche Kauffrequenz der Kunden. Diese individuelle Kauffrequenz kann nun beispielsweise im E-Mail-Marketing und bei dem Versand von Werbepost berücksichtigt werden.

Tabelle 15 zeigt die Quantile der Kaufkurven analog zu Abbildung 21. Die farbige Markierung anlog einer Ampel der Tabelle spiegelt den unterschiedlichen Handlungsbedarf und die Dringlichkeit von Marketingmaßnahmen wider. In der grünen Phase, hier gewählt bis $Q_{0.3}$ haben höchstens 30% der Kunden bis zu dem angegebenen Tag gekauft: Der Handlungsbedarf ist gering.

Tabelle 15: Quantile der Kauffunktion getrennt nach Scoreklasse mit farbiger Markierung nach 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 |

Betrachtet man die Gruppe mit der geringsten Kauffrequenz, also Kunden der Scoreklasse 10, so haben nach 44 Tagen bereits die Hälfte dieser Kunden einen Folgekauf getätigt. Hat ein Kunde bis zu diesem Zeitpunkt noch nicht gekauft, so ist der Handlungsbedarf für Marketingmaßnahmen erhöht.

Im roten Bereich haben mehr als 60% der Kunden einer Scoreklasse bereits einen Folgekauf getätigt.

Welche Maßnahmen geeignet sind und wie die Versandhäufigkeit von Werbemitteln gestaltet wird, ist zielgruppen- und branchenabhängig und muss individuell definiert werden. Die folgende Liste soll exemplarisch Möglichkeiten aufzeigen, welche Maßnahmen in den unterschiedlichen Phasen einer Scoreklasse zu definieren sind. Die hier genannten möglichen Steuergrößen einer Marketingmaßnahme im Dialogmarketing sind angelehnt an [LPZG15].

Mögliche Steuergrößen einer Marketingmaßnahme:

- Rabatt/ Gutscheine: Rabatte lassen sich in der grünen Phase sparsam einsetzen und sollten mit steigender Gefahr der Abwanderung erhöht werden, um Kunden

zum Kauf zu animieren. Einen Rabatt zu früh zu gewähren, obwohl der entsprechende Kunde auch ohne Rabatt noch kaufen würde, bedeutet Umsatzeinbußen. Besser ist es, die Rabatthöhe an die Dringlichkeit der Reaktivierung anzupassen.

- Frequenz: Die Newsletter-Frequenz lässt sich je nach Ampelphase anpassen. Auch könnten in der roten Phase gesonderte inhaltliche Reaktivierungsmaßnahmen und Rückgewinnugnsmaßnahmen ("Wir vermissen Sie") verwendet werden. Bei Scoreklasse 10 wäre eine solche Maßnahme beispielsweise nach 95 Tagen erforderlich.

- Gutscheinbedingungen: Bedingungen eines Gutscheins wie der Mindestbestellwert lassen sich an die Dringlichkeit der Reaktivierung eines Kunden anpassen. Ebenso kann die Gültigkeitsdauer von Gutscheinen eingeschränkt oder verlängert werden.

Gesondert betrachtet werden sollten die Zeitpunkte und Scoreklassen, wenn Quantile der Kauffunktionen nicht mehr berechnet werden können. Bei Klasse 1 bis 6 sind nicht alle Quantile berechenbar. In Tabelle 15 sind diese grau hinterlegt. Nach 327 Tagen haben 30 % der Kunden in Scoreklasse 1 gekauft. Die Zahl der Kunden die innerhalb des Beobachtungszeitraums von einem Jahr noch kaufen ändert sich nicht mehr merklich, da $Q_{0,4}$ nicht mehr erreicht wird.

Zusammenfassend lässt sich sagen: Sind die Quantile der Kauffunktionen nicht mehr berechenbar so erfolgt mit hoher Wahrscheinlichkeit im Beobachtungszeitraum kein Folgekauf und Reaktivierungsmaßnahmen sind erforderlich.

Tätigt ein Kunde einen Folgekauf, kann sein Score neu berechnet werden und er wird gegebenenfalls in eine andere Scoreklasse eingeordnet.

### 4.4.7 Validierung des Modells

Das Modell wurde mit 80% der Daten entwickelt und soll nun mit verbleibenden 20% der Testdaten validiert werden. Listing 7 zeigt das Vorgehen zur Validierung.

```
1  test_monat <- test[test$Monat == "01 Jan", ]
2
3  # Score berechnen mit Klassengrenzen
4  test_monat$score <- predict(fit, newdata = test_monat, type="lp") # lp =
       Lineare Vorhersage fuer Testdaten
5  test_monat$kl <- cut(test_monat$score, klassengrenzen, include.lowest=
       TRUE, labels = FALSE) # Definition der Klassen auf Basis der
       Klassengrenzen des Modells
6
7  scores_test <- test_monat %>%
8  group_by(kl) %>%
9  summarise(max = max(score), mw = mean(score))
10
11 fit_test <- survfit(Surv(tage365, status) ~ kl, data = test_monat)
12
13 grenzen_test <- quantile(fit_test, probs = c(0.1 ,0.2, 0.3, 0.4, 0.5,
14 0.6, 0.7, 0.8), conf.int = FALSE) # Quantile der Klassen der Testdaten
15
16 # Daten fuer Plot speichern
17 # Merkmal "dat" definiert ob Datensatz aus Test oder Training.
18 data_plot <- rbind(test_monat[, c("tage365", "status", "kl", "dat")],
19 train_monat[, c("tage365", "status", "kl", "dat")])
20
21 fit_test <- survfit(Surv(tage365, status) ~ kl + dat, data = data_plot)
22
23 ggsurvplot(fit_test,
24 data = test_monat,
25 fun = "event") # Grafik mit Gegenueberstellung von Modell und Testdaten.
```

Listing 7: Validierung des Modells mithilfe von Quantilen und Kauffunktionen mit Gegenüberstellung von Cox-Regressionsmodell und Testdaten (Januar)

Mithilfe der `predict()` Funktion wird auf Basis der in Kapitel 4.4.5 berechneten Parameterschätzungen der Score und die entsprechende Klassierung und somit die Scoreklassen für die Testdaten berechnet. Tabelle 16 zeigt die Quantile der Kauffunktion der Testdaten und des Modells. Aus Gründen der Übersichtlichkeit werden nur $Q_{0,2}$, $Q_{0,4}$, $Q_{0,6}$ gegenübergestellt.

In Abbildung 22 sind die Kauffunktionen der Testdaten dem Modell, entwickelt mit den Trainingsdaten gegenübergestellt. Es zeigt sich, dass die Kauffunktionen der Testdaten gut durch das Modell beschrieben werden. Die Kurven liegen aufeinander und lassen sich nur schwerlich unterscheiden. Für Scoreklasse 3 zeigen sich im Bereich der 200

Tabelle 16: Gegenüberstellung der Quantile aus dem Modell und den Testdaten

| | $Q_{0,2}$ | | $Q_{0,4}$ | | $Q_{0,6}$ | |
|---|---|---|---|---|---|---|
| Scoreklasse | Modell | Test | Modell | Test | Modell | Test |
| 1 | 174 | 172 | - | - | - | - |
| 2 | 112 | 113 | 280.5 | 274 | - | - |
| 3 | 101 | 102 | 226 | 232 | - | - |
| 4 | 77 | 76 | 159 | 155 | 316 | 314 |
| 5 | 64 | 66 | 129 | 130 | 251 | 253 |
| 6 | 55 | 53 | 106 | 107 | 186 | 185 |
| 7 | 46 | 44 | 90 | 88 | 150 | 147 |
| 8 | 41 | 42 | 78 | 77 | 123 | 120 |
| 9 | 33,5 | 33 | 61 | 59 | 95 | 95 |
| 10 | 19 | 19 | 35 | 35 | 58 | 56 |

Tage kleinere Unterschiede in den Kauffunktionen: Hier wird der Anteil der Kunden, die bis zu 200 Tagen kaufen, etwas unterschätzt. Ähnliches zeigt sich bei Scoreklasse 7 ab 300 Tagen.

Tabelle 17: Gegenüberstellung der Kaufwahrscheinlichkeiten nach einem, drei und sechs Monaten aus dem Modell und den Testdaten (in %)

| | 1M | | 3M | | 6M | |
|---|---|---|---|---|---|---|
| Scoreklasse | Modell | Test | Modell | Test | Modell | Test |
| 1 | 4,2 | 4,5 | 15,1 | 15,6 | 20,8 | 21,0 |
| 2 | 4,7 | 5,1 | 20,9 | 20,9 | 30,5 | 30,4 |
| 3 | 5,1 | 5,1 | 23,6 | 23,4 | 35,3 | 34,1 |
| 4 | 7,1 | 7,0 | 31,8 | 32,6 | 44,4 | 45,5 |
| 5 | 8,4 | 7,5 | 38,5 | 37,9 | 52,3 | 51,8 |
| 6 | 9,5 | 9,2 | 44,8 | 44,8 | 59,2 | 59,3 |
| 7 | 11,6 | 11,9 | 52,2 | 53,0 | 66,3 | 67,0 |
| 8 | 13,4 | 13,2 | 59,6 | 60,1 | 73,1 | 73,1 |
| 9 | 17,7 | 18,5 | 69,9 | 70,5 | 82,1 | 81,9 |
| 10 | 35,1 | 34,3 | 87,0 | 86,7 | 93,3 | 92,4 |

Tabelle 17 sind die Kaufwahrscheinlichkeiten nach einem, drei und sechs Monaten der Testdaten den Kaufwahrscheinlichkeiten aus dem Modell gegenübergestellt. Es wird eine gute Schätzung erreicht: Die Wahrscheinlichkeiten unterscheiden sich nur gering. In Scoreklasse 1 liegt die Kaufwahrscheinlichkeit nach 6 Monaten bei nur 21,0% (Modell: 20,8%). Hingegen liegt die Wahrscheinlichkeit in Scoreklasse 10 bis zu diesem Zeitpunkt bereits gekauft zu haben bei 92,4% (Modell: 93,3%).

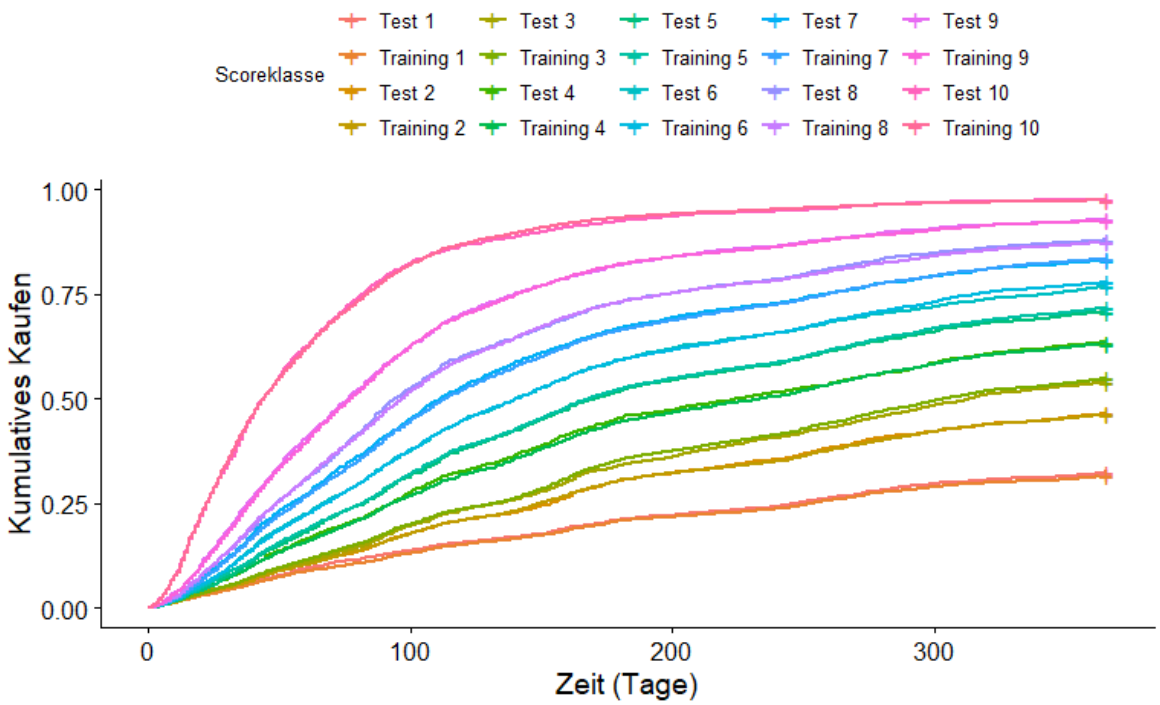



Abbildung 22: Kaplan-Meier-Schätzer für die Warengruppen 1 bis 6 für Januar

Ergänzend zum Vergleich von Trainigns-und Testdaten werden nun noch die Cox-Snell-Residuen zur Beurteilung der Güte des Januarmodells herangezogen. Listings 8 zeigt die Berechnung der Residuen und die Erstellungen von Abbildung 23.

```
# Cox-Snell Residuen berechnen
# Berechnung mithilfe der Martingale-Residuen
resid_cs <- train_monat$status - residuals(fit, type = "martingale")

# Schaetzer der kumulativen Hazard Funktion mit Fleming-Harrington
resid.surv <- survfit(Surv(resid_cs, status) ~1, ctype = 2, data = train_
    monat)
resid <- data.frame(resid.surv$surv, resid.surv$time)
colnames(resid) <- c("surv", "time")

# Relative Haeufigkeit Datenpunkte
# Datenpunkt kleiner 2.5
textlinks <- round(prop.table(table(resid$time<2.5))[2]*100,1)
# Datenpunkt groesser 2.5
textrechts <- round(prop.table(table(resid$time<2.5))[1]*100,1)

# Cox-Snell Residuenplot
ggplot(resid, aes(time, -log(surv))) + geom_point(size = 0.01) +
          xlab("Cox-Snell Residuen") + ylab("Kumulatives Risiko") +
          theme_classic() + geom_abline(slope = 1) +
          geom_vline(xintercept = 2.5, col = "red", lty = 5) +
```

```
21          annotate(geom="text", x=1, y=3.5, label=paste0(textlinks, "%
                der Daten"), color="red") +
22          annotate(geom="text", x=4, y=1.5, label=paste0(textrechts, "%
                der Daten"), color="red")
```

Listing 8: Beurteilung der Güte der Modellanpassung mit Cox-Snell-Residuen

Abbildung 23 zeigt den zugehörigen Plot. Wenn das Cox-Modell eine gute Anpassung der Daten liefert, zeigen die Daten eine gerade Linie durch den Ursprung mit Steigung 1 (vgl. Kapitel 2.4.3).

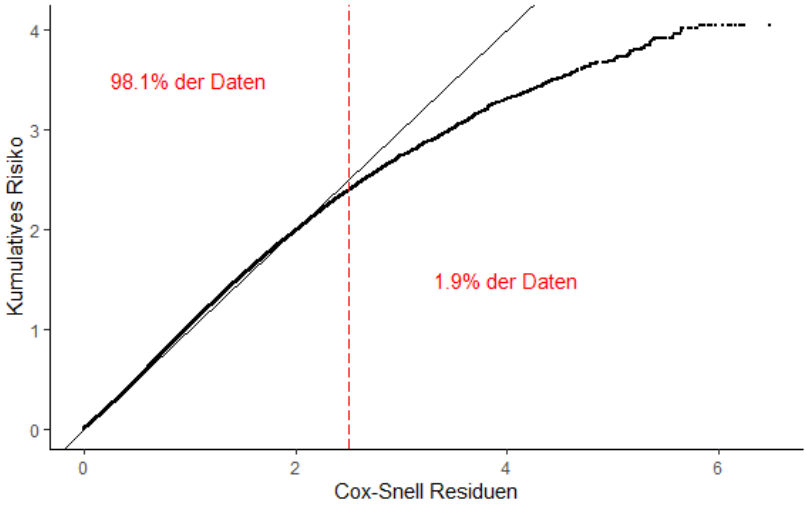



Abbildung 23: Cox-Snell-Residuen zur Beurteilung der Güte der Schätzung

In der Abbildung zeigen sich deutliche Abweichungen von der Geraden. Aufgrund der Größe des Datensatzes lässt sich jedoch graphisch nicht erkennen, für wie viele der Datenpunkte sich eine Abweichung von der Geraden ergibt. Betrachtet man die Fallzahlen hinter den Residuen wird ersichtlich, dass für 98,1% der Daten das Modell eine gute Anpassung liefert. 1,9% der Daten weichen von der Geraden ab.

Zusätzlich muss beachtet werden, dass der Plot der Cox-Snell-Residuen auf den Trainingsdaten und nicht auf den Testdaten basiert. Aufgrund des Ursprungs der Cox-Regression in der Medizin ist die Validierung von Testdaten in den R Paketen `survival` und `survminer` nicht verfügbar. Die Unterteilung der Daten zur Entwicklung von Modellen in Training-, Test- und Validierungsdatensatz werden jedoch eher im Bereich des maschinellen Lernens angewandt. Da der Plot nur ergänzend aufgeführt ist, wurde auf die Entwicklung entsprechender Methoden für die Testdaten in R verzichtet.

### 4.4.8 Ausblick: Validierung des Modells mit KPIs

Die Testphase sollte nach Entwicklung des Modells erfolgen. Sie besteht darin, dass Modell und den Maßnahmenkatalog für eine Zufallstichprobe an Kunden für einen zu definierenden Zeitraum zu testen und den Erfolg der Maßnahmen zu messen. Im Rahmen dieser Arbeit kann die Testphase zeitlich nicht mehr erfolgen.

Dabei bietet es sich an, den Erfolg mithilfe von Key Performance Indicators (kurz: KPIs) zu untersuchen. KPIs sind in der Betriebswirtschaft Kennzahlen zur Messung des Erfolgs von Unternehmensprozessen, Projekten oder Maßnahmen. Je nach Geschäftsbereich werden unterschiedliche KPIs definiert und spielen eine Rolle bei der Entscheidungsfindung [WV13].

Eine Übersicht möglicher KPIs des Controllings im E-Mail-Marketing und für die Gestaltung von Werbepost im Allgemeinen (hier: Print) zum Testen der Marketingmaßnahmen zeigt Tabelle 18 und ist angelehnt an [Kre18].

Tabelle 18: Auswahl von Key Performance Indicators im Dialogmarketing

| KPI | Marketingkanal | Beschreibung |
|---|---|---|
| Geöffnet-Rate | E-Mail | Prozentualer Anteil an Kunden, die einen E-Mail-Newsletter oder eine Werbe-E-Mail geöffnet haben im Verhältnis zu allen Empfängern |
| Klick-Rate | E-Mail | Prozentualer Anteil an Kunden die einen E-Mail Newsletter geöffnet haben und einen Link der E-Mail angeklickt haben im Verhältnis zu allen Newsletter-Empfängern |
| Conversion Rate | E-Mail | Prozentualer Anteil an Kunden, die einen Newsletter geöffnet, angeklickt und infolgedessen einen Kauf getätigt haben. |
| Gutschein-Rate | E-Mail & Print | Anteil an verwendeten Gutscheinen im Verhältnis zu versendeten Gutscheinen |
| Umsatzerlös | E-Mail & Print | Durchschnittlich erzielter Umsatz eines Kunden nach Abzug von eingelösten Rabatten und Gutscheinen |

| | | |
|---|---|---|
| Bestell-Rate | E-Mail & Print | Anzahl an Bestellungen im Verhältnis zu verschickter Werbepost |
| Abmelde-Rate | E-Mail | Der Anteil an Kunden, die sich aus dem Newsletterverteiler abgemeldet haben im Verhältnis zu den Newsletter-Empfängern. Diese Kennzahl ist im E-Mail-Marketing von besonderer Bedeutung: Hat sich ein Kunde erstmal abgemeldet, kann er nicht mehr über diesen Kanal reaktiviert werden. |

Die KPIs, berechnet für die Test-Kundengruppe werden den KPIs der Kontroll-Kundengruppe nach Abschluss der Testphase gegenübergestellt um den Erfolg der Maßnahmen zu überprüfen.

# 5 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. Statt der Überlebenszeit kann im Marketing die Zeit bis zum Folgekauf eines Kunden betrachtet werden. Vorgestellt wurden der Kaplan-Meier-Schätzer, der Log-Rank-Test und die Cox-Regression. Die Survival Analyse mit der Zielgröße Zeit bis Folgekauf hilft zu verstehen, welche Merkmale das Kaufintervall eines Kunden beeinflussen.

Zudem wurde ein Verfahren entwickelt, das die Bildung von Kundengruppen auf Basis ihres Kaufintervalls erlaubt. Das Verfahren beschreibt den Ablauf einer möglichen Analyse des Kaufintervalls von Kunden beginnend mit einer bivariaten Analyse über die multivariate Modellbildung bis hin zur Entwicklung der Scoreklassen und Validierung derselben. Die Scoreklassen, die dazu dienen, Kunden mit kurzem Kaufintervall von Kunden mit längerem Kaufintervall zu unterscheiden, basieren auf der Vorhersage des Cox-Regressionsmodells.

Im praktischen Teil der Arbeit wurde das entwickelte Verfahren umgesetzt. Dabei wurden auf Grund von Saisoneffekten und Datenstruktur 12 Modelle für Verkaufsdaten eines Versandhändlers von einem Jahr entwickelt. Um die Modelle validieren zu können, erfolgte eine Aufteilung in Trainings- und Testdaten. Für die Testdaten zeigte sich in allen 12 Modellen eine gute Anpassung. Die Modelle lieferten eine gute Vorhersage der Wahrscheinlichkeit, bis zu einem gewissen Zeitpunkt zu kaufen. Zudem wurden Möglichkeiten aufgezeigt, die Kenntnis der Kaufintervalle der Scoreklassen zu nutzen, um die Frequenz und den Inhalt von Werbemaßnahmen zu optimieren.

Neben der Validierung und Betrachtung der Güte des Modells mithilfe der Cox-Snell-Residuen bietet es sich an, den Erfolg, der aus dem Modell abgeleiteten und auf die Scoreklassen zugeschnittenen Werbemaßnahmen mithilfe von Key Performance Indikatoren zu überprüfen. Dies war im Rahmen dieser Arbeit zeitlich nicht mehr möglich.

In der Analyse wurden die Daten eines Geschäftsjahres einbezogen. Die Modelle den Daten des Folgejahres gegenüberzustellen, wäre eine weitere Möglichkeit der Validierung. Auch dies konnte im Rahmen dieser Arbeit zeitlich nicht mehr erfolgen, da die entsprechenden Daten noch nicht erhoben sind. Effekte, die innerhalb des Geschäftsjahres auftreten und somit Bestandteile des Modells sind, jedoch im Folgejahr nicht reproduzierbar sind, lassen sich mit der Aufteilung in Test & Trainingsdaten nicht identifizieren. Ein Testen der Modelle im Folgejahr sollte demnach erfolgen.
Weiterhin kann sich das Kaufverhalten von Kunden über längere Zeit hinweg verändern. Die berechneten Quantile der Kauffunktionen basieren auf dem Kaufverhalten der Kunden des untersuchten Geschäftsjahres. Werbemaßnahmen werden auf Basis dieser

Quantile optimiert. Ändert sich jedoch das Verhalten einer Scoreklasse, das Kaufintervall einer Scoreklasse wird demnach länger oder kürzer, so liefern die Modelle keine zufriedenstellenden Ergebnisse mehr. Aus diesem Grund müssen die Modelle in regelmäßigen Abständen aktualisiert werden. Sie liefern keine allgemein gültigen Aussagen über das Kundenverhalten, sondern spiegeln das Kaufverhalten der Kunden zum aktuellen Zeitpunkt wider. Ändert sich das Kaufverhalten der Kunden, so müssen auch die Modelle entsprechend angepasst werden.

Bei der Survival Analyse handelt es sich um ein umfangreiches Teilgebiet der Statistik. Im Rahmen dieser Arbeit wurde ein Teil der möglichen Verfahren vorgestellt und angewendet. Bei den vorgestellten Verfahren handelt es sich um nicht-parametrische bzw. semi-parametrische Verfahren. Doch auch parametrische Ansätze zur Modellierung von Kauffunktionen sind möglich.
Mögliche Modelle, die den Einfluss von Parametern auf die Zeit bis zum Folgekauf untersuchen, sind das exponentielle Regressionsmodell und das Weibull Regressionsmodell. Als weitere Verfahren, um Kaufintervalle zu analysieren, seien ATF-Modelle (Accelerated-Failure-Time) genannt.
Eine weitere Möglichkeit der Anwendung der Survival Analyse im Marketing sind baumbasierte Ansätze. Die sogenannten Survival Trees sind u.a. in den R Paketen `rpart` oder `MST` implementiert. Auch Ensemblemethoden wie Random Survival Trees sind in R verfügbar.

Ein Vorteil der in dieser Arbeit verwendeten Verfahren liegt in deren Verbreitung: Sie sind in gängiger Statistik Software implementiert und in der Literatur gut beschrieben. Da es sich um nicht-parametrische bzw. semi-parametrische Verfahren handelt, sind keine Annahmen über die Verteilung der Ereigniszeit erforderlich. Damit sind nur wenige Voraussetzungen vor Anwendung der Verfahren zu überprüfen.

Abschließend lässt sich sagen, dass die Survival Analyse diverse Möglichkeiten bietet, um Marketingmaßnahmen zu optimieren.

# Literaturverzeichnis

[BHM86] Blossfeld, Hans-Peter ; Hamerle, Alfred ; Mayer, Karl U.: *Ereignisanalyse: statistische Theorie und Anwendung in den Wirtschafts-und Sozialwissenschaften*. Bd. 569. Campus, 1986

[Bre70] Breslow, Norman: A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. In: *Biometrika* 57 (1970), Nr. 3, S. 579–594

[BS11] Bortz, J. ; Schuster, C.: *Statistik für Human- und Sozialwissenschaftler: Limitierte Sonderausgabe*. Springer Berlin Heidelberg, 2011 (Springer-Lehrbuch). – ISBN 9783642127700

[Cox72] Cox, David R.: Regression models and life-tables. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34 (1972), Nr. 2, S. 187–202

[GT94] Grambsch, Patricia M. ; Therneau, Terry M.: Proportional hazards tests and diagnostics based on weighted residuals. In: *Biometrika* 81 (1994), Nr. 3, S. 515–526

[HDE19] HDE: *Umsatz durch E-Commerce (B2C) in Deutschland in den Jahren 1999 bis 2018 sowie eine Prognose für 2019 (in Milliarden Euro)*. `https://de.statista.com/statistik/daten/studie/3979/umfrage/e-commerce-umsatz-in-deutschland-seit-1999`. Version: 20. Mai, 2019. – Gesehen am 20. Januar 2020

[HL99] Hosmer, D.W. ; Lemeshow, S.: *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley, 1999 (A Wiley-Interscience publication). – ISBN 0471154105

[HS18] Hedderich, J. ; Sachs, L.: *Angewandte Statistik: Methodensammlung mit R*. Springer Berlin Heidelberg, 2018. – ISBN 9783662566572

[KM58] Kaplan, Edward L. ; Meier, Paul: Nonparametric estimation from incomplete observations. In: *Journal of the American statistical association* 53 (1958), Nr. 282, S. 457–481

[KM06] Klein, J.P. ; Moeschberger, M.L.: *Survival Analysis: Techniques for Censored and Truncated Data*. Springer New York, 2006 (Statistics for Biology and Health). – ISBN 9780387216454

[KP80] Kalbfleisch, J.D. ; Prentice, R.L.: *The statistical analysis of failure time data*. Wiley, 1980 (Wiley series in probability and mathematical statistics: Applied probability and statistics). – ISBN 9780471055198

[Kre18] Kreutzer, Ralf T.: *E-Mail-Marketing kompakt*. Springer, 2018

[LPZG15] Linnhoff-Popien, Claudia ; Zaddach, Michael ; Grahl, Andreas: *Marktplätze im Umbruch: Digitale Strategien für Services im Mobilen Internet*. Springer-Verlag, 2015

[Sch82] Schoenfeld, David: Partial Residuals for The Proportional Hazards Regression Model. In: *Biometrika* 69 (1982), Nr. 1, S. 239–241

[Sta17] Statista: *In welchen Fällen bestellen Sie Newsletter ab?* https://de.statista.com/statistik/daten/studie/712711/umfrage/gruende-fuer-abmeldungen-von-newslettern-in-deutschland/. Version: 31. Mai, 2017. – Gesehen am 20. Januar 2020

[WV13] WIRTSCHAFTSLEXIKON, GABLER ; Verlag, Springer G.: Gabler Wirtschaftslexikon. In: *Stichwort: KPI, online: https://wirtschaftslexikon.gabler.de/definition/key-performance-indicator-kpi-52670/version-275788 (Abruf zuletzt: 06.01.2020)* (2013)

[ZBH11] Zwiener, Isabella ; Blettner, M ; Hommel, Gerhard: Überlebenszeitanalyse. Teil 15 der Serie zur Bewertung wissenschaftlicher Publikationen. In: *Deutsches Ärzteblatt* 108 (2011), Nr. 10, S. 163–169

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Listingverzeichnis

# Anhang

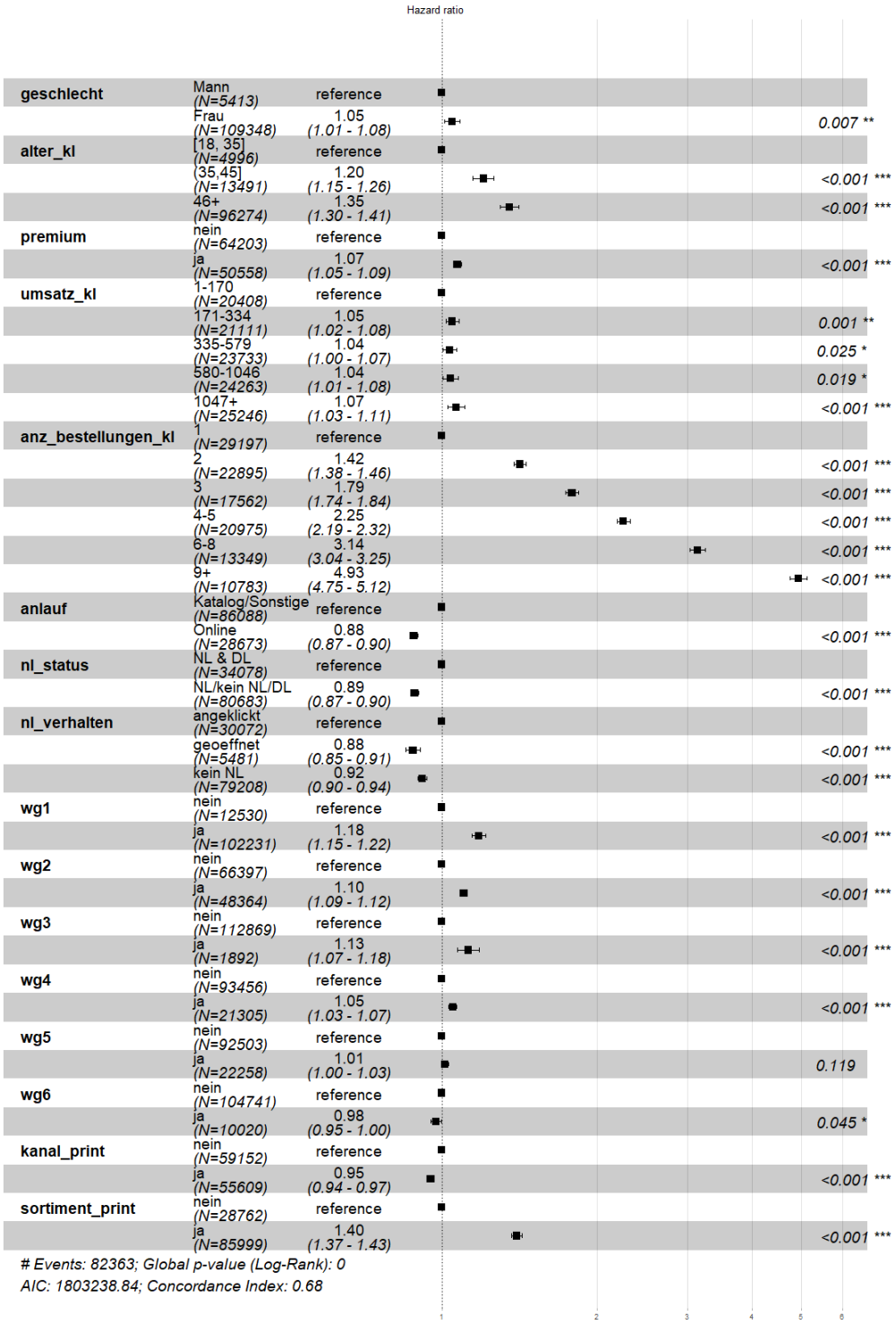
## A Übersicht der Variablen im Datensatz

- ID: definiert einen Kunden
- Kaufdatum: Datum des letzten Einkaufs im Monat
- Monat: Monat des letzten Kauf
- Folgebestellung: Folgebestellung nach Kaufdatum
- Tage: Anzahl der Tage bis Kauf
  Abgeleitet aus Folgebestellung - Kaufdatum
- Status: beschreibt, ob Ereignis (hier: Folgebestellung) eingetreten ist oder nicht
  1 = Folgebestellung 2 = Zensiert
- Geschlecht: 1 = Frau, 2 = Mann
- Alter in Jahren
- Premium: definiert einen Premiumkunden 1 = ja 2 = nein
- Umsatz: Umsatz in Euro der letzten 12 Monate
- AnzBestellungen: Anzahl der Bestellungen der letzten 12 Monate
- Erstkontakt: beschreibt den Kanal des Erstkontakts
  1 = Katalog 2 = Online 3 = Sonstige
- NLStatus: Newsletter Status des Kunden zum aktuellen Zeitpunkt (Export der Daten)
  beschreibt, welche Art von Newsletter der Kunde abonniert hat
  1 = DL (Dienstleitung) 2 = DL & NL (Newsletter) 3 = Nur NL / kein NL
  Dienstleitstung beschreibt hier Informationen zur Bestellung (z. B. Bestellbestätigung, Mahnung, etc.)
- NLKlicken: Newsletter Klickverhalten der vergangenen 12 Monate
  1 = angeklickt 2 = geöffnet 3 = kein NL erhalten 4 = nicht geöffnet
- WG1 – WG6: beschreibt, ob Kunde in den letzten 12 Monaten in Warengruppe 1 bis 6 gekauft hat
  1 = ja 2 = nein
- KanalOnline: mindestens einmal in den letzten 12 Monaten online bestellt
  1 = ja 2 = nein

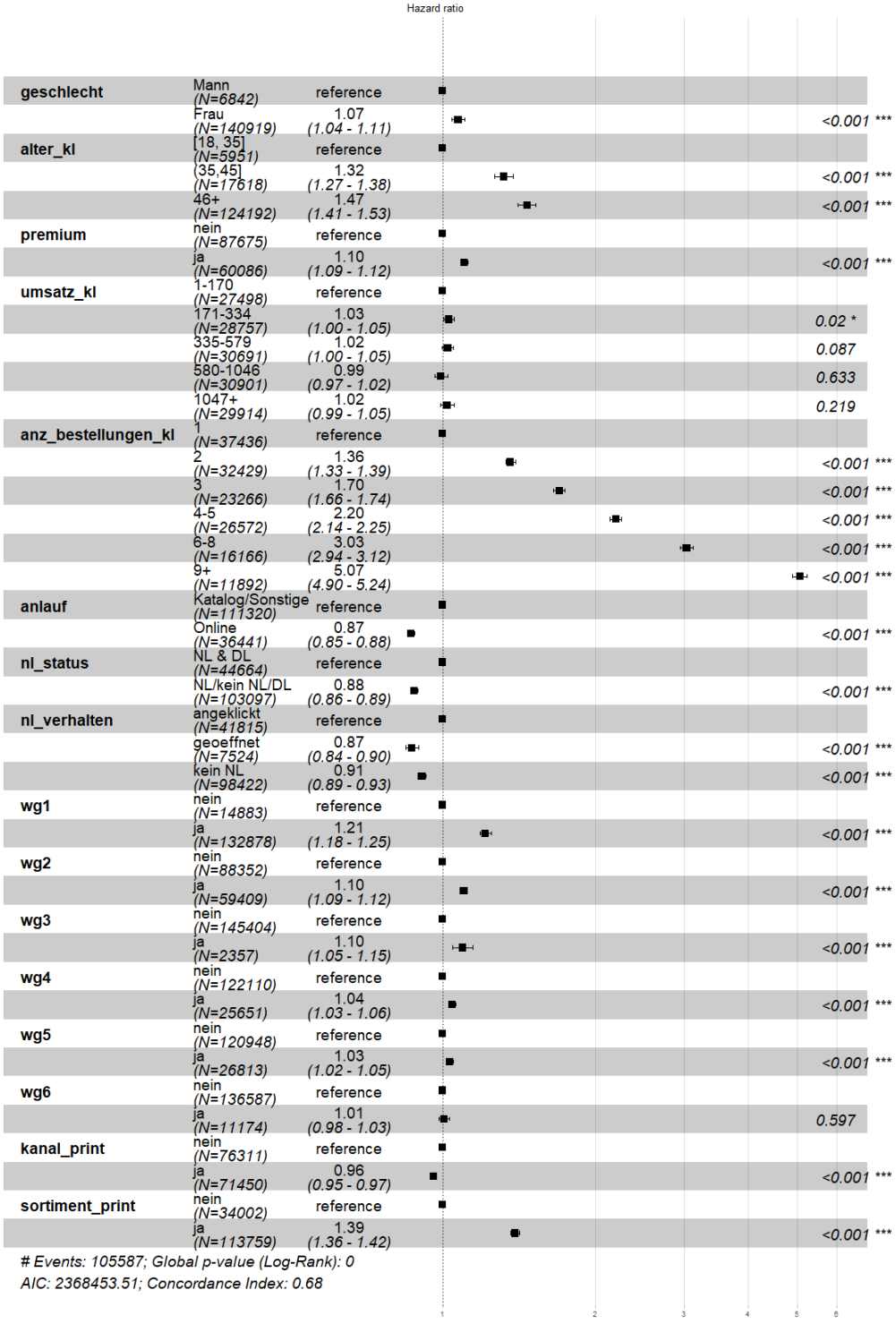
- KanalPrint: mindestens einmal in den letzten 12 Monaten im Print Katalog bestellt (über Telefon oder Bestellformular)
  1 = ja 2 = nein
- SortimentOnline: mindestens ein Artikel des Onlinesortiments in den letzten 12 Monaten bestellt
- SortimentPrint: mindestens ein Artikel des Katalogs in den letzten 12 Monaten bestellt

# B Wald-Diagramme für Parameterschätzungen der Cox-Regression für die Monate Februar bis Dezember

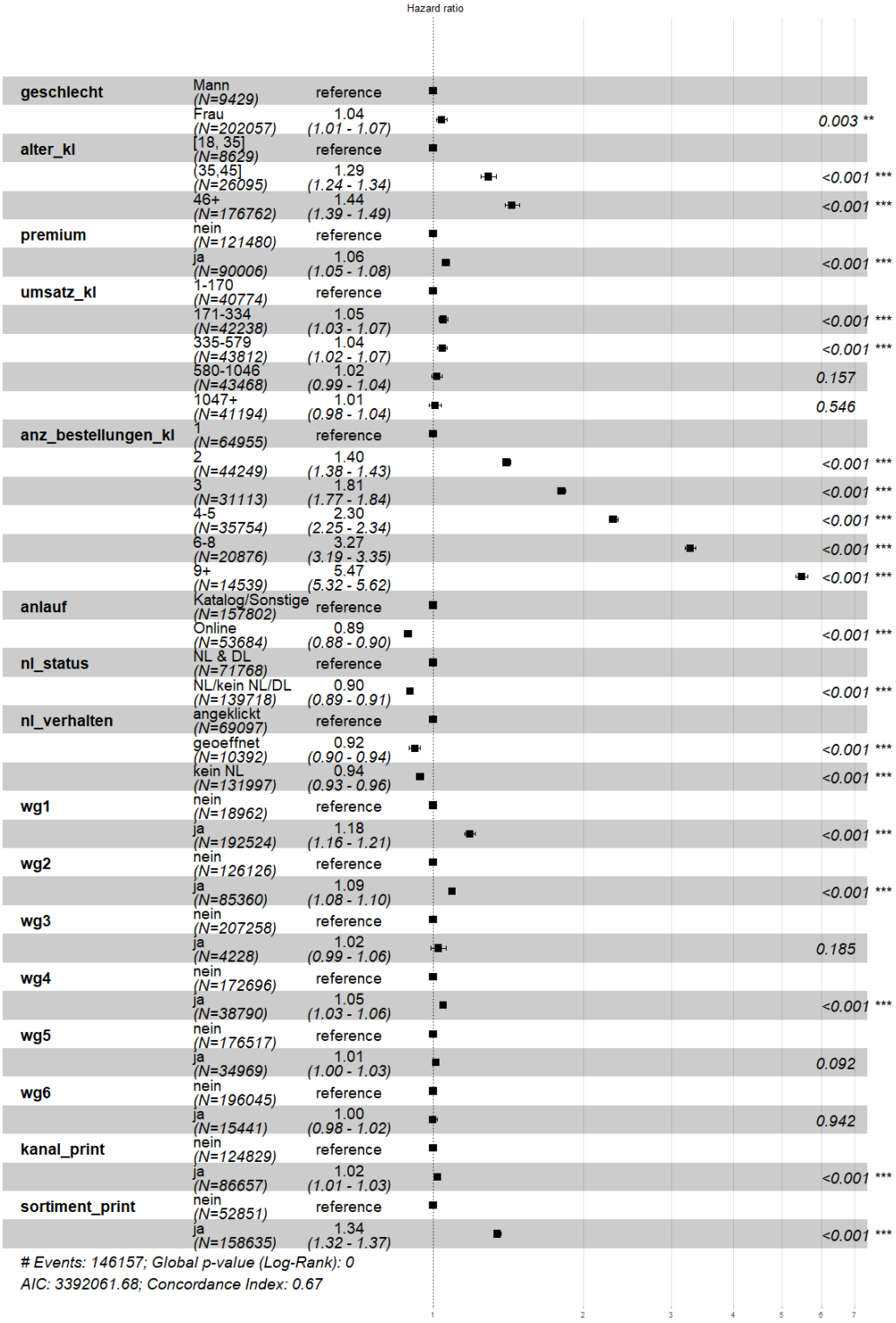
## Februar



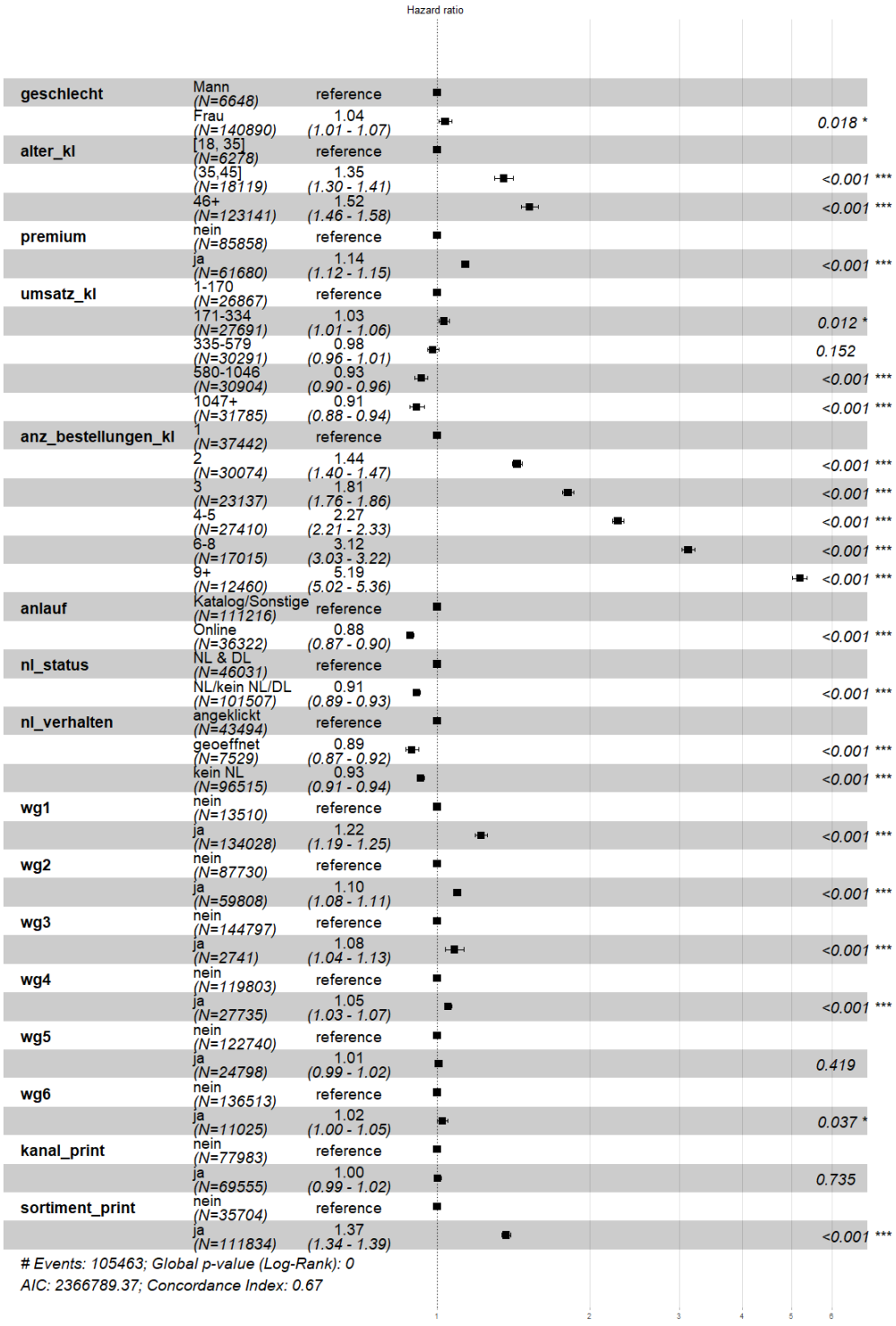
## März


Hazard ratio
geschlecht
Mann (N=6842) reference
Frau (N=140919) 1.07 (1.04 - 1.11) <0.001 ***
alter_kl
[18, 35] (N=5951) reference
(35,45] (N=17618) 1.32 (1.27 - 1.38) <0.001 ***
46+ (N=124192) 1.47 (1.41 - 1.53) <0.001 ***
premium
nein (N=87675) reference
ja (N=60086) 1.10 (1.09 - 1.12) <0.001 ***
umsatz_kl
1-170 (N=27498) reference
171-334 (N=28757) 1.03 (1.00 - 1.05) 0.02 *
335-579 (N=30691) 1.02 (1.00 - 1.05) 0.087
580-1046 (N=30901) 0.99 (0.97 - 1.02) 0.633
1047+ (N=29914) 1.02 (0.99 - 1.05) 0.219
anz_bestellungen_kl
1 (N=37436) reference
2 (N=32429) 1.36 (1.33 - 1.39) <0.001 ***
3 (N=23266) 1.70 (1.66 - 1.74) <0.001 ***
4-5 (N=26572) 2.20 (2.14 - 2.25) <0.001 ***
6-8 (N=16166) 3.03 (2.94 - 3.12) <0.001 ***
9+ (N=11892) 5.07 (4.90 - 5.24) <0.001 ***
anlauf
Katalog/Sonstige (N=111320) reference
Online (N=36441) 0.87 (0.85 - 0.88) <0.001 ***
nl_status
NL & DL (N=44664) reference
NL/kein NL/DL (N=103097) 0.88 (0.86 - 0.89) <0.001 ***
nl_verhalten
angeklickt (N=41815) reference
geoeffnet (N=7524) 0.87 (0.84 - 0.90) <0.001 ***
kein NL (N=98422) 0.91 (0.89 - 0.93) <0.001 ***
wg1
nein (N=14883) reference
ja (N=132878) 1.21 (1.18 - 1.25) <0.001 ***
wg2
nein (N=88352) reference
ja (N=59409) 1.10 (1.09 - 1.12) <0.001 ***
wg3
nein (N=145404) reference
ja (N=2357) 1.10 (1.05 - 1.15) <0.001 ***
wg4
nein (N=122110) reference
ja (N=25651) 1.04 (1.03 - 1.06) <0.001 ***
wg5
nein (N=120948) reference
ja (N=26813) 1.03 (1.02 - 1.05) <0.001 ***
wg6
nein (N=136587) reference
ja (N=11174) 1.01 (0.98 - 1.03) 0.597
kanal_print
nein (N=76311) reference
ja (N=71450) 0.96 (0.95 - 0.97) <0.001 ***
sortiment_print
nein (N=34002) reference
ja (N=113759) 1.39 (1.36 - 1.42) <0.001 ***
# Events: 105587; Global p-value (Log-Rank): 0
AIC: 2368453.51; Concordance Index: 0.68
1
2
3
4
5
6

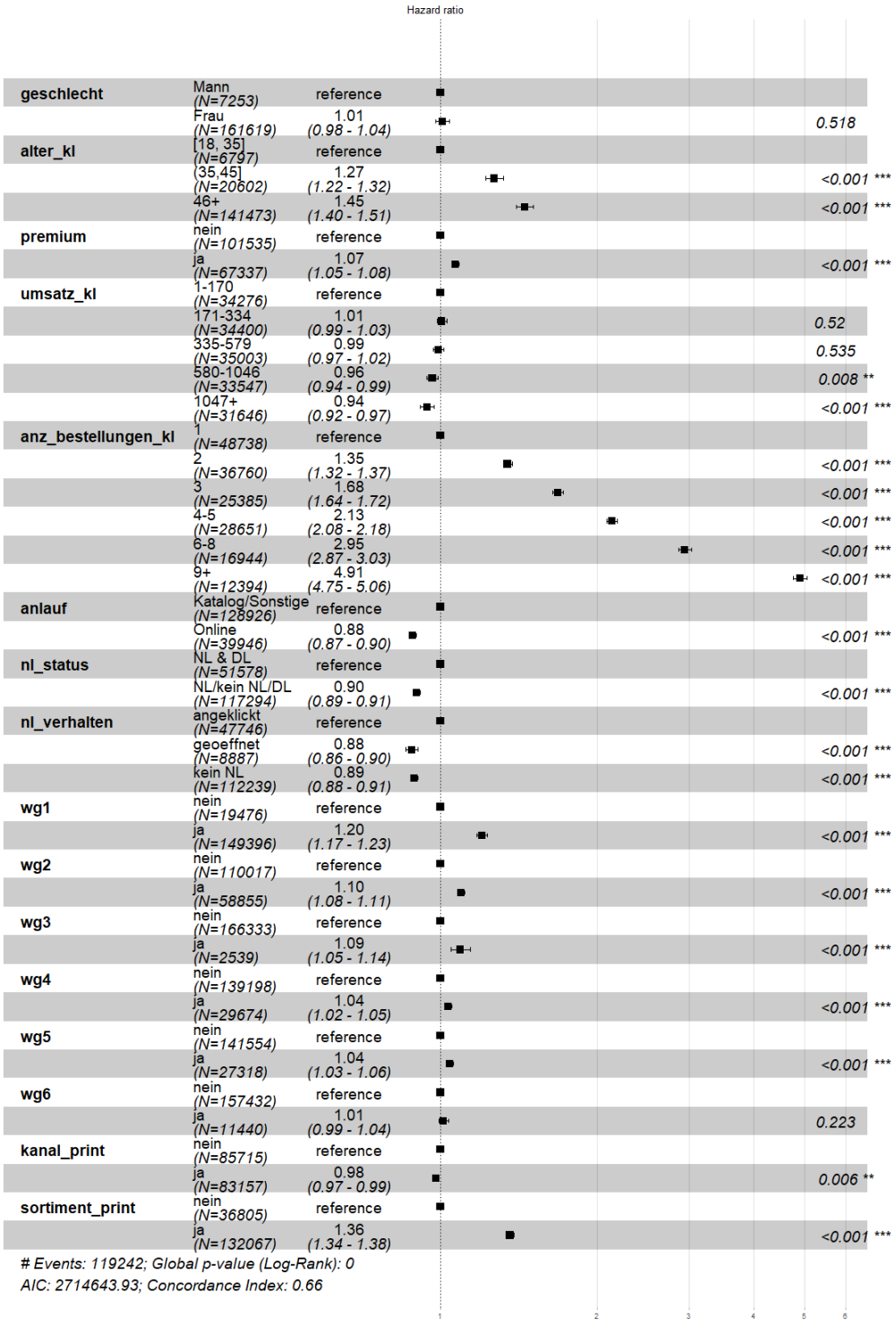
## April


Hazard ratio
geschlecht
Mann (N=9429) reference
Frau (N=202057) 1.04 (1.01 - 1.07) 0.003 **
alter_kl
[18, 35] (N=8629) reference
[35,45] (N=26095) 1.29 (1.24 - 1.34) <0.001 ***
46+ (N=176762) 1.44 (1.39 - 1.49) <0.001 ***
premium
nein (N=121480) reference
ja (N=90006) 1.06 (1.05 - 1.08) <0.001 ***
umsatz_kl
1-170 (N=40774) reference
171-334 (N=42238) 1.05 (1.03 - 1.07) <0.001 ***
335-579 (N=43812) 1.04 (1.02 - 1.07) <0.001 ***
580-1046 (N=43468) 1.02 (0.99 - 1.04) 0.157
1047+ (N=41194) 1.01 (0.98 - 1.04) 0.546
anz_bestellungen_kl
1 (N=64955) reference
2 (N=44249) 1.40 (1.38 - 1.43) <0.001 ***
3 (N=31113) 1.81 (1.77 - 1.84) <0.001 ***
4-5 (N=35754) 2.30 (2.25 - 2.34) <0.001 ***
6-8 (N=20876) 3.27 (3.19 - 3.35) <0.001 ***
9+ (N=14539) 5.47 (5.32 - 5.62) <0.001 ***
anlauf
Katalog/Sonstige (N=157802) reference
Online (N=53684) 0.89 (0.88 - 0.90) <0.001 ***
nl_status
NL & DL (N=71768) reference
NL/kein NL/DL (N=139718) 0.90 (0.89 - 0.91) <0.001 ***
nl_verhalten
angeklickt (N=69097) reference
geoeffnet (N=10392) 0.92 (0.90 - 0.94) <0.001 ***
kein NL (N=131997) 0.94 (0.93 - 0.96) <0.001 ***
wg1
nein (N=18962) reference
ja (N=192524) 1.18 (1.16 - 1.21) <0.001 ***
wg2
nein (N=126126) reference
ja (N=85360) 1.09 (1.08 - 1.10) <0.001 ***
wg3
nein (N=207258) reference
ja (N=4228) 1.02 (0.99 - 1.06) 0.185
wg4
nein (N=172696) reference
ja (N=38790) 1.05 (1.03 - 1.06) <0.001 ***
wg5
nein (N=176517) reference
ja (N=34969) 1.01 (1.00 - 1.03) 0.092
wg6
nein (N=196045) reference
ja (N=15441) 1.00 (0.98 - 1.02) 0.942
kanal_print
nein (N=124829) reference
ja (N=86657) 1.02 (1.01 - 1.03) <0.001 ***
sortiment_print
nein (N=52851) reference
ja (N=158635) 1.34 (1.32 - 1.37) <0.001 ***
# Events: 146157; Global p-value (Log-Rank): 0
AIC: 3392061.68; Concordance Index: 0.67
1
2
3
4
5
6
7

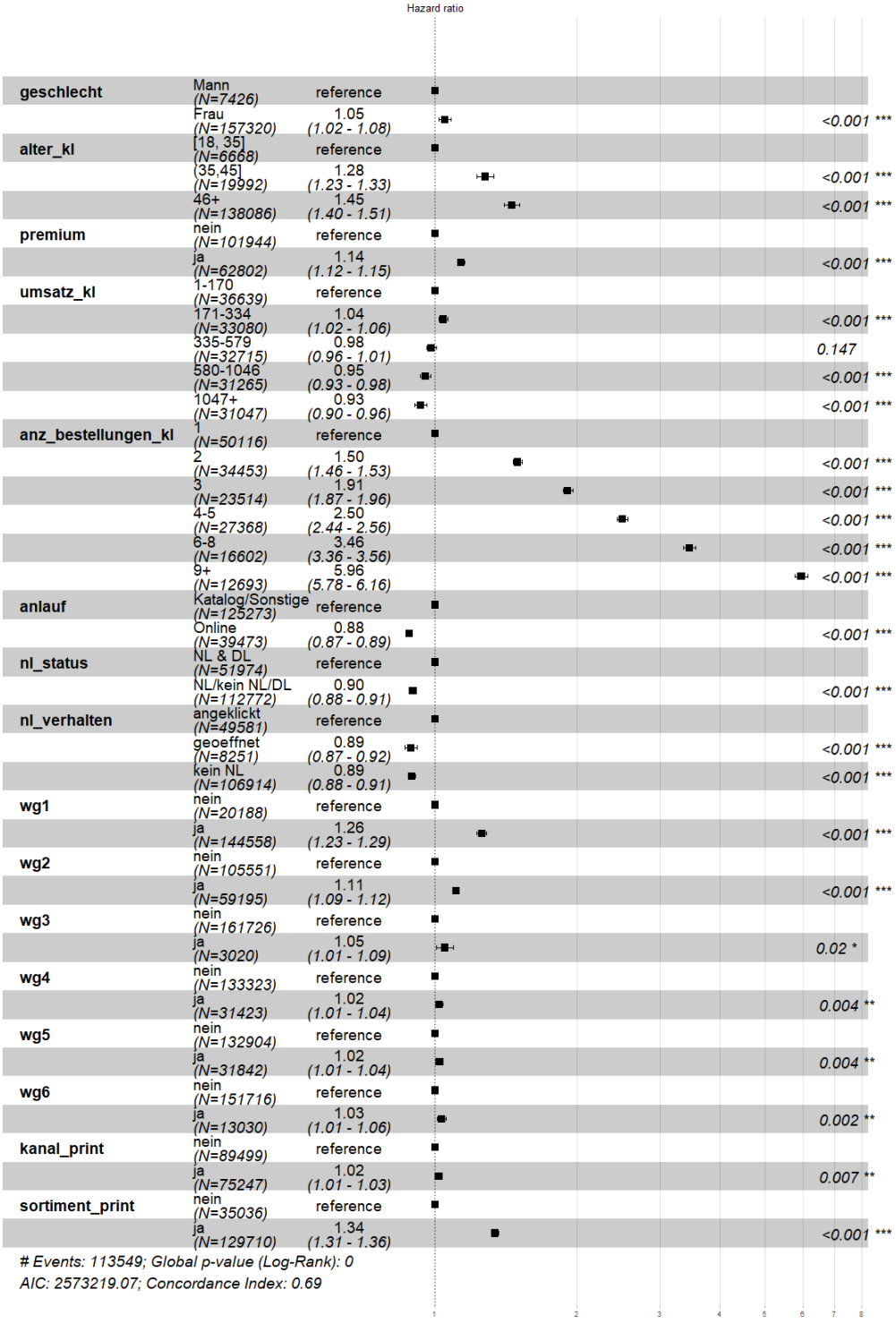
## Mai


Hazard ratio
geschlecht
Mann (N=6648) reference
Frau (N=140890) 1.04 (1.01 - 1.07) 0.018 *
alter_kl
[18, 35] (N=6278) reference
(35,45] (N=18119) 1.35 (1.30 - 1.41) <0.001 ***
46+ (N=123141) 1.52 (1.46 - 1.58) <0.001 ***
premium
nein (N=85858) reference
ja (N=61680) 1.14 (1.12 - 1.15) <0.001 ***
umsatz_kl
1-170 (N=26867) reference
171-334 (N=27691) 1.03 (1.01 - 1.06) 0.012 *
335-579 (N=30291) 0.98 (0.96 - 1.01) 0.152
580-1046 (N=30904) 0.93 (0.90 - 0.96) <0.001 ***
1047+ (N=31785) 0.91 (0.88 - 0.94) <0.001 ***
anz_bestellungen_kl
1 (N=37442) reference
2 (N=30074) 1.44 (1.40 - 1.47) <0.001 ***
3 (N=23137) 1.81 (1.76 - 1.86) <0.001 ***
4-5 (N=27410) 2.27 (2.21 - 2.33) <0.001 ***
6-8 (N=17015) 3.12 (3.03 - 3.22) <0.001 ***
9+ (N=12460) 5.19 (5.02 - 5.36) <0.001 ***
anlauf
Katalog/Sonstige (N=111216) reference
Online (N=36322) 0.88 (0.87 - 0.90) <0.001 ***
nl_status
NL & DL (N=46031) reference
NL/kein NL/DL (N=101507) 0.91 (0.89 - 0.93) <0.001 ***
nl_verhalten
angeklickt (N=43494) reference
geoeffnet (N=7529) 0.89 (0.87 - 0.92) <0.001 ***
kein NL (N=96515) 0.93 (0.91 - 0.94) <0.001 ***
wg1
nein (N=13510) reference
ja (N=134028) 1.22 (1.19 - 1.25) <0.001 ***
wg2
nein (N=87730) reference
ja (N=59808) 1.10 (1.08 - 1.11) <0.001 ***
wg3
nein (N=144797) reference
ja (N=2741) 1.08 (1.04 - 1.13) <0.001 ***
wg4
nein (N=119803) reference
ja (N=27735) 1.05 (1.03 - 1.07) <0.001 ***
wg5
nein (N=122740) reference
ja (N=24798) 1.01 (0.99 - 1.02) 0.419
wg6
nein (N=136513) reference
ja (N=11025) 1.02 (1.00 - 1.05) 0.037 *
kanal_print
nein (N=77983) reference
ja (N=69555) 1.00 (0.99 - 1.02) 0.735
sortiment_print
nein (N=35704) reference
ja (N=111834) 1.37 (1.34 - 1.39) <0.001 ***
# Events: 105463; Global p-value (Log-Rank): 0
AIC: 2366789.37; Concordance Index: 0.67
1
2
3
4
5
6

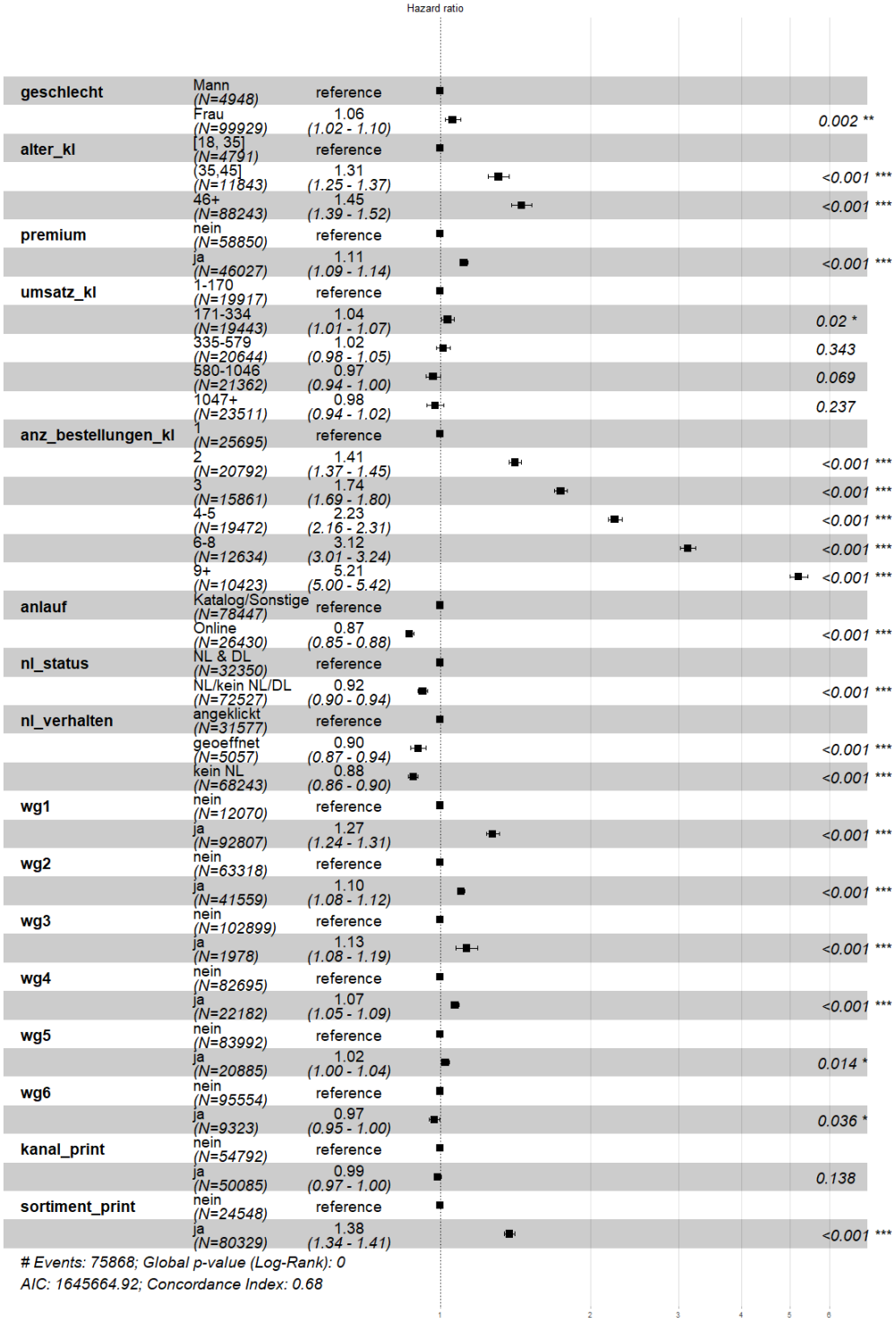
## Juni


Hazard ratio
geschlecht
Mann (N=7253) reference
Frau (N=161619) 1.01 (0.98 - 1.04) 0.518
alter_kl
[18, 35] (N=6797) reference
(35,45] (N=20602) 1.27 (1.22 - 1.32) <0.001 ***
46+ (N=141473) 1.45 (1.40 - 1.51) <0.001 ***
premium
nein (N=101535) reference
ja (N=67337) 1.07 (1.05 - 1.08) <0.001 ***
umsatz_kl
1-170 (N=34276) reference
171-334 (N=34400) 1.01 (0.99 - 1.03) 0.52
335-579 (N=35003) 0.99 (0.97 - 1.02) 0.535
580-1046 (N=33547) 0.96 (0.94 - 0.99) 0.008 **
1047+ (N=31646) 0.94 (0.92 - 0.97) <0.001 ***
anz_bestellungen_kl
1 (N=48738) reference
2 (N=36760) 1.35 (1.32 - 1.37) <0.001 ***
3 (N=25385) 1.68 (1.64 - 1.72) <0.001 ***
4-5 (N=28651) 2.13 (2.08 - 2.18) <0.001 ***
6-8 (N=16944) 2.95 (2.87 - 3.03) <0.001 ***
9+ (N=12394) 4.91 (4.75 - 5.06) <0.001 ***
anlauf
Katalog/Sonstige (N=128926) reference
Online (N=39946) 0.88 (0.87 - 0.90) <0.001 ***
nl_status
NL & DL (N=51578) reference
NL/kein NL/DL (N=117294) 0.90 (0.89 - 0.91) <0.001 ***
nl_verhalten
angeklickt (N=47746) reference
geoeffnet (N=8887) 0.88 (0.86 - 0.90) <0.001 ***
kein NL (N=112239) 0.89 (0.88 - 0.91) <0.001 ***
wg1
nein (N=19476) reference
ja (N=149396) 1.20 (1.17 - 1.23) <0.001 ***
wg2
nein (N=110017) reference
ja (N=58855) 1.10 (1.08 - 1.11) <0.001 ***
wg3
nein (N=166333) reference
ja (N=2539) 1.09 (1.05 - 1.14) <0.001 ***
wg4
nein (N=139198) reference
ja (N=29674) 1.04 (1.02 - 1.05) <0.001 ***
wg5
nein (N=141554) reference
ja (N=27318) 1.04 (1.03 - 1.06) <0.001 ***
wg6
nein (N=157432) reference
ja (N=11440) 1.01 (0.99 - 1.04) 0.223
kanal_print
nein (N=85715) reference
ja (N=83157) 0.98 (0.97 - 0.99) 0.006 **
sortiment_print
nein (N=36805) reference
ja (N=132067) 1.36 (1.34 - 1.38) <0.001 ***
# Events: 119242; Global p-value (Log-Rank): 0
AIC: 2714643.93; Concordance Index: 0.66
1
2
3
4
5
6

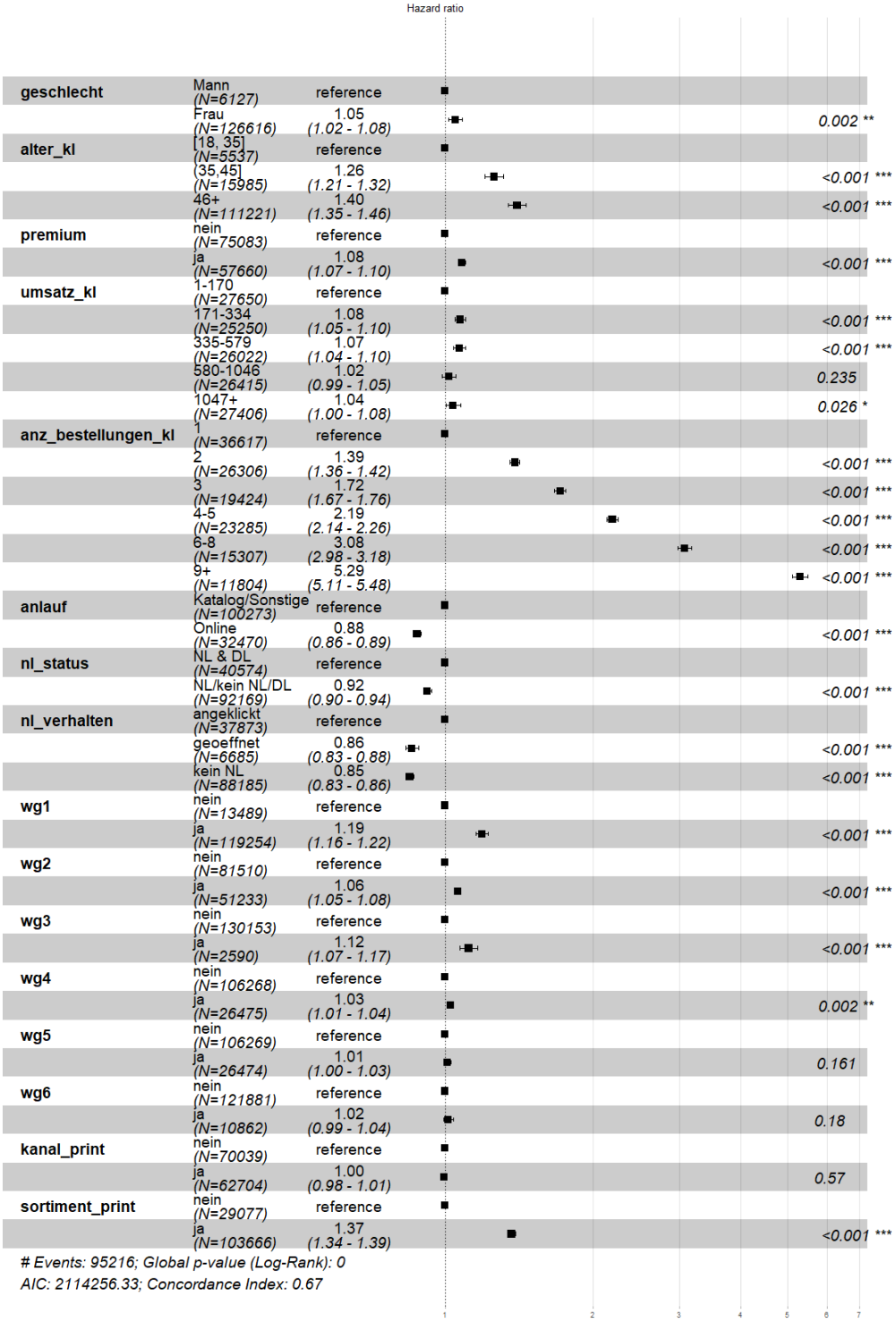
## Juli


Hazard ratio
geschlecht
Mann (N=7426) reference
Frau (N=157320) 1.05 (1.02 - 1.08) <0.001 ***
alter_kl
[18, 35] (N=6668) reference
(35,45] (N=19992) 1.28 (1.23 - 1.33) <0.001 ***
46+ (N=138086) 1.45 (1.40 - 1.51) <0.001 ***
premium
nein (N=101944) reference
ja (N=62802) 1.14 (1.12 - 1.15) <0.001 ***
umsatz_kl
1-170 (N=36639) reference
171-334 (N=33080) 1.04 (1.02 - 1.06) <0.001 ***
335-579 (N=32715) 0.98 (0.96 - 1.01) 0.147
580-1046 (N=31265) 0.95 (0.93 - 0.98) <0.001 ***
1047+ (N=31047) 0.93 (0.90 - 0.96) <0.001 ***
anz_bestellungen_kl
1 (N=50116) reference
2 (N=34453) 1.50 (1.46 - 1.53) <0.001 ***
3 (N=23514) 1.91 (1.87 - 1.96) <0.001 ***
4-5 (N=27368) 2.50 (2.44 - 2.56) <0.001 ***
6-8 (N=16602) 3.46 (3.36 - 3.56) <0.001 ***
9+ (N=12693) 5.96 (5.78 - 6.16) <0.001 ***
anlauf
Katalog/Sonstige (N=125273) reference
Online (N=39473) 0.88 (0.87 - 0.89) <0.001 ***
nl_status
NL & DL (N=51974) reference
NL/kein NL/DL (N=112772) 0.90 (0.88 - 0.91) <0.001 ***
nl_verhalten
angeklickt (N=49581) reference
geoeffnet (N=8251) 0.89 (0.87 - 0.92) <0.001 ***
kein NL (N=106914) 0.89 (0.88 - 0.91) <0.001 ***
wg1
nein (N=20188) reference
ja (N=144558) 1.26 (1.23 - 1.29) <0.001 ***
wg2
nein (N=105551) reference
ja (N=59195) 1.11 (1.09 - 1.12) <0.001 ***
wg3
nein (N=161726) reference
ja (N=3020) 1.05 (1.01 - 1.09) 0.02 *
wg4
nein (N=133323) reference
ja (N=31423) 1.02 (1.01 - 1.04) 0.004 **
wg5
nein (N=132904) reference
ja (N=31842) 1.02 (1.01 - 1.04) 0.004 **
wg6
nein (N=151716) reference
ja (N=13030) 1.03 (1.01 - 1.06) 0.002 **
kanal_print
nein (N=89499) reference
ja (N=75247) 1.02 (1.01 - 1.03) 0.007 **
sortiment_print
nein (N=35036) reference
ja (N=129710) 1.34 (1.31 - 1.36) <0.001 ***
# Events: 113549; Global p-value (Log-Rank): 0
AIC: 2573219.07; Concordance Index: 0.69
1
2
3
4
5
6
7
8

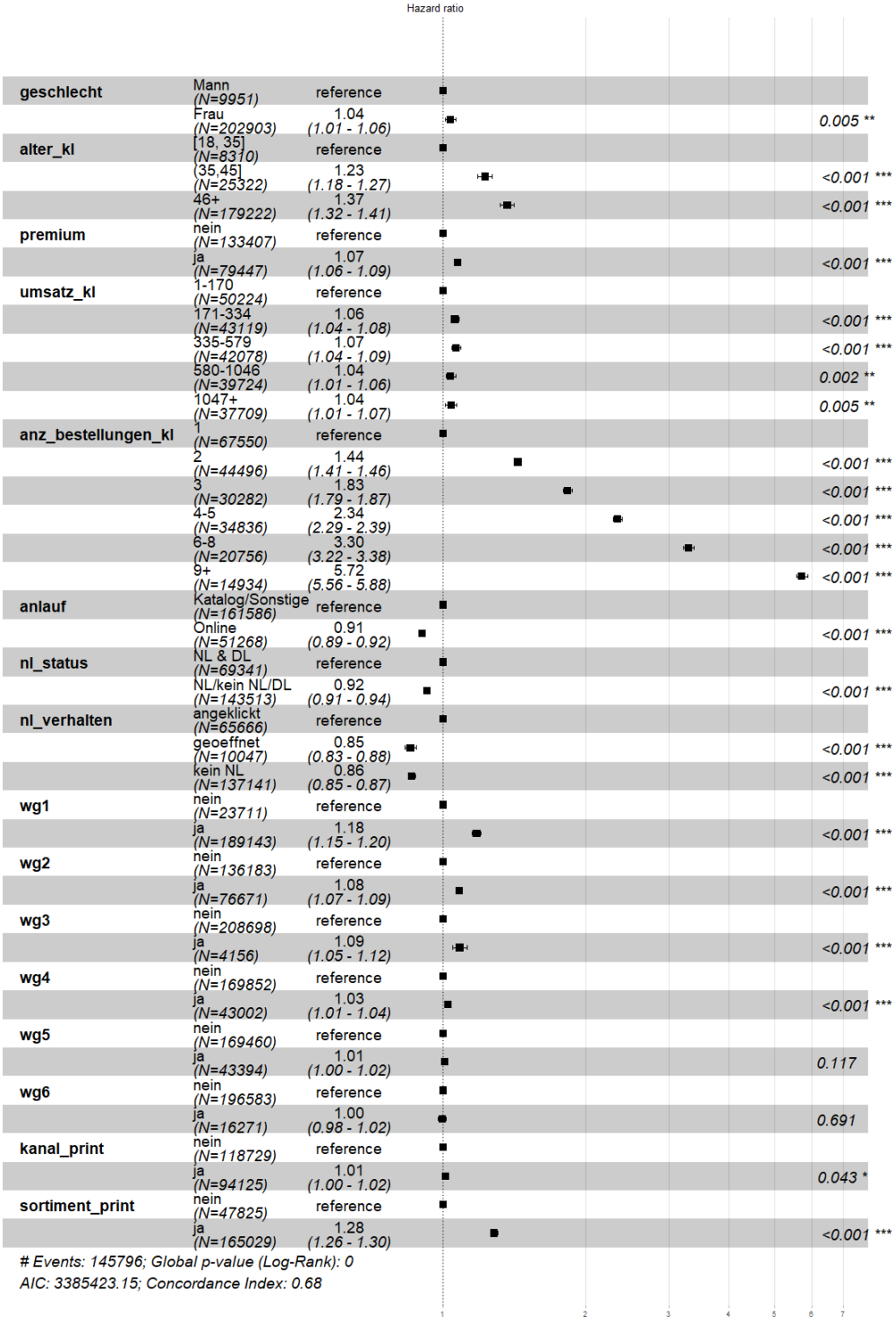
## August


Hazard ratio
geschlecht
Mann (N=4948) reference
Frau (N=99929) 1.06 (1.02 - 1.10) 0.002 **
alter_kl
[18, 35] (N=4791) reference
(35,45] (N=11843) 1.31 (1.25 - 1.37) <0.001 ***
46+ (N=88243) 1.45 (1.39 - 1.52) <0.001 ***
premium
nein (N=58850) reference
ja (N=46027) 1.11 (1.09 - 1.14) <0.001 ***
umsatz_kl
1-170 (N=19917) reference
171-334 (N=19443) 1.04 (1.01 - 1.07) 0.02 *
335-579 (N=20644) 1.02 (0.98 - 1.05) 0.343
580-1046 (N=21362) 0.97 (0.94 - 1.00) 0.069
1047+ (N=23511) 0.98 (0.94 - 1.02) 0.237
anz_bestellungen_kl
1 (N=25695) reference
2 (N=20792) 1.41 (1.37 - 1.45) <0.001 ***
3 (N=15861) 1.74 (1.69 - 1.80) <0.001 ***
4-5 (N=19472) 2.23 (2.16 - 2.31) <0.001 ***
6-8 (N=12634) 3.12 (3.01 - 3.24) <0.001 ***
9+ (N=10423) 5.21 (5.00 - 5.42) <0.001 ***
anlauf
Katalog/Sonstige (N=78447) reference
Online (N=26430) 0.87 (0.85 - 0.88) <0.001 ***
nl_status
NL & DL (N=32350) reference
NL/kein NL/DL (N=72527) 0.92 (0.90 - 0.94) <0.001 ***
nl_verhalten
angeklickt (N=31577) reference
geoeffnet (N=5057) 0.90 (0.87 - 0.94) <0.001 ***
kein NL (N=68243) 0.88 (0.86 - 0.90) <0.001 ***
wg1
nein (N=12070) reference
ja (N=92807) 1.27 (1.24 - 1.31) <0.001 ***
wg2
nein (N=63318) reference
ja (N=41559) 1.10 (1.08 - 1.12) <0.001 ***
wg3
nein (N=102899) reference
ja (N=1978) 1.13 (1.08 - 1.19) <0.001 ***
wg4
nein (N=82695) reference
ja (N=22182) 1.07 (1.05 - 1.09) <0.001 ***
wg5
nein (N=83992) reference
ja (N=20885) 1.02 (1.00 - 1.04) 0.014 *
wg6
nein (N=95554) reference
ja (N=9323) 0.97 (0.95 - 1.00) 0.036 *
kanal_print
nein (N=54792) reference
ja (N=50085) 0.99 (0.97 - 1.00) 0.138
sortiment_print
nein (N=24548) reference
ja (N=80329) 1.38 (1.34 - 1.41) <0.001 ***
# Events: 75868; Global p-value (Log-Rank): 0
AIC: 1645664.92; Concordance Index: 0.68
1
2
3
4
5
6

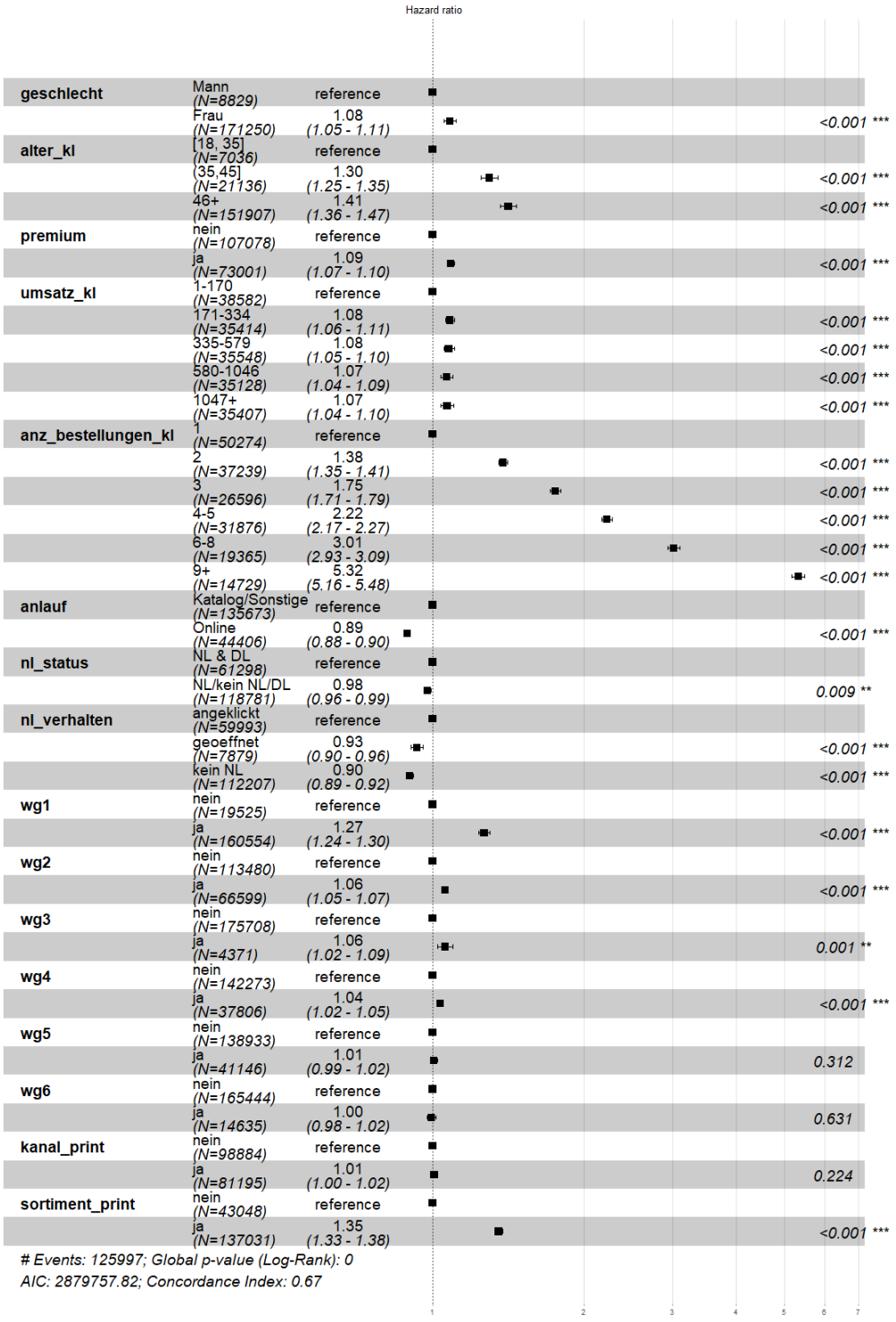
## September


Hazard ratio
geschlecht
Mann (N=6127) reference
Frau (N=126616) 1.05 (1.02 - 1.08) 0.002 **
alter_kl
[18, 35] (N=5537) reference
(35,45] (N=15985) 1.26 (1.21 - 1.32) <0.001 ***
46+ (N=111221) 1.40 (1.35 - 1.46) <0.001 ***
premium
nein (N=75083) reference
ja (N=57660) 1.08 (1.07 - 1.10) <0.001 ***
umsatz_kl
1-170 (N=27650) reference
171-334 (N=25250) 1.08 (1.05 - 1.10) <0.001 ***
335-579 (N=26022) 1.07 (1.04 - 1.10) <0.001 ***
580-1046 (N=26415) 1.02 (0.99 - 1.05) 0.235
1047+ (N=27406) 1.04 (1.00 - 1.08) 0.026 *
anz_bestellungen_kl
1 (N=36617) reference
2 (N=26306) 1.39 (1.36 - 1.42) <0.001 ***
3 (N=19424) 1.72 (1.67 - 1.76) <0.001 ***
4-5 (N=23285) 2.19 (2.14 - 2.26) <0.001 ***
6-8 (N=15307) 3.08 (2.98 - 3.18) <0.001 ***
9+ (N=11804) 5.29 (5.11 - 5.48) <0.001 ***
anlauf
Katalog/Sonstige (N=100273) reference
Online (N=32470) 0.88 (0.86 - 0.89) <0.001 ***
nl_status
NL & DL (N=40574) reference
NL/kein NL/DL (N=92169) 0.92 (0.90 - 0.94) <0.001 ***
nl_verhalten
angeklickt (N=37873) reference
geoeffnet (N=6685) 0.86 (0.83 - 0.88) <0.001 ***
kein NL (N=88185) 0.85 (0.83 - 0.86) <0.001 ***
wg1
nein (N=13489) reference
ja (N=119254) 1.19 (1.16 - 1.22) <0.001 ***
wg2
nein (N=81510) reference
ja (N=51233) 1.06 (1.05 - 1.08) <0.001 ***
wg3
nein (N=130153) reference
ja (N=2590) 1.12 (1.07 - 1.17) <0.001 ***
wg4
nein (N=106268) reference
ja (N=26475) 1.03 (1.01 - 1.04) 0.002 **
wg5
nein (N=106269) reference
ja (N=26474) 1.01 (1.00 - 1.03) 0.161
wg6
nein (N=121881) reference
ja (N=10862) 1.02 (0.99 - 1.04) 0.18
kanal_print
nein (N=70039) reference
ja (N=62704) 1.00 (0.98 - 1.01) 0.57
sortiment_print
nein (N=29077) reference
ja (N=103666) 1.37 (1.34 - 1.39) <0.001 ***
# Events: 95216; Global p-value (Log-Rank): 0
AIC: 2114256.33; Concordance Index: 0.67
1
2
3
4
5
6
7

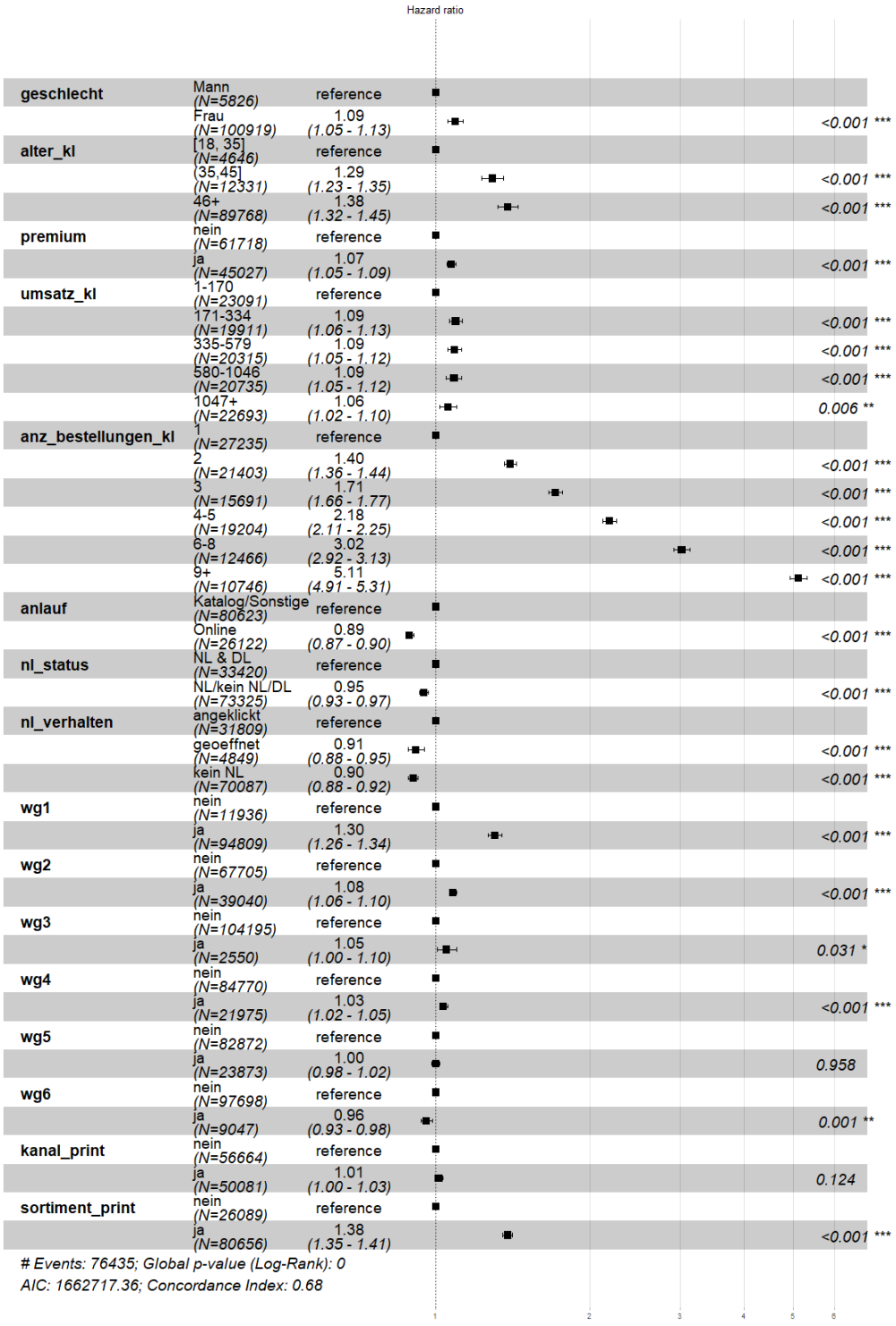
## Oktober


Hazard ratio
geschlecht
Mann (N=9951) reference
Frau (N=202903) 1.04 (1.01 - 1.06) 0.005 **
alter_kl
[18, 35] (N=8310) reference
(35,45] (N=25322) 1.23 (1.18 - 1.27) <0.001 ***
46+ (N=179222) 1.37 (1.32 - 1.41) <0.001 ***
premium
nein (N=133407) reference
ja (N=79447) 1.07 (1.06 - 1.09) <0.001 ***
umsatz_kl
1-170 (N=50224) reference
171-334 (N=43119) 1.06 (1.04 - 1.08) <0.001 ***
335-579 (N=42078) 1.07 (1.04 - 1.09) <0.001 ***
580-1046 (N=39724) 1.04 (1.01 - 1.06) 0.002 **
1047+ (N=37709) 1.04 (1.01 - 1.07) 0.005 **
anz_bestellungen_kl
1 (N=67550) reference
2 (N=44496) 1.44 (1.41 - 1.46) <0.001 ***
3 (N=30282) 1.83 (1.79 - 1.87) <0.001 ***
4-5 (N=34836) 2.34 (2.29 - 2.39) <0.001 ***
6-8 (N=20756) 3.30 (3.22 - 3.38) <0.001 ***
9+ (N=14934) 5.72 (5.56 - 5.88) <0.001 ***
anlauf
Katalog/Sonstige (N=161586) reference
Online (N=51268) 0.91 (0.89 - 0.92) <0.001 ***
nl_status
NL & DL (N=69341) reference
NL/kein NL/DL (N=143513) 0.92 (0.91 - 0.94) <0.001 ***
nl_verhalten
angeklickt (N=65666) reference
geoeffnet (N=10047) 0.85 (0.83 - 0.88) <0.001 ***
kein NL (N=137141) 0.86 (0.85 - 0.87) <0.001 ***
wg1
nein (N=23711) reference
ja (N=189143) 1.18 (1.15 - 1.20) <0.001 ***
wg2
nein (N=136183) reference
ja (N=76671) 1.08 (1.07 - 1.09) <0.001 ***
wg3
nein (N=208698) reference
ja (N=4156) 1.09 (1.05 - 1.12) <0.001 ***
wg4
nein (N=169852) reference
ja (N=43002) 1.03 (1.01 - 1.04) <0.001 ***
wg5
nein (N=169460) reference
ja (N=43394) 1.01 (1.00 - 1.02) 0.117
wg6
nein (N=196583) reference
ja (N=16271) 1.00 (0.98 - 1.02) 0.691
kanal_print
nein (N=118729) reference
ja (N=94125) 1.01 (1.00 - 1.02) 0.043 *
sortiment_print
nein (N=47825) reference
ja (N=165029) 1.28 (1.26 - 1.30) <0.001 ***
# Events: 145796; Global p-value (Log-Rank): 0
AIC: 3385423.15; Concordance Index: 0.68
1
2
3
4
5
6
7

## November


Hazard ratio
geschlecht
Mann (N=8829) reference
Frau (N=171250) 1.08 (1.05 - 1.11) <0.001 ***
alter_kl
[18, 35] (N=7036) reference
(35,45] (N=21136) 1.30 (1.25 - 1.35) <0.001 ***
46+ (N=151907) 1.41 (1.36 - 1.47) <0.001 ***
premium
nein (N=107078) reference
ja (N=73001) 1.09 (1.07 - 1.10) <0.001 ***
umsatz_kl
1-170 (N=38582) reference
171-334 (N=35414) 1.08 (1.06 - 1.11) <0.001 ***
335-579 (N=35548) 1.08 (1.05 - 1.10) <0.001 ***
580-1046 (N=35128) 1.07 (1.04 - 1.09) <0.001 ***
1047+ (N=35407) 1.07 (1.04 - 1.10) <0.001 ***
anz_bestellungen_kl
1 (N=50274) reference
2 (N=37239) 1.38 (1.35 - 1.41) <0.001 ***
3 (N=26596) 1.75 (1.71 - 1.79) <0.001 ***
4-5 (N=31876) 2.22 (2.17 - 2.27) <0.001 ***
6-8 (N=19365) 3.01 (2.93 - 3.09) <0.001 ***
9+ (N=14729) 5.32 (5.16 - 5.48) <0.001 ***
anlauf
Katalog/Sonstige (N=135673) reference
Online (N=44406) 0.89 (0.88 - 0.90) <0.001 ***
nl_status
NL & DL (N=61298) reference
NL/kein NL/DL (N=118781) 0.98 (0.96 - 0.99) 0.009 **
nl_verhalten
angeklickt (N=59993) reference
geoeffnet (N=7879) 0.93 (0.90 - 0.96) <0.001 ***
kein NL (N=112207) 0.90 (0.89 - 0.92) <0.001 ***
wg1
nein (N=19525) reference
ja (N=160554) 1.27 (1.24 - 1.30) <0.001 ***
wg2
nein (N=113480) reference
ja (N=66599) 1.06 (1.05 - 1.07) <0.001 ***
wg3
nein (N=175708) reference
ja (N=4371) 1.06 (1.02 - 1.09) 0.001 **
wg4
nein (N=142273) reference
ja (N=37806) 1.04 (1.02 - 1.05) <0.001 ***
wg5
nein (N=138933) reference
ja (N=41146) 1.01 (0.99 - 1.02) 0.312
wg6
nein (N=165444) reference
ja (N=14635) 1.00 (0.98 - 1.02) 0.631
kanal_print
nein (N=98884) reference
ja (N=81195) 1.01 (1.00 - 1.02) 0.224
sortiment_print
nein (N=43048) reference
ja (N=137031) 1.35 (1.33 - 1.38) <0.001 ***
# Events: 125997; Global p-value (Log-Rank): 0
AIC: 2879757.82; Concordance Index: 0.67
1
2
3
4
5
6
7

## Dezember


Hazard ratio
geschlecht
Mann (N=5826) reference
Frau (N=100919) 1.09 (1.05 - 1.13) <0.001 ***
alter_kl
[18, 35] (N=4646) reference
(35,45] (N=12331) 1.29 (1.23 - 1.35) <0.001 ***
46+ (N=89768) 1.38 (1.32 - 1.45) <0.001 ***
premium
nein (N=61718) reference
ja (N=45027) 1.07 (1.05 - 1.09) <0.001 ***
umsatz_kl
1-170 (N=23091) reference
171-334 (N=19911) 1.09 (1.06 - 1.13) <0.001 ***
335-579 (N=20315) 1.09 (1.05 - 1.12) <0.001 ***
580-1046 (N=20735) 1.09 (1.05 - 1.12) <0.001 ***
1047+ (N=22693) 1.06 (1.02 - 1.10) 0.006 **
anz_bestellungen_kl
1 (N=27235) reference
2 (N=21403) 1.40 (1.36 - 1.44) <0.001 ***
3 (N=15691) 1.71 (1.66 - 1.77) <0.001 ***
4-5 (N=19204) 2.18 (2.11 - 2.25) <0.001 ***
6-8 (N=12466) 3.02 (2.92 - 3.13) <0.001 ***
9+ (N=10746) 5.11 (4.91 - 5.31) <0.001 ***
anlauf
Katalog/Sonstige (N=80623) reference
Online (N=26122) 0.89 (0.87 - 0.90) <0.001 ***
nl_status
NL & DL (N=33420) reference
NL/kein NL/DL (N=73325) 0.95 (0.93 - 0.97) <0.001 ***
nl_verhalten
angeklickt (N=31809) reference
geoeffnet (N=4849) 0.91 (0.88 - 0.95) <0.001 ***
kein NL (N=70087) 0.90 (0.88 - 0.92) <0.001 ***
wg1
nein (N=11936) reference
ja (N=94809) 1.30 (1.26 - 1.34) <0.001 ***
wg2
nein (N=67705) reference
ja (N=39040) 1.08 (1.06 - 1.10) <0.001 ***
wg3
nein (N=104195) reference
ja (N=2550) 1.05 (1.00 - 1.10) 0.031 *
wg4
nein (N=84770) reference
ja (N=21975) 1.03 (1.02 - 1.05) <0.001 ***
wg5
nein (N=82872) reference
ja (N=23873) 1.00 (0.98 - 1.02) 0.958
wg6
nein (N=97698) reference
ja (N=9047) 0.96 (0.93 - 0.98) 0.001 **
kanal_print
nein (N=56664) reference
ja (N=50081) 1.01 (1.00 - 1.03) 0.124
sortiment_print
nein (N=26089) reference
ja (N=80656) 1.38 (1.35 - 1.41) <0.001 ***
# Events: 76435; Global p-value (Log-Rank): 0
AIC: 1662717.36; Concordance Index: 0.68
1
2
3
4
5
6

# C Deskriptive Beschreibung der Scoreklassen für Januar

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Geschlecht | Mann | 11.1 | 5.4 | 3.2 | 5.1 | 4.9 | 3.8 | 4.4 | 3.6 | 3.5 | 4.2 |
| | Frau | 88.9 | 94.6 | 96.8 | 94.9 | 95.1 | 96.2 | 95.6 | 96.4 | 96.5 | 95.8 |
| Alter | [18,35] | 21.1 | 4.6 | 3.1 | 4.3 | 2.6 | 2.3 | 1.4 | 1.1 | 1 | 1.2 |
| | (35,45] | 22.8 | 15.5 | 10.3 | 13.2 | 11.5 | 9.7 | 9.7 | 8.4 | 8 | 8.3 |
| | 46+ | 56.1 | 80 | 86.6 | 82.6 | 85.9 | 88 | 88.9 | 90.4 | 91 | 90.4 |
| Umsatz | 1-170 | 73.9 | 59.4 | 48.6 | 23.7 | 8.5 | 1.8 | 0.4 | 0.1 | 0 | 0 |
| | 171-334 | 17.9 | 25 | 31 | 37.8 | 30.9 | 25 | 14.9 | 4.6 | 0.8 | 0 |
| | 335-579 | 6.1 | 11 | 14.5 | 23.6 | 36.2 | 35.3 | 33.9 | 21.6 | 11.7 | 1.5 |
| | 580-1046 | 1.8 | 4 | 4.9 | 11.8 | 19.2 | 27.8 | 35.4 | 39.9 | 35.8 | 13.6 |
| | 1047+ | 0.3 | 0.6 | 1 | 3.1 | 5.2 | 10.1 | 15.4 | 33.9 | 51.7 | 84.9 |
| Anzahl | 1 | 98.1 | 90.6 | 78.3 | 15.8 | 0.1 | 0 | 0 | 0 | 0 | 0 |
| Best. | 2 | 1.9 | 8.9 | 20.2 | 71.8 | 71.6 | 15.4 | 0.5 | 0 | 0 | 0 |
| | 3 | 0 | 0.5 | 1.4 | 11.2 | 23.7 | 69.1 | 33.9 | 3.6 | 0 | 0 |
| | 4-5 | 0 | 0 | 0.1 | 1.2 | 4.5 | 15.1 | 63.5 | 81.6 | 10.1 | 0 |
| | 6-8 | 0 | 0 | 0 | 0 | 0.1 | 0.4 | 2.2 | 14.8 | 88.3 | 7.4 |
| | 9+ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.7 | 92.6 |
| Anlauf | Kat./So. | 32.2 | 69.8 | 81.1 | 73.8 | 80.6 | 82 | 83.1 | 84.8 | 85.1 | 81.2 |
| | Online | 67.8 | 30.2 | 18.9 | 26.2 | 19.4 | 18 | 16.9 | 15.2 | 14.9 | 18.8 |
| WG 1 | nein | 53 | 33.4 | 8.9 | 13 | 5 | 2.7 | 1.6 | 0.9 | 0.3 | 0.3 |
| | ja | 47 | 66.6 | 91.1 | 87 | 95 | 97.3 | 98.4 | 99.1 | 99.7 | 99.7 |
| WG 2 | nein | 90.9 | 87.2 | 86.1 | 76.7 | 67.8 | 61.8 | 55.1 | 37.8 | 29.4 | 16.1 |
| | ja | 9.1 | 12.8 | 13.9 | 23.3 | 32.2 | 38.2 | 44.9 | 62.2 | 70.6 | 83.9 |
| WG 3 | nein | 99.2 | 99.6 | 99.6 | 99.4 | 99.2 | 99 | 98.8 | 98.1 | 96.9 | 93.3 |
| | ja | 0.8 | 0.4 | 0.4 | 0.6 | 0.8 | 1 | 1.2 | 1.9 | 3.1 | 6.7 |
| WG 4 | nein | 93.6 | 93.6 | 93.3 | 90.1 | 86.3 | 85.2 | 83.4 | 74.6 | 71.2 | 57.7 |
| | ja | 6.4 | 6.4 | 6.7 | 9.9 | 13.7 | 14.8 | 16.6 | 25.4 | 28.8 | 42.3 |
| WG 5 | nein | 84.7 | 86.1 | 91.6 | 84.7 | 82.7 | 81.2 | 79.5 | 74.4 | 72 | 64.7 |
| | ja | 15.3 | 13.9 | 8.4 | 15.3 | 17.3 | 18.8 | 20.5 | 25.6 | 28 | 35.3 |
| WG 6 | nein | 93.6 | 94.6 | 96.3 | 93.5 | 93 | 92.6 | 90.7 | 89.6 | 86.9 | 84.3 |
| | ja | 6.4 | 5.4 | 3.7 | 6.5 | 7 | 7.4 | 9.3 | 10.4 | 13.1 | 15.7 |
| Sortiment | nein | 94.4 | 56.9 | 20.5 | 29.6 | 12.7 | 10.3 | 5.8 | 4.3 | 1.9 | 2.2 |
| Print | ja | 5.6 | 43.1 | 79.5 | 70.4 | 87.3 | 89.7 | 94.2 | 95.7 | 98.1 | 97.8 |
| Kanal | nein | 89.8 | 69.9 | 43.8 | 50.2 | 45.6 | 40.7 | 39.7 | 46 | 43.7 | 49.9 |
| Print | ja | 10.2 | 30.1 | 56.2 | 49.8 | 54.4 | 59.3 | 60.3 | 54 | 56.3 | 50.1 |

# D R Code für Log-Rank-Test bei unterschiedlicher Stichprobengröße

```
library(survival)
library(survminer)

set.seed(123)
nlogrank <- function(n){

          # Datensatz Gruppe1
          Time <- rexp(n, 1/180)
          Status <- rep(1, n)
          Gruppe <- rep("g1", n)
          df_g1 <- cbind.data.frame(Time, Status, Gruppe)

          # Datensatz Gruppe2
          df_g2 <- df_g1
          df_g2$Time <- df_g2$Time + 0.5
          df_g2$Gruppe <- 'g2'

          df <- rbind(df_g1, df_g2)

          # Log-Rank-Test
          survdiff(Surv(Time, Status) ~ Gruppe, data = df)

          }

nlogrank(1000)
nlogrank(10000)
nlogrank(100000)
nlogrank(500000)
nlogrank(1000000)
nlogrank(2500000)
nlogrank(5000000)
```

# E Kaplan-Meier-Kurven von Sortiment Online und Katalog Online für 12 Monate