Statistisches Consulting

Master Studiengang Statistik
Institut für Statistik
Ludwig-Maximilians-Universität München

Online-Marketing der Interhyp AG
Analyse von Tracking-Daten

Daniel Fuckner d.fuckner@gmx.de
Markus Vogler markus@vogler-lindau.de

Projektpartner:
Nicola Gries (Interhyp AG)

Betreuer:
Dr. Fabian Scheipl

München, 26.08.2014

Abstract

Die Interhyp AG ist Vermittler für private Baufinanzierungen. Das primäre Ziel des Marketing der Interhyp AG ist die Kundenakquise. Da etwa 80% aller Kundenanträge online abgeschickt werden, liegt der Fokus auf dem Online-Marketing, das über verschiedene Kampagnen verfügt. Beispiele sind Kooperationen mit anderen Unternehmen, bezahlte Anzeigen bei Suchmaschinen oder Bannerschaltungen.
Die Refined Labs GmbH ist verantwortlich für das Online-Tracking der Werbekampagnen der Interhyp AG. Durch Online-Tracking werden die Werbekontakte eines potentiellen Kunden zu einem sogenannten Funnel zusammengefasst. Am Ende eines jeden Funnels steht das Ausfüllen eines Onlineantrages oder der Abbruch. Man spricht von konvertierten beziehungsweise nicht-konvertierten Funnels.
In dieser Arbeit werden die Tracking-Daten zunächst anhand deskriptiver Analysen vorgestellt. Außerdem werden Methodik und Ergebnisse eines zeitdiskreten Survival-Modells und eines Sequential Pattern Mining-Algorithmus beschrieben. Mit Hilfe des zeitdiskreten Survival-Modells, das mit Stochastic Gradient Boosting geschätzt wurde, konnten Zusammenhänge zwischen den erklärenden Variablen und der Wahrscheinlichkeit für eine Konvertierung gefunden werden. Häufige Abfolgen von Kampagnen in den konvertierten und nicht-konvertierten Funnels wurden mit dem Sequential Pattern Mining-Algorithmus entdeckt. Zudem wurden die Funnels in Form eines Netzwerkes visualisiert. Die Daten für dieses Netzwerk und ein Programm, dass die interaktive Betrachtung ermöglicht, ist im elektronischen Anhang enthalten.

Inhaltsverzeichnis

1 Einleitung

Die Interhyp AG ist Vermittler für private Baufinanzierungen. Das heißt, sie wählt aus einem Angebot von verschiedenen Darlehensgebern die optimale Finanzierungsstruktur für einen Kunden aus. Das Unternehmen wurde 1999 von den ehemaligen Goldman-Sachs-Bankern Robert Haselsteiner und Marcus Wolsdorf gegründet. Sechs Jahre später eröffnete die Interhyp AG erste Niederlassungen und konnte gleichzeitig den erfolgreichsten deutschen Börsengang des Jahres verzeichnen. Nach weiteren drei Jahren erfolgte die Übernahme durch ING DIRECT, der weltweilt größten und erfolgreichsten Direktbanken-Gruppe. Heute ist die Interhyp AG der größte Vermittler für private Baufinanzierungen in Deutschland, wurde acht mal in Folge als "Bester Baufinananzierer"(Zeitschrift €, Ausgabe 08/2013) ausgezeichnet und verfügt über mehr als 60 Beratungsstandorte mit über 1.000 Mitarbeitern.
Das primäre Ziel des Marketing der Interhyp AG ist die Kundenakquise. Da etwa 80% aller Kundenanträge online abgeschickt werden, liegt der Fokus der Marketing-Abteilung auf dem Online-Marketing, das über verschiedene Kanäle verfügt. Beispiele sind Kooperationen mit anderen Unternehmen wie Immobilienscout24, bezahlte Anzeigen bei Suchmaschinen, Newsletter oder diverse Bannerschaltungen. Durch Online-Tracking können die Werbekontakte eines potentiellen Kunden mit der Interhyp AG zusammengefasst werden. So entsteht ein sogenannter Funnel, wie es in Abbildung 1 skizziert ist. Jeder potentielle Kunde hat einen oder mehrere aufeinanderfolgende Kontaktpunkte, wobei jeder Kontaktpunkt die genaue Zeit des Kontaktes sowie die Art des Kontaktes, das heißt die Information über welche Kampagne es zu dem Kontakt kam, enthält. Am Ende eines jeden Funnels steht der Abbruch der Beobachtung oder im Idealfall das Ausfüllen eines Onlineantrages durch den Kunden.

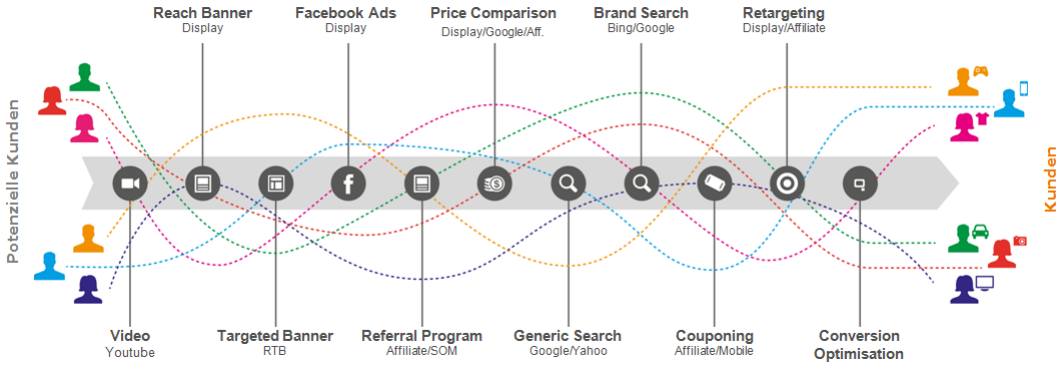

Abbildung 1: Entstehung eines Funnels (Quelle: Interhyp AG)

Die Refined Labs GmbH ist auf dem Gebiet des Online-Marketing spezialisiert und verantwortlich für das Online-Tracking der Werbekampagnen der Interhyp AG. Ein Funnel beginnt mit dem ersten Online-Werbekontakt eines potentiellen Kunden mit der Interhyp AG und der damit einhergehenden Erstellung eines Cookies. So können alle weiteren Werbekontakte dem potentiellen Kunden eindeutig zugewiesen werden. Das

Tracking endet sobald der potentielle Kunde einen Onlineantrag versendet und damit zum Kunden wird. In diesem Fall spricht man von einem konvertierten Funnel. Wird innerhalb von 90 Tagen kein Onlineantrag versendet, so wird das Cookie nicht weiter verfolgt und der Funnel wird als nicht-konvertiert bezeichnet.
Primäres Ziel dieses Projektes ist das Entdecken von Unterschieden zwischen konvertierten und nicht-konvertierten Funnels. Um dieser Fragestellung gerecht zu werden, wurde ein zeitdiskretes Survival-Modell mittels Stochastic Gradient Boosting geschätzt und ein Sequential Pattern Mining-Algorithmus angewendet. Außerdem wurden die Funnels anhand eines Netzwerkes visualisiert.
Die Arbeit ist wie folgt gegliedert. In Kapitel 2 und 3 werden die Datenaufbereitung und die Variablen erklärt. Daraufhin wird die Methodik des Survival-Modells in Kapitel 4, der Sequential Pattern Mining-Algorithmus in Kapitel 5 und das Netzwerk in Kapitel 6 erläutert. Die Ergebnisse dieser Methoden werden in Kapitel 7 vorgestellt und abschließend erfolgt eine Zusammenfassung in Kapitel 8 und die Beschreibung des elektronischen Anhangs in Kapitel 9.

2 Datenlage

Die Daten wurden von der Refined Labs GmbH als SQL-Dump bereitgestellt, der eine Größe von circa 13 Gigabyte hat. Die MySQL-Datenbank enthält die vier Tabellen *project_ out*, *redirects_ short*, *searchFunnel* und *stage2_ transactionHandling*. Mit Hilfe der vorhanden Informationen in *searchFunnel* und *stage2_ transactionHandling* konnten die Kontaktpunkte in *redirects_ short* in konvertierte und nicht-konvertierte Funnels unterteilt werden. In *projects_ out* sind die Kampagnen in Form einer Baumstruktur organisiert. In Absprache mit der Interhyp AG wurden 17 Kategorien ausgewählt, die sich auf den ersten drei Ebenen dieser Baumstruktur befinden. Anhand von IDs wurde jedem Kontaktpunkt eine dieser Kategorien zugewiesen.

ID	Campaign	Transaction	Position	...
1	Affiliate - Partnerprogramm	0	1	...
1	SEM - Brand	0	2	...
1	Direct	0	3	...
1	Direct	1	4	...
2	Display	0	1	...
2	SEM - Generisch	0	2	...
2	Social Media	0	3	...

Tabelle 1: Beispiel für einen Auszug aus der Datenbank

Tabelle 1 enthält ein Datenbeispiel mit den Spalten *ID*, *Campaign*, *Transaction* und *Position*. Das Beispiel enthält zwei verschiedene *IDs*, das heißt zwei Funnels, wobei der

erste vier und der zweite drei Kontaktpunkte hat. Die Spalte *Campaign* enthält die Kampagne der Kontaktpunkte. *Transaction* ist eine binäre Variable und gibt an, ob der Kunde konvertiert ist, wobei der Wert 1 nur für den letzten Kontaktpunkt vor der Konvertiertung angenommen wird. Das heißt bei *ID* 1 handelt es sich um einen konvertierten und bei *ID* 2 um einen nicht-konvertierten Funnel. Die *Position* gibt die Nummer des Kontaktpunktes an und reicht damit von 1 bis zur Länge des Funnels.
Nun sollen noch die verschiedenen Kategorien des Features *campagin*, das heißt die verschiedenen Werbeformen betrachtet werden. Tabelle 2 enthält Erklärungen der 17 verwendeten Kategorien.

Kampagne	**Beschreibung**
Affiliate - Partnerprogramm	Partner, die Werbemittel einbinden
Affiliate - Rest	Partner, die Zinsvergleich bereitstellen
Direct	Direkte Eingabe von *www.interhyp.de*
Display	Bannerschaltungen
E-Mailing	Mails an Interessenten, die schon einen Antrag o.ä. gestellt haben
Generic	Unbezahlter Link
Kooperationen - Focus Kooperationen - Immonet Kooperationen - Immoscout24 Kooperationen - Immowelt Kooperationen - Rest	Individuelle Zusammenarbeit mit größeren Partnern
Newsletter	Regelmäßige Rundschreiben
SEM - Brand SEM - Remarketing SEM - Generisch	Bezahlte Suchergebnisse
SEO	Unbezahlte Suchergebnisse
Social Media	*facebook* und *gutefrage.net*

Tabelle 2: Beschreibung der Kampagnen

Affiliate - Partnerprogramm sind Partner, die von der Interhyp AG bereitgestellte Werbemittel wie Rechner, Logo oder Banner einbinden. Partner, die einen Zinsvergleich bereitstellen, welcher das Zinsangebot der Interhyp AG mit deren Wettbewerbern im Vergleich dargestellt, gehören *Affiliate - Rest* an. *Direct* bedeutet, dass ein potentieller Kunde im Browser direkt *www.interhyp.de* eingibt und *Display* sind Bannerschaltungen auf diversen Webseiten. *E-Mailing* umfasst Mails an Interessenten, die bereits einen Antrag gestellt oder ein Infopaket angefordert haben. Wenn ein potentieller Kunde über einen unbezahlten Link zur Interhyp AG gelangt, so wird der Kontaktpunkt der Kampagne *Generic* zugewiesen. *Kooperationen* sind individuelle Zusammenarbeiten mit größeren Partnern, die je nach Vertrag verschiedene Werbemittel auf ihrer Seite einbinden

und *Newsletter* sind regelmäßige Rundschreiben. *SEM* sind bezahlte Suchergebnisse, wobei *Brand* bedeutet, dass die Suchanfrage das Wort *Interhyp* beinhaltete, *Generisch* bedeutet, dass etwas wie *Baufinanzierung* oder ähnliches gesucht wurde und *Remarketing* bedeutet, dass der potentielle Kunde bereits zuvor auf der Seite der Interhyp AG war und deshalb erneut eine Werbeeinblendung der geschaltet wurde. Unbezahlte Suchergebnisse werden *SEO* zugewiesen und *Social Media* umfasst Werbung auf *Facebook* und *gutefrage.net*.
Ein Kontaktpunkt mit einer der 17 Kampagnen tritt entweder in der Form eines *Clicks* oder eines *Views* auf. Man spricht von einem *Click*, wenn der potentielle Kunde tatsächlich etwas angeklickt hat und ein *View* wird getrackt, wenn ein Banner oder ähnliches lediglich gesehen, aber nicht angeklickt wird. An dieser Stelle wirft die Datenerhebung allerdings ein Problem für die statistischen Analysen auf. Die *Views* werden für alle konvertierten Funnels gespeichert, für die nicht-konvertierten Funnels allerdings nur, wenn diese bei einem anderen Kunden der Refined Labs GmbH konvertieren. Dass heißt, es ist eine systematische Veränderung der Daten gegeben. Deshalb besteht keine Möglichkeit die *Views* in statistische Analysen, die konvertierte und nicht-konvertierte Funnels vergleichen, einzubeziehen. Die *Views* werden lediglich in Kapitel 3 in einigen Plots betrachtet, die nur konvertierte Funnels enthalten, und von den weiteren Analysen ausgeschlossen.
Nach der Vorverarbeitung der Daten, die mit *MySQL 6.1* und *R 3.0.2* [23] durchgeführt wurde, liegen $297,963$ *Clicks* für die konvertierten und $9,550,802$ *Clicks* für die nicht-konvertierten Funnels vor. Zum Handhaben der großen Datenmenge wurden die Pakete *data.table* [6] und *plyr* [28] verwendet. Eine nähere Beschreibung der in *R* erstellten Features erfolgt in Kapitel 3.

3 Deskriptive Analyse

Tabelle 3 enthält eine Übersicht mit den erzeugten Variablen. Diese werden in diesem Kapitel näher betrachtet. Als Position wird im folgenden die Nummer eines Kontaktpunktes innerhalb eines Funnels bezeichnet. Das heißt für jeden Funnel nimmt der erste Kontakt die Position 1 an. Die Abbildungen wurden mit dem *R*-Paket *ggplot2* [27] erzeugt.

$clickCount \in \mathbb{N}$	Anzahl an *Clicks* bis zur aktuellen Position
$hasClicked \in \{0, 1\}$	Dummyvariable, die angibt ob vor der aktuellen Position schon geklicked wurde (1) oder nicht (0).
$campaign$	Kampagne der aktuellen Position
$campaignLast$	Kampagne der vorherigen Position
$campaignLast2$	Kampagne der vorletzten Position
$weekday \in \{Montag, ...\}$	Wochentag des Kontaktes
$hour \in \{0, 1, \ldots, 23\}$	Uhrzeit des Kontaktes
$timeSinceLast \in \mathbb{R}$	Zeitdifferenz zwischen aktueller und vorheriger Position
$timeSinceFirst \in \mathbb{R}$	Zeitdifferenz zwischen aktueller und erster Position
$freq \in \mathbb{R}$	Frequenz der Kontaktpunkte in einem Funnel

Tabelle 3: Variablenbeschreibung

3.1 Views in den konvertierten Funnels

clickCount

Aufgrund der in Kapitel 2 beschriebenen Problematik bei der Datenerhebung der *Views* können die Variablen *clickCount* und *hasClicked* nur in den konvertierten Funnels betrachtet werden.

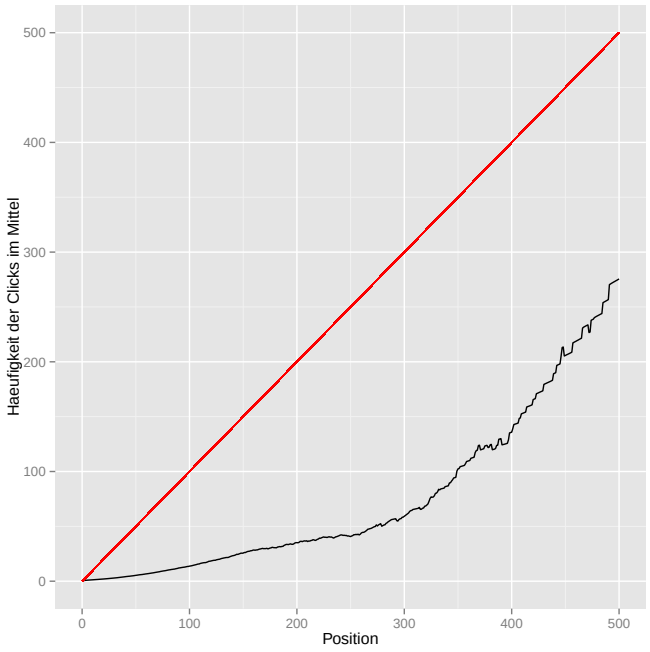

Abbildung 2: Häufigkeit der Clicks im Mittel für jede Position

In Abbildung 2 ist der *clickCount* dargestellt. Auf der x-Achse ist die Position aufgetra-

gen und auf der y-Achse der *clickCount*, das heißt die Häufigkeit der *Clicks* gemittelt über alle konvertierten Funnels für jede Position. Die rote Diagonale wäre erreicht, wenn die konvertierten Funnels nur aus *Clicks* bestehen würden. Es ist zu erkennen, dass die Linie mit der Position ansteigt. An Position 100 ist die mittlere Anzahl der Clicks 13.7, das heißt im Mittel bestehen die ersten 100 Kontakte eines Funnels aus 14 *Clicks* und 86 *Views*. Die Anzahl der *Views* übersteigt die Anzahl der *Clicks* also deutlich.

hasClicked

Die Variable *hasClicked* (siehe Abbildung 3) gibt für jede Position den Anteil der Funnels an, die bis dorthin mindestens einen *Click* enthalten. Dieser Wert nimmt zwischen Position 1 und Position 7 ab. Dies ist dadurch zu erklären, dass es viele Funnels gibt, die *Clicks* enthalten und deren Länge kleiner als 6 ist. Sobald diese Funnels beendet sind, werden sie an der nächsten Position selbstverständlich nicht mehr berücksichtigt. Ab Position 7 steigt die Kurve dann bis zur 1 an. Ein Wert von 1 bedeutet, dass alle Funnels bereits einen *Click* hatten.

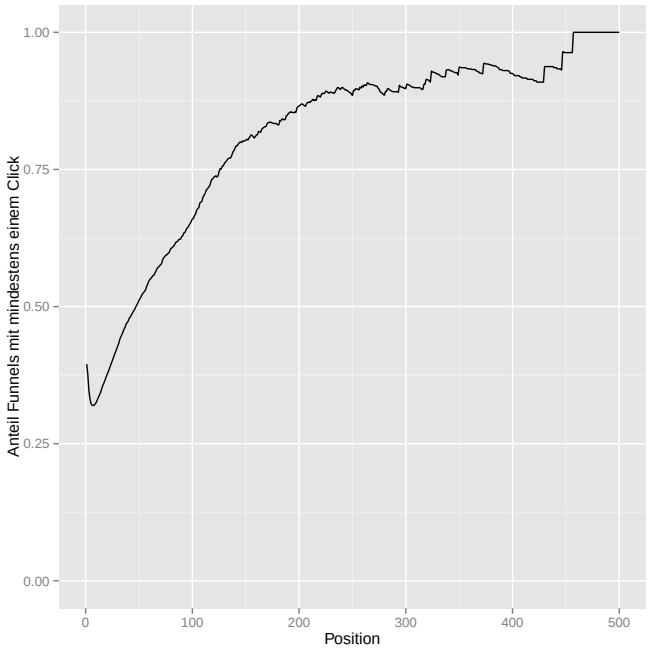

Abbildung 3: Anteil der Funnels mit mindestems einem Click für jede Position

campaign

In Kapitel 2 wurde bereits eine Übersicht über die Kampagnen gegeben. Abbildung 4 enthält die Verteilung dieser 17 Kategorien in den konvertierten Funnels, das heißt auf der x-Achse ist die relative Häufigkeit aufgetragen und auf der y-Achse die Kategorien. Die orangefarbigen Balken repräsentieren die Verteilung in den konvertierten Funnels

nur mit *Clicks*, wie sie auch in den späteren Analysen verwendet werden. Die blauen Balken enthalten *Clicks* und *Views*.
Es ist zu erkennen, dass die Kampagne *Display* bei den Funnels mit *Views* 84% der gesamten Kontaktpunkte ausmacht. Das heißt die Bannerschaltungen überwiegen deutlich und ansonsten hat nur *Direct* einen Anteil von über 5%.
Werden die *Views* nicht berücksichtigt so verteilen sich die Kampagnen besser. *Display* macht jetzt weniger als 10% aus und *Direct* ist mit über 35% die am häufigsten auftretende Kampagne. Außerdem haben auch *SEO*, *SEM - Generisch*, *SEM - Brand*, *Kooperationen - Immoscout24* und *Affiliate - Partnerprogramm* einen Anteil von über 5%. Die restlichen Kampagnen, besonders *Social Media*, machen nur einen kleinen Teil der Daten aus.

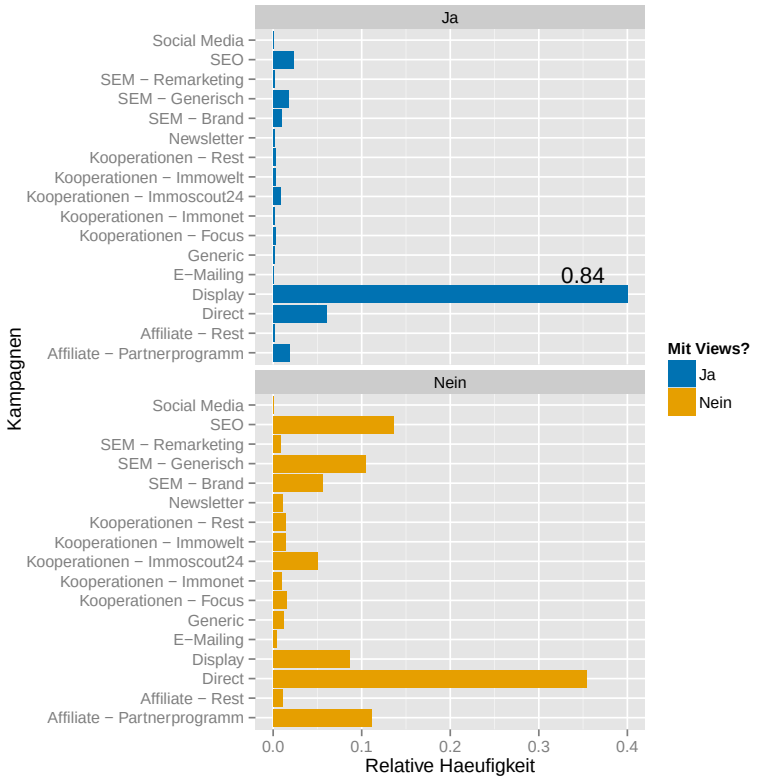

Abbildung 4: Kampagnen der konvertierten Funnels mit und ohne Views

3.2 Vergleich von konvertierten und nicht-konvertierten Funnels

Nachdem bis hierhin nur die konvertierten Funnels mit Augenmerk auf den *Views* betrachtet wurden, sollen in diesem Kapitel die konvertierten und nicht-konvertierten Funnels miteinander verglichen werden. Das heißt, die *Views* werden von nun an nicht mehr in die Analysen mit einbezogen.

weekday

Die Variable *Weekday* gibt an, an welchem Wochentag ein Konaktpunkt aufgetreten ist. Abbildung 5 enthält diesbezüglich Histogramme. Die orangefarbigen Balken entsprechen den konvertierten und die blauen Balken den nicht-konvertierten Funnels. Für weitere Plots in diesem Kapitel gelten die selben Farben. Außerdem ergeben die blauen und orangefarbigen Balken erneut aufsummiert jeweils eins, das heißt sie spiegeln die Verteilung wieder.
Es ist zu erkennen, dass die Häufigkeit der Kontaktpunkte von Montag bis Samstag sinkt. Dieser Trend ist in den konvertierten Funnels etwas stärker. Dort sinkt die relative Häufigkeit von 0.17 am Montag auf 0.09 am Samstag. Bei den nicht-konvertierten Funnels sinkt die relative Häufigkeit von 0.15 am Montag auf 0.11 am Samstag. Außerdem ist der Sonntag bei den nicht-konvertierten mit Abstand der stärkste Tag, während in den konvertierten Funnels der Montag etwas stärker ist als der Sonntag.

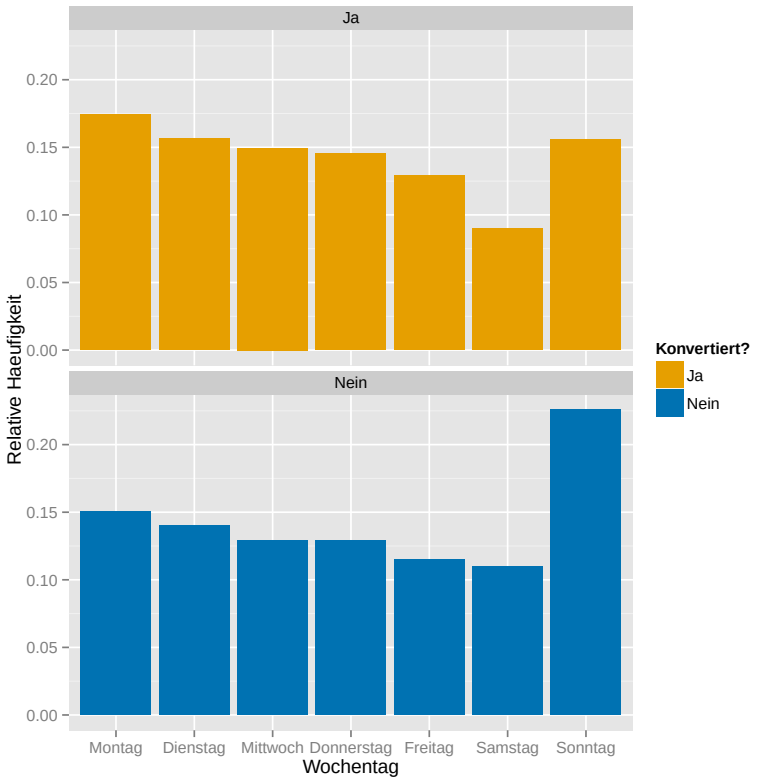

Abbildung 5: Wochentage der Kontaktpunkte

hour

Analog zu der Abbildung mit den Wochentagen, enthält Abbildung 6 Informationen zu der Uhrzeit der Kontaktpunkte in konvertierten und nicht-konvertierten Funnels. Hierfür wurden die Minutenangaben der Uhrzeit jeweils abgeschnitten, so dass *hour* nur die

Werte $0, 1, ..., 23$ annehmen kann.

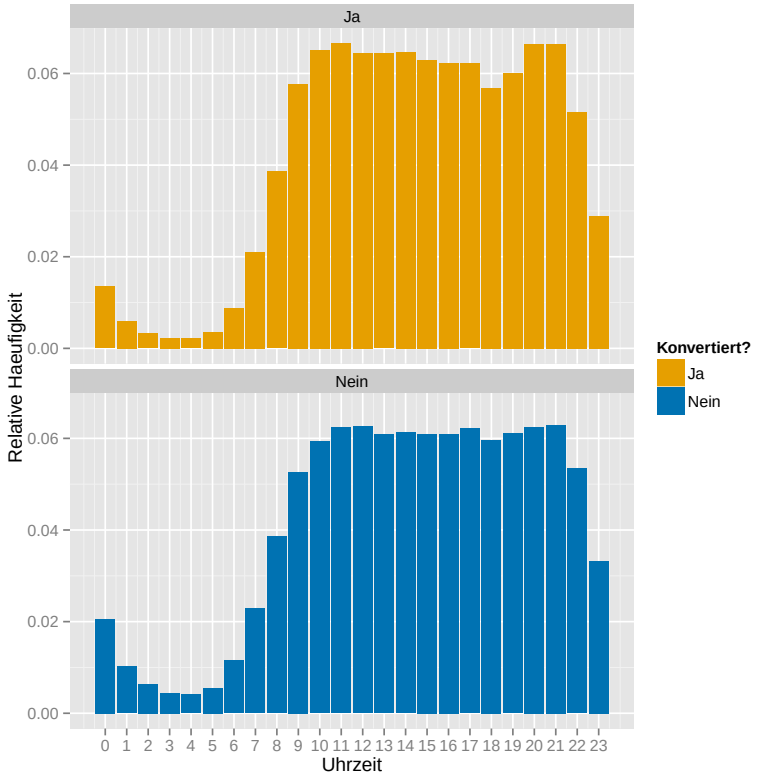

Abbildung 6: Uhrzeit der Kontaktpunkte

Die Verteilungen in konvertierten und nicht-konvertierten Funnels sind sich sehr ähnlich, so dass keine deutlichen Unterschiede erkennbar sind. Insgesamt kann man zusammenfassen, dass in der Nacht zwischen zwei und sechs Uhr sehr wenige Kontaktpunkte stattfinden. Ab sechs Uhr steigt die relative Häufigkeit der Kontaktpunkte bis circa elf Uhr an und dann bleibt sie konstant bis circa 21 Uhr. Daraufhin fällt die relative Häufigkeit wieder ab. Dies ist lediglich darauf zurück zu führen, dass nachts weniger Menschen online sind.

campaign

Wie bereits in Kapitel 3.1 werden die verschiedenen Kampagnen betrachtet. Allerdings werden an dieser Stelle die konvertierten und nicht-konvertierten Funnels miteinander verglichen, wobei die *Views* nicht berücksichtigt werden (siehe Abbildung 7). Das heißt, die Verteilung der konvertierten Funnels ohne *Views* aus Kapitel 3.1 entspricht der Verteilung der konvertierten Funnels in diesem Abschnitt. Eine nähere Beschreibung der unterschiedlichen Kampagnen ist in Abbildung 2 zu finden.
Die Verteilung in den konvertierten Funnels wurde bereits in Kapitel 3.1 etwas näher

beleuchtet. Während dort *Direct* mit Abstand die stärkste Kampagne ist, sind in den nicht-konvertierten Funnels *Affiliate - Partnerprogramm* und *Display* die stärksten Kategorien und *Direct* ist lediglich drittstärkste mit einer relativen Häufigkeit von 0.123. Wie in den konvertierten Funnels haben hier auch *SEO*, *SEM - Generisch* und *Kooperationen - Immoscout24* einen Anteil von über 5%. *SEM - Brand* tritt in den konvertierten Funnels deutlich häufiger auf als in den nicht-konvertierten. Für die restlichen Kampagnen liegen insgesamt wenige Daten vor.

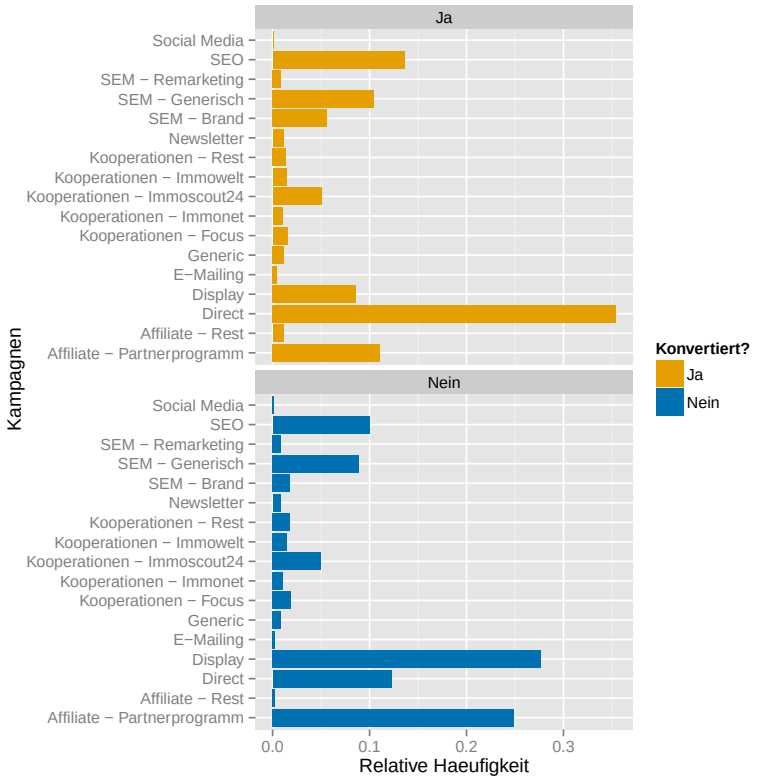

Abbildung 7: Kampagnen der konvertierten und nicht-konvertierten Funnels

funnelLength

Die *funnelLength* gibt die Anzahl der Kontaktpunkte eines Funnels an. In Abbildung 8 sind nur diejenigen Funnels dargestellt, deren Länge 20 Kontaktpunkte nicht überschreitet, da es, relativ gesehen, sehr wenige längere Funnels gibt.
Von den nicht-konvertierten Funnels haben 75% nur einen Kontaktpunkt. Von dort nimmt die relative Häufigkeit der Funnels mit steigender Länge sehr schnell ab. Von den konvertierten Funnels haben 41% nur einen Kontaktpunkt. Das heißt, relativ betrachtet, gibt es dort mehr Funnels mit mehreren Kontaktpunkten. Allerdings gibt es insgesamt deutlich mehr nicht-konvertierte als konvertierte Funnels, so dass die absoluten Anzahl für die nicht-konvertierten stets größer ist. Der Mittelwert beziehungsweise der Median der Länge der Funnels ist bei den konvertierten Funnels 4.1 beziehungsweise

2 und bei den nicht-konvertierten 1.66 beziehungsweise 1.

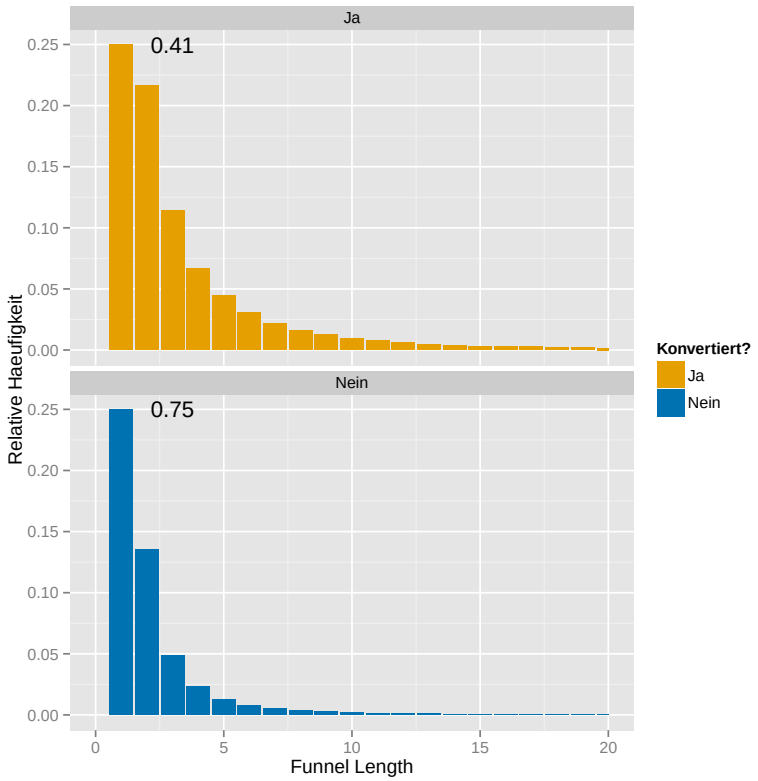

Abbildung 8: Länge der Funnels in konvertierten und nicht-konvertierten Funnels

timeSinceFirst

Das Feature *timeSinceFirst* gibt für jede Position die verstrichene Zeit seit dem ersten Kontaktpunkt an. In Abbildung 9 wird dieses nur für die letzte Position der Funnels geplottet, so dass die Gesamt-Beobachtungsdauer der Funnels betrachtet wird, das heißt die verstrichene Zeit zwischen dem ersten und dem letzten Kontaktpunkt eines jeden Funnels. Auf der x-Achse ist die Beobachtungsdauer in Tagen von 0 bis 50 aufgetragen. Hier gestaltet sich ein ähnliches Bild wie in Abbildung 8. Allerdings muss berücksichtigt werden, dass die *timeSinceFirst* für die erste Position nicht existiert und somit hier nicht berücksichtigt wird. Von den nicht-konvertierten Funnels haben 58% eine Beobachtungsdauer von weniger als einem Tag. Längere Beobachtungsdauern treten deutlich seltener auf. Von den konvertierten Funnels haben 33% eine Beobachtungsdauer von weniger als einem Tag und längere Beobachtungsdauern treten, relativ betrachtet, häufiger auf als in den nicht-konvertierten Funnels.
Auffällig sind außerdem die Hückel, die im Rhythmus von sieben Tagen auftreten. Dies ist darauf zurückzuführen, dass Sonntag und Montag die zwei Tage mit den häufigsten Kontakten sind und beispielsweise Bannerschaltungen am Wochenende besonders häufig eingesetzt werden. Bei den konvertierten Funnels liegt der Mittelwert der Beobachtungs-

dauer bei 21.1 Tagen und bei den nicht-konvertierten Funnels bei 5.3 Tagen.

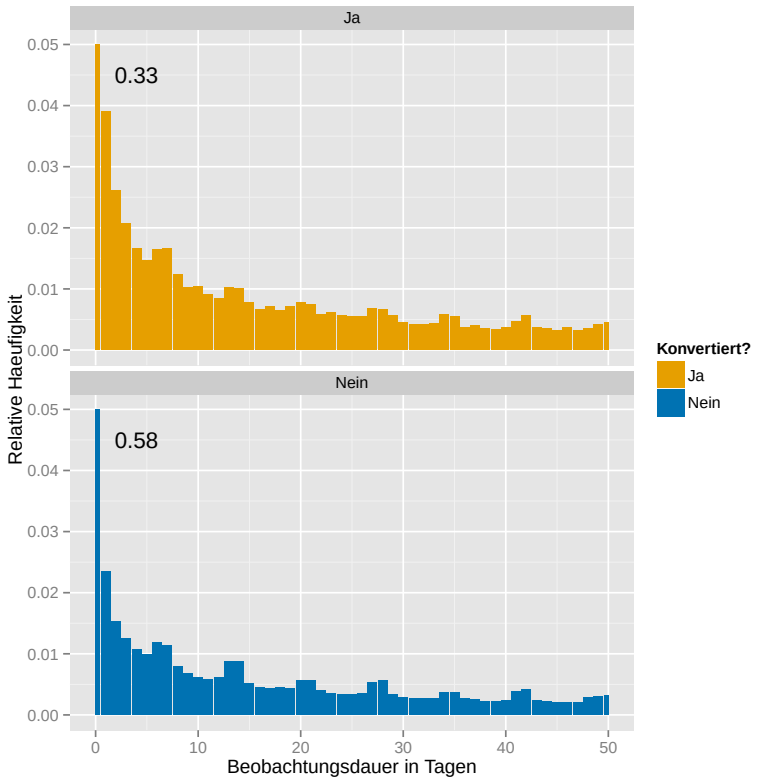

Abbildung 9: Beobachtungsdauer in Tagen der konvertierten und nicht-konvertierten Funnels

timeSinceLast

Die Variable *timeSinceLast* gibt die verstrichene Zeit zwischen zwei aufeinander folgenden Kontaktpunkte an. Diese ist in Abbildung 10 abgebildet, wobei nun alle Kontaktpunkte berücksichtigt werden und nicht nur der letzte, wie es bei *timeSinceFirst* der Fall war.
Die relative Häufigkeit der Abstände, die kürzer als ein Tag sind ist bei den nicht-konvertierten Funnels höher als bei den konvertierten. Ansonsten sind die Werte bei den konvertierten Funnels höher, wobei wieder ein Abfall mit der Zeit und eine wöchentlich Periodizität zu erkennen sind.
Da hier nur die Verteilungen jeweils innerhalb der konvertierten und nicht-konvertierten Funnels verglichen werden, ist *timeSinceLast* keine geeignetes Maß zum Vergleich der Frequenzen der Kontaktpunkte. Dafür wurde die Variable *freq* erzeugt, die im nächsten Abschnitt beschrieben wird.

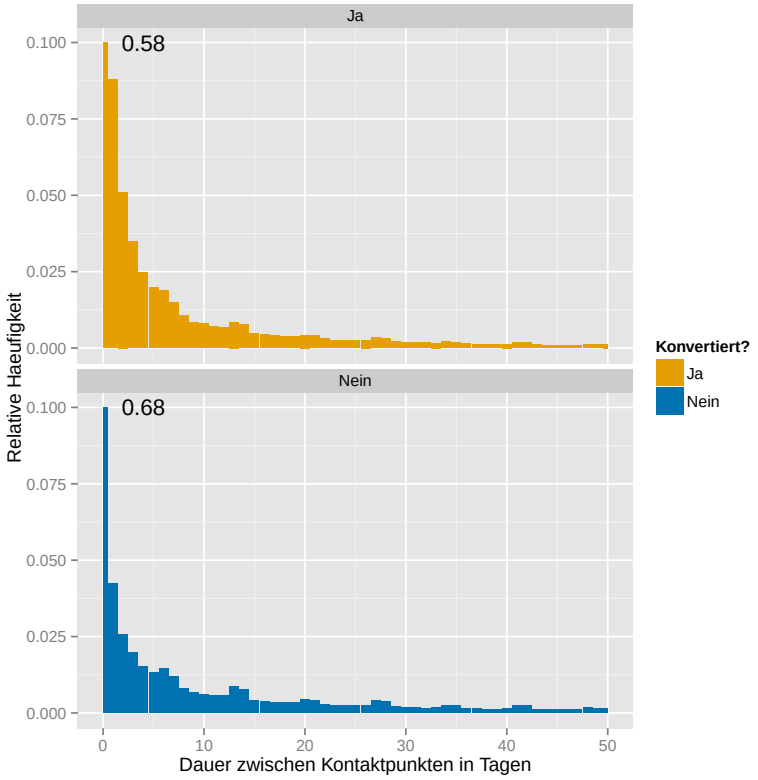

Abbildung 10: Dauer zwischen zwei Kontaktpunkten in den konvertierten und nicht-konvertierten Funnels

freq

Die Frequenz wird wie folgt berechnet. Die Daten werden dahingehend gefiltert, dass für jeden Funnel nur der letzte Kontaktpunkt vorhanden ist. Die Variable *timeSinceFirst* gibt somit wieder die Beobachtungsdauer des gesamten Funnels an. Daraufhin wird die Länge des Funnels (siehe Abbildung 8) durch die Beobachtungsdauer in Stunden geteilt, so dass eine Größe ensteht, die angibt, wieviel Kontaktpunkte der jeweilige Funnel pro Stunde hatte. Diese Frequenz wird in Abbildung 11 abgebildet, wobei auf der x-Achse die Länge der Funnels zwischen vier und 25 aufgetragen ist. Für Funnels mit einem Kontaktpunkt existiert offentsichtlich keine Frequenz und die Längen zwei und drei werden nicht mit abgebildet, da die Frequenzen dort verhältnismäßig groß sind. Außerdem sind einige Boxplots nach oben hin abgeschnitten, da die Grenze der y-Achse auf 0.15 gesetzt wurde, damit die Boxplots besser sichtbar sind, wobei der Median, das heißt der schwarze Balken in den Boxplots, für jeden Boxplot erkennbar bleibt. Die orangefarbigen Boxplots entsprechen den Frequenzen der konvertierten und die blauen den Frequenzen der nicht-konvertierten Funnels.
Für die Funnel Länge zwei liegt der Median der Frequenzen bei 1.22 in den konvertierten und bei 19.89 in den nicht-konvertierten Funnels sowie für die Länge drei bei 0.021 beziehungsweise 0.348.

Insgesamt ist zu erkennen, dass die Frequenzen in den nicht-konvertierten Funnels höher zu sein scheint. Das heißt in denjenigen Funnels die zu keiner Konvertierung führen liegen die Kontaktpunkte näher beieinander, während sie in den konvertierten Funnels mehr über die Zeit verteilt sind.

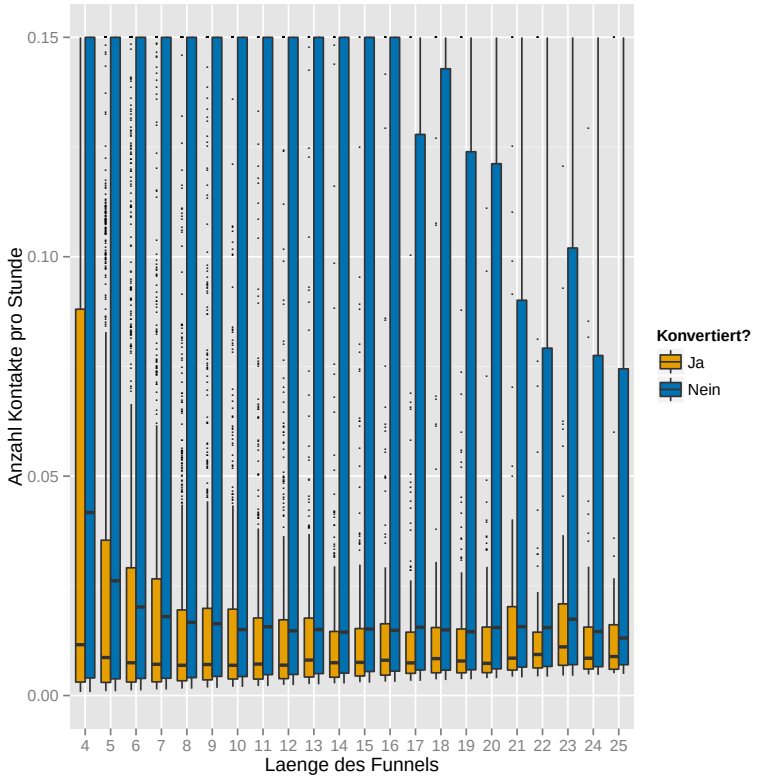

Abbildung 11: Frequenz der Kontaktpunkte in konvertierten und nicht-konvertierten Funnels

4 Zeitdiskretes Survival-Modell

4.1 Lebensdauer-Modell

Aufgrund der in Kapitel 2 beschriebenen Datenlage erscheint die Anwendung eines Modells aus dem Feld der Lebensdaueranalyse intuitiv. Es wird die Zeit bis zu einem Ereignis betrachtet, welches in diesem Fall das Ausfüllen eines Online-Antrages ist. Der potentielle Kunde befindet sich während der Beobachtungsspanne im transienten Zustand bis er durch die Konvertierung in den absorbierenden Zustand wechselt, an dem die Beobachtung endet. Tritt am Ende der Beobachtung keine Konvertierung ein, so spricht man von einer Rechtszensierung.
Die Position gibt die Nummer des Kontaktpunktes an und bildet die Zeitachse des Modells. Das heißt, es handelt sich um ein zeitdiskretes Modell und die Zielvariable y_{ip}

(1) nimmt den Wert Eins an, wenn Kunde i an der Position p konvertiert ist. Für alle vorherigen Positionen eines konvertierten Funnels und für alle Positionen eines nicht-konvertierten Funnels nimmt y_{ip} den Wert Null an. N_p ist die Anzahl der Beobachtungen an Position p. Diese nimmt mit steigendem p ab, da in jeder Position Funnels konvertieren oder Beobachtungen ohne Konvertierung enden. Deshalb wird das Modell nur auf die ersten 25 Positionen angewendet, da für spätere Positionen nicht ausreichend konvertierte Funnels vorliegen.

$$y_{ip} = \begin{cases} 1 & \text{Beobachtung } i \text{ konvertiert an Position } p \\ 0 & \text{sonst} \end{cases}, \; p = 1, ..., 25, \; i = 1, ..., N_p \tag{1}$$

Das Modell schätzt die Hazardrate λ_{ip} (2), das heißt die Wahrscheinlichkeit, dass Beobachtung i an Position p konvertiert unter der Bedingung, dass die Länge des Funnels von Beobachtung i größer oder gleich p ist, was lediglich bedeutet, dass für Beobachtung i an Position p überhaupt noch ein Kontaktpunkt vorliegt. Außerdem wird auf die Features x_{ip} bedingt, die später noch näher erläutert werden.

$$\lambda_{ip} = P(y_{ip} = 1 | funnelLength_i \geq p, x_{ip}) \tag{2}$$

Die Hazardrate wird mittels eines Logit-Modells (3-4) mit der Zielvariable y_{ip} an jeder Position p seperat geschätzt. Die Annahmen des Modells sind, dass die $y_{ip}|x_{ip}$ unabhängig Bernoulli-verteilt sind mit der Hazardrate λ_{ip} als Parameter und der Erwartungswert wird anhand der Responsefunktion h mit der Prädiktorfunktion f_p verknüpft.

$$y_{ip}|x_{ip} \overset{ind}{\sim} Bin(1, \lambda_{ip}) \tag{3}$$

$$E(y_{ip}|x_{ip}) = P(y_{ip} = 1|x_{ip}) = \lambda_{ip} = h(f_p(x_{ip})) = \frac{\exp(f_p(x_{ip}))}{1 + \exp(f_p(x_{ip}))} \tag{4}$$

Aus diesen Annahmen lassen sich Likelihood (5) und Log-Likelihood (6) des Modells ableiten.

$$L(\lambda_{ip}) = \prod_{i=1}^{N_p} \lambda_{ip}^{y_{ip}} (1 - \lambda_{ip})^{1 - y_{ip}} \tag{5}$$

$$\begin{aligned} l(\lambda_{ip}) = \ln(L(\lambda_{ip})) &= \sum_{i=1}^{N_p} (y_{ip} \ln(\lambda_{ip}) + (1 - y_{ip}) \ln(1 - \lambda_{ip})) \\ &= \sum_{i=1}^{N_p} (y_{ip} f_p(x_{ip}) - \ln(1 + \exp(f_p(x_{ip})))) \end{aligned} \tag{6}$$

Damit ergibt sich der binomielle Verlust (7) aus der negativen Log-Likelihood und das Logit-Modell ist lösbar durch die Minimierung dieses Verlusts.

$$L(y_{ip}, f_p(x_{ip})) = -\sum_{i=1}^{N_p} (y_{ip} f_p(x_{ip}) + \ln(1 + \exp(f_p(x_{ip})))) \tag{7}$$

Um ein gutes Prognose-Modell zu entwickeln, wird eine Ensemble-Methode angewendet, die im nächsten Abschnitt vorgestellt wird.

4.2 Stochastic Gradient Boosting

Algorithmus

Stochastic Gradient Boosting ist eine Ensemble-Methoden, die durch mehrfache Anwendung des sogenannten Basis-Lerners ein Ensemble von Schätzern für eine Prognosefunktion liefert. Durch Aggregation der Schätzer erhält man die endgültige Prognosefunktion. Ein sehr beliebter Basis-Lerner sind Stümpfe, das heißt Bäume mit nur einem Split. Einige Vorteile von Bäumen sind, dass sie mit kategoriellen Features, Ausreißern und fehlenden Werten umgehen können. Außerdem wird der schwachen Prognoseleistung von Bäumen durch die Kombination mit Boosting entgegen gewirkt.
Gesucht ist also eine Prognosefunktion, die den Erwartungswert einer Verlustfunktion minimiert. Als Verlustfunktion wird der binomielle Verlust (7) verwendet, wobei sich die Prädiktorfunktion wie folgt ergibt.

$$\begin{aligned} f(x_{ip}) =& \text{offset}(\hat{\lambda}_{i,p-1}) + \\ & f_{weekday,p}(\text{weekday}_{ip}) + \\ & f_{hour,p}(\text{hour}_{ip}) + \\ & f_{campaign,p}(\text{campaign}_{ip}) + \\ & f_{campaignLast,p}(\text{campaign}_{i,p-1}) + \\ & f_{campaignLast2,p}(\text{campaign}_{i,p-2}) + \\ & f_{timeSinceLast,p}(\text{timeSinceLast}_{ip}) + \\ & f_{timeSinceFirst,p}(\text{timeSinceFirst}_{ip}) \end{aligned} \tag{8}$$

Der Prädiktor ist also eine additive Funktion von Treppenfunktionen der sieben verwendeten Features und einem offset. Hier sei nochmal darauf hingewiesen, dass Features wie *hasClicked* oder *clickCount* nicht verwendet werden können, da die Views aufgrund der Problematik der Datenerhebung nicht berücksichtigt werden. Die verwendeten Einflussgrößen sind also der Wochentag und die Stunde des jeweiligen Kontaktpunktes, die Kampagne des aktuellen und der letzten zwei Kontakte, sowie die Dauer seit dem vorherigen Kontaktpunkt und die Gesamtdauer seit dem ersten bis zum jetzigen Kontakt. Die Funktionen $f_{.,p}$ können theoretisch für verschiedene Positionen komplett unterschiedliche Formen annehmen. Es sei ausdrücklich darauf hingewiesen, dass $f_{campaignLast,p}(campaign_{i,p-1})$ und $f_{campaign,p-1}(campaign_{i,p-1})$ zwei unterschiedliche Funktionen sind. Erstere gibt den Einfluss der Art des vorherigen Kontaktpunktes auf die Konvertierungswahrscheinlichkeit an der Position p wieder und zweitere den Einfluss der Art des aktuellen Kontaktpunktes auf die Konvertierungswahrscheinlichkeit an der Position $p-1$. Unter der Annahme, dass der Einfluss der Features an den verschiedenen Positionen ähnlich ist, fließt zusätzlich noch die Vorhersage des Modells der vorherigen Position als offset mit ein. Dadurch konnten die Ergebnisse besonders an späteren Positionen, an denen weniger Daten vorhanden sind, deutlich verbessert werden. Ergebnisse ohne offset sind im elektronischen Anhang zu finden.
Der offset fällt für Position eins weg, da es noch keine Vorhersagen eines vorherigen Modells gibt. Außerdem sind für Position eins die Features *campaignLast*, *campaignLast2*,

timeSinceLast und *timeSinceFirst* offentsichtlich noch nicht vorhanden, so dass diese ebenfalls wegfallen. An Position zwei ist *campaignLast2* noch nicht vorhanden und *timeSinceLast* und *timeSinceFirst* sind hier identisch, so dass nur eine der beiden berücksichtigt wird. Ab Position Drei ergibt sich der Prädiktor dann exakt so wie in (8) dargestellt. Algorithmus 1 enthält Pseudo-Code, der das grobe Vorgehen beim Gradient Boosting

Algorithmus 1 Gradient Boosting

Setze Startwert für $f_{0p}(x_{ip})$
for $m = 1 : n.trees$ **do**
  Setze $\lambda_{ip}(x_{ip}) = \frac{\exp(f_{m-1,p}(x_{ip}))}{1+\exp(f_{m-1,p}(x_{ip}))}$
  for $i = 1 : N_p$ **do**
    $r_{imp} = -\frac{\partial L(y_{ip}, f_{m-1,p}(x_{ip}))}{\partial f_{m-1,p}(x_{ip})} = y_{ip} - \lambda_{ip}(x_{ip})$
  end for
  Fitte Stumpf: $\theta_{mp} = \arg\min_\theta \sum_{i=1}^{N_p}(r_{imp} - h(x_{ip}, \theta))^2$
  Line-Search: $\beta_{mp} = \arg\min_\beta \sum_{i=1}^{N_p} L(y_{ip}, f_{m-1,p}(x_{ip}) + \beta h(x_{ip}, \theta_{mp}))$
  Update: $f_{mp}(x_{ip}) = f_{m-1,p}(x_{ip}) + \beta_{mp} h(x_{ip}, \theta_{mp})$
end for

erläutern soll. Dieser muss für jedes $p = 1, ..., 25$ durchgeführt werden. Zunächst muss ein Startwert des Prädiktors an Position p festgelegt werden. Daraufhin werden folgende Schritte für $m = 1$ bis *n.trees* iteriert. Zunächst werden die Hazardraten λ_{ip} durch Rücktransformation der Prädiktorfunktion aus der vorherigen Iteration berechnet. Für jede Beobachtung i werden dann die Pseudo-Residuen r_{imp} berechnet, die die Richtung des negativen Gradienten angeben. Die Pseudo-Residuen entsprechen also der Richtung des steilsten Abstiegs der Verlustfunktion. An die Pseudo-Residuen wird ein Basis-Lerner $h(x_{ip}, \theta_{mp})$, in diesem Fall ein Entscheidungsbaum, so angepasst, dass er den negativen Gradienten so gut wie möglich approximiert. Bildlich gesprochen wird das Modell also in die Richtung der größten Verringerung des Verlusts verschoben. Da als Basis-Lerner Stümpfe verwendet werden, wird die Verbesserung des Modells in einem Boosting-Schritt durch nur eines der sieben Features erklärt. Per Line-Search wird daraufhin die optimale Schrittweite β_{mp} berechnet und das Modell wird geupdatet mit $f_{mp}(x_{ip}) = f_{m-1,p}(x_{ip}) + \beta_{mp} h(x_{ip}, \theta_{mp})$. Nach *n.trees* Iterationen endet der Algorithmus. Durch die Verringerung des Verlusts in jeder Iteration kann es vor allem bei einer hohen Anzahl von Iterationen zu Overfitting kommen. Deshalb wird mittels Kreuzvalidierung die optimale Anzahl an Iterationen ausgewählt. Das heißt $\hat{f}(x_{ip}) = f_{m\_opt,p}(x_{ip})$ wird als Ergebnis verwendet. Das Ensemble der Splits bis zur optimalen Iterationsanzahl bildet für jedes der Feautures eine Treppenfunktion.

Parameter des Modells

Das Modell wurde mit dem Paket *gbm* [21] berechnet und mit Hilfe der R-Pakete *foreach* [3] und *doSNOW* [2] parallelisiert. Die Modelle für die einzelnen Positionen können allerdings nur dann parallelisiert werden, wenn kein offset benützt wird. Für die Anwendung des Modells müssen noch einige Parameter eingestellt werden.

Zunächst wurden die Daten in Trainings- und Testdaten aufgeteilt, sodass Trainings- und Testdaten jeweils die Hälfte der gesamten Daten ausmachen. Da die Anzahl der nicht-konvertierten Funnels deutlich überwiegt und einige Kampagnen vergleichsweise selten in den Daten auftreten, wurden die Trainingsdaten stratifiziert bezüglich konvertierter beziehungsweise nicht-konvertierter Funnel und der Variable *Campaign* gezogen. Außerdem wurde die Länge der Funnels bei der Ziehung berücksichtigt, so dass Trainings- und Testdaten positionsübergreifend klar getrennt sind und das Verhältnis von eins zu eins trotzdem eingehalten wird. Das heißt, dass eine Beobachtung, die an Position 1 in den Trainingsdaten enthalten ist auch an allen späteren Positionen in den Trainingsdaten ist. Das selbe gilt für die Testdaten. Dies ist wichtig, damit Trainings- und Testdaten bei der Anwendung des offsets nicht vermischt werden. Anhand der Trainingsdaten wurde das Modell gefittet. Die Testdaten dienen der späteren Bewertung der Prognosegüte des Modells.
Die maximale Anzahl der Bäume *n.trees* wurde gleich 3000 gesetzt, wobei für die Präsentation der Ergebnisse die optimale Anzahl an Bäumen mittels 5-facher Kreuzvalidierung gewählt wurde. Dafür werden die Trainingsdaten in fünf gleich große Partitionen aufgeteilt. Dann wird der Gradient Boosting-Algorithmus fünf mal angewendet, wobei jeweils eine Partition der Daten außen vorgelassen wird. Anhand dieser Partition wird dann der binomielle Verlust nach jeder Iteration mit der aktuellen Prognosefunktion berechnet. Das führt für jede Iteration zu fünf Schätzern für den binomiellen Verlust. Diese fünf Schätzer werden für jede Iteration gemittelt und die Iteration mit dem geringsten binomiellen Verlust wird als optimale Iterationsanzahl gewählt. Daraufhin wird der Gradient Boosting-Algorithmus auf alle fünf Partitionen angewendet.
Zusätzlich zu der Beschränkung der Iterationen, wird dem Overfitting auch durch einen Shrinkage-Parameter μ entgegen gewirkt. Dieser bewirkt, dass nicht die optimale Schrittweite β_{mp} in jeder Iteration gegangen wird, sondern nur ein Bruchteil dieser Schrittweite. Es wird also durch $f_{mp}(x_{ip}) = f_{m-1,p}(x_{ip}) + \mu\beta_{mp}h(x_{ip}, \theta_{mp})$ geupdatet, wobei μ gleich 0.01 gewählt wurde, was einem üblichen Wert für diesen Parameter entspricht. Bei der Wahl von μ und *n.trees* muss stets die Rechenzeit im Auge behalten werden.
Wie bereits erwähnt, wurden als Basis-Lerner Stümpfe gewählt. Das wurde durch die Festlegung von *interaction.depth* auf 1 realisiert. Mit einer höheren *interaction.depth* ließen sich auch Interaktionen modellieren. Die Stärke von solchen Interaktionen lassen sich beispielsweise mit Friedmans h-Statistik [12] untersuchen. Die Einführung von Interaktionen führt bei den vorliegenden Daten aber zu einer Verschlechterung der Prognosegüte. Bei Interesse sind Ergebnisse diesbezüglich im elektronischen Anhang zu finden.
Der bis hierhin beschriebene Algorithmus entspricht lediglich dem Gradient Boosting. Wenn man in jeder Iteration allerdings nur einen Teil der Daten verwendet, spricht man vom Stochastic Gradient Boosting. Diese Anpassung führt meist zur einer Verbesserung der Ergebnisse [10], wobei noch nicht geklärt ist, wie genau es zu dieser Verbesserung kommt. In diesem Modell wurde die *bag.fraction*, das heißt der Anteil der verwendeten Daten in jedem Schritt, auf 0.5 gesetzt. Wird die *bag.fraction* auf 1 gesetzt, was dazu führt das in jedem Schritt die kompletten Trainingsdaten verwendet werden, so sind die Ergebnisse bei den gegebenen Daten nahezu identisch.

Output

Als Ergebnis ist für jede Beobachtung i und jede Position p ein Wert der Prädiktorfunktion $\hat{f}_p(x_{ip})$ gegeben. Daraus können die Hazardraten durch Rücktransformation (9) berechnet werden. Diese geben die Wahrscheinlichkeit für Beobachtung i an Position p zu konvertieren an.

$$\hat{\lambda}_{ip} = \frac{\exp(\hat{f}_p(x_{ip}))}{1 + \exp(\hat{f}_p(x_{ip}))} \tag{9}$$

Ein weiterer Kennwert, der die Prognosefunktion beschreibt, ist die Relative Wichtigkeit der Einflussgrößen [11]. Sie gibt die Stärke des Einflusses der Features auf die Prognoseregel an. $\hat{I}^2_{jp}(m)$ (10) ist die Verbesserung, die an Position p in Iteration m durch die Variable j realisiert wird, wobei $\hat{i}_{mp}1_{jmp}$ eben diese Verbesserung ist und 1_{jmp} die Indikatorfunktion, die angibt, ob in Iteration m das Feature j als Splitvariable genutzt wurde. $\hat{I}^2_{jp}$ (11) mittelt für jedes Feature j über alle Iterationen beziehungsweise Bäume und die Relative Wichtigkeit $\hat{I}_{jp}$ (12) von Feature j ergibt sich dann aus der Wurzel dieses Wertes. Die Wichtigkeitswerte werden so skaliert, dass sie aufsummiert über alle Features den Wert 100 ergeben.

$$\hat{I}^2_{jp}(m) = \hat{i}_{mp}1_{jmp} \tag{10}$$

$$\hat{I}^2_{jp} = \frac{1}{M}\sum_{m=1}^{n.trees} \hat{I}^2_{jp}(m) \tag{11}$$

$$\hat{I}_{jp} = \sqrt{\hat{I}^2_{jp}} \tag{12}$$

Neben der Information zur Stärke des Einflusses eines Features, interessiert natürlich auch die Art dieses Einflusses, das heißt der marginale Effekt eines Features [11]. Das Ziel ist somit ein Plot, der den Zusammenhang zwischen dem Feature j und der Prognosefunktion $\hat{f}_p$ an Position p darstellt. Da die Prognosefunktion von mehreren Features abhängt, müssen die marginalen Effekte berechnet werden. Das Feature x_{jp} ist eine Teilmenge aller Features $\{x_{1p}, ..., x_{np}\}$ und $x_{\setminus j,p}$ sei dessen Komplement. Damit kann die Prognosefunktion als $\hat{f}_p(x_p) = \hat{f}_p(x_{jp}, x_{\setminus j,p})$ geschrieben werden und wenn man auf bestimmte Werte von $x_{\setminus j,p}$ bedingt, kann die Prognosefunktion als Funktion von lediglich x_{jp} betrachtet werden.

$$\hat{f}_{x_{\setminus j,p}}(x_{jp}) = \hat{f}(x_{jp}|x_{\setminus j,p}) \tag{13}$$

Die Form von $\hat{f}_{x_{\setminus j,p}}(x_{jp})$ hängt von dem für $x_{\setminus j,p}$ gewählten Wert ab. Wenn diese Abhängigkeit nicht zu stark ist, dann ist (14) eine gute Zusammenfassung des marginalen Effektes von x_{jp} auf die Prognosefunktion. Dabei ist $p_{\setminus j,p}(x_{\setminus j,p})$ die marginale Dichte von $x_{\setminus j,p}$ und $p(x_p)$ ist die gemeinsame Dichte der Features.

$$\bar{f}_{jp}(x_{jp}) = E_{x_{\setminus j,p}}(\hat{f}_p(x_p)) = \int \hat{f}_p(x_{jp}, x_{\setminus j,p}) p_{\setminus j,p}(x_{\setminus j,p})\, \mathrm{d}x_{\setminus j,p}. \tag{14}$$

$$p_{\setminus j,p}(x_{\setminus j,p}) = \int p(x_p)\, \mathrm{d}x_{\setminus j,p}. \tag{15}$$

Die marginale Dichte kann aus den Trainingsdaten geschätzt werden, so dass (14) zu (16) wird.

$$\bar{f}_{jp}(x_{jp}) = \frac{1}{N} \sum_{i=1}^{N_p} \hat{f}_p(x_{jp}, x_{i,\setminus j,p}) \tag{16}$$

Die marginalen Effekte werden wiederum zu den Hazardraten rücktransformiert.
Zur Bewertung der Prognosegüte wird das Modell auf die Testdaten angewendet, so dass für jede Beobachung und jede Position der in den Testdaten eine Hazardrate vorliegt. Da die nicht-konvertierten Funnels in den Daten deutlich überwiegen, sind diese Wahrscheinlichkeiten sowohl für konvertierte als auch für nicht-konvertierte Funnels sehr niedrig. Deshalb kommt der Bayes-Klassifikator zur Beurteilung der Prognosegüte nicht in Frage, da dieser schlicht alle Beobachtungen als nicht-konvertiert vorhersagen würde. Es muss also ein geeigneter Schwellenwert gefunden werden, um die Prognosegüte des Modells zu bewerten. Wenn die Hazardrate einer Beobachtung größer ist als der Schwellenwert, so wird diese als konvertiert vorhergesagt und sonst als nicht-konvertiert. Dafür werden die Kenngrößen Sensitivität und Spezifität benötigt. Die Sensitivität ist die Richtig Positiv Rate, das heißt der Anteil der wirklich Konvertierten unter allen, die als konvertiert vorhergesagt werden, gegeben ein spezifischer Schwellenwert. Die Spezifität ist die Richtig Negativ Rate, das heißt der Anteil der Nicht-Konvertierten unter allen, die als nicht-konvertiert vorhergesagt werden. Diese Kennwerte wurden für ein Gitter von potentiellen Schwellenwerten berechnet.
Die ROC (Receiver Operating Characteristics)-Kurve erlaubt die simultane Betrachtung dieser Schwellenwerte. Diese bildet die Falsch Positiv Rate, das heißt 1 minus Spezifität, auf der x-Achse und die Sensitivität auf der y-Achse ab. Die ROC-Kurve hat ihren Ursprung im Punkt $(0, 0)$ und steigt von dort monoton an bis $(1, 1)$. Wenn Falsch Positiv Rate und Richtig Positiv Rate für alle Schwellenwerte gleich ist, so kann das Modell nicht zwischen konvertiert und nicht-konvertiert unterscheiden und die ROC-Kurve ist die Diagonale zwischen $(0, 0)$ und $(1, 1)$. Ein perfektes Modell, dass alle Beobachtungen korrekt zuweist geht durch den Punkt $(0, 1)$. Je größer die Fläche zwischen der Diagonalen und der ROC-Kurve ist, desto besser ist somit auch das Modell.
An dieser Stelle kommt die AUC (Area Under the Curve) ins Spiel. Sie ist definiert als die Fläche unter der ROC-Kurve (17). Wenn das Modell nicht zwischen konvertiert und nicht-konvertiert unterscheiden kann, so ist die AUC gleich 0.5 und die ROC ist, wie bereits erwähnt die Diagonale. Ein AUC von 1 entspricht einem perfekten Modell. Die AUC wird interpretiert als die Wahrscheinlichkeit, dass bei einer konvertierten Beobachtung die Hazardrate größer ist als bei einer nicht-konvertierten Beobachtung.

$$AUC = \int_0^1 ROC(t)\,\mathrm{d}t. \tag{17}$$

5 Sequential Pattern Mining

5.1 Überblick

Sequential Pattern Mining entdeckt häufige Subsequenzen, das heißt Teilfolgen, in Datenbanken. Sogenannte Sequenz-Datenbanken bestehen aus Transaktionen, die jeweils eines oder mehrere Items enthalten, welche der Zeit nach geordnet sind. Bei den gegebenen Daten ist eine Transaktion ein Kontaktpunkt, wobei das Item die Kampagne des Kontaktpunktes ist. Ein Funnel setzt sich aus mehreren Transaktionen beziehungsweise Kontaktpunkten zusammen.
Ein Anwendungsfeld ist die Warenkorbanalyse. Angenommen es wird das Kaufverhalten in einem Supermarkt einen Monat lang beobachtet, dann könnte [Kunde 1, <(Brot, Milch), (Brot, Milch, Tee), (Zucker), (Milch, Salz)>]; [Kunde 2, <(Brot), (Milch, Tee)>] eine Beispiel-Datenbank sein. Kunde 1 war vier mal im beobachteten Monat im Supermarkt einkaufen, wobei Kunde 2 nur zweimal einkaufen war. Der Kunde kann nur eins oder auch mehrere Items pro Besuch einkaufen. Im Falle von mehreren Items spricht man von Itemsets.
Web Usage Mining ist das am weitesten verbreitete Anwendungsfeld von Sequential Pattern Mining in der Literatur ([18, 26, 13]). Unter der Annahme, dass ein Internetnutzer nur eine Webseite an einem Zeitpunkt aufrufen kann, bestehen die Sequenzen nur aus einzelnen Items und enthalten keine Itemsets. Ist also eine Menge von Items $I = \{a, b, c, d, e\}$ gegeben, die beispielsweise verschiedene Webseiten repräsentieren, so könnte eine Datenbank mit zwei Nutzern folgendermaßen aussehen: [Kunde 1, <abedcab>]; [Kunde 2, <edcaa>] ([19]).
In den letzten zwei Jahrzehnten wurden im Forschungsfeld des Sequential Pattern Mining eine Vielzahl von Algorithmen entwickelt. Beispiele sind der HVSM- [24], der PrefixSpan- [22], der FS-Miner [7] und der PLWAP-Algorithmus [8]. Folglich stellt sich die Frage, welcher Algorithmus aus dieser großen Auswahl hier verwendet werden soll. Es gibt einige Arbeiten (beispielsweise [25, 9, 20, 15]), die Sequential Pattern Mining-Algorithmen, die auf Online-Tracking-Daten anwendbar sind, miteinander vergleichen. Diese konzentrieren sich allerdings nur auf die Techniken und Theorien, auf denen die jeweiligen Algorithmen basieren und bewerten sie anhand ihrer Leistung bei unterschiedlichen Datenstrukturen. Auf die Bewertung der eigentlichen Ergebnisse der Algorithmen wird dabei nicht eingegangen. Deshalb kann an dieser Stelle reinen Gewissens der SPADE-Algorithmus [30] verwendet werden, der der einzige Sequential Pattern Mining-Algorithmus ist, der in R implementiert ist. Außerdem schneidet er im Vergleich zu anderen Algorithmen stets solide ab.

5.2 Problemstellung

Das Problem wurde erstmals im Jahre 1995 in der Arbeit *Mining Sequential Patterns* [1] von Agrawal & Srikant vorgestellt. Es ist eine Datenbank D von Transaktionen gegeben, wobei jede Transaktion eine ID, den Zeitpunkt der Transaktion und die dazugehörigen Items enthält. Für jede ID können mehrere Transaktionen existieren, allerdings nie zwei

oder mehr Transaktionen mit dem selben Zeitpunkt.
Ein Itemset $i = (i_1 i_2 ... i_m)$ ist eine nichtleere Menge von Items i_j und eine Sequenz $s = s_1 \rightarrow s_2 \rightarrow ... \rightarrow s_n$ eine geordnete Liste von Itemsets s_j. Ohne Beschränkung der Allgemeinheit wird angenommen, dass die Menge der Items auf eine Menge von fortlaufenden ganzen Zahlen abgebildet wird.
Eine Sequenz $a_1 \rightarrow a_2 \rightarrow ... \rightarrow a_n$ heißt Subsequenz einer anderen Sequenz $b_1 \rightarrow b_2 \rightarrow ... \rightarrow b_m$ wenn ganze Zahlen $i_1 < i_2 < ... < i_n$ existieren, so dass $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$, ..., $a_n \subseteq b_{i_n}$ gilt. Beispielsweise ist $< (1)(58)(27) >$ Subsequenz von $< (15)(248)(358)(27) >$, da $(1) \subseteq (15)$, $(58) \subseteq (358)$ und $(27) \subseteq (27)$ aber $< (27) >$ keine Subsequenz von $< (2)(7) >$ und umgekehrt. Eine Sequenz heißt maximal in einer Menge von Sequenzen, wenn sie von keiner Sequenz in dieser Menge Subsequenz ist.
Die Transaktionen einer ID können als Sequenz verstanden werden. Jede Transaktion entspricht dabei einem Itemset und die nach den Zeitpunkten $T_1, T_2, ..., T_n$ geordneten Transaktionen entsprechen einer Sequenz $itemset(T_1) \rightarrow itemset(T_2) \rightarrow ... \rightarrow itemset(T_n)$. Der Support einer Sequenz s ist der Anteil der IDs, die s unterstützen, das heißt deren Sequenz Subsequenz von s sind. Eine Sequenz mit k Items wird auch k-Sequenz genannt.
Das Problem des Sequential Pattern Mining besteht darin, in einer Datenbank D unter allen existierenden Sequenzen die maximalen Sequenzen zu finden, deren support größer oder gleich einem festgelegten minimalen Support ist.

5.3 SPADE

Im Folgenden wird der SPADE Algorithmus [30] vorgestellt, der in dem *R*-Paket *arulesSequences* [5] implementiert ist. Dieser sucht häufige Sequenzen von Itemsets, das heißt er kann auch mit Daten umgehen, die mehrere Items pro Zeitpunkt enthalten. Da ein Benutzer nur eine Webseite zum exakt selben Zeitpunkt besuchen kann, ist diese Eigenschaft in diesem Fall nicht relevant.

ID	Position	Anzahl Items	Items
1	1	1	A
1	2	1	D
1	3	1	A
1	4	1	B
2	1	1	C
2	2	1	B
3	1	1	D
4	1	1	A
4	2	1	D
4	3	1	B

Tabelle 4: Beispiel für eine Input-Datenbank für den SPADE-Algorithmus

Die Daten werden für den Algorithmus vertikal bezüglich der IDs angeordnet, so dass diese sich in der ersten Spalte der Datenbank befinden. Die zweite Spalte enthält die Zeit des Kontaktes, die dritte Spalte die Anzahl der Items zu diesem Zeitpunkt und die vierte Spalte die eigentlichen Items. Bei den gegebenen Daten sind die Zeitpunkte die Positionen und damit pro ID von eins bis zur Anzahl der Positionen für die ID durchnummeriert. Die Anzahl der Items in der dritten Spalte ist immer gleich eins und die vierte Spalte enthält pro Zeile nur ein Item (siehe Tabelle 4).

Berechnung des Supports

In diesem Abschnitt wird beschrieben, wie der Support einer Sequenz berechnet wird. Tabelle 5 enthält die ID-Listen der Items aus den Daten aus Tabelle 4. Item A beispielsweise, tritt für ID 1 an Position 1 und 3 und für ID 4 an Position 1 auf.

A		B		C		D	
ID	Position	ID	Position	ID	Position	ID	Position
1	1	1	4	2	1	1	2
1	3	2	2			3	1
4	1	4	3			4	2

Tabelle 5: ID-Listen der Items

Jede Sequenz ist eine zeitabhängige Verknüpfung von Items und der Support der Sequenz kann mittels einer zeitabhängigen Verknüpfung der ID-Listen jedes Items in der Sequenz berechnet werden. In Tabelle 6 ist diese Vorgehen beispielhaft für die Sequenz $A \rightarrow D \rightarrow B$ dargestellt. Den Support ergibt sich dann, indem die Anzahl der einzigartigen IDs, die die Sequenz enthalten, durch die Anzahl aller IDs in den Daten geteilt wird.

A		AD			ADB			
ID	Pos(A)	ID	Pos(A)	Pos(D)	ID	Pos(A)	Pos(D)	Pos(B)
1	1	1	1	2	1	1	2	4
1	3	4	1	2	4	1	2	3
4	1							

Tabelle 6: Temporale Verknüpfung

Werden für jedes Item einer Sequenz die Zeitpunkte zusammengefasst, wie in Tabelle 6 dargestellt, so wird sehr viel Speicherplatz verbraucht. Um diesem Problem entgegen zu wirken, wird das Korollar ausgenutzt, das besagt, dass eine Sequenz der Länge k immer aus der Kombination ihrer lexikographisch ersten zwei Subsequenzen der Länge $(k-1)$ gebildet werden kann. Damit werden für jede Sequenz die ID-Spalte und nur eine Spalte

für die Zeitpunkte benötigt.
Dies führt zu einer Speicherreduktion da die zwei ersten Sequenzen der Länge $(k-1)$, X_1 und X_2, einer Sequenz X einen Prefix der Länge $(k-2)$ teilen. Damit sind auch die Zeitpunkte im Prefix gleich und X_1 und X_2 unterscheiden sich lediglich bezüglich des letzten Items.
In dem Beispiel mit $X = (A \to D \to B)$ entsteht X durch den temporalen join von $X_1 = (A \to D)$ und $X_2 = (A \to B)$. X_1 entsteht durch das Wegfallen des letzten Items von X und X_2 durch das Wegfallen des vorletzten Items. Im Falle längerer Sequenzen wird dieses Vorgehen rekursiv durchgeführt bis die einzelnen Items übrig bleiben. Sei X eine Subsequenz von Y, so gilt, dass die Mächtigkeit der ID-Liste von Y kleiner oder gleich der Mächtigkeit der ID-Liste von X sein muss. Das führt zu schnellen Joins und schnellen Berechnungen des Supports.

Prefix-basierte Klassen

Sei $p : (S, N) \to S$ eine Funktion, wobei S die Menge der Sequenzen und N eine Menge von nichtnegativen natürlichen Zahlen sind. $p(X, k) = X[1 : k]$ gibt den Prefix von X der Länge k zurück. Die Äquivalenz Relation θ_k wird wie folgt definiert. Für alle $X, Y \in S$ ist X mit Y verwandt unter θ_k, in Zeichen $X \equiv_{\theta_k} Y$, genau dann wenn $p(X, k) = p(Y, k)$. Das heißt, zwei Sequenzen sind in der selben Klasse bezüglich θ_k, wenn sie den selben Prefix der Länge k teilen.

Suchstrategien

Breadth-First und Depth-First sind effiziente Suchstrategien, um häufige Sequenzen innerhalb von Eltern-Klassen zu finden. Beide basieren auf der rekursiven Zerlegung der Eltern-Klassen in kleinere Klassen die anhand der Äquivalenz Relation θ_k gebildet werden und dem resultierenden Verbund von Äquivalenzklassen.
Bei der Breadth-First-Suche (BFS) wird der Verbund der Äquivalenklassen, die durch die rekursive Anwendung von θ_k entstehen, bottom-up untersucht. Das heißt alle Klassen einer Ebene werden verarbeitet, bevor man zur nächsten Ebene vorgeht. Der Vorteil dieser Vorgehensweise ist, dass man beispielsweise die Menge der 2-Sequenzen kennt bevor die 3-Sequenzen gebildet werden. DFS verbraucht weniger Arbeitsspeicher als BFS.
Bei der Depth-First-Suche (DFS) werden alle Äquivalenzklassen eines Pfades bearbeitet, bevor man zum nächsten Pfad übergeht. In R sind beide Suchstrategien implementiert, wobei in dieser Arbeit DFS verwendet wurde.

Verknüpfung von ID-Listen

Hier soll die Verknüpfung von zwei ID-Listen etwas genauer erklärt werden. Sei $P = [A \to B]$ eine Äquivalenzklasse mit den Elementen $\{P \to B, P \to C, P \to E\}$. Da bei Online-Tracking-Daten keine Itemsets möglich sind, reduziert sich die Anzahl der mögliche Fälle bei der Verknüpfung. Werden $P \to B$ und $P \to C$ verknüpft, so gibt es nur zwei mögliche Ergebnisse, $P \to B \to C$ oder $P \to C \to B$. Ein Spezialfall ist die Verknüpfung von $P \to B$ mit sich selber. Dabei kann nur die Sequenz $P \to B \to B$

entstehen.

$P \rightarrow B$		$P \rightarrow C$		$P \rightarrow B \rightarrow C$		$P \rightarrow C \rightarrow B$	
ID	Position	ID	Position	ID	Position	ID	Position
1	2	1	7	1	7	8	5
1	3	1	8	1	8	8	8
1	4	3	2	8	3	13	5
4	6	5	7	8	4	13	7
7	4	8	3	8	5		
8	2	8	4	8	8		
8	3	8	5				
8	5	8	8				
8	8	11	3				
13	5	13	2				
13	7	16	8				
15	6	20	2				
17	2						
20	2						

Tabelle 7: Verknüpfung von ID-Listen

Tabelle 7 enthält ein Beispiel zur Verknüpfung von den ID-Listen von $P \rightarrow B$ und $P \rightarrow C$. Um die ID-Liste von $P \rightarrow B \rightarrow C$ zu ermitteln muss überprüft werden, ob es zeitliche Beziehungen gibt. Das heißt, für jedes $(ID_i, Position_i)$-Paar aus der ID-Liste von $P \rightarrow B$ wird überprüft, ob es in der ID-Liste von $P \rightarrow C$ ein $(ID_j, Position_j)$-Paar gibt mit $Position_j > Position_i$ und $ID_i = ID_j$. Wenn dies der Fall ist, so enthält ID_i die Sequenz $P \rightarrow B \rightarrow C$ und das Paar $(ID_i, Position_k)$ wird dessen ID-Liste hinzugefügt. Die ID-Liste von $P \rightarrow C \rightarrow B$ wird analog ermittelt.
Da nur Sequenzen innerhalb einer Klasse, die den selben Prefix besitzen, verknüpft werden, muss nur die Position des letzten Items berücksichtigt werden.

Der Algorithmus

Algorithmus 2 enthält die oberflächliche Struktur von SPADE. Die Hauptschritte sind die Berechnung der häufigen 1-Sequenzen und 2-Sequenzen, die Zerlegung in Prefix-basierte Eltern-Äquivalenzklassen und die Berechnung aller anderen häufigen Sequenzen mittels BFS oder DFS. Im Folgenden sollen die einzelnen Schritte etwas genauer erläutert werden.
Aus der vertikalen Datenbank (Tabelle 4) erhält man die ID-Listen (Tabelle 5). Der Support der 1-Sequenzen entspricht dann jeweils der Anzahl der unterschiedlichen IDs in den ID-Listen. Wenn der Support für ein Item größer als der minimal geforderte Support ist, so wird dieses Item als häufige 1-Sequenz abgespeichert. Dieser Vorgang benötigt nur einen Scan der Datenbank.

Algorithmus 2 SPADE(min_sup, D)

$F_1 = \{frequent\ items\ or\ 1\text{-}sequences\}$
$F_2 = \{frequent\ 2\text{-}sequences\}$
$\epsilon = \{equivalence\ classes\ [X]_{\theta_1}\}$
for all [X] $\in \epsilon$ **do**
  Enumerate-Frequent-Seq([X])
end for

Sei $N = |F_1|$ die Anzahl der häufigen Items und A die mittlere ID-Listen Größe in Bytes. Ein naiver Ansatz ist, alle $\binom{N}{2}$ ID-List-Verknüpfungen durchzuführen und dann jeweils den Support zu berechnen. Dafür müssten $A * N * (N-1)/2$ Bytes gelesen werden, was ungefähr $N/2$ Datenscans entspricht.
Eine effizientere Alternative ist eine nahtlose Transformation der vertikalen ID-Listen in eine horizontale Datenbank. Tabelle 8 stellt diese Transformation für das Beispiel aus Tabelle 5 dar. Die (ID, Position)-Paare kategorisiert nach Items werden zu (Item, Position)-Paaren kategorisiert nach ID transformiert. Anhand der horizontalen Datenbank kann F_2 einfach berechnet werden. Es wird eine Liste aller 2-Sequenzen in jeder ID-Liste gebildet und in einem zweidimensionalen Array werden die Anzahlen der 2-Sequenzen geupdated.

ID	(Item, Position)-Paare
1	(A,1)(A,3)(B,4)(D,2)
2	(B,2)(C,2)
3	(D,1)
4	(A,1)(B,3)(D,2)

Tabelle 8: Transformation in horizontale Datenbank

Daraufhin wird die Menge ϵ aller Äquivalenzklassen $[X]_{\theta_1}$ gebildet. Das heißt alle 2-Sequenzen, die das selbe erste Item haben, werden zu einer Klasse zusammengefasst.
Im letzten Schritt werden dann für alle Klassen $[X]$ aus ϵ die häufigen Sequenzen ermittelt. Dies geschieht durch die Verknüpfung der ID-Listen von allen Elementen aus $[X]$ und der Überprüfung, ob diese den minimalen Support erfüllen. Die als häufig ermittelten Sequenzen liefern die ID-Listen zur Verknüpfung auf der nächsten Ebene. Dieser Prozess wird fortgesetzt bis alle häufigen Sequenzen ermittelt wurden. Dieser Suchvorgang kann mittels BFS oder DFS bewerkstelligt werden.

Pruning

Der SPADE-Algorithmus unterstützt Pruning von Sequenzen. Sei α_1 das erste Element einer Sequenz α. Vor der Verknüpfung einer neuen k-Sequenz β, wird überprüft, ob alle k Subsequenzen der Länge $(k-1)$ von β häufig sind. Ist dies nicht der Fall, kann β keine

häufige Sequenz sein und wird nicht weiter berücksichtigt. Ein Problem dabei ist, dass nur $(k-1)$ der k Subsequenzen in der selben Klasse wie β sind.
Man betrachte das folgende Beispiel mit der 4-Sequenz $\beta = (A \to B \to C \to D)$. Die 4 Subsequenzen der Länge 3 sind $(A \to B \to C)$, $(A \to B \to D)$, $(A \to C \to D)$ und $(B \to C \to D)$, wobei letztere nicht der selben Klasse angehört wie β. Die ersten drei Subsequenzen liegen in der Klasse $[A]$ und die letzte in der Klasse $[B]$. Wenn die Klasse $[B]$ bereits bearbeitet wurde, liegen alle Informationen zum Pruning bereits vor. Ansonsten kann kein Pruning durchgeführt werden, wobei partielles Pruning, dass nur auf den Subsequenzen aus Klasse $[A]$ basiert möglich ist.
Ein Nachteil des Pruning ist, dass alle häufigen Sequenzen der vorherigen Ebenen in einer Art Hashtabelle gespeichert werden müssen. Obwohl diese Art von Pruning eine große Anzahl von potentiellen Sequenzen ausschließen kann, haben Simulationsstudien [30] gezeigt, dass dadurch die Laufzeit nicht verbessert wird. Das liegt vor allem daran, dass die Verknüpfung von ID-Listen besonders für längere Sequenzen sehr schnell ist. Das heißt, dass die Verknüpfung ungefähr gleich schnell ist wie das Überprüfen, ob alle Subsequenzen häufig sind. Zudem kann bei einer großen Anzahl von häufigen Sequenzen der virtuelle Arbeitsspeicher beim Laden eben dieser überschritten werden. Deshalb ist das Pruning von Sequenzen in R nicht implementiert.

6 Visualisierung der Daten anhand eines Netzwerkes

Während Sequential Pattern Mining lediglich häufige Sequenzen in den Daten findet, sollen die Daten zusätzlich noch in Form eines Netzwerkes dargestellt werden. Dies ermöglicht die Visualisierung des kompletten Datensatzes und liefert weitere Informationen bezüglich von Mustern in den Daten.
Zum bessern Verständnis sollen an dieser Stelle einige elementare Grundlagen der Graphentheorie vorgestellt werden. Ein geordneter Graph $G = (V, E)$ besteht aus einer Menge V von Knoten und einer Menge E von Kanten. Eine Kante $e_i \in E$ besteht aus einem geordneten Paar von zwei Knoten (v_j, v_k), wobei $v_j, v_k \in V$. Das heißt eine Kante stellt die geordnete Verbindung zwischen zwei Knoten dar [17, S. 16].
Das Netzwerk ist wie folgt strukturiert. Ausgehend von einem Startpunkt führen 17 Kanten zu den 17 Kampagnen der ersten Position. Diese 17 Kampagnen entsprechen jeweils einem Knoten. Die Namen der Knoten setzen sich aus dem Namen der Kampagne und _1 für Position 1 zusammen. Von jeder Kampagne führt dann eine Kante zu $Success\_1$ für die Funnels, die nach einem Kontakt konvertiert sind sowie eine Kante zu $Fail\_1$ für die Funnels, deren Beobachtung nach einem Konaktpunkt ohne Konvertierung abbricht. Außerdem führen von jeder der 17 Kampagnen der ersten Position jeweils 17 Kanten zu den 17 Kampagnen der zweiten Position. Die Knoten der zweiten Position setzen sich wiederum aus dem Namen der Kampagne und _2 zusammen. Von dort führen wieder Kanten zu $Success\_2$, $Fail\_2$ und den Kampagnen der dritten Position. Dieses Prinzip setzt sich für die weiteren Positionen fort.
Diese Kanten werden bezüglich der Anzahl der Nutzer gewichtet. Anhand imaginärer Zahlen soll hier das Prinzip kurz dargestellt werden. Wenn 50000 Nutzer als ersten Kon-

takt *Direct* haben, so hat der Knoten *Direct*_1 ein Gewicht von 50000. Diese verteilen sich nun auf die Knoten, die mit *Direct*_1 durch eine Kante verbunden sind. Wenn die Kante zwischen *Direct*_1 und *Direct*_2 beispielsweise ein Gewicht von 10000 hat, so haben 10000 Nutzer als ersten und zweiten Kontakt *Direct*. Wenn die Kante zwischen *Direct*_1 und *Success*_1 ein Gewicht von 500 hat, so bedeutet das, dass 500 Nutzer als ersten Kontakt *Direct* haben und danach direkt konvertieren.
Um aussagekräftigere Ergebnisse zu erzielen, wurden für die Gewichtung der Kanten nicht die absoluten Anzahlen gewählt. Stattdessen wurden zwei Strategien verfolgt, die im folgenden als relative Ausgänge beziehungsweise relative Eingänge bezeichnet werden. Für die relativen Ausgänge werden die Kanten mit den relativen Häufigkeiten gewichtet, wobei die zugrundeliegende Menge die Summe aller Nutzer ist, die einen Knoten verlassen. Das heißt die Summe der Gewichte aller Kanten, die einen Knoten verlassen ergeben in der Summe 1. Wenn nun das Gewicht der Kante zwischen *Direct*_2 und *Fail*_2 0.2 entspricht, so sind 20% der Nutzer, die als zweiten Kontakt *Direct* haben, nach dem zweiten Kontaktpunkt konvertiert. Diese Gewichtung erlaubt somit eine relative Betrachtung der Kanten die eine Kampagne verlassen.
Für die relativen Eingänge werden die Kanten ebenfalls mit den relativen Häufigkeiten gewichtet, wobei die zugrundeliegende Menge nun die Summe aller Nutzer ist, die in einen Knoten gehen. Betrachtet man beispielsweise *Success*_2, so kann man erkennen, aus welchen Kampagnen der zweiten Position sich die konvertierten Funnels der Länge 2 zusammen setzen. Hat die Kante zwischen *Direct*_2 und *Success*_2 nun einen Gewicht von 0.2, so haben 20% der konvertierten Funnels der Länge 2 als letzten Kontaktpunkt vor der Konvertiertung *Direct*. Somit sind zwei verschiedene Interpretationsmöglichkeiten durch die unterschiedliche Gewichtung der Kanten gegeben.
Das Netzwerk mit den gewichteten Kanten wird in *R* mit dem Paket *rgexf* [29] erzeugt und als *gexf*-Datei exportiert. Daraufhin wird die Datei in das Open-Source Programm *Gephi* [4] geladen. Dort werden die Daten dann als Graph dargestellt. Abbildung 12 zeigt die räumliche Anordnung der Knoten und Kanten nach dem ersten Einlesen der *gexf*-Datei in *Gephi*, die noch willkürlich ist. Die farbigen Punkte sind die Knoten und die Linien die Kanten.
Für die Berechnung der räumlichen Anordnung der Knoten und Kanten stehen in *Gephi* einige Algorithmen zur Verfügung. Diese berechnen die Anordung der Knoten anhand der Anziehungs- beziehungsweise Abstoßungskraft der Knoten, die aus den relativen Häufigkeiten resultieren. Stärker verbundene Knoten liegen damit näher beieinander als schwach verbundene. Das heißt Knoten, die überhaupt keine Verbindung enthalten, liegen weiter auseinander. Ein Beispiel wäre *Direct*_1 und *Direct*_3, da man von der ersten Position nicht direkt zur dritten Position springen kann, sondern zunächst einen Kontaktpunkt an der zweiten Position benötigt. Durch die Anwendung eines Algorithmus ergibt sich somit eine lineare Struktur, die die Positionen aneinander reiht.
Für diese Arbeit wurden der Algorithmus *Force Atlas 2* [16] und der Algorithmus nach *Yifan Hu* [14] verwendet, wobei *Yifan Hu* das Netzwerk besser in einzelne Ebenen einteilen kann. Das heißt die Positionen sind hier räumlich deutlicher getrennt. Für die Präsentation der Ergebnisse in Kapitel 7.3 wurde die räumliche Anordnung der Knoten allerdings noch manuell bearbeitet, um die Ergebnisse anschaulicher zu machen.

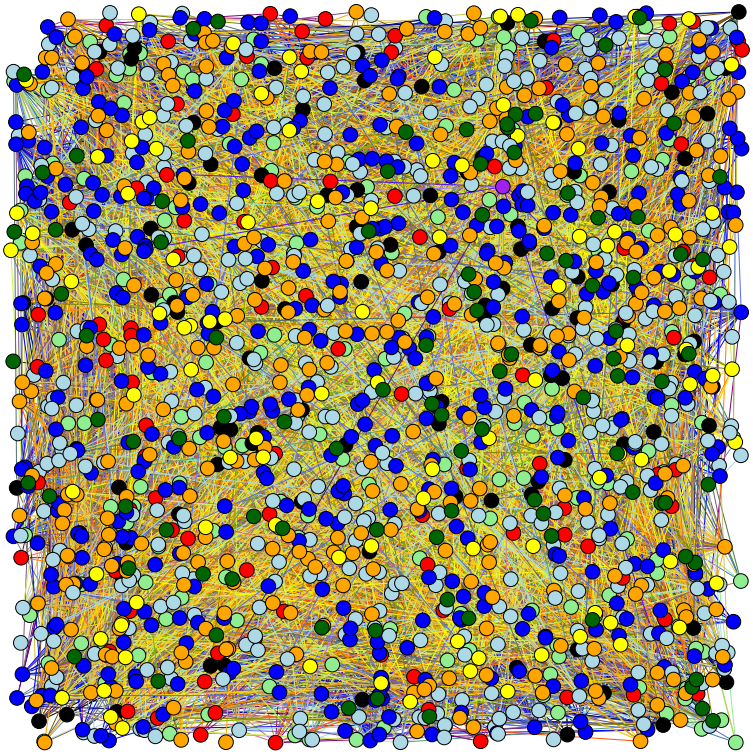

Abbildung 12: Anordnung der Knoten und Kanten direkt nach dem Einlesen in Gephi

Das Netzwerk kann in *Gephi* interaktiv bearbeitet werden. Die Daten sowie das Programm sind im elektronischen Anhang enthalten und werden in Kapitel 9 näher beschrieben.

7 Ergebnisse

7.1 Zeitdiskretes Survival-Modell

Die in diesem Kapitel präsentierten Ergebnisse basieren jeweils auf der bestmöglichen Anzahl an Iterationen. Diese wurden mittels 5-facher Kreuzvalidierung ermittelt und sind in Abbildung 13 dargestellt. Auf der x-Achse ist die Position von 1 bis 25 aufgetragen und auf der y-Achse die Optimale Iterationsanzahl, die durch $n.trees = 3000$ nach oben begrenzt ist. Es ist zu erkennen, dass für die Positionen 1 bis 4 über 2500 Stümpfe für die Ergebnisse verwendet werden. Dann fällt die Kurve sehr schnell ab bis sie sich bei ungefähr 500 Iterationen einpendelt. Dieser Abfall ist dadurch zu erklären, dass mit steigender Position die Datenmenge sinkt.

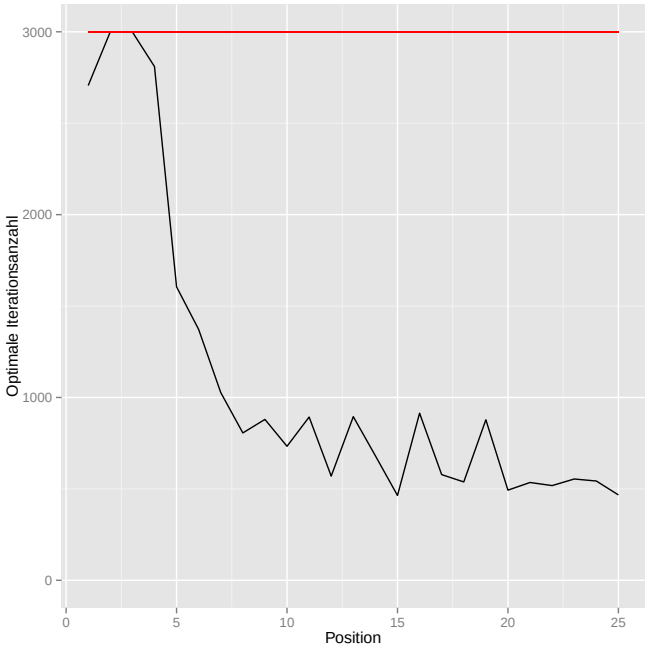

Abbildung 13: Optimale Iterationsanzahl des Stochastic Gradient Boosting

Der relative Einfluss der Features gibt dessen Wichtigkeit bei der Erstellung der Prädiktorfunktion an und ist in Abbildung 14 dargestellt. Auf der x-Achse ist erneut die Position aufgetragen und auf der y-Achse der relative Einfluss. Der relative Einfluss aller Features an einer Position ergibt summiert jeweils 100 und die Balken sind der Größe nach geordnet, das heißt die wichtigsten Variablen sind unten abgebildet. Die Farben der Balken werden in der Legende rechts neben der Abbildung erläutert.
An Position 1 sind nur drei Features vorhanden, wobei *Campaign*, das heißt die Art des Kontaktpunktes, für circa 90% der Minimierung der Verlustfunktion verantwortlich ist. Die Variablen *Hour* und *Weekday* haben kaum Einfluss. An Position 2 ist die Kampagne immer noch das stärkste Feature, wobei hier auch die Kampagne des ersten Kontaktpunktes und die Zeit, die seit dem ersten Kontaktpunkt vergangen ist, Einfluss haben.
An den Positionen 8 und 22 hat die Kampagne des letzten Kontaktpunktes den stärksten Einfluss und damit selbstverständlich auch einen höheren Einfluss als die Kampagne des aktuellen Kontaktpunktes, die insbesondere an den höheren Positionen oft nur noch einen geringen Einfluss hat. Mit steigender Position wird auch die Kampagne des vorletzten Kontaktes wichtiger und ist manchmal das wichtigste Feature. Die Zeit, die seit dem vorherigen Kontaktpunkt vergangen ist, ist bereits an Position 3 zweitwichtigstes Feature und spielt auch für die folgenden Positionen eine Rolle. Später nimmt die Wichtigkeit dieses Features allerdings ab und die Zeit, die seit dem ersten Kontaktpunkt vergangen ist, wird bedeutend wichtiger. Die Features *Hour* und *Weekday* spielen positionsübergreifend kaum eine Rolle.
Nachdem die Wichtigkeit der Variablen für die Klassifizierung in konvertierte und nichtkonvertierte Funnel präsentiert wurde, ist nun die Art dieses Einflusses interessant.

Dieser wird im folgenden für einige Positionen dargestellt, wobei die Variablen dafür nach positionsspezifischer Wichtigkeit ausgesucht werden und das Augenmerk vor allem auf die niedrigeren Positionen fällt, da dort mehr Daten vorhanden sind. Folglich werden die Features *Hour* und *Weekday* garnicht präsentiert, das sie kaum einen Einfluss haben. Die Kampagne des vorletzten Kontaktpunktes *CampaignLast2* wird an dieser Stelle ebenfalls nicht berücksichtigt, da die Ergebnisse für *Campaign* beziehungsweise *CampaignLast* oft sehr ähnlich sind. Im elektronischen Anhang sind die marginalen Effekte aller Features für jeweils jede Position zu finden.

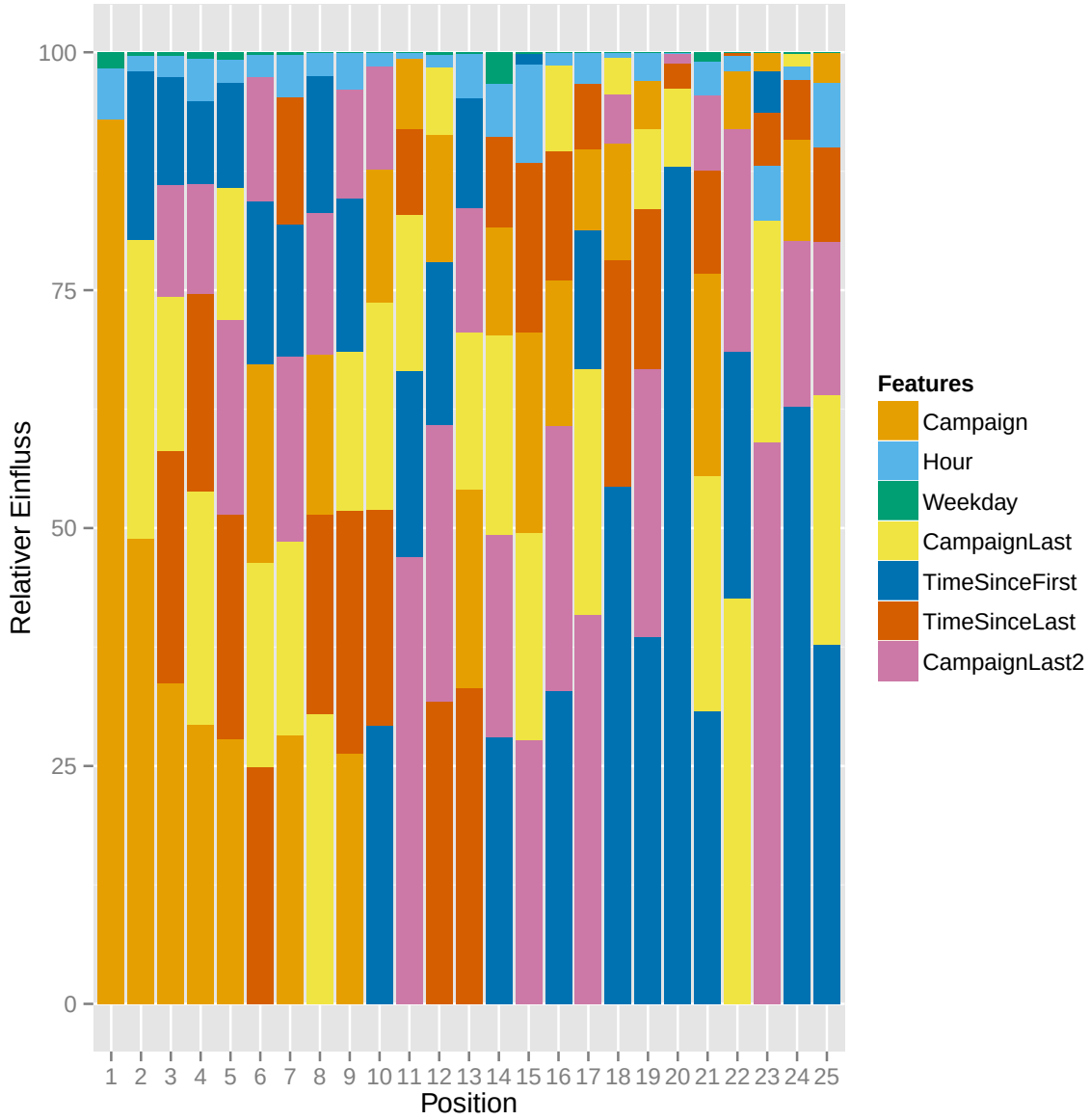

Abbildung 14: Wichtigkeit der Variablen

In Abbildung 15 sind die marginalen Effekte der Variable *Campaign* für die Positio-

nen eins bis vier, von links oben nach rechts unten, aufgetragen. *Campaign* ist an den ersten vier Positionen das wichtigste Feature. Mit steigender Position verliert die Variable deutlich an Einfluss. Auf der x-Achse ist die Art des aktuellen Kontaktpunnktes aufgetragen und auf der y-Achse der Marginale Effekt.
An Position eins haben die Kampagnen *Affiliate - Rest*, *E-Mailing* und *SEM - Brand* im Vergleich zu den restlichen Kampagnen einen starken, positiven Einfluss auf die Konvertierungswahrscheinlichkeit.

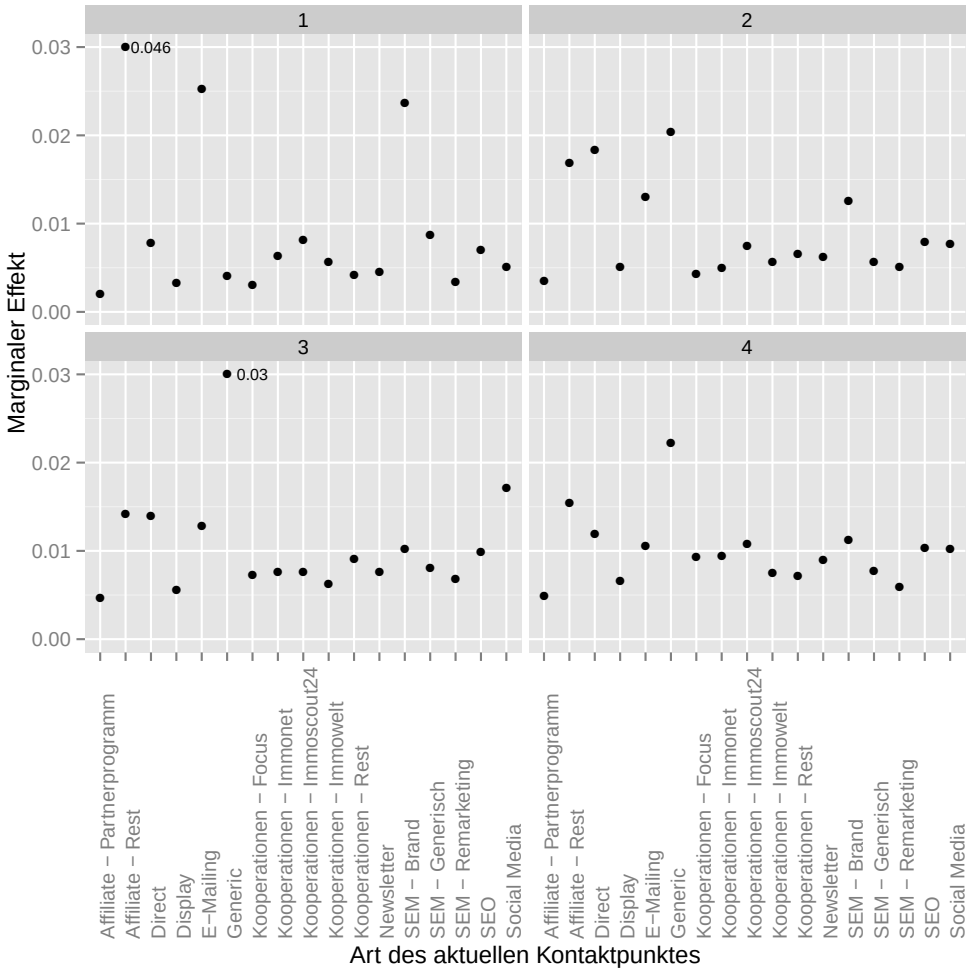

Abbildung 15: Marginaler Effekt des Features *Campaign* an den Positionen 1, 2, 3 und 4

Das heißt das Bereitstellen von Zinsvergleichen, welche das Zinsangebot der Interhyp AG mit deren Wettbewerbern im Vergleich darstellt, durch die Partner unter *Affiliate - Rest* scheint oft an Position 1 bereits zur Konvertierung zu führen. Die Partner unter *Affiliate - Partnerprogramm*, die hauptsächlich Rechner der Interhyp AG auf ihren Seiten einbinden, teilweise aber auch Banner schalten oder Verlinkungen in Texten unterbringen, haben an Position 1 einen deutlich geringeren Effekt. *E-Mailing* sind Mails,

die an Interessenten, die schon einen Antrag gestellt haben oder ein Infopaket angefordert hatten, versendet werden. Da dieses Klientel sich offentsichtlich bereits intensiv mit einer Baufinanzierung beschäftigt hat, macht es Sinn, dass diese Mails schon nach wenigen Kontakten häufig zu einer Konvertierung führen und dieser Effekt für spätere Positionen schwächer ist. SEM sind bezahlte Suchergebnisse, hauptsächlich auf Google. *SEM - Brand* bedeutet, dass der Suchbegriff das Wort *Interhyp* enthielt. Dieser Effekt ist deutlich größer als für *SEM - Generisch* beziehungsweise *SEM - Remarketing*, wobei ersteres bedeutet, dass etwas wie *Baufinanzierung* gesucht wurde und zweiteres, dass der potentielle Kunde bereits zuvor auf der Seite von Interhyp war und deshalb nochmal eine Einblendung mit Werbung der Interhyp AG bekommen hat.

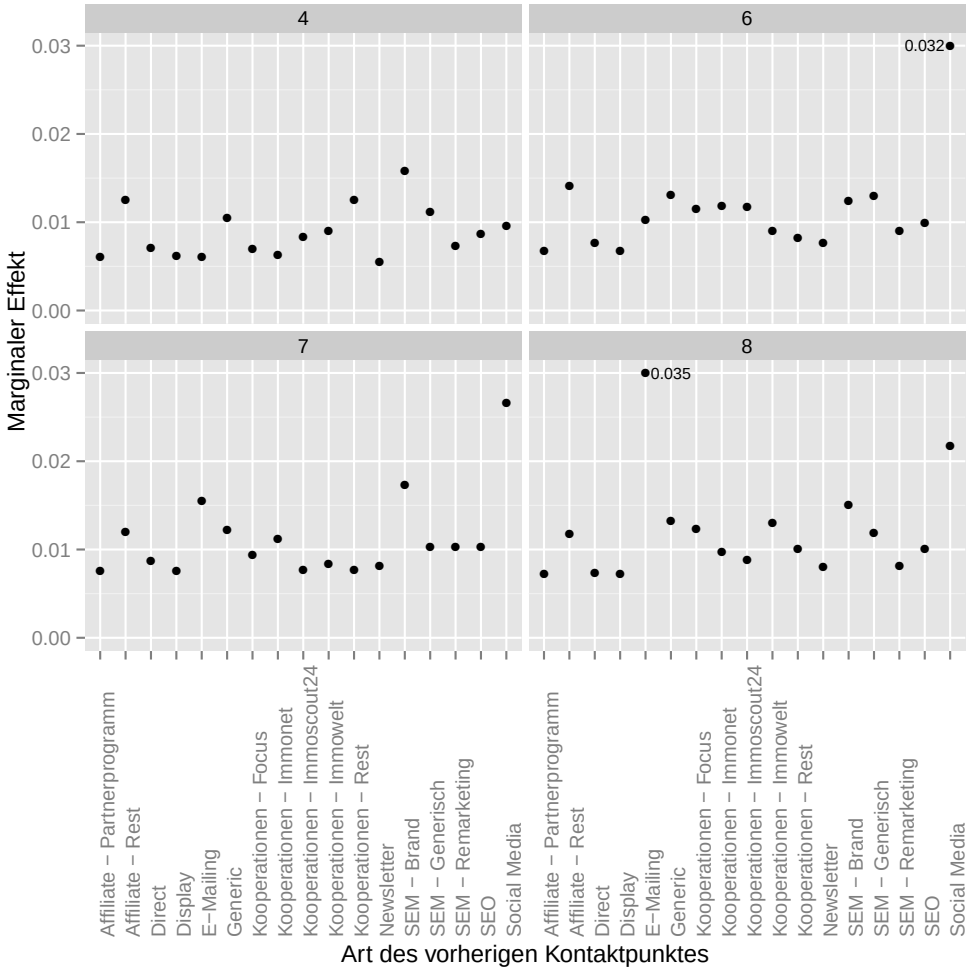

Abbildung 16: Marginaler Effekt des Features *CampaignLast* an den Positionen 4, 6, 7 und 8

An Position 2 stechen neben *Affiliate - Rest*, *SEM - Brand* und *E-Mailing* auch *Direct* und *Generic* durch ihre Marginalen Effekte leicht hervor. *Direct* bedeutet, dass jemand im Browser direkt *www.interhyp.de* eingegeben hat und *Generic*, dass jemand über einen

unbezahlten Link zur Interhyp kam. Für die Positionen 3 und 4 ist vor allem *Generic* wichtig. *Display*, *Social Media*, *SEO* sowie die Kooperationen spielen, neben den bereits erwähnten, keine große Rolle für die ersten vier Positionen.
An den Positionen 4, 6 und 7 ist *CampaignLast* das zweitwichtigste Feature sowie an Position 8 das wichtigste. Dessen marginalen Effekte sind in Abbildung 16 dargestellt. Hier heben sich teilweise die selben Kategorien hervor, wobei die Unterschiede zwischen den Kampagnen kleiner sind. Überraschend ist, dass hier auch *Social Media* einen starken Effekt hat. Allerdings liegen hierfür kaum Daten vor, wie in Abbildung 7 zu erkennen ist.

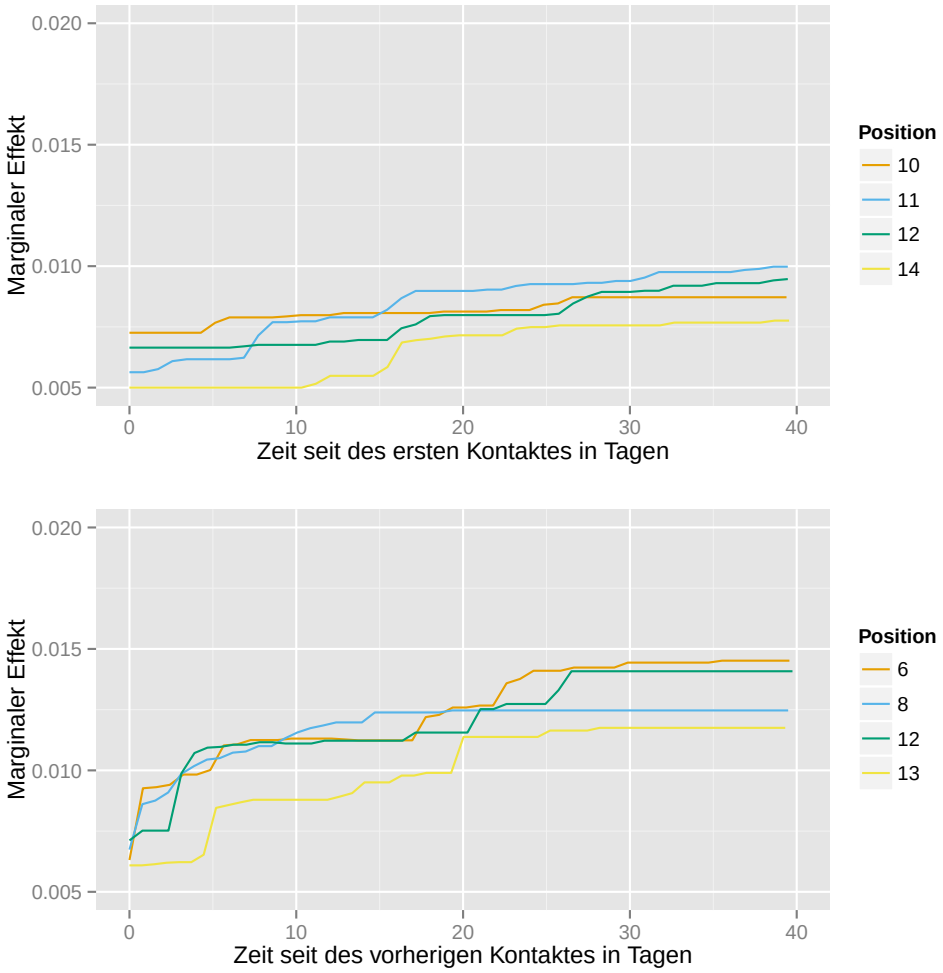

Abbildung 17: Marginaler Effekt des Features *TimeSinceFirst* an den Positionen 10, 11, 12 und 14 (oben) und des Features *TimeSinceLast* an den Positionen 6, 8, 12 und 13 (unten)

Das Feature *TimeSinceFirst* ist an den Positionen 10 und 14 das wichtigste sowie an Position 11 und 12 das zweit- beziehungsweise drittwichtigste. *TimeSinceLast* ist an den Positionen 6, 12 und 13 das wichtigste Feature und an Position 8 das zweitwichtigste

Feature. Die marginalen Effekte für diese Positionen sind in Abbilung 17 dargestellt, wobei diese an anderen Positionen ähnlich aussehen. Auf der x-Achse ist die Zeit in Tagen aufgetragen und auf der y-Achse die Marginalen Effekte. Für beide Features und alle Positionen steigen die marginalen Effekte mit der Zeit, wobei der Effekt bei *TimeSinceLast* etwas größer ist.
Dass die Konvertierungswahrscheinlichkeit bei zeitlich längeren Funnels höher ist, könnte man dadurch erklären, dass die Entscheidung für den Bau eines Hauses oder ähnlichem und das damit verbundene Ausfüllen eines Online-Antrages viel Zeit in Anspruch nimmt. Dass für die Zeit seit des vorherigen Kontaktes der selbe Effekt auftritt, könnte man ähnlich erklären. Bei der Entscheidung für eine Baufinanzierung sind viele Dinge zu beachten. So erscheint es plausibel, dass sich interessierte Kunden eine längere Zeit auch offline mit dem Thema beschäftigen und später wieder online den Kontakt zur Interhyp AG suchen.
Bis hierhin wurden die Ergebnisse des, auf den Trainingsdaten geschätzten, Modells vorgestellt. Nun soll das Modell auf die Testdaten angewendet werden, um die Prognosegüte zu beurteilen.

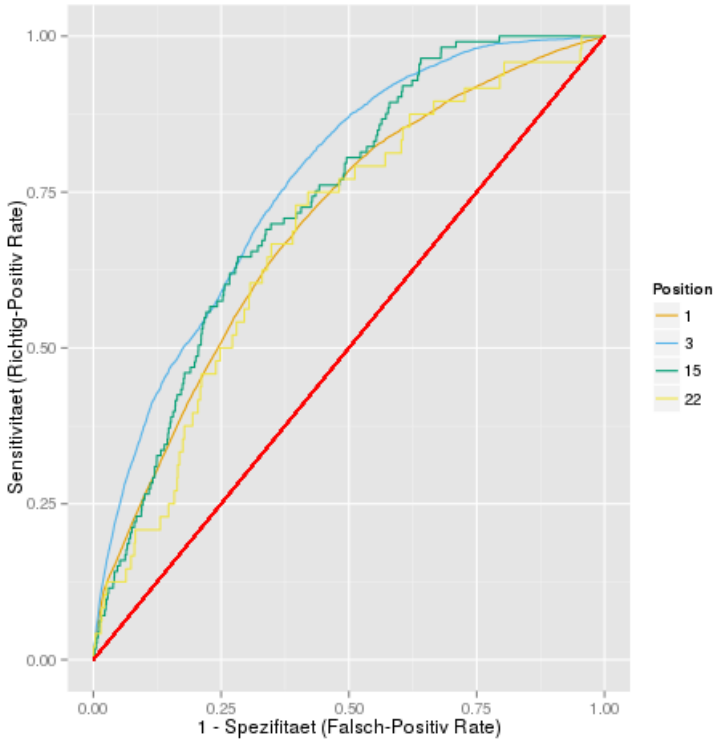

Abbildung 18: ROC-Kurve für die Positionen 1, 3, 15 und 22

Wenn man die geschätzte Prognosefunktion auf die Testdaten anwendet, bekommt man für jede Position und jede ID aus den Testdaten die Hazardrate, das heißt die Wahrscheinlichkeit an einer bestimmten Position zu konvertieren. Diese sind sowohl für konvertierte als auch nicht-konvertierte Funnels sehr niedrig, da die absolute Anzahl an

nicht-konvertierten Funnels deutlich überwiegt. Um die Prognosegüte des Modells näher zu betrachten wurde deshalb für jede Position eine ROC-Kurve berechnet. Diese sind in Abbildung 18 für vier Positionen beispielhaft dargestellt.
Auf der x-Achse ist der Anteil der nicht-konvertierten Funnels, die als konvertiert vorhergesagt wurden aufgetragen und auf der y-Achse der Anteil der konvertierten Funnels, die auch als konvertiert vorhergesagt wurden. Je weiter die ROC-Kurve also oberhalb der roten Diagonalen liegt, desto besser kann das Modell zwischen konvertierten und nicht-konvertierten Funnels trennen. Für Position 3 liegt die Kurve für alle Punkte überhalb der Kurve von Position 1. Dies ist dadurch zu erklären, dass an Position 3 mehr Informationen zur Erstellung des Modells verwendet werden als an Position 1. Für Position 15 und 22 ist das Modell schlechter als an Position 3. Die Kurven für die späteren Positionen sind deutlich rauher. Dies ist darauf zurück zu führen, dass dort weniger Datenpunkte vorhanden sind. Durch das Einbringen des Offsets kann die Prognoseleistung des Modells allerdings halbwegs konstant gehalten werden.
Dies ist in Abbildung 19 noch deutlicher zu erkennen. Hier ist die Fläche unterhalb der ROC-Kurve (AUC) für jede Position abgebildet. Dieser Wert fällt zwischen Position 15 und 22 unter 0.75 ab, hält sich ansonsten aber relativ konstant auf 0.75. Das heißt die Wahrscheinlichkeit, dass bei einer konvertierten Beobachtung die Hazardrate größer ist als bei einer nicht-konvertierten Beobachtung ist positionsübergreifend circa 0.75.

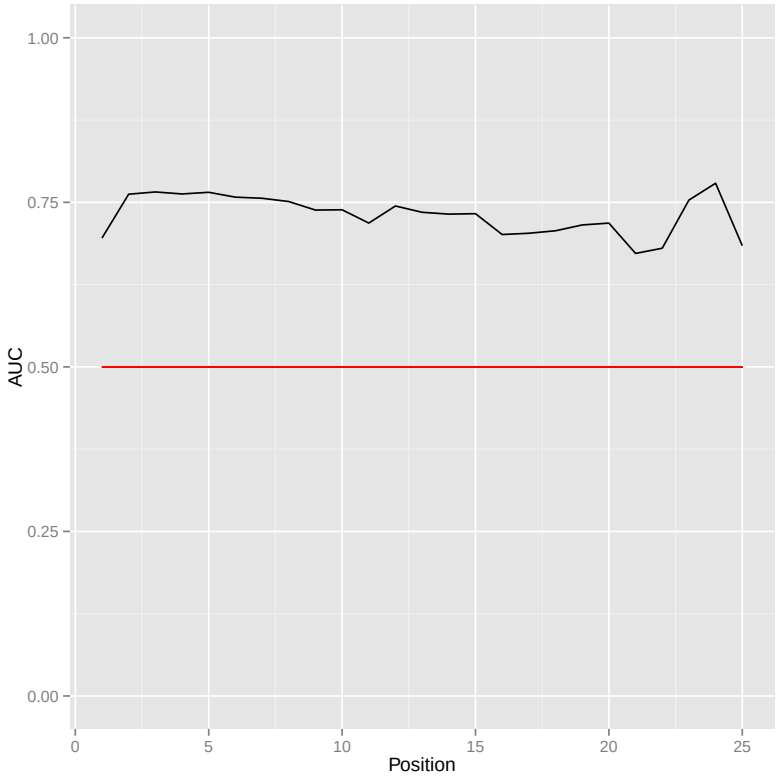

Abbildung 19: AUC für alle Positionen

7.2 Sequential Pattern Mining

Der in Kapitel 5 beschriebene Algorithmus wurde zunächst seperat auf alle konvertierten sowie nicht-konvertierten Funnels angewendet. Dabei wurden Funnels mit der Länge eins ausgeschlossen, da in diesen offentsichtlich keine interessanten Sequenzen enthalten sein können. Der minimale Support wurde auf 0.05 gesetzt, so dass nur Sequenzen, die in mindestens 5% aller Funnels vorkommen, als Ergebnis ausgegeben werden.

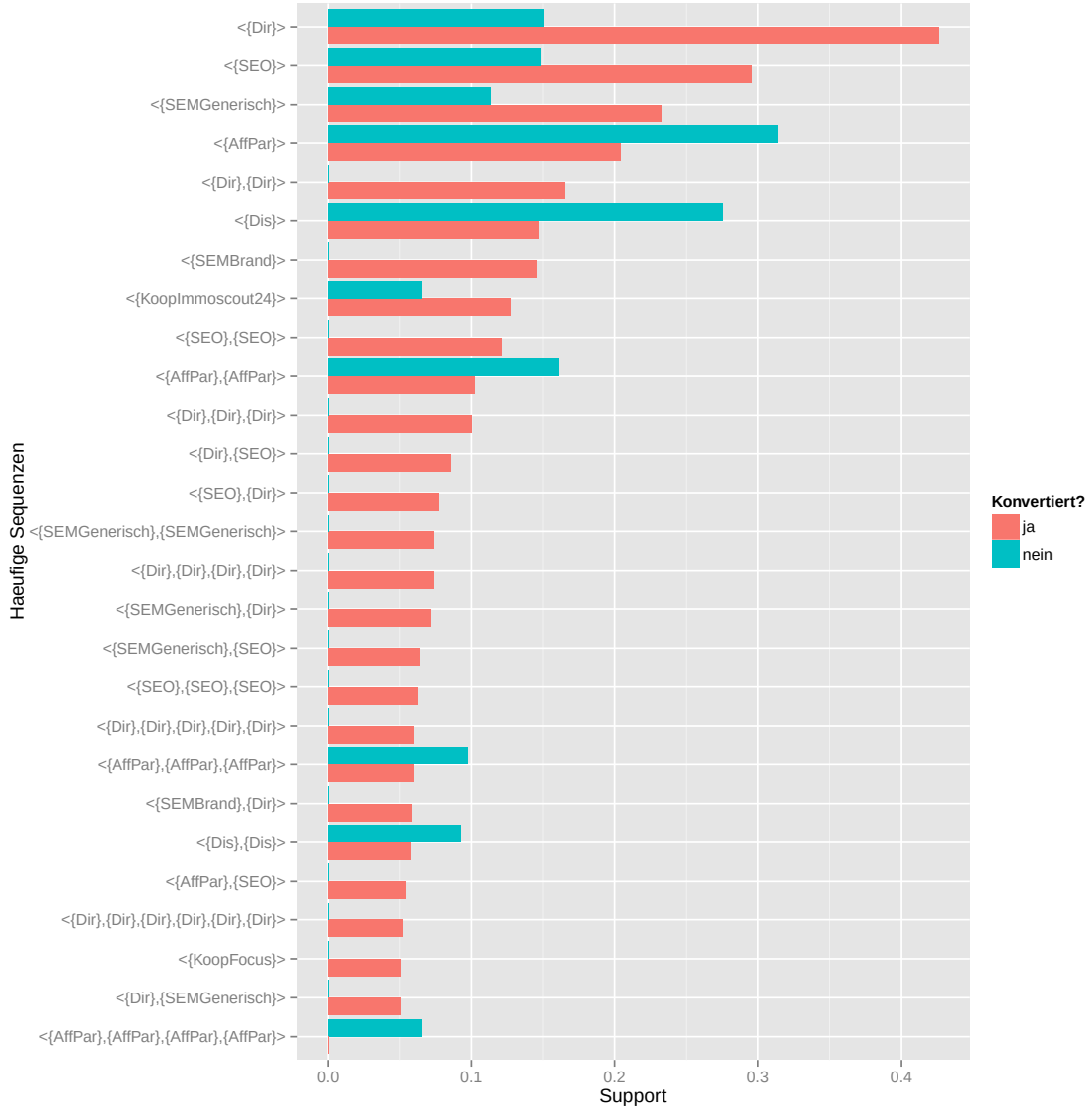

Abbildung 20: Häufige Sequenzen in den konvertierten und nicht-konvertierten Funnels

In Abbildung 20 sind diese Sequenzen geplottet, wobei auf der x-Achse der Suppport aufgetragen ist und die Sequenzen anhand des Supports der Sequenzen in den konvertierten Funnels absteigend geordnet sind. An den Stellen, wo keine Balken geplottet sind,

war der Support kleiner als 0.05. Die Namen der Kampagnen sind teilweise abgekürzt, um die Darstellung zu verbessern. *AffPar* steht für *Affiliate - Partnerprogramm*, *Dir* für *Direct* und *Dis* für *Display*. Außerdem ist zu beachten, dass die Sequenzen nicht exakt in der dargestellten Weise in den Funnels vorkommen müssen, sondern auch Abstände zwischen den aufeinanderfolgenden Kampagnen erlaubt sind.

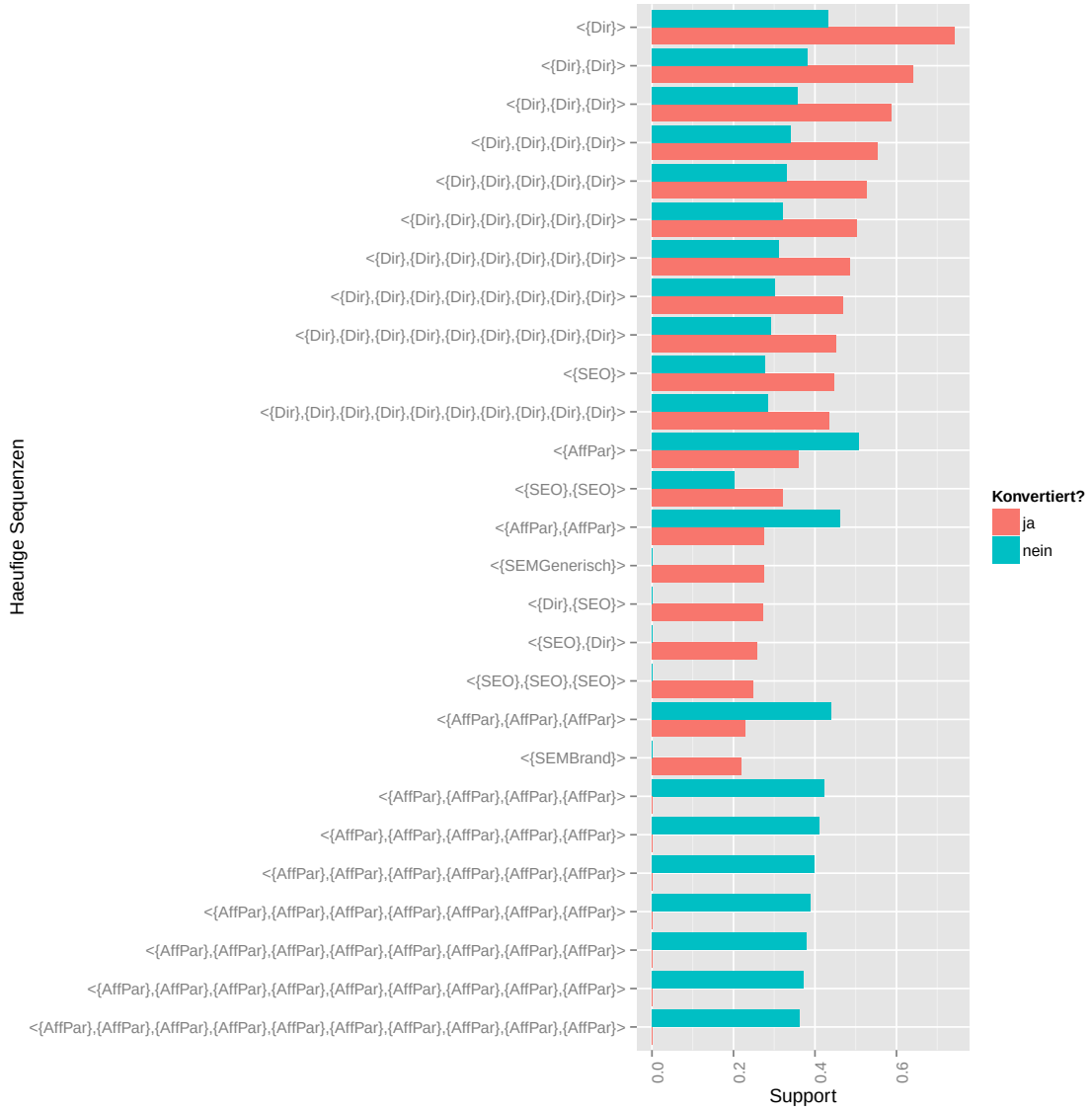

Abbildung 21: Häufige Sequenzen in den konvertierten und nicht-konvertierten Funnels mit mindestens 15 Kontakten

Der Support von Sequenzen der Länge eins, wie *<{Dir}>* oder *<{SEO}>*, geben lediglich an, wie groß der Anteil der Funnels ist, die diese Kampagne mindestens einmal enthalten. Interessanter sind Sequenzen, die mindestens zwei Kampagnen enthalten. Hier fällt auf, dass Sequenzen mit wiederholtem *Direct*-Kontakt in den konvertierten Funnels

stärker sind. Umgekehrt sind Sequenzen mit wiederholtem *Affiliate - Partnerprogramm*-Kontakt in den nicht-konvertierten Funnels stärker. Allerdings haben die Sequenzen insgesamt einen sehr geringen Support. Das liegt vor allem daran, dass die Daten zu großem Teil aus sehr kurzen Funnels bestehen (siehe Abbildung 8).
Deshalb wurde der SPADE-Algorithmus erneut seperat auf konvertierte und nicht-konvertierte Funnel angewendet, wobei dieses mal nur Funnels verwendet wurden, die eine Mindestlänge von 15 haben. Der minimale Support wurde auf 0.2 erhöht, um die Anzahl an häufigen Sequenzen zu beschränken. Die Ergebnisse sind in Abbildung 21 dargestellt.
Die Sequenzen haben jetzt teilweise einen Support, der größer als 0.5 ist, das heißt mehr als die Hälfte der Funnels beinhalten diese Sequenz. Hier ist noch deutlicher zu erkennen, dass die *Direct*-Sequenzen in den konvertierten Funnels stärker sind als in den nicht-konvertierten Funnels. Die *Affiliate - Partnerprogramm*-Sequenzen haben in den nicht-konvertierten Funnels einen Support von ungefähr 40% und in den konvertierten Funnels einen Support von unter 20%.

7.3 Netzwerk

Mit dem Programm *Gephi* ist es möglich interaktiv mit dem Netzwerk zu arbeiten. Eine komplette Analyse des Netzwerkes würde den Rahmen dieses Berichtes sprengen. Deshalb soll hier nur beispielhaft, anhand einiger Grafiken, dargestellt werden, welche Ergebnisse man aus dem Netzwerk ziehen kann. Dafür wurde ein Ausschnitt an Position 2 gewählt. Eine genauere Anleitung zur Verwendung von *Gephi* ist zudem in Kapitel 9 zu finden.

Relative Ausgänge

Zunächst werden die relativen Ausgänge präsentiert, das heißt die Gewichte aller Kanten, die einen Knoten verlassen, ergeben in der Summe 1.
Abbildung 22 enthält den Ausschnitt des Netzwerkes an Position 2 mit den relativen Ausgängen. Die Knoten sind farblich codiert und mit Labels versehen. Am rechten Bildrand ist noch eine Kampagne der ersten Position *SEO_1* zu erkennen. Von dort und den anderen Kampagnen der ersten Position führen die Kanten zu den Kampagnen der zweiten Position in der Mitte der Abbildung. Für Funnels, die mehr als zwei Kontaktpunkte haben gehen die Kanten von diesen Kampagnen zu den Knoten der dritten Positionen jenseits des linken Bildrandes. Die Kampagne *Affiliate - Partnerprogramm_3* ist noch zu erkennen.
Funnels, die lediglich über zwei Kontaktpunkte verfügen, enden in *Fail_2* im Falle einer Nicht-Konvertierung beziehungsweise in *Success_2* im Falle einer Konvertierung.

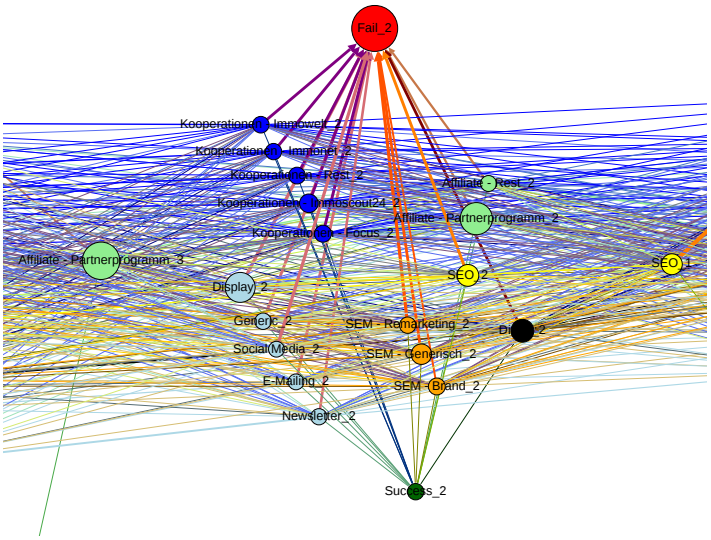

Abbildung 22: Auschnitt des Netzwerkes an Position 2 mit relativen Ausgängen

Die Dicke der Kanten hängt von deren Gewichtung ab. Abbildung 23 enthält den selben Ausschnitt, wobei nun ein Filter von 0.02 auf die Kanten angewendet wurde. Somit sind nur noch Kanten abgebildet, deren Gewicht größer als 0.02 ist. Durch das Auswählen des Knoten *Success_ 2* werden alle Knoten hervorgehoben, die nach der Anwendung des Filters noch durch eine Kante mit *Success_ 2* verbunden sind.

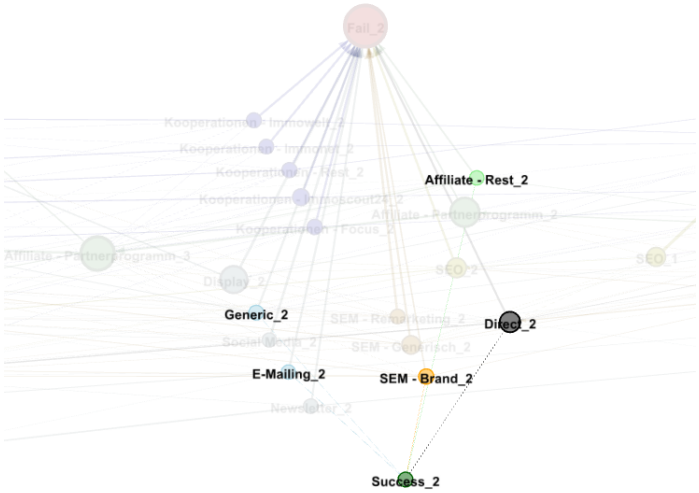

Abbildung 23: Auschnitt des Netzwerkes an Position 2 mit relativen Ausgängen, Filter 0.02 und Fokus auf den konvertierten Funnels

Das heißt über 2% aller Nutzer, die als zweiten Kontaktpunkt *Direct* haben, konvertieren an der zweiten Position. Selbiges gilt für *Affiliate - Rest*, *Generic*, *SEM - Brand*

und *E-Mailing*. Durch Veränderung der Filter und weiterer Einstellungen können die Ergebnisse auch noch genauer betrachtet werden. Wie bereits erwähnt, ist in Kapitel 9 ein Tutorial dazu vorhanden.
Interessant ist an dieser Stelle, dass die fünf hervorgehobenen Kampagnen gleichzeitig die fünf stärksten Kampagnen bei den Marginalen Effekten des Features *Campaign* an Position 2 sind. Das heißt die Ergebnisse des Survival-Modells können durch die Betrachtung des Netzwerkes überprüft und bestätigt werden.
Abbildung 24 enthält erneut den selben Ausschnitt des Netzwerkes, wobei nun ein Filter von 0.5 angewendet wurde und der Fokus auf den nicht-konvertierten Funnels liegt. Von den hervorgehobenen Knoten enden also jeweils mindestens die Hälfte der Beobachtungen ohne Konvertierung.
Analog zu der Betrachtung von *Success_2* oder *Fail_2* können auch die Kampagnen durch Auswählen näher betrachtet werden. Durch die positionsübergreifende Betrachtung können zudem die häufigen Sequenzen, die mittels des Sequential-Pattern-Mining Algorithmus ermittelt wurden, erkannt werden.

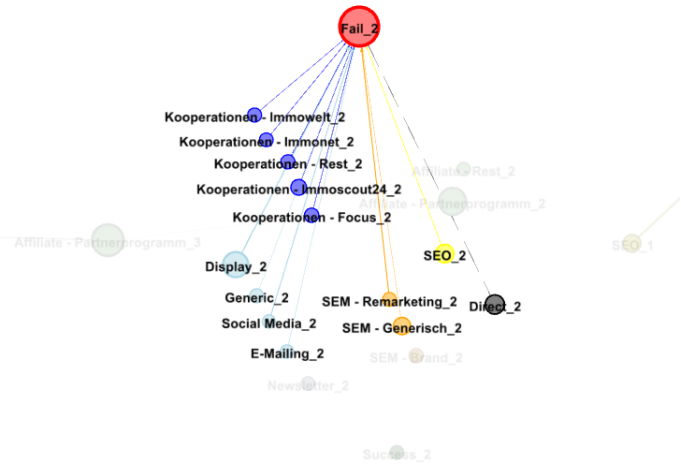

Abbildung 24: Auschnitt des Netzwerkes an Position 2 mit relativen Ausgängen, Filter 0.5 und Fokus auf den nicht-konvertierten Funnels

Relative Eingänge

In diesem Abschnitt wird erneut Position 2 betrachtet, wobei die Kanten nun anhand der relativen Eingänge gewichtet sind. Das heißt die Gewichte aller Kanten, die in einen bestimmten Knoten gehen, ergeben aufsummiert 1. Abbildung 25 enthält den Ausschnitt des Netzwerkes. Am rechten beziehungsweise linken Bildrand ist erneut eine Kampagne von Position 1 beziehungsweise 3 zu erkennen. Die Farben und Beschriftungen der Knoten sind die selben wie in Abbildung 22.

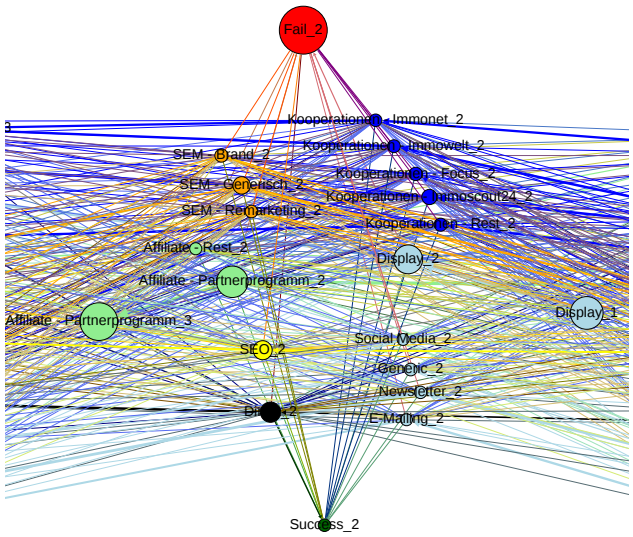

Abbildung 25: Auschnitt des Netzwerkes an Position 2 mit relativen Eingängen

Auf diesen Ausschnitt wurde ein Filter von 0.1 angewendet, wobei der Fokus auf den konvertierten Funnels liegt. Der resultierende Graphh ist in Abbildung 26 zu sehen.

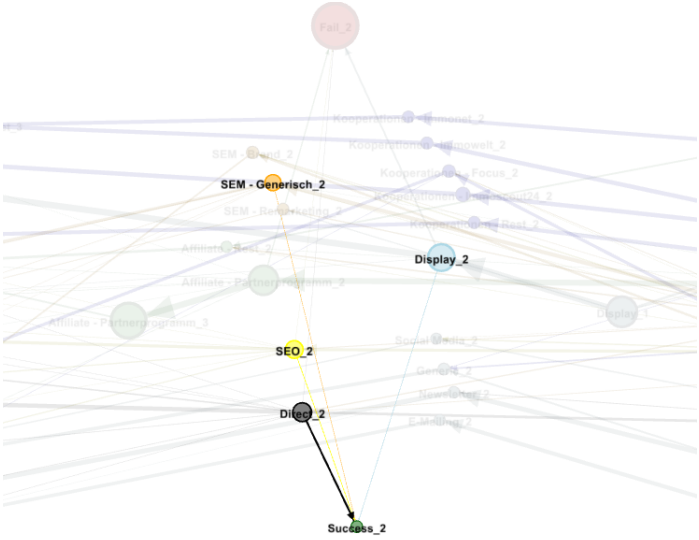

Abbildung 26: Auschnitt des Netzwerkes an Position 2 mit relativen Eingängen, Filter 0.1 und Fokus auf den konvertierten Funnels

Die hervorgehobenen Knoten *SEM-Generisch_ 2*, *Display_ 2*, *SEO_ 2* und *Direct_ 2* machen jeweils mehr als 10% der konvertierten Funnels aus, die aus zwei Kontaktpunkten bestehen. Das heißt hier kann nun betrachtet werden, aus welchen Kampagnen des letzten Kontaktpunktes sich die Menge der konvertierten beziehungsweise nicht-konvertierten Funnels zusammensetzen. Zudem ist es auch möglich eine Kampagne zu

markieren, so dass erkennbar wird, aus welchen Kampagnen der vorherigen Position, sich diese zusammensetzt. Im Gegensatz dazu, war die Interpretation im vorherigen Kapitel eine andere. Dort wurde betrachtet, wohin die Nutzer von einem Knoten aus gehen.
Abbildung 27 enthält den selben Ausschnitt mit dem selben Filter von 0.1, wobei nun die nicht-konvertierten Funnels betrachtet werden. Das heißt von den Funnels, die nach zwei Kontaktpunkten ohne Konvertierung abbrechen, haben jeweils mindestens 10% als letzten Kontaktpunkt *Affiliate - Partnerprogramm_ 2*, *Display_ 2*, *SEO_ 2* oder *Direct_ 2*.

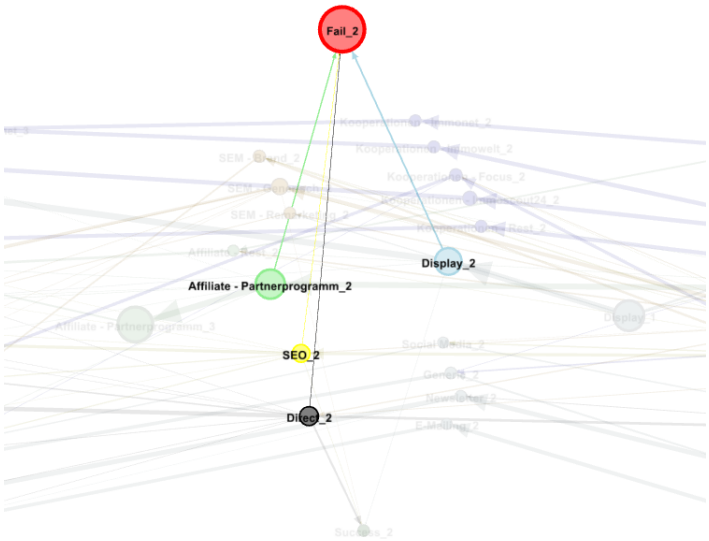

Abbildung 27: Auschnitt des Netzwerkes an Position 2 mit relativen Eingängen, Filter 0.1 und Fokus auf den nicht-konvertierten Funnels

8 Zusammenfassung

An dieser Stelle sollen die angewendeten Methoden und Ergebnisse noch einmal kurz zusammengefasst werden.
Es wurde ein zeitdiskretes Survival-Modell aufgestellt, dass für jede Beobachtung an jeder Position die Konvertierungswahrscheinlichkeit schätzt. Dieses Modell kann als Klassifikator in konvertierte und nicht-konvertierte Funnels fungieren. Die Prognosegüte konnte vor allem durch die Anwendung von Stochastic Gradient Boosting und die Einführung des Offsets verbessert werden. Außerdem liefern die Marginalen Effekte der Features eine gute Interpretationsmöglichkeit der Prädiktorfunktion.
Die wichtigsten Ergebnisse an dieser Stelle sind, dass die Kampagnen *Affiliate - Rest*, *E-Mailing*, *SEM - Brand*, *Direct* und *Generic* im Vergleich zu den restlichen Kampagnen besonders gut für die Konvertierung funktionieren. Außerdem haben Funnels mit einer längeren Beobachtungsspanne und einer niedrigeren Frequenz von Kontaktpunkten eine höhere Konvertierungswahrscheinlichkeit.
Die Einführung von Interaktionen in das Modell hat zu einer Verschlechterung der Prognoseleistung geführt und wurde deshalb nicht weiter verfolgt. Außerdem konnten die

Views aufgrund der Datenlage in den statistischen Analysen nicht berücksichtigt werden.
Mit dem Sequential Pattern Mining-Algorithmus sollten häufige Sequenzen in den Funnels entdeckt werden. Von den resultierenden Ergebnissen sind die *Direct*-Sequenzen in den konvertierten und die *Affiliate - Partnerprogramm*-Sequenzen in den nicht konvertierten Funnels erwähnenswert. Allerdings ist zu berücksichtigen, dass zwischen den einzelnen Elementen der Sequenzen auch andere Kampagnen erlaubt sind. Ein weiterer Ansatz wäre nur diejenigen Funnels zu berücksichtigen, die die Sequenzen in der exakten Reihenfolge ohne Lücken vorweisen. Bei der gegebenen Datenlage würde das allerdings zu einer drastischen Reduzierung des Supports der Sequenzen führen, zumal der Support der häufigen Sequenzen aufgrund der überwiegenden Anzahl an kurzen Funnels ohnehin nicht sehr groß ist. Um interessantere Ergebnisse zu erlangen, wurden die Daten deshalb noch in Form eines Netzwerkes visualisiert.
Dieses Netzwerk erlaubt die Visualisierung der gesamten Daten. Außerdem können durch die Gewichtung der Kanten mit den relativen Ein- und Ausgängen wertvolle Informationen aus den Daten gezogen werden. Mit Hilfe des, in diesem Bericht bereitgestellten, Tutorials (siehe Kapitel 9) ist ein interaktives Arbeiten mit dem Netzwerk sowie eine weiterführende Erforschung der Daten möglich.

9 Elektronischer Anhang

9.1 Verzeichnisstruktur

Abbildung 28 zeigt die Verzeichnisstruktur des elektronischen Anhangs. Dieser soll hier noch erklärt werden.
Die rohen Daten, die von der Refined Labs GmbH bereitgestellt wurden, sind im Ordner *raw_data* enthalten und im Ordner *r_data* sind die vorverarbeiteten Daten im *RData*-Format abgespeichert. Der Ordner *sql_scripts* enthält die SQL Skripte, die für die Vorverarbeitung der Daten verwendet wurden. Die R Skripte zur weiteren Vorverarbeitung der Daten, zur deskriptiven Analyse, zur Schätzung des zeitdiskreten Survival-Modells mittels Stochastic Gradient Boosting, zum Sequential Pattern Mining sowie zur Erstellung des Netzwerkes sind im Ordner *r_scripts* gespeichert. Der Unterordner *support_functions* enthält zudem einige Hilfsfunktionen.
Die Ergebnisse sind in dem Ornder *r_results* gespeichert. Der Unterordner *descriptive* enthält die in Kapitel 3 vorgestellten Grafiken als *pdf*-Dateien, wobei die Namen der Dateien aus den Namen der Features resultieren.
Der Ordner *spm* enthält die Ergebnisse des Sequential Pattern Mining-Algorithmus im *RData*-Format sowie als Plot, wie in Kapitel 7.2 vorgestellt. Die Dateien *spadeFail_all.RData* und *spadeSucc_all.RData* sind die Ergebnisse des SPADE-Algorithmus in den nicht-konvertierten und konvertierten Funnels, wobei die kompletten Daten berücksichtigt wurden, und *spm_all.pdf* ist die Visualisierung der Ergebnisse in Form eines Barplots. Außerdem wurde der Algorithmus nur auf Funnels mit mindestens 10 beziehungsweise 15 Kontaktpunkten angewendet. Die Namen dieser Ergebnis-Dateien sind

analog mit den Endungen *min10* beziehungsweise *min15*.

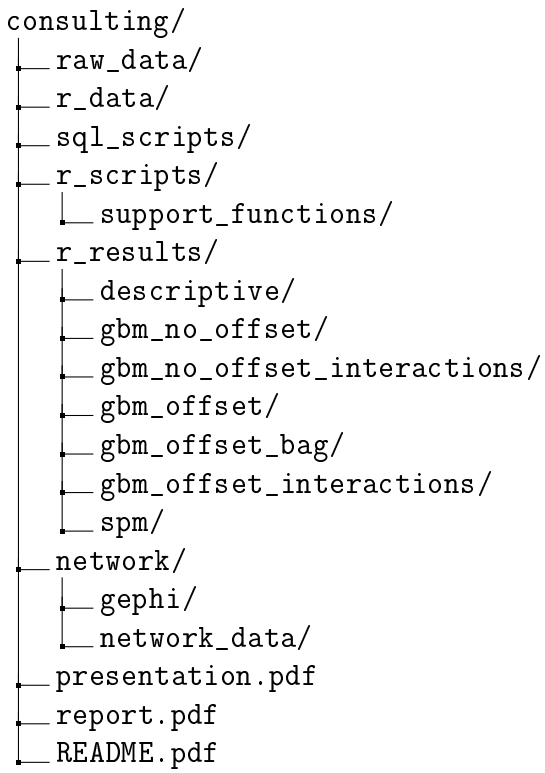

Abbildung 28: Verzeichnisstruktur des Projekts

Der Ordner *gbm_offset* enthält die Ergebnisse des Survival-Modells, die in diesem Bericht vorgestellt wurden. Die Dateien *dataTrain.RData* und *dataTest.RData* enthalten die Trainings- und Testdaten. Die Datei *model.RData* enthält das eigentliche Ergebnis des Modells und *predTrain.RData* und *predTest.RData* enthalten die Vorhersagen auf den Daten anhand des Modells. Die Vorhersagen auf den Trainingsdaten wurden lediglich für den Offset verwendet und die Vorhersagen auf den Testdaten wurden zur Bewertung der Prognosegüte des Modells verwendet. Außerdem enthält der Ordner einige Plots zur Visualisierung der Ergebnisse. Unter anderem sind hier die marginalen Effekte aller Features für jede Position enthalten.
Zusätzlich zu den Ergebnissen, die in diesem Bericht vorgestellt wurden, sind noch Ergebnisse des Survival-Modells mit anderen Parametereinstellungen im Ordner *r_results* bereitgestellt. Diese Parametereinstellungen haben allerdings stets zu schlechteren Resultaten geführt als die Ergebnisse in dem Ordner *gbm_offset*. Der Ordner *gbm_no_offset* enthält Ergebnisse des Modells ohne Offset und *gbm_no_offset_interactions* Ergebnisse ohne Offset und mit Interaktionen, das heißt einer *interaction.depth* von 2. Für den Ordner *gbm_offset_bag* wurde lediglich die *bag.fraction* auf 1 gesetzt und die Ergebnisse sind nahezu identisch zu *gbm_offset*. Ergebnisse mit Offset und Interaktionen sind im Ordner *gbm_offset_interactions* zu finden.
Für die Ergebnisse des Netzwerkes wurde ein seperater Ordner *network* angelegt. Der Un-

terordner *gephi* enthält die Installationsdateien für das Open Source-Programm *Gephi*, dass für das Arbeiten mit dem Netzwerk benötigt wird. Die *dmg*-Datei ist für Computer mit dem Betriebssystem *Mac OSX*, die *gz*-Datei ist für Linux und die *exe*-Datei für Windows. Alternativ können die Installationsdateien auch auf *https://gephi.github.io/* heruntergeladen werden.
Nach der Installation von *Gephi* kann das Netzwerk in das Programm gelesen werden. Der Ordner *network_ data* enthält dafür sechs Dateien, die in Kapitel 9.2 und Tabelle 9 näher beschrieben werden. Außerdem wird im nächsten Kapitel beschrieben, wie man mit diesen Dateien in *Gephi* arbeiten kann.
Neben den bereits beschriebenen Ordnern sind im Hauptverzeichnis noch die Folien des Vortrages vom 12.08.2014 im Institut für Statistik, dieser Bericht sowie dieser elektronische Anhang als *README*-Datei gespeichert.

9.2 Tutorial zu Gephi

Dieses Kapitel enthält ein kurzes Tutorial zum Arbeiten mit dem Open Source-Programm *Gephi*. Wie bereits erwähnt, sind die Installationsdateien im Unterordner *gephi* enthalten.

relative_ausgaenge.gexf	Unbearbeitetes Netzwerk mit den ersten 100 Positionen; Gewichtung der Kanten: relative Ausgänge
relative_ausgaenge_pos_10.gexf	Unbearbeitetes Netzwerk mit den ersten 10 Positionen; Gewichtung der Kanten: relative Ausgänge
relative_ausgaenge_pos_10_Layout.gephi	Bearbeitetes Netzwerk mit den ersten 10 Positionen und Layout; Gewichtung der Kanten: relative Ausgänge
relative_eingaenge.gexf	Unbearbeitetes Netzwerk mit den ersten 100 Positionen; Gewichtung der Kanten: relative Eingänge
relative_eingaenge_pos_10.gexf	Unbearbeitetes Netzwerk mit den ersten 10 Positionen; Gewichtung der Kanten: relative Eingänge
relative_eingaenge_pos_10_Layout.gephi	Bearbeitetes Netzwerk mit den ersten 10 Positionen und Layout; Gewichtung der Kanten: relative Eingänge

Tabelle 9: Übersicht der Gephi Dateien

Das Program liest *gexf*-Dateien ein und erstellt daraus ein Netzwerk aus Knoten und Kanten. Die Dateien für dieses Tutorial wurden in *R* mit den Skripten *gexf.R* und

gexf_support.R erstellt, die in dem Ornder *r_scripts* zu finden sind. Mit den Skripten können verschiedene Parameter und Attribute übergeben werden, wie zum Beispiel die farbliche Kodierung der Knoten und die Gewichtung der Kanten. Ein Netzwerk, dass in *Gephi* bearbeitet wurde, kann als *gephi*-Datei abgespeichert werden. Dieses Format kann ausschlieslich in Gephi eingelesen werden. In diesen Dateien sind typischerweise Layouts, das heißt die Anordnung der Knoten, und Filter für die Kanten gespeichert. Tabelle 9 gibt eine Übersicht über die *gexf*- und *gephi*-Dateien, die im elektronischen Anhang bereits vorhanden sind und mit Hilfe derer das Tutorial beschrieben wird. Für die zwei *gephi*-Dateien wurde der *ForceAtlas2*-Algorithmus angewendet und die Anordnung der Knoten daraufhin noch manuell bearbeitet.
Nach der Installation und dem Öffnen des Programms, können über *File* → *Open*, die Dateien geöffnet werden. Abbildung 29 enthält einen Screenshot des Workspaces nach dem Laden von *relative_ausgaenge.gexf*. Im lilafarbenen Kasten links oben sind die einzelnen Bearbeitungsmodi von *Gephi* auswählbar. *Overview* ist dafür gedacht den Graphen zu bearbeiten und mit einem Layout zu versehen. Im *Data Laboratory* können einzelne Knoten und Kanten bearbeitet und neue Attribute hinzufügt werden. Die *Preview* liefert eine Ansicht der Datei, wie sie im Falle eines Exports aussieht.

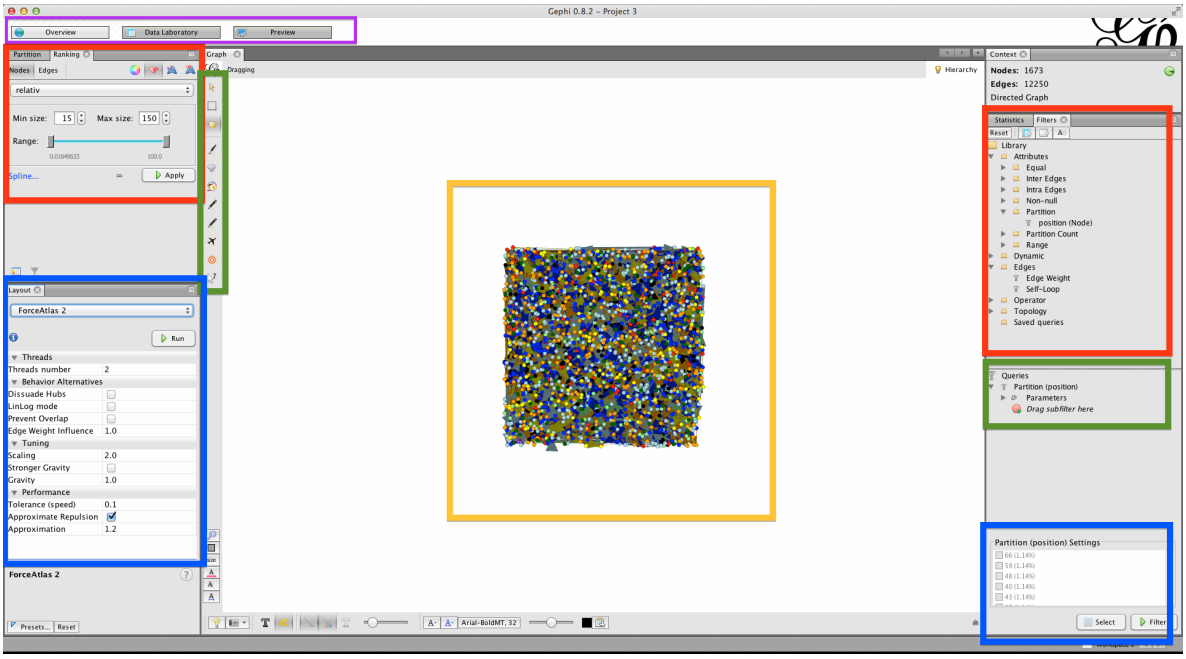

Abbildung 29: Workspace in Gephi

Gelb markiert, in der Mitte der Abbildung, ist das Netzwerk zu sehen. Diese viereckige Struktur besitzt der Graph nach dem Einlesen einer *gexf*-Datei. Anhand der Optionen in dem blauen Kasten auf der linken Seite, kann das Netzwerk bearbeitet werden. Hier kann der gewünschte Algorithmus ausgewählt werden, der die räumliche Anordnung der Knoten berechnet. Zusätzlich können die Parameter des Algorithmus verändert werden. Der Parameter *Edge Weight Influence* steuert den Einfluss der Gewichtung der Kanten

auf die Anordnung der Knoten und mit *Scaling* kann die Größe des Graphen festgelegt werden. *Approximate Repulsion* regelt die Konvergenz des Graphen und sollte abgewählt werden, um einen übersichtlicheres Netzwerk zu bekommen. Das Aktivieren von *Prevent Overlap* bewirkt, dass die Knoten nicht überlappen. Durch Klicken auf die Schaltfläche *Run* kann der Algorithmus gestartet und gestoppt werden. Durch das Ausprobieren von verschiedenen Algorithmen mit verschiedenen Einstellungen können schnell gute Ergebnisse erzielt werden.
Um das Layout weiter zu verändern, kann in dem linken roten Kasten die Größe der Kanten und Knoten eingestellt werden. Dafür muss der rote Diamant angeklickt und ein Gewichtungsattribut ausgewählt werden, welches in diesem Fall die Variable *relativ*, das heißt die relative Häufigkeit der Knoten in einer Position ist. Im grünen Feld daneben befinden sich die Optionen für die Maus, beispielsweise die Einstellung, dass man Knoten manuell im Graphen bewegen und positionieren kann. Aber auch die letzte Option ist hilfreich, mit der man durch einen Klick auf einen Knoten, dessen Informationen sehen kann.

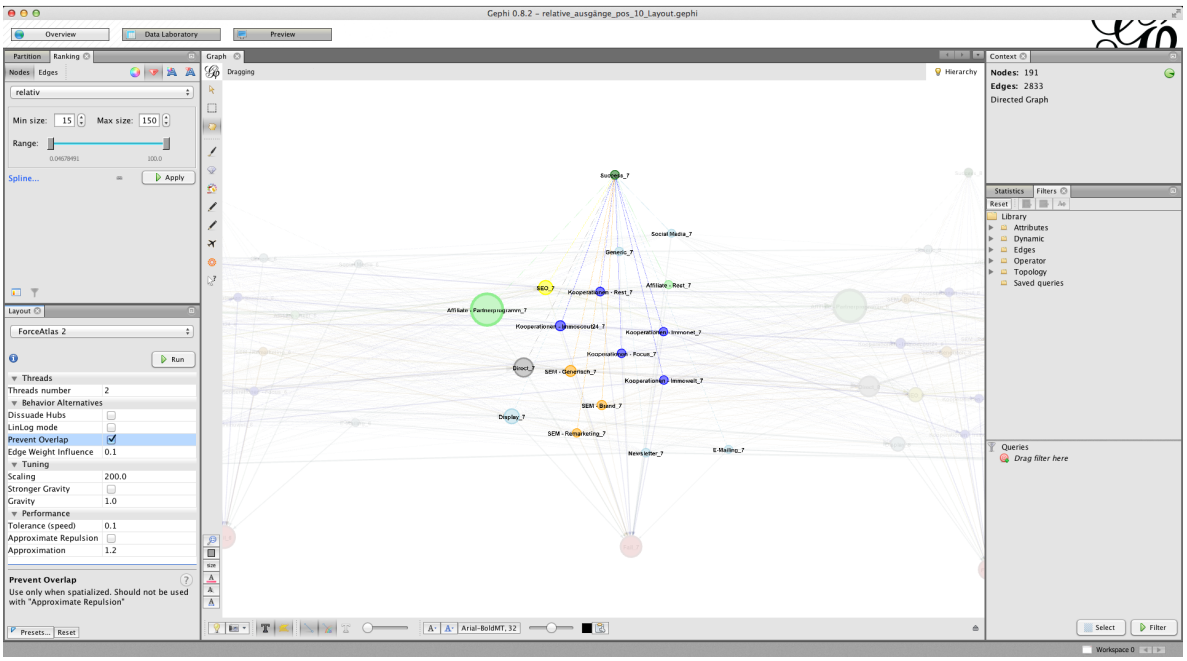

Abbildung 30: Nähere Betrachtung einer Position in Gephi

Im roten Feld auf der rechten Seite kann man *Statistics* und *Filters* anwenden. Unter *Filters* kann bezüglich der übergebenen Attribute gefiltert werden. Bei *Partition* befindet sich die Position als Attribut, mit der eingestellt werden kann welche Positionen im Graphen angezeigt werden. Dieser Filter muss in den grünen Kasten gezogen werden, um die Auswahlmöglichkeiten im blauen Kasten zu bekommen. Hier können dann die gewünschten Position ausgewählt werden und mittels dem Button *Filter* aktiviert werden. Desweiteren kann auch einen *subfilter*, also ein zweiter Filter angewendet werden, beispielsweise das *Edge Weight*, das heißt die Gewichtung der Kanten. Generell empfiehlt

es sich, den Graphen zunächst zu filtern und dann den Layout Algorithmus anzuwenden, da sich dieser je nach Filterung anpasst. Weitere hilfreiche Features des Programms befinden sich in der Fußzeile des Workspaces, beispielsweise das Ein- oder Ausblenden der Knotenlabels mithilfe des großen T oder die Screenshot Funktion.
Um das Netzwerk näher zu analysieren, kann durch Scrollen in den Graphen gezoomt und durch Halten der rechten Maustaste und gleichzeitiges Bewegen der Maus navigiert werden. Durch das Bewegen das Mauszeigers auf einen Knoten, werden außerdem alle Knoten, die mit eben diesem verbunden sind, hervorgehoben. In Abbildung 30 ist dies an Position 7 dargestellt. Hier wurde der Knoten *Success_ 7* angewählt und es ist zu erkennen, dass dieser Knoten mit allen Kampagnen der siebten Position verbunden ist. Durch die Anwendung eines Filters für die Kanten, fallen einige dieser Verbindungen weg und es können interessante Ergebnisse erzielt werden.
Um ein bearbeitetes Netzwerk als Bilddatei abzuspeichern, muss es exportiert werden. Dafür muss man links oben den Reiter *Preview* auswählen. Nützliche Einstellungen beim Export sind im lilafarbenen Feld in Abbildung 31 zu finden. *Rescale Weight* sorgt dafür, dass die Kanten nicht zu dick sind und das Abwählen von *Curved* macht das Netzwerk übersichtlicher. Um Änderungen sichtbar zu machen, muss der Button *Refresh* in der grünen Box angeklickt werden. Außerdem kann der Graph hier dann als *SVG*-, *PDF*- oder *PNG*-Datei abgespeichert werden.

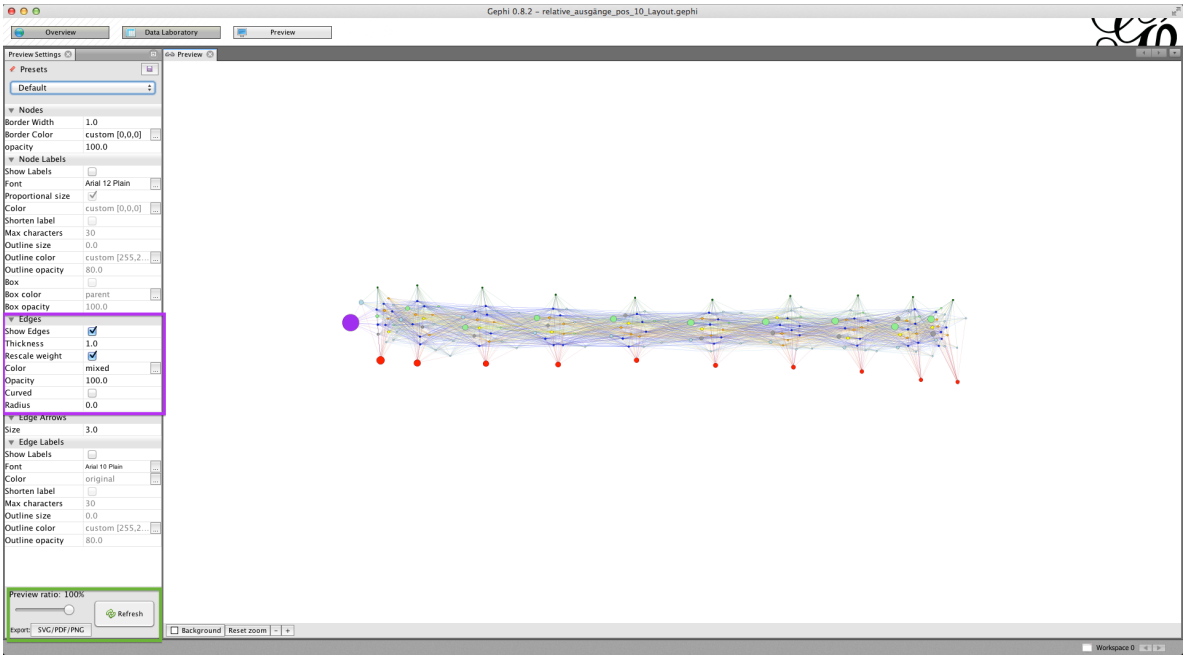

Abbildung 31: Export in Gephi

Literatur

[1] R. Agrawal & R. Srikant. „Mining sequential patterns“. In: *Proceedings of the 11th Conference on Data Engineering (ICDE'95)* (1995), S. 3–14.

[2] Revolution Analytics & Steve Weston. *doSNOW: Foreach parallel adaptor for the snow package*. R package version 1.0.12. 2014. URL: `http://CRAN.R-project.org/package=doSNOW`.

[3] Revolution Analytics & Steve Weston. *foreach: Foreach looping construct for R*. R package version 1.4.2. 2014. URL: `http://CRAN.R-project.org/package=foreach`.

[4] M. Bastian, S. Heymann & M. Jacomy. „Gephi: An Open Source Software for Exploring and Manipulating Networks“. In: *Proceedings of the Third International ICWSM Conference* (2009), S. 361–362.

[5] Christian Buchta, Michael Hahsler & with contributions from Daniel Diaz. *arulesSequences: Mining frequent sequences*. R package version 0.2-6. 2014. URL: `http://CRAN.R-project.org/package=arulesSequences`.

[6] M Dowle et al. *data.table: Extension of data.frame*. R package version 1.9.2. 2014. URL: `http://CRAN.R-project.org/package=data.table`.

[7] M. El-Sayed, C. Ruiz & E. A. Rundensteiner. „FS-Miner: Efficient and incremental mining of frequent sequence patterns in web logs“. In: *Proceedings of the 6th Annual ACM International Workshop on Web Information and Data Management* (2004), S. 128–135.

[8] C. I. Ezeife, Y. Lu & Y. Liu. „PLWAP sequential mining: Open source code“. In: *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementation* (2005), S. 26–35.

[9] F. M. Facca & P. L. Lanzi. „Recent developments in web usage mining research“. In: *Proceedings of the 5th International Conference on Data Warehousing and Knowledge Discovery* (2003).

[10] J. H. Friedman, T. Hastie & R. Tibshirani. „Additive Logistic Regression: a Statistical View of Boosting“. In: *Annals of Statistics* 28.2 (2000), S. 337–374.

[11] J. H. Friedman. „Greedy Function Approximation: A Gradient Boosting Machine“. In: *Annals of Statistics* 29.5 (2001), S. 1189–1232.

[12] J. H. Friedman & B. E. Popescu. „Predictive Learning via Rule Ensembles“. In: *The Annals of Applied Statistics* 2.3 (2008), S. 916–954.

[13] B. Goethals. „Frequent set mining“. In: *The Data Mining and Knowledge Discovery Handbook* (2005), S. 377–397.

[14] Yifan Hu. *Efficient and High Quality Force-Directed Graph Drawing*. URL: `http://yifanhu.net/PUB/graph_draw_small.pdf`.

[15] R. Iváncsy & I. Vajk. „Frequent pattern mining in web log data“. In: *Acta Polytech. Hungarica* 3.1 (2006), S. 77–90.

[16] M. Jacomy et al. *ForceAtlas2, A Graph Layout Algorithm for Handy Network Visualization*. 2011. URL: http://webatlas.fr/tempshare/ForceAtlas2_Paper.pdf.

[17] Eric D. Kolaczyk. *Statistical Analysis of Network Data. Methods and Models*. New York: Springer, 2009.

[18] Y. Lu & C. I. Ezeife. „Position coded pre-order linked WAP-tree for web log sequential pattern minings". In: *Proceedings of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (2003), S. 337–349.

[19] Nizar R. Mabroukeh & C. I. Ezeife. „A Taxonomy of Sequential Pattern Mining Algorithms". In: *ACM Computing Surveys* 43.1 (2010), 3:1–3:41.

[20] F. Masseglia, M. Teisseire & P. Poncelet. „Sequential pattern mining: A survey on issues and approaches". In: *Encyclopedia of Data Warehousing and Mining* (2005), S. 1–14.

[21] Greg Ridgeway with contributions from others. *gbm: Generalized Boosted Regression Models*. R package version 2.1. 2013. URL: http://CRAN.R-project.org/package=gbm.

[22] J. Pei et al. „PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth". In: *Proceedings of the International Conference on Data Engineering* (2001), S. 215–224.

[23] R Development Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2012. URL: http://www.R-project.org/.

[24] S. Song, H. Hu & S. Jin. „HVSM: A new sequential pattern mining algorithm using bitmap representation". In: *Advanced Data Mining and Applications* 3584 (2005), S. 455–463.

[25] J. Srivastava et al. „Web usage mining: Discovery and applications of usage patterns from Web data". In: *ACM SIGKDD* 2.1 (2000), S. 12–23.

[26] J. Wang & J. Han. „BIDE: Efficient mining of frequent closed sequences". In: *Proceedings of the 20th International Conference on Data Engineering* (2004), S. 79–90.

[27] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN: 978-0-387-98140-6. URL: http://had.co.nz/ggplot2/book.

[28] Hadley Wickham. „The Split-Apply-Combine Strategy for Data Analysis". In: *Journal of Statistical Software* 40.1 (2011), S. 1–29. URL: http://www.jstatsoft.org/v40/i01/.

[29] George Vega Yon, Jorge Fábrega Lacoa & Joshua B. Kunst. *rgexf: An R package to build GEXF graph files*. R package version 0.14.3.11. 2014. URL: http://CRAN.R-project.org/package=rgexf.

[30] M. J. Zaki. „SPADE: An efficient algorithm for mining frequent sequences“. In: *Mach. Learn.* 42 (2001), S. 31–60.

Tabellenverzeichnis

Abbildungsverzeichnis