

Attribution 2.0

Von einfachen Heuristiken zu optimalen datengetriebenen Modellen

Dr. Steffen Wagner



Noch vor wenigen Jahren stellte der Wechsel von ‘‘Last-Contact’’ zur ‘‘Badewanne’’ eine kleine Revolution im Online-Marketing dar. Doch beide Verfahren sind reine Heuristiken und damit willkürliche Regelwerke, die zwar auf Daten angewendet werden, aber nicht von diesen lernen. Die nächste größere Evolutionsstufe stellen die datengetriebenen Attributionsmodelle dar, die heute State of the Art sind. Mit zunehmender Verbreitung der datengetriebenen Ansätze, wächst die Zahl der angebotenen Modelle. Schnell sehen sich Marketer auf der Suche nach einem zu ihnen passenden Attributionsmodell mit einer überfordernden Anzahl statistischer Modelle konfrontiert. Es fällt zunehmend schwerer, diesen Dschungel zu durchblicken, die Modelle hinsichtlich ihrer Stärken und Schwächen zu bewerten und sich für das passende Modell zu entscheiden.

Die Güte der Attribution entscheidet maßgeblich über die Effizienz des Online-Marketings und ist somit zentral für die Wettbewerbsfähigkeit der Unternehmen. Um das zur Verfügung stehende Budget sinnvoll einzusetzen, ist es notwendig den Status Quo der Attribution aufzubrechen und den Weg zu bereiten für die Attribution 2.0. Dies bedeutet die Abkehr von klassischen Heuristiken (Last-Contact oder der Badewanne) aber auch die Weiterentwicklung und Individualisierung von datengetriebenen Verfahren. Zu Letzteren gehören spieltheoretische Ansätze, Machine Learning Algorithmen und Regressionsansätze. Die Frage nach dem passenden Modell lässt sich dabei nicht pauschal beantworten. Die Verfahren variieren in Bezug auf Komplexität und Wartungsaufwand erheblich. Welches Verfahren das richtige ist, hängt vom Budget, den personellen Ressourcen, den technischen Gegebenheiten und den Analyse-Anforderungen ab. Für einen kleinen Shop mit geringem Budget nimmt

die Umstellung von der Badewanne auf ein einfaches datengetriebenes Modell einen weit größeren Anteil der finanziellen Ressourcen in Anspruch als bei einem großen Shop. Für einen großen Shop mit einer breit angelegten Online-Marketing-Strategie wiederum, sind die Anforderungen jedoch auch ganz andere. Hier lohnt sich ein komplexeres Modell unter Einbeziehung Shop-spezifischer Metriken zur weiteren Erhöhung der Modellgüte. Das ermöglicht z.B. auch verlässliche Kohortenanalysen mit kleineren Fallzahlen und die schnelle Bewertung neuer Marketingpartner.

Das vorliegende Paper bringt Licht ins Dickicht der datengetriebenen Attributionsverfahren. Es geht auf Grundlagen der Attribution ein und stellt die gängigsten Ansätze samt ihrer Stärken und Schwächen vor. Dabei bietet es eine Orientierungshilfe im komplexen Markt der Attributionsmodelle und zeigt auf, wohin die Entwicklung in der Zukunft gehen wird.

Zielstellung der Attribution

Hauptsächlich Online-Shops (aber nicht nur diese) haben hohe monatliche Kosten für Online-Werbung und die Optimierung der Sichtbarkeit ihrer Landingpage. Das Geld fließt in verschiedene Marketingkanäle, wie Affiliate, Newsletter, Retargeting, SEM, SEO, etc. Innerhalb dieser Kanäle finden sich häufig Sub-Kanäle, z.B. in Form verschiedener Anbieter oder einer Unterteilung zwischen Brand- und Non-Brand Keywords bei SEO oder SEM. Darunter kommen weitere Ebenen bis hinab zu Kampagnen und Keywords/Creatives. Dies ergibt ein großes Potpurri an Möglichkeiten Geld auszugeben. Hierbei lassen sich die horizontale Allokation und die vertikale Allokation unterscheiden (siehe Abbildung 1).



Abbildung 1: Horizontale versus vertikale Budget-Allokation

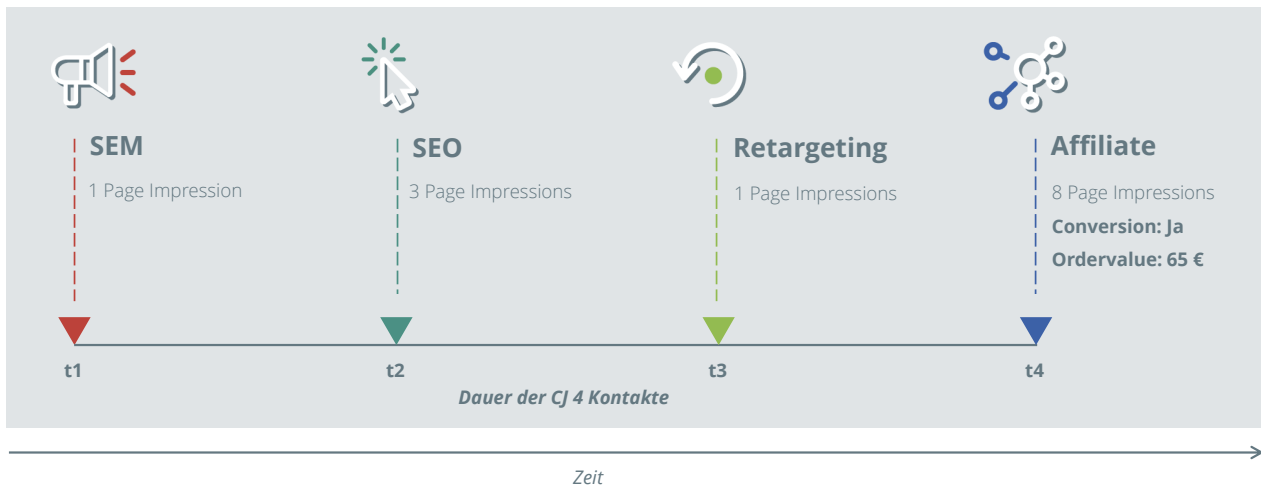


Abbildung 2: Ein Beispiel für eine Customer Journey

Die Frage ist, wo ein zusätzlicher Euro eingesetzt werden sollte, damit er den größten Nutzen (den höchsten Deckungsbeitrag, die meisten Conversions, Klicks oder Views) bringt. Genau diese Frage beantwortet die Attribution im Zusammenspiel mit der aktuellen Budgetverteilung.

Das grundlegende Konzept bei der Attribution basiert auf der Analyse der Customer Journey (im Folgenden kurz "CJ"). Eine Customer Journey ist die Reise des Users, der über verschiedene Kanäle oft mehrfach auf die Landingpage gelangt (z.B. über eine Suche bei Google oder über den Klick auf ein Display-Banner) und am Ende eine definierte Aktion (sog. "Conversion", in der Praxis meist der Kauf) durchführt oder eben auch nicht. Steht am Ende der CJ die Conversion, so wird von einer erfolgreichen CJ gesprochen.

Ein Beispiel für eine CJ ist in der vorhergehenden Grafik abgebildet (Abbildung 2). Hier sieht ein User zunächst eine (zu seiner Suchanfrage passende) Werbung und besucht erstmalig die Landingpage durch einen Klick auf das Banner. Nach einiger Zeit sucht er das Produkt z.B. über eine Suchmaschine und gelangt erneut auf die Landingpage. Einige Tage später gelangt er über eine Retargeting-Kampagne ein drittes Mal auf die Landingpage, um dann letztendlich über einen Affiliate-Partner ein viertes Mal auf die Landingpage zu kommen und das Produkt zu kaufen. Es stellt sich bei einer solchen CJ nun die Frage, welchen Anteil die einzelnen Kanäle an der Conversion und dem damit erzielten Umsatz (genauer: Deckungsbeitrag, im Folgenden kurz "DB") haben.

Datengetriebene Attributionsmodelle beantworten diese Frage über den modellgestützten Vergleich von erfolgreichen und erfolglosen CJs. Daraus lässt sich ableiten, welche Faktoren für den Erfolg ausschlaggebend

sind. Den an der erfolgreichen CJ beteiligten Kanälen wird der generierte Deckungsbeitrag dann anteilig - gemäß ihrem Beitrag zum Erfolg - zugeschrieben ("attribuiert"). Am Ende jeder Planungsperiode, typischerweise monatlich, findet im Rahmen der Budgetoptimierung ein Abgleich zwischen attribuierten Deckungsbeiträgen und dem Budget der Kanäle/Sub-Kanäle/... statt. Ziel ist dabei, das Budget optimal auf die Kanäle aufzuteilen und so mit gegebenem Budget möglichst viele Conversions und hohe Deckungsbeiträge zu erzielen. Dies kann ggf. unter der Berücksichtigung von Nebenbedingungen, wie z.B. einer gewissen Präsenz durch Branding-Kampagnen, geschehen.

Daten

Datengetriebene Verfahren lernen aus den Daten. Die Ergebnisse der Attribution können daher immer nur so gut wie die Datenbasis sein. Der Qualität der Daten sollte daher besondere Aufmerksamkeit geschenkt werden, da sich Probleme mit der Datenqualität unmittelbar auf die Güte und Verlässlichkeit der Ergebnisse auswirken. Grundlage für die Attribution sind üblicherweise Tracking-Daten (On-Site Tracking), Daten externer Partner (z.B. Display-Werbung), Warenkorb Daten und Daten aus dem internen Datawarehouse (z.B. Kunden-, Bestell- und Produktinformationen). Es können und sollten also auch CRM-Daten (wie Neukunde vs. Bestandskunde, Retouren, Customer-Lifetime-Value, usw.) explizit berücksichtigt werden:

Auch wenn viele Tracking-Dienstleister nicht gerne darüber sprechen - in der Praxis sollte man sich im Rahmen der Sicherung der Datenqualität zumindest mit folgenden Punkten auseinandersetzen:



Abbildung 3: Viele Datenquellen können zu einer Attribution beitragen

CrossDevice

Wechseln User während ihrer CJ mehrfach das Device (z.B. Computer auf der Arbeit und zu Hause, Smartphone, Tablet etc.) fragmentiert die CJ. Die einzelnen Fragmente können i.d.R. nur dann zusammengeführt werden, wenn sich der User von jedem Gerät mit einem Account einloggt. Dies ist allerdings zumeist nur beim Kauf der Fall.

Löschen von Cookies

Um einzelne Besuche eines Users (vom selben Device) auch einer CJ zuordnen zu können, werden im Tracking i.d.R. Cookies eingesetzt. Löscht der User die Cookies oder hat er seinen Browser so konfiguriert, dass dieser Cookies automatisch verwirft, zerfällt die betreffende CJ im Tracking in Teilstücke. Zwar gibt es weitere Methoden zur Wiedererkennung von Usern (genauer: der verwendeten Devices), diese sind jedoch oft datenschutzrechtlich bedenklich oder weisen gravierende Schwächen bei stark standardisierten Geräten (z.B. iPhones) auf.

Sichtbarkeit bei Bannern

Beim Einsatz von Display-Bannern kann von einer Werbewirkung überhaupt nur dann ausgegangen werden, wenn das Banner im auf dem Display sichtbaren Bereich einer Website angezeigt wurde. Befindet sich ein Banner bspw. im unteren Teil einer Website und der User scrollt nicht bis zum Ende, ist eine Beeinflussung des Userverhaltens durch das Banner ausgeschlossen. Obwohl es technisch möglich wäre, dies zu erfassen, liefern viele Werbenetzwerke eine solche Information nicht mit, da sie dem Eigeninteresse an der Berechnung der Views zuwider läuft.

Fehlende Daten zu AdImpressions

Einige bedeutende Werbepartner (z.B. Google und Facebook) erlauben die Erfassung der Einblendung von Bannern nicht oder nur ausgewählten zertifizierten Partnern. Über die Dashboards von Google AdWords oder Facebook stehen dann lediglich aggregierte Daten über die Anzahl der Einblendungen und Click-Through-Raten, z.B. auf Basis von Placements, zur Verfügung. Es kann aber nicht bestimmt werden, ob einzelne Nutzer ein Banner gesehen haben und wie sie auf dieses Banner reagiert haben.

Unvollständige On-Site-Metriken

Für die innerhalb einer Session zuletzt aufgerufene Seite kann die Verweildauer nicht zuverlässig bestimmt werden. Die Verwendung von Javascript zum regelmäßigen Senden von Events an den Server, ist als Lösung unzureichend, da der Mechanismus einfach vom User umgangen werden kann. Bei Visits mit vielen Seitenaufrufen fällt diese Unzulänglichkeit der Daten kaum ins Gewicht. Sie kann für einen Blog allerdings zum Problem werden, wenn User direkt auf einen Beitrag kommen und keine weiteren Interaktionen stattfinden. Es kann dann nicht unterschieden werden, ob die Seite sofort verlassen wurde, weil der Inhalt nicht den Erwartungen entsprach (sog. "Bounce") oder der Artikel in Ruhe gelesen wurde und der User erst anschließend die Seite verlassen hat.

CRM-Daten

Viele CRM-Daten lassen sich nur verknüpfen, wenn sich ein User eingeloggt hat. Dies geschieht regelmäßig bei erfolgreichen CJs - nämlich mit dem Kauf. Bei erfolglo-

sen CJs hingegen loggt sich nur ein kleiner Teil der User auf der Landingpage ein, so dass hier z.B. nicht zwischen (nicht eingeloggt) Bestandskunden und echten Neukunden unterschieden werden kann.

Pflege von Metadaten

Ein weiterer Fallstrick lauert bei der Wartung der Kanalstrukturen. Wird bei der datengetriebenen Attribution explizit auf die in der CJ auftretenden Kanäle abgestellt (z.B. bei spieltheoretischen Ansätzen und Lebenszeit-Modellen), so ist eine über die Zeit einheitliche Bezeichnung der Kanäle extrem wichtig. Ändern sich die Bezeichnungen über die Zeit (z.B. Umbenennung von 'SEM' in 'SEA'), so muss dies im der Attribution vorgelagerten Datenmanagement berücksichtigt werden. Dies erfordert die lückenlose (in der Realität leider nicht immer gegebene) Kommunikation der Änderung an alle beteiligten Instanzen.

Änderungen an der Website

Gleiches gilt für grundlegende Änderungen an der Landingpage oder dem Tagging von Seitenkategorien (z.B. Produktkataloge, Produktdetailseiten, Hilfeseiten). Während grundlegende Änderungen meist kommuniziert werden, wird dies bei kleineren - aber ggf. relevanten - Änderungen häufig vergessen. Speziell bei größeren Shops mit verteilten Zuständigkeiten müssen klare Regeln definiert werden, wer bei einer Änderung an der Website zu benachrichtigen ist.

Die erwähnten Punkte sind keinesfalls als abschließende Liste zu verstehen. Sie stellen lediglich Beispiele für regelmäßig auftretende Schwierigkeiten dar. Insgesamt ist die Qualität von On-Site und Off-Site Daten häufig deutlich geringer als bei CRM-Daten. Da sich viele der oben genannten Probleme (derzeit) nicht grundsätzlich vermeiden lassen, sollten Attributionsverfahren mit fehlerhaften Daten umgehen können. Die Fähigkeit eines statistischen Modells - trotz problematischer Daten - noch näherungsweise optimale Ergebnisse zu liefern, wird Robustheit genannt. Im Bezug auf die Realität der Attribution ist diese Eigenschaft von zentraler Bedeutung!

Auswahl des Analysezeitraums

Datengetriebene Modelle können immer nur so gut sein wie die Daten, auf denen die Modelle geeicht wurden. Neben der Beachtung der im vorherigen Abschnitt beschriebenen Probleme, kommt daher der Wahl des Zeitraums eine größere Bedeutung zu. Abhängig von der Branche kann die Dauer der CJs deutlich variieren. Während in einigen Branchen 90% der erfolgreichen CJs

kürzer als ein bis zwei Wochen sind, können es in anderen Branchen mit längeren Entscheidungsprozessen durchaus drei bis sechs Wochen sein. Ist der für die Analyse gewählte Zeitraum zu kurz, werden viele CJs an den Rändern des Zeitfensters abgeschnitten, weil sie nur zu Teilen im Beobachtungszeitraum liegen. So erscheinen CJs am Beginn des Zeitfensters zu kurz, weil vorhergehende Kontakte nicht exportiert wurden. Am Ende des Betrachtungszeitraums besteht hingegen das Problem, dass CJs noch nicht abgeschlossen sind. Hierbei ist insbesondere problematisch, wenn CJs als erfolglos gewertet werden, die in der näheren Zukunft noch bei nachfolgenden Kontakten mit einer Conversion abgeschlossen werden. Um den Anteil dieser Problemfälle zu reduzieren, sollte mindestens ein Zeitfenster von 12 Wochen (ggf. länger) als Basis für die Analyse dienen. Weiterhin ist darauf zu achten, dass in diesem Zeitraum keine grundlegenden Änderungen an der Struktur der Marketingkanäle, dem Tracking und der Landingpage vorgenommen wurden, damit die Übertragbarkeit der Ergebnisse gewährleistet ist.

Datengetriebene Verfahren

Die klassischen Heuristiken (Last-Contact oder die Badewanne) haben in der Attribution ausgedient und werden durch datengetriebene Modelle ersetzt. Prinzipiell wird die Ausprägung einer binären Variable (Conversion: ja/nein) modelliert, d.h. systematische Unterschiede zwischen erfolgreichen und erfolglosen CJs werden identifiziert und analysiert. Das macht die Attributionsmodellierung zu einem Discrete-Choice-Problem. Nach Identifikation dieser Unterschiede lässt sich analysieren, welche Kontakteigenschaften zu einer positiven Kaufentscheidung beitragen. Auf Basis der ermittelten Kontakteigenschaften lässt sich der Beitrag jedes einzelnen Kontakts einer erfolgreichen CJ zur Kaufentscheidung bestimmen. Dies ermöglicht die Attribution des Umsatzes bzw. des Deckungsbeitrags.

In der Praxis haben sich dazu folgende Verfahren etabliert:

- Spieltheoretische Modelle (Shapley-Value)
- Machine Learning Algorithmen
- Statistische Regressionsansätze
 - Logistische Regression
 - Bayesianische Modelle

Spieltheoretische Modelle (Shapley-Value)

Die in der Attribution verwendeten spieltheoretischen Modelle basieren auf dem Ansatz von Lloyd Shapley.

Abstrakt gesprochen betrachtet Shapley eine Koalition von Spielern, die gemeinsam auf ein Ergebnis hinwirken. Der Ansatz ermöglicht es nun den Beitrag des einzelnen Spielers zu diesem Ergebnis zu bewerten. Der Ansatz greift dazu auf den marginalen Beitrag jedes Spielers zum Gesamtergebnis zurück. Wie hätte das Ergebnis ohne den betrachteten Spieler ausgesehen? Die Differenz zum erzielten Ergebnis wird dann als marginaler Beitrag dieses Spielers aufgefasst.

Übertragen auf die Attribution treten die an einer CJ beteiligten Kanäle an die Stelle der Spieler. Das untersuchte Ergebnis ist der erfolgreiche oder erfolglose Ausgang der CJ. Das Verfahren berücksichtigt ausschließlich die an einer CJ beteiligten Kanäle und deren Reihenfolge. In der Praxis wird die Reihenfolge der Kanäle üblicherweise lediglich für relativ kurze CJs mit max. vier bis fünf Kontakten berücksichtigt, da bei längeren Ketten die Anzahl der Abfolgekombinationen sehr schnell sehr groß und damit numerisch zu aufwändig wird. Weitere, über die Kanalinformation hinausgehende Informationen können nicht berücksichtigt werden. Zwar ist der Ansatz in Bezug auf die Bewertung der Kanäle flexibler als gängige Regressionsansätze, dafür aber inkompatibel zu allen Erkenntnissen über den hohen Stellenwert der User spezifischen Engagement Metriken (s.u.). Fürsprecher des spieltheoretischen Ansatzes führen als Argument gerne die hohe Robustheit des Shapley-Konzepts ins Feld und begründen dies damit, dass beim Shapley-Ansatz immer nur bezüglich der Kanalabfolge "ähnliche" CJs miteinander verglichen werden.

Machine Learning

Machine Learning Algorithmen haben ihre Wurzeln überwiegend in der Informatik, wo sie für die Erkennung von Kategorien und Mustern entwickelt wurden. Entsprechend ihrer Entwicklungsgeschichte handelt es sich bei den Algorithmen i.d.R. um eine Black-Box, die keine Insights über die Wirkungszusammenhänge zwischen verwendeten Metriken und identifizierter Kategorie erlaubt. Die Optimierung auf den zur Eichung des Modells verwendeten Datensatz birgt die Gefahr des "Overfittings". Das bedeutet, dass die im Trainingsdatensatz geltenden Gesetzmäßigkeiten perfekt abgebildet werden. Wird der ermittelte Regelsatz dann auf neue Datensätze und zukünftige CJs angewendet, besteht die Gefahr, dass hier veränderte Gesetzmäßigkeiten gelten und daher das Modell eine schlechte Anpassung besitzt. Ob dies gegeben ist, lässt sich anhand der "Out-of-sample-Güte" messen. Die "Sinnhaftigkeit" der für die Klassifizierung genutzten Regeln kann hingegen nicht geprüft werden. In der Online-Attribution konnten sich Machine Learning-Ansätze nur begrenzt durchsetzen.

Statistische Regressionsansätze

Die statistischen Regressionsansätze im Kontext der Attribution gehen weit über die multiple lineare Regression hinaus. Vielmehr handelt es sich um die Klasse der generalisierten linearen Modelle, die ein sehr flexibles und bewährtes Spektrum an Modellen umfasst, welche die Modellierung der Beziehungen zwischen einer abhängigen und (nahezu) beliebig vielen erklärenden Variablen ermöglichen. Es stehen weiterhin Techniken zur Variablenselektion, Modellierung nichtlinearer Zusammenhänge, Gewichtung und Ausbalancierung der Datenbasis sowie zahlreiche Gütemaße zur Verfügung. Im Rahmen der Modellbildung können Hypothesen über vermutete Einflussfaktoren getestet und Wirkungszusammenhänge identifiziert werden. Regressionsansätze öffnen also die Black-Box und ermöglichen ein Verständnis der Gesetzmäßigkeiten, die eine Conversion begünstigen. Bezüglich der Page-Impressions (PI) könnte ein gefundener Zusammenhang z.B. so aussehen: "bis zu einem Sättigungsbereich von 20 PIs erhöht jede zusätzliche PI bei einem Visit die Chance auf eine Conversion um 5%". Darüber hinaus lassen sich die gefundenen Zusammenhänge bezüglich ihrer Plausibilität überprüfen und anschaulich interpretieren.

Logistische Regression

Bei der logistischen Regression wird das Eintreten eines Ereignisses (binäre abhängige Variable) durch die erklärenden Variablen modelliert. In Bezug auf die Attribution bedeutet dies, dass der erfolgreiche oder erfolglose Abschluss einer CJ durch Metriken wie das On-Site-Verhalten des Users (Time on Site, Page Impressions), die Anzahl der erfolgten Kontakte, die in der CJ beobachteten Kanäle und weitere Variablen erklärt wird. Ein Nachteil der Attributionsmodellierung auf Basis eines logistischen Modells besteht darin, dass sich die Conversion nur auf Ebene der CJ beobachten lässt. Für die Modellierung bedeutet dies, dass Metriken, die auf Ebene einzelner Kontakte vorliegen (z.B. die Anzahl der Seitenaufrufe pro Session) über alle Kontakte der CJ aggregiert werden müssen. Hierbei wird z.B. die Summe, das arithmetische Mittel oder das Maximum über alle beobachteten Werte als Maßzahl verwendet.

Bayesianische Modelle

Bayesianische Modelle sind im eigentlichen Sinne keine eigene Modellklasse, sondern vielmehr eine Erweiterung bestehender (Regressions-)Ansätze um eine bayesianische Komponente. Anzutreffen sind sie in der Praxis z.B. als Erweiterung der logistischen Regression. Der Bayesianische Ansatz erlaubt es, neben den Daten und der darin enthaltenen Information, zusätzliches "Vorwissen" in das Modell aufzunehmen. Genutzt werden Baye-

Ansatz	Berücksichtigte CJ Information	Validierung mittels Prognose möglich	Inhaltliche Interpretation des Wirkungsmechanismus möglich
Spieltheoretische Ansätze	Abfolge Werbekanäle	✗	✗
Machine Learning Algorithmen	Vollständig	✓	✗
Regressionsansätze	Vollständige und statistische Validierung	✓	✓

Abbildung 4: Vergleich datengetriebener Attributionsansätze

sianische Modelle klassischerweise vor allem dort, wo die Fallzahlen knapp sind. Dies ist - zumindest bei Websites mit hohem Traffic - in der Attribution meist nicht der Fall. Bei geringen Fallzahlen oder komplexen Modellen mit vielen Einflussfaktoren, kann die Erweiterung um eine bayesianische Komponente jedoch durchaus Vorteile bringen. Zu beachten ist dabei aber immer, dass die hinzugezogene Vorabinformation kritisch auf Plausibilität geprüft werden muss: ist sie nicht stimmig, liefert das Modell verfälschte Ergebnisse.

Vergleich der Ansätze

Die datengetriebenen Ansätze lassen sich durch die Beantwortung drei wesentlicher Fragen miteinander vergleichen:

1. Welcher Teil der in den CJ Daten enthaltenen Information kann bei der Attribution berücksichtigt werden?
2. Besteht die Möglichkeit im Rahmen einer Prognose die Modellierung zu validieren?
3. Erlaubt die Modellierung die Identifikation von Wirkungsmechanismen und fördert so ein mikroskopisches Verständnis der Kaufentscheidung?

Genutzte Information

Die Menge der vom Modell genutzten Information beim spieltheoretischen Ansatz ist im Vergleich zu Machine Learning und Regressionsansätzen gering. Berücksichtigt wird ausschließlich die konkrete Abfolge

der Werbekanäle in den CJs. Daher wird dieses Konzept i.d.R. nur für kurze CJs eingesetzt. Weitere Metriken wie das On-Site-Verhalten der User fließen bei der Berechnung des Shapley-Values nicht mit ein. Das hat zur Folge, dass der Beitrag eines Kontakts zum Kauf Erfolg identisch beurteilt wird, unabhängig davon, ob der User sich intensiv (lange Verweildauer, viele Seitenaufrufe) On-Site mit dem Angebot auseinandersetzt oder nicht. In der Praxis sind jedoch gerade diese Engagement Metriken sehr bedeutsam. Hintergrund ist die Erkenntnis, dass interessierte User eher kaufen als welche, die wenig intrinsisches Kaufinteresse zeigen. Da diese Regel auf grundlegendem Konsumentenverhalten beruht, sind die darauf basierenden Modelle in der Praxis sehr stabil. Analysen zeigen bspw., dass entsprechende Zusammenhänge sogar für das vierte Quartal des Jahres gelten, welches durch das Weihnachtsgeschäft von erhöhten Kosten für Werbung und höherer Nachfrage auf der Endkundenseite geprägt ist. Die Fähigkeit mehr als nur die Kanalinformation zu berücksichtigen, ist daher ein klarer Vorteil von Machine Learning Algorithmen und Regressionsansätzen. Allerdings ist zu beachten, dass bei allen Verfahren bei denen die Untersuchungseinheit die komplette CJ ist, der Analyse vorgelagert immer eine Aggregation über die Kontakte der CJ vorgenommen werden muss, was zwangsläufig mit einem Informationsverlust einhergeht (siehe Abbildung 5).

Mit dieser Schwierigkeit müssen Machine Learning Algorithmen und Regressionsansätze gleichermaßen zurecht kommen. Regressionsansätze bieten aufgrund ihrer statistischen Natur allerdings zusätzlich die Möglichkeit, die Signifikanz der berücksichtigten Variablen zu bestimmen und somit die Variablenauswahl statistisch zu steuern.

Einzelne Kontakte									
	Cookie ID	Visit ID	Start	End	PIs	Kanal	...	Conversion	
CJ 1	1014029144759194112	1052358947148965888	2014-09-08 20:54:21	2014-09-08 21:06:34	6	SEA	...	0	
	1014029144759194112	1952373370890660096	2014-09-08 21:08:41	2014-09-08 21:15:18	5	SEO	...	0	
	1014029144759194112	1955016269621012480	2014-09-10 16:54:10	2014-09-10 16:58:12	5	Display	...	0	
	1014029144759194112	1955028424915331328	2014-09-10 17:06:14	2014-09-10 17:16:35	6	Retargeting	...	0	
CJ 2	1030232047379288704	1935083979319191808	2014-08-27 22:53:13	2014-08-27 22:54:46	711	SEA	...	0	
	1030232047379288704	1940531378682503424	2014-08-31 17:04:43	2014-08-31 17:08:55	5	Display	...	0	
	1030232047379288704	1945841695205533696	2014-09-04 09:00:02	2014-09-04 09:00:39	24	SEA	...	0	
	1030232047379288704	1983984641373013504	2014-09-30 16:50:09	2014-09-30 16:49:34	20	SEA	...	0	
	1030232047379288704	1984003268008647936	2014-09-30 16:50:09	2014-09-30 17:03:28	20	Affiliate	...	1	
Aggregation									
	Cookie ID	Summe PI	Kontakte	Dauer	...	Conversion			
CJ 1	1014029144759194112	22	4	1.85	...	0			
CJ 2	1030232047379288704	67	5	33.76	...	1			

Abbildung 5: Aggregation der Kontakte einer CJ

Validierung der Modelle

Das eigentliche Attributionsproblem besteht darin, dass die Beiträge der einzelnen Kontakte zu dem erfolgreichen Ausgang der Customer Journey nicht direkt messbar sind. Es stellt sich also die Frage, inwieweit theoretisch motivierte Attributionskonzepte bezüglich ihrer Praxistauglichkeit validiert werden können. Da den Machine Learning Algorithmen und den Regressionsansätzen die Modellierung des Wirkungszusammenhangs zwischen Zielgröße und Customer Journey-Daten zugrunde liegt, können zukünftige Kaufentscheidungen vorhergesagt werden. Ein verlässliches Maß für die Güte dieser Verfahren ist damit die Fähigkeit zukünftige Kaufentscheidungen korrekt zu prognostizieren. An die Besonderheiten des Shops angepasste Regressionsansätze setzen in dieser Hinsicht Maßstäbe und erreichen ein AUC (Area under the Curve) von 0.75 bis 0.85¹. Dies stellt nicht nur eine Bestätigung der identifizierten Wirkungszusammenhänge dar, sondern unterstreicht auch das Potenzial dieser Verfahren im Bereich Predictive Analytics (s.u.).

¹Findet - analog zum spieltheoretischen Konzept - eine Beschränkung auf die Verwendung der Kanalinformation bei der Modellierung statt, sinkt die Prognosegüte der Regressionsverfahren um über 40%. Das bedeutet, dass die zusätzlich im Regressionsmodell berücksichtigten Metriken äußerst relevant für die Modellierung der Kaufentscheidung sind und die Attribution - wenn das Augenmerk auf der Modellgüte liegt - nicht ausschließlich auf der Abfolge der Kanäle beruhen sollte.

Inhaltliche Interpretation des Wirkungsmechanismus

Der spieltheoretische Ansatz verwendet ein fixes Kalkül zur Quantifizierung der marginalen Beiträge einzelner Kanäle zum Käuferfolg. Dies hat zur Folge, dass im Rahmen dieses Ansatzes keine darüber hinausgehenden Insights in den Kaufentscheidungsprozess gewonnen werden können. Bei Machine Learning Algorithmen wiederum handelt es sich aufgrund ihrer Entwicklungsgeschichte sehr häufig um Black-Box-Ansätze, die zwar die Zusammenhänge zwischen erklärenden und abhängiger Variable identifizieren, deren Parameter sich aber inhaltlich nicht (z.B. neuronale Netze) oder nur sehr schwer (z.B. Random Forests) interpretieren lassen. Regressionsmodelle hingegen haben den Vorteil, dass sich die identifizierten Wirkungszusammenhänge in Form direkt interpretierbarer Koeffizienten und der dazugehörigen Information zur statistischen Verlässlichkeit angeben und überprüfen lassen.

Das linke Panel (Abbildung 6) zeigt beispielhaft den quantifizierten Einfluss der Time on Site auf die Wahrscheinlichkeit, eine CJ mit einem Kauf abzuschließen. Verweildauern von fünf bis zehn Minuten resultieren in einer hohen Kaufwahrscheinlichkeit, längere Verweildauern zeigen einen abnehmenden Impact. Das rechte Panel zeigt (ebenfalls beispielhaft), dass eine mittlere Anzahl von Kontakten optimal in Bezug auf die Kaufwahrscheinlichkeit ist, wohingegen längere Ketten auf unentschlossene Kunden hindeuten.

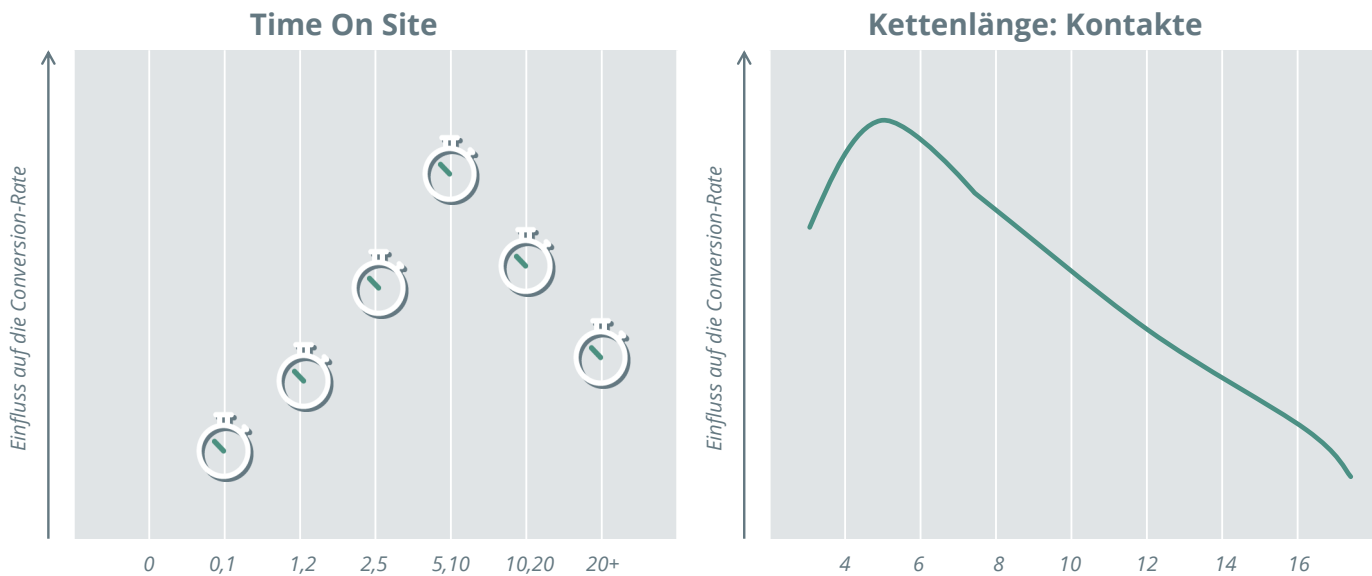


Abbildung 6: Der Einfluss der auf der Website verbrachten Zeit auf die Conversion-Rate

Attribution 2.0

Im Anschluss an die ausführliche Diskussion des Status Quo der datengetriebenen Ansätze stellt sich die Frage nach Perspektiven für die Weiterentwicklung der Attributionsmodellierung. Großes Potenzial besitzen die Survival- oder Überlebenszeitmodelle, eine weitere Spielart aus dem Feld der statistischen Regressionsansätze. Sie vereinbaren die der Regression eigene Robustheit und Interpretierbarkeit mit der Fähigkeit, Daten unmittelbar auf Ebene der einzelnen Kontakte zu nutzen. Dadurch kann die Aggregation und damit der wesentliche Nachteil bei der Modellierung mittels logistischer Regression entfallen. Die Modelle besitzen analog zu den spieltheoretischen Ansätzen ebenfalls die Möglichkeit die Kanalabfolge zu berücksichtigen - ein potenzieller Hebel um die Modellgüte zu erhöhen.

Die Survival- oder Überlebenszeitmodelle haben ihren Ursprung in der Biometrie. Analysiert wird die Dauer bis zum Eintritt eines bestimmten Ereignisses - in der klinischen Statistik oftmals die Zeit bis zum Tod. Im Rahmen der CJ Analysen wird analog die Dauer bis zum möglicherweise erfolgreichen Abschluss der CJ untersucht. Damit passt diese Modellklasse ideal auf die Datenstruktur von CJs: Für die Modellierung wird lediglich die Information verwendet, dass bis zu einem bestimmten Zeitpunkt noch keine Conversion erfolgt ist. Die oben beschriebenen Datenprobleme bei der Wahl des Analysezeitraums mit abgeschnittenen CJs gehören damit der Vergangenheit an. Da gerade an Methoden der Biometrie, die regelmäßig bei der Zulassung von Medikamenten und der Beurteilung der Wirksamkeit

von Therapien zum Einsatz kommen, höchste methodische Ansprüche gestellt werden, gelten die Survival Modelle als sehr gut erprobt und zuverlässig.

Zentrale Größe der Lebenszeitanalyse in der Attribution ist die zeitaufgelöste Conversion-Rate. Im Rahmen der Modellierung wird ermittelt, inwieweit die Historie der CJ (also die bis zu einem bestimmten Zeitpunkt stattgefundenen Kontakte und deren Eigenschaften) die Conversion-Rate beeinflusst. Dabei wird die Information zu jedem einzelnen Werbekontakt im Modell berücksichtigt und eine Aggregation über alle Kontakte der CJ entfällt (s.o.). Bei den Informationen kann es sich sowohl um die Kanalinformation eines jeden Kontakts (analog zum spieltheoretischen Ansatz) als auch um beliebige weitere Metriken (z.B. Engagement, View-Kontakte) handeln.

Der Survival-Ansatz kann ebenfalls zur Prognose des zukünftigen Kaufverhaltens verwendet werden und lässt sich somit out-of-sample hinsichtlich der Modellgüte beurteilen. Die geschätzten Wirkungszusammenhänge lassen sich überprüfen und im Rahmen einer inhaltlichen Interpretation plausibilisieren. Eine typische Fragestellung der Lebenszeitanalyse ist die Modellierung der Kaufwahrscheinlichkeit in Abhängigkeit der Tageszeit zu der der erste Kontakt stattgefunden hat. Gleichzeitig ist der Einfluss der seit dem ersten Kontakt vergangenen Zeit auf die Kaufwahrscheinlichkeit von Interesse. Der Survival-Ansatz ermöglicht die Identifikation und Visualisierung der Interaktion zwischen beiden Größen hinsichtlich ihrer Wirkung auf die Kaufwahrscheinlichkeit und ermöglicht so differenziertere Aussagen:

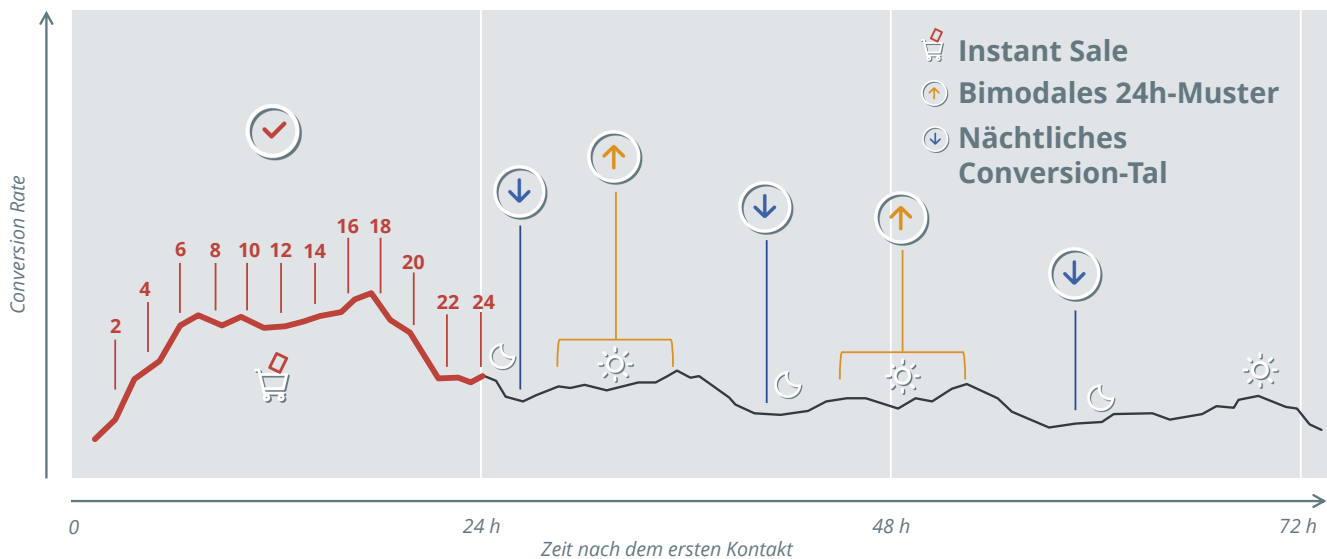


Abbildung 7: Conversion Rate in Abhängigkeit der Tageszeit nach dem ersten Kontakt

Zum einen lässt sich ein klares Muster für diejenigen Cjs erkennen, die tagsüber beginnen. Der erfolgreiche Abschluss ist hier direkt nach dem ersten Kontakt am wahrscheinlichsten und man beobachtet einen Instant Sale. Wird die CJ nicht zeitnah abgeschlossen, fällt die Wahrscheinlichkeit ab und ist in den Nachtstunden am niedrigsten (nächtliches Conversion-Tal). Eine erhöhte Conversion-Rate ist alle 24h zu verzeichnen. CJs, deren erster Kontakt in der Nacht stattfindet, weisen ein abweichendes Muster auf. Hier besteht sowohl nach relativ kurzer Zeit direkt am nächsten Morgen als auch erst wieder spät in der darauffolgenden Nacht eine erhöhte Kaufbereitschaft. Das klare Muster der nächtlichen Täler wird aufgebrochen und gleicht eher einem bimodalen 24h Muster. Perspektivisch kann dieses Wissen zur individuellen zeitlichen Aussteuerung von Display-Einblendungen im Real-Time-Bidding (RTB) genutzt werden und verspricht Einsparungspotenzial gegenüber pauschalen zeitbasierten Regeln (z.B. keine Werbung zwischen 0-7 Uhr).

Abbildung 8 zeigt beispielhaft, wie die im Rahmen einer Survival-Analyse gewonnenen Ergebnisse bei der Attribuierung des Umsatzes berücksichtigt werden. Dargestellt ist der Beitrag der verschiedenen Kontakte zum erfolgreichen Abschluss der CJ. Ergänzend ist zu jedem Kontakt die unterschiedliche Gewichtung der Einflussfaktoren ersichtlich:

Fazit und Ausblick

Die Attribution bildet die Grundlage einer effizienten Online-Marketing-Strategie und beeinflusst dadurch unmittelbar das erwirtschaftete Ergebnis. Die immer noch von einigen Unternehmen eingesetzten heuristischen Attributionsmodelle sind nicht mehr zeitgemäß. Wesentlich ist, den Schritt von der Heuristik hin zu einem datengetriebenen dynamischen Attributionsmodell zu gehen. Auf welches der oben vorgestellten Modelle dabei die Entscheidung fällt ist dabei zunächst nachrangig. Auf Ebene der Kanäle (horizontale Allokation) fallen die Unterschiede zwischen den Verfahren meist gering aus. Typisch sind hier Abweichungen bis zu +/- 5%. Erst wenn die Anforderungen spezieller sind, z.B. wenn verlässliche Ergebnisse auch in kleinen Subgruppen (z.B. kurzfristig bei der Beurteilung eines neuen Banners) benötigt werden, kommt der Modellgüte eine erhöhte Bedeutung zu. Die jüngere Entwicklung zeigt, dass sich insbesondere die Modelle aus der Familie der Regressionsansätze durch Shop-spezifische Erweiterungen optimieren lassen. Dabei ist im Einzelfall auch anhand des zu steuernden Budgets zu entscheiden, bis zu welchem Punkt eine Optimierung sinnvoll ist.

Das Survival Modell stellt einen datengetriebenen Ansatz dar, der ohne einen Informationsverlust durch Aggregation auskommt. So lassen sich sowohl die Kanalinformation eines jeden Kontakts nutzen (Stärke des spieltheoretischen Ansatzes) als auch beliebige weitere Metriken (z.B. Engagement, View-Kontakte, etc.) im Rahmen einer Regressionsmodellierung nutzen.

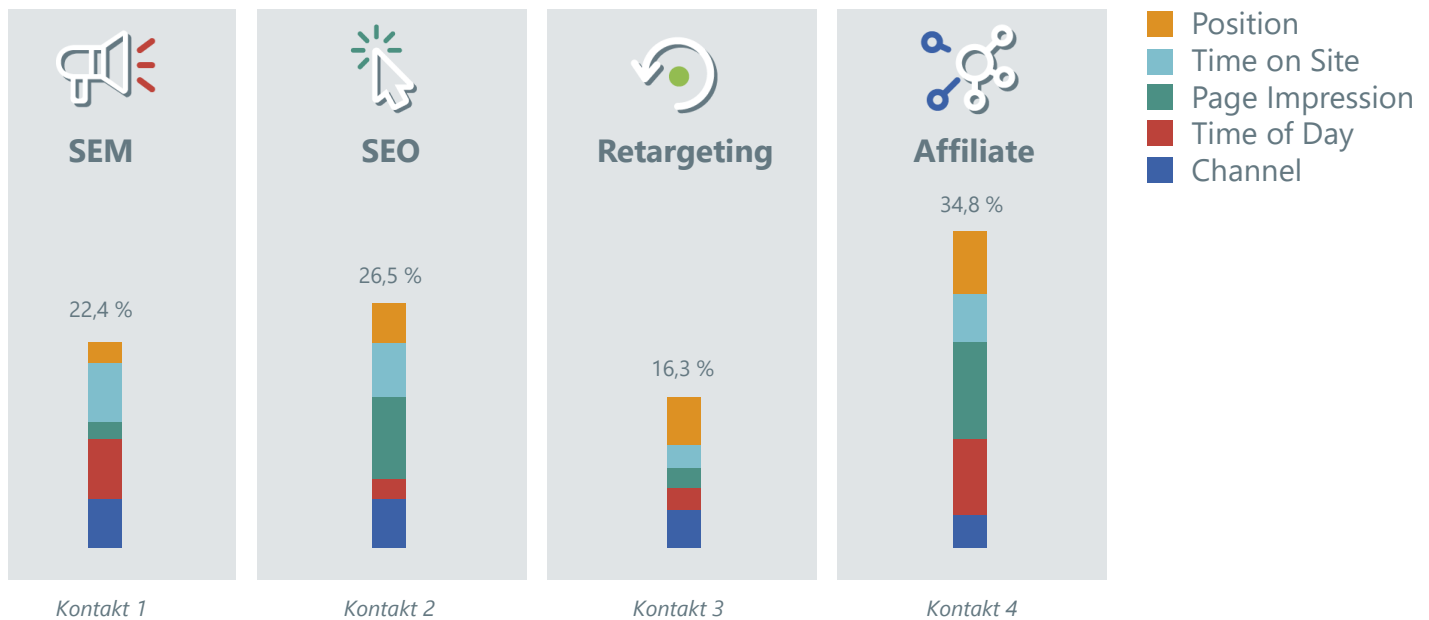


Abbildung 8: Beitrag der verschiedenen Kontakte zum erfolgreichen Abschluss der CJ

Erweiterungen

Bei der Erweiterung der Online-Attribution stehen zwei Themen im Fokus: Die Einbeziehung von Offline-Kontakten (hauptsächlich TV-Werbung und der Einfluss von postalisch verschickten Katalogen) und die Nutzung der Daten und Modelle für den Bereich Predictive-Analytics.

Die Nutzung von CJ Daten birgt neben der datengetriebenen Attribution großes Potenzial im Bereich Predictive Analytics. Die in den Daten enthaltene Information lässt sich nicht nur rückwirkend für die Attribution, sondern auch in Echtzeit für die Prognose des Kundenverhaltens verwenden. Das Konzept der Predictive Customer Segmentation prognostiziert die Kaufwahrscheinlichkeit einzelner Kunden mit Hilfe von Lebenszeitmodellen. Die auf Basis dieser Prognose identifizierten Kundensegmente können dann mittels differenzierter Marketingmaßnahmen - ggf. auch in Echtzeit über RTB - angesprochen werden. Die Kombination aus der Prognose der Kaufwahrscheinlichkeiten und dem Prognose-basierten Marketing ermöglichen eine deutliche Steigerung des ROI.

Ein gewisses Mindestbudget für Offline-Werbung vorausgesetzt, funktionieren sowohl die Berücksichtigung von TV-Werbung als auch die Erkennung von Katalog-Einfluss bei Online-Bestellungen mittlerweile recht zuverlässig. Analog zur Online-Attribution findet beim TV-Impact gerade die Ablösung früher heuristischer

Ansätze, die ausschließlich auf einer simplen Baseline-Substraction basieren, durch fundierte datengetriebene Algorithmen statt (vgl.: Best Practice TV-Tracking: Why a simple baseline correction is not enough!).

Die Strategie einiger Shops beinhaltet die gezielte Ansprache bestimmter Bestandskunden durch Katalogsendungen. Nicht selten ist das Budget für den Katalogversand erheblich. Entsprechend groß ist der Hebel, der sich aus der gezielten Aussteuerung von Online- und Katalog-Budget ergibt. Die häufig praktizierte Strategie, jeden online erzielten Verkauf vollständig dem Katalog zu attribuieren, sofern dem Kauf der Erhalt eines Katalogs innerhalb eines definierten Zeitfensters vorausging, ist irreführend. Einerseits zeigen Analysen von CJ Daten, dass sich die Verhaltensmuster typischer Katalogkunden von denen klassischer Pure-Online-Kunden unterscheiden: Erstere informieren sich primär außerhalb des Internets. Die CJs von Katalogkunden sind daher im Durchschnitt kürzer, enthalten weniger Pls und beinhalten einen größeren Anteil an Kontakten aus den Kanälen "Direct" und "SEO/SEM Brand".

Zudem unterscheidet sich häufig die Zusammensetzung der Warenkörbe. Diese Muster finden sich jedoch nicht bei allen Online-Kunden, die einen Katalog erhalten haben. Dies legt den Verdacht nahe, dass einige Kunden den Katalog nicht beachten. In diesen Fällen hat der Katalog soviel Relevanz für die Kaufentscheidung, wie ein Display-Banner außerhalb des sichtbaren Bereichs des Bildschirms. Dieser Verdacht lässt sich durch Ergebnisse

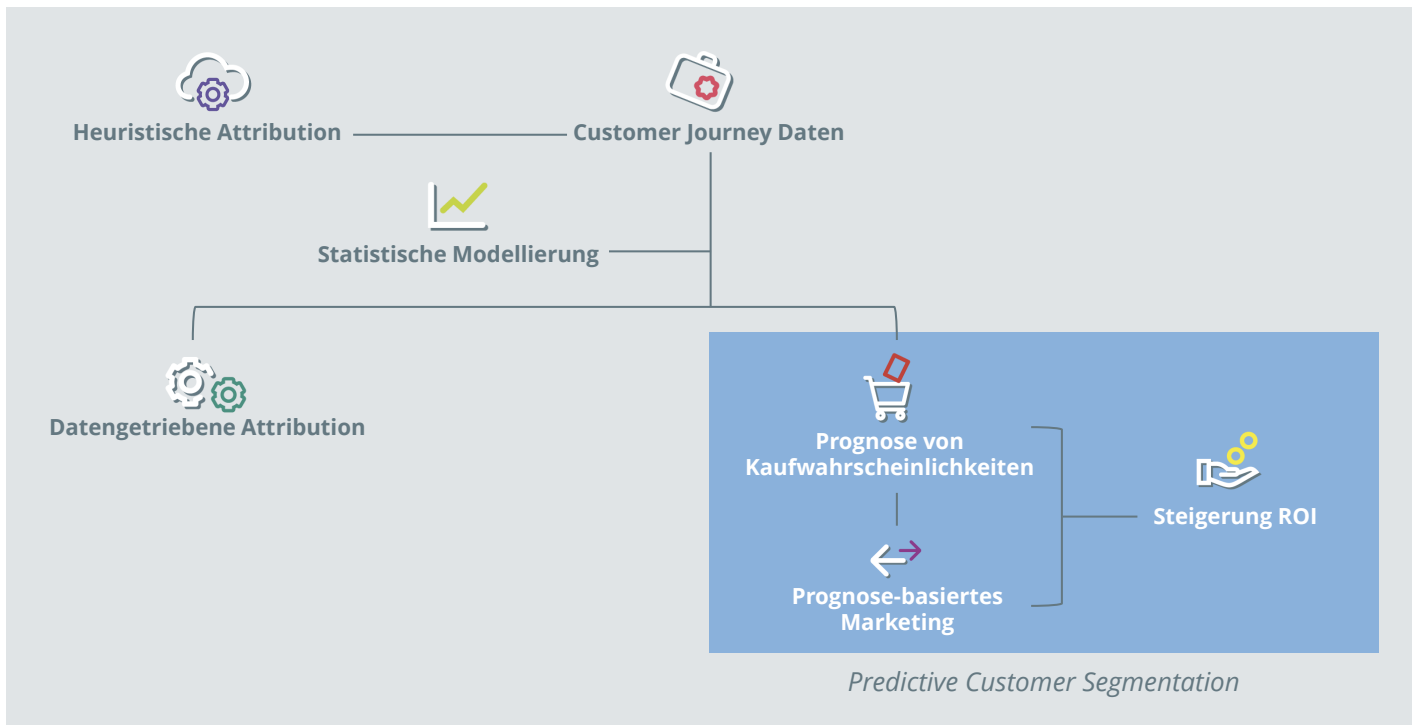


Abbildung 9: Potenzial der Nutzung von CJ Daten

von flankierenden Bestellabschlussbefragungen untermauern, nach denen bei bis zu einem Drittel der Katalogempfänger der Erhalt des Katalogs irrelevant für den Kauf war. Intelligente datengetriebene Modelle sind in der Lage diese Bestellungen zu identifizieren - der Hebel in Bezug auf die effiziente Verwendung des Offline-Budgets ist offensichtlich.

Ihr Ansprechpartner





Dr. Steffen Wagner

Steffen ist Mitgründer von INWT. Er beschäftigt sich schwerpunktmäßig mit den Themen Predictive Analytics, Online Marketing und Customer Relationship Management. Der promovierte Physiker gibt als Dozent im 'Joint Masters Program of Statistics' in Berlin Einblicke in seine Tätigkeit als Data Scientist.


Kontakt

- **Tel.:** +49 30 1208231-58
- **E-Mail:** steffen.wagner@inwt-statistics.de

 **INWT Statistics GmbH**
Hauptstraße 8
Meisenbach Höfe, Aufgang 3a
10827 Berlin

 +49 30 1208231-0

 info@inwt-statistics.de

 www.inwt-statistics.de

