

planung & analyse

Zeitschrift für Marktforschung und Marketing www.planung-analyse.de
Eine Marke der dfv Mediengruppe

6/2015 D11700F

Datenschutz

EU-Verordnung
ante portas

Special

Forschung für
B2B-Märkte

Schwerpunkt Kommunikation

Um jeden
Preis
auffallen



Treiberanalysen im Fitness-Check

Ein Vergleichsstudie für die Marktforschungspraxis

Herauszufinden, welche Faktoren die Kundenzufriedenheit oder auch den Absatz eines Unternehmens beeinflussen, ist vielleicht das zentrale Kernanliegen der Marktforschung. Hierfür werden Treiberanalysen genutzt. Im Laufe der Zeit ist eine ganze Reihe an verschiedenen Varianten und Algorithmen entstanden, die hier einem systematischen Vergleich unterzogen werden.

Die Ergebnisse von umfangreichen Simulationsstudien und praktischen Erfahrungen des TNS Infratest Advanced Analytics-Teams können einen Beitrag zur Objektivierung der Diskussion leisten, welches Treiberverfahren das „beste“ ist. Diese gewinnt durch die Tatsache an Schärfe, dass unter dem Begriff der Treiberanalyse grundsätzlich zwei verschiedene Fragestellungen oder Zielsetzungen verborgen sind:

► Wie stark beeinflusst eine Treibervariable die Zielvariable? Dies ist die Frage nach der Wichtigkeit (Importance) der Treiber. Hier gilt es also, die Effektstärken einzelner Treiber möglichst unverfälscht zu ermitteln.

► Welche Veränderung der Zielvariablen wird durch eine Veränderung der Treibervariablen ausgelöst? Hierbei geht es stärker um die Prognosewirkung (Predictive Power).

Gleichzeitig warten Daten in der Marktforschungspraxis mit problematischen Strukturen wie fehlenden Werten oder Multikollinearität (Treibervariablen korrelieren substantiell miteinander) auf. Die einzelnen Verfahren sind hierfür unterschiedlich gut geeignet. Was also zeichnet ein gutes oder geeignetes Verfahren der Treiberanalyse aus? Wir legen unserem Verfahrenvergleich folgende Prüfpunkte zugrunde:

1. Inhaltliche Zielsetzung: Steht die Wirkung (Importance) einzelner Treibervariablen oder Prognose von Veränderungen im Fokus?

2. Umgang mit fehlenden Werten: Ist ein Verfahren in der Lage mit fehlenden Werten umzugehen oder müssen je Datenfall immer vollständige Informationen vorliegen?

3. Beherrschung von Multikollinearität: Kann ein Verfahren auch bei moderaten bis hohen Interkorrelationen der Treibervariablen plausible Wirkungen der Treiber abschätzen?

4. Fallzahlen: Welche minimale Fallzahl ist nötig und wie gut funktioniert das Verfahren auch bei kleinem n ?

5. Rechenaufwand: Nimmt die Rechenzeit mit zunehmender Modellkomplexität, das heißt

mit steigender Anzahl an Treibervariablen, stark zu, so dass das Verfahren sogar mit heutigen Rechnern nicht sinnvoll anwendbar ist?

6. Abhängigkeit von Startlösungen: Sind Ergebnisse vergleichbar, das heißt, liefert jeder Analysedurchlauf mit denselben Daten immer reliable Ergebnisse oder variieren diese?

7. Validität: Kann das Verfahren „echte“ Zusammenhänge und Wirkungseinflüsse, die in der Realität bestehen, richtig abbilden?

In unserem Vergleich (siehe Kasten) haben wir uns auf Verfahren konzentriert, die entweder allgemein als Standard gelten, das Kernproblem Multikollinearität gut handhaben können oder hin und wieder in der Diskussion stehen (Random Forest, Correlated Component Regression, Shapley Value, Kruskal). Abbildung 1 zeigt einen knappen Überblick über Fähigkeiten und Eigenschaften der Verfahren hinsichtlich der Qualitätsmerkmale, die wir gerade definiert haben.

Welche Verfahren kommen zu ähnlichen Ergebnissen?

Aus einer ersten Analyse erkennen wir, das Rennen um das „beste“ Treiberanalyseverfahren ist eher ein Rennen von Verfahrensgruppen als einzelner Verfahren. Hierfür untersuchten wir reale Datensätze aus den Branchen Consumer, Automobil und Verkehr sowie Finanzforschung mit Stichproben unterschiedlichster Größe (zwischen $n=769$ und $n=16.333$) sowie mit jeweils zwischen 16 und 23 Treibervariablen. Auch die abhängigen Variablen hatten unterschiedliche Inhalte (Markenstärke, NPS, Zufriedenheit, Image, Wiederkauf). Zur Ähnlichkeitsbeurteilung wurden alle Treiberanalyseverfahren auf alle Datensätze angewendet und die mittleren Abweichungen der errechneten Treiberwichtigkeiten untereinander ermittelt. Abbildung 2 zeigt diese relativen Abweichungen. Grundsätzlich ist festzustellen, dass die Verfahren mit mittleren Abweichungen der relativen

► Die Autoren



Dr. Markus Eberl (oben), Senior Director; **André Hahn** (Mitte links), Associate Director; **Dr. Daniel Mühlhaus** (Mitte rechts), Associate Director; **Steffen Seiffarth**, (unten) Associate Director; **Dr. Corina Wagner** Senior Consultant. Alle Autoren arbeiten im Bereich Applied Marketing Science (AMS) bei TNS Infratest. AMS ist das Advanced Analytics-Kompetenzzentrum von TNS in Deutschland und der methodische Forschungs- und Entwicklungsbereich.

Verfahren der Treiberanalyse

Bivariate Korrelation (CORR): Die Korrelation ist ein – (meist) auf den Wertebereich von -1 bis +1 normiertes – Maß für den linearen Zusammenhang zweier Merkmale. Das Vorzeichen gibt dabei die Richtung, die betragsmäßige Abweichung von Null die Stärke des Zusammenhangs an. Je nach Datenlage gibt es unterschiedliche Definitionen von Korrelationskoeffizienten, wie etwa die Korrelation nach (Bravais-) Pearson, die Spearman-Rangkorrelation, den Tanimoto-Koeffizienten, oder Kendalls Tau. Allen gemeinsam ist jedoch, dass sie nicht in der Lage sind, relative Wichtigkeiten mehrerer Treiber simultan zu quantifizieren.

Multiple lineare Regression (Ordinary Least Squares – OLS): OLS schätzt den linearen Einfluss jedes Prädiktors. Standardisierte Koeffizienten (beta) eliminieren unterschiedliche Messdimensionen und eignen sich damit quadriert und relativ zur Bestimmung der relativen Wichtigkeit (siehe Grömping 2006, Bocken 2007).

Correlated Component Regression (CCR): Hier werden k korrelierte Komponenten (jeweils als Linearkombination der Prädiktoren) erstellt und über eine Regression der Zielvariable auf die Komponenten dann die Treiberwirkungen ermittelt. Drei bis vier Komponenten gelten als ausreichend für gute Vorhersagen (Magidson 2010).

Relative Weight Analyse nach Johnson: In dem auf Johnson (2000) zurückgehenden Verfahren werden durch Transformationen Variablen gebildet, die maximal mit den ursprünglichen Prädiktoren korrelieren, untereinander aber unkorreliert sind. Die relativen Gewichte ergeben sich aus der Multiplikation der Koeffizienten einer Regression der abhängigen Variable auf diese transformierten Variablen mit den Koeffizienten einer Regression des betrachteten Prädiktors auf die transformierten Variablen (siehe Tonidandel und LeBreton 2011).

Shapley Value Regression: Grömping (2006) hat auf der Basis der Arbeiten von Lindeman, Merenda und Gold (1980) ein Maß für die relative Wichtigkeit von Prädiktoren entwickelt, das sich aus der mittleren Verbesserung des R² bei Einführung des jeweils betrachteten Prädiktors zu allen mög-

lichen Kombinationen der anderen Treiber errechnet. Der mittlere R²-Zuwachs kann als Maß der relativen Einflussstärke verwendet werden. Aus den ermittelten Maßen für relative Wichtigkeiten („Shapley Values“) lassen sich über die Lösung eines nicht-lineares Gleichungssystems neue Regressionskoeffizienten ermitteln, welche zur Prognose einsetzbar sind.

Genizi Coefficients: Die Idee dieses Verfahrens von Genizi (1993) liegt darin, die untereinander korrelierten unabhängigen Variablen einer linearen Transformation zu unterziehen, welche zu einer gleichen Anzahl von möglichst unkorrelierten Prädiktoren der unabhängigen Variablen führt. Diese sollen einen geringstmöglichen quadrierten Abstand zu den ursprünglichen unabhängigen Variablen aufweisen. Aus den Transformationskoeffizienten sowie den Regressionskoeffizienten der transformierten Prädiktoren werden dann die Wichtigkeiten der Prädiktoren bzw. Prognosekoeffizienten geschätzt.

Kruskal Importance (KI): Kruskals (1987) „Averaging over Orderings“ gibt als Wichtigkeit eines Prädiktors das Mittel der quadrierten Partialkorrelationen an. Um Rechenzeiten zu verkürzen, wird in der Regel ein reduzierter Ansatz verwendet, der lediglich Partialkorrelationen erster und zweiter Ord-

nung in die Berechnung einbezieht. Dieser Ansatz funktioniert bei den üblichen Anwendungen in der Marktforschungspraxis auch mit bis zu 100 Treibervariablen gut und liefert Ergebnisse, die mit dem „vollständigen“ Kruskal-Verfahren nahezu identisch sind. Bei „Kruskal 2.0“ handelt es sich um eine von den Autoren dieses Beitrags entwickelte Modifikation des Kruskal-Ansatzes. Sie ist eine Mischform aus Kruskal Importance-Analyse und Shapley Value Regression. Analog zu diesen beiden Ansätzen werden auch hier für Kombinationen an Treibern Wirkungseffekte ermittelt. Da jeder Subanalyse eine Regression zugrunde liegt, sind Beta-Koeffizienten für Prognosen verfügbar. Als weitere Variante ist mit Kruskal Light (KIL) ein stark vereinfachter, robuster Ansatz ableitbar, der die mittleren Partialkorrelationsbeiträge erster Ordnung verwendet.

Random Forest: Bei diesem Verfahren wird eine große Zahl an Entscheidungsbäumen (viele Bäume = Wald) geschätzt. Für jeden Knoten innerhalb eines Baumes wird derjenige Treiber aus der Menge an Treibern ausgewählt, der die höchste Varianzaufklärung zeigt (Breimann 2001). Die anhand von verschieden, hierarchisch angeordneten Knoten zusammengesetzten Bäume sind dann die Grundlage für die Abschätzung der Wichtigkeiten der Treiber (Liaw und Wiener 2002).

Verfahren	Inhaltliche Zielsetzung			Datenstruktur			Effizienz
	Relative Wichtigkeiten	Prognose	Fehlende Werte	Multikollinearität	Geringe Fallzahl	Max. Itemzahl (Rechenzeiten)	
Bivariate Korrelation (CORR)	x	x	✓	x	✓	∞	✓
Lineare Regression (OLS)	✓	✓	x	x	○	∞	✓
Correlated Component Reg (CCR)	✓	✓	x	○	○	∞	✓
Johnson's Relative Weights (JOHNS)	✓	✓	x	✓	○	∞	✓
Shapley Value Regression (SV)	✓	✓	x	✓	○	20	✓
Genizi (GEN)	✓	✓	x	✓	○	∞	✓
Kruskal Importance (KI)	✓	x	✓	✓	○	100	✓
Kruskal 2.0 (KI2.0)	✓	✓	○	✓	○	150	✓
Kruskal Light (KIL)	✓	x	✓	✓	✓	∞	✓
Random Forest (RF)	✓	✓	x	✓	○	50	x

Goegnet ✓
Machbar ○
Problematisch x

© Eberl et.al.: planung & analyse 6/2015

Abbildung 1: Übersicht der untersuchten Verfahren.

	OLS	CCR	JOHNS	KI	KI2.0	SV	GEN	RF	KIL	CORR
OLS	0,00	0,04	0,17	0,19	0,21	0,18	0,19	0,19	0,23	0,23
CCR	0,04	0,00	0,16	0,20	0,21	0,18	0,19	0,20	0,24	0,23
JOHNS	0,17	0,16	0,00	0,04	0,07	0,04	0,04	0,09	0,10	0,10
KI	0,19	0,20	0,04	0,00	0,03	0,02	0,03	0,08	0,07	0,06
KI2.0	0,21	0,21	0,07	0,03	0,00	0,03	0,04	0,08	0,04	0,04
SV	0,18	0,18	0,04	0,02	0,03	0,00	0,01	0,08	0,06	0,06
GEN	0,19	0,19	0,04	0,03	0,04	0,01	0,00	0,08	0,06	0,06
RF	0,19	0,20	0,09	0,08	0,08	0,08	0,08	0,00	0,09	0,09
KIL	0,23	0,24	0,10	0,07	0,04	0,06	0,06	0,09	0,00	0,01
CORR	0,23	0,23	0,10	0,06	0,04	0,06	0,06	0,09	0,01	0,00

© Eberl et.al.; planung & analyse 6/2015

► **Kurzfassung** Die Identifikation von Faktoren, die einen starken Einfluss auf Zielgrößen wie die Kundenzufriedenheit oder den Markterfolg eines Unternehmens aufweisen, steht seit jeher im Kern der Marktforschung. Das Spektrum an bestehenden Varianten der Treiberanalyse ist groß und wächst stetig. Brauchbare Verfahrensvergleiche sind jedoch nur unzureichend verfügbar, sodass sich hier oftmals treue Verfahren-Anhänger standhaften Ablehnern gegenüberstehen. Der vorliegende Beitrag nimmt einen umfangreichen und systematischen Vergleich aktuell diskutierter Verfahren der Treiberanalyse vor. Anhand einer Simulationsstudie sowie anhand realer Daten wird aufgezeigt, unter welchen Bedingungen welches Verfahren geeignet ist, womit ein erster Beitrag zu mehr Objektivität und Übersicht in diesem höchstrelevanten Bereich der Marktforschung geleistet wird.

► **Abstract** Identifying those levers that impact KPIs like customer satisfaction or success in the marketplace has always been a key question for market research as a discipline. However, there is a myriad of possible techniques and algorithms when it comes to this key driver analysis. In addition to this ever growing amount of possible techniques (from regression to SEM to Genizi to Kruskal – to name only a few), there is a lack of helpful and practitioner-oriented advice which of these to choose for which kind of research and data. It seems like often this question is more a matter of personal taste and gut feeling than systematic research on characteristics of the various algorithms. This paper presents an overview of those techniques, discusses their pros and cons and provides empirical comparisons of the behavior of those algorithms when it comes to everyday market research data. We do this in the context of real-world data as well as a comprehensive simulation study that provides clear guidance which algorithm is fittest in the race to the title of best driver analysis technique.

Abbildung 2: Mittlere Abweichungen der relativen Wichtigkeiten (Hervorhebungen: „Cluster“ von Verfahren, die untereinander nur geringe Abweichungen zeigen).

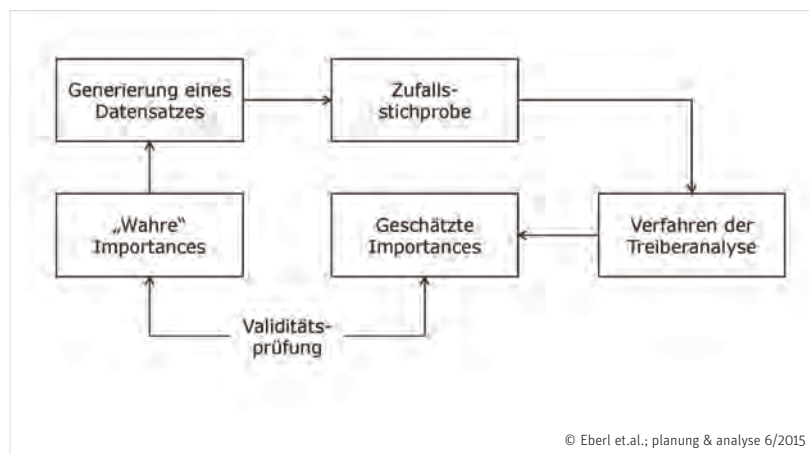
Wichtigkeiten im Bereich von teilweise mehr als drei Prozentpunkten bei den realen Datensätzen zu durchaus unterschiedlichen Ergebnissen führen. Wir entdecken aber vier Gruppen mit jeweils untereinander stabilen Aussagen: Regressionsverfahren (OLS, CCR), partitionierende Verfahren (JOHNS, KI, SV, GEN), Random Forest (RF) und Korrelationsverfahren (CORR und KIL).

Welche Verfahren können reale Zusammenhänge abbilden?

Die Analyse realer Daten gibt zwar Aufschluss über die Ähnlichkeit der Verfahrensergebnisse, sie kann aber natürlich keine Aussage darüber liefern, welches Verfahren unter welchen Bedingungen die Realität am besten beschreibt. Hierzu benötigen wir simulierte Datensätze, bei denen bekannt ist, welche

Ergebnisse „wahr“ sind, das heißt die tatsächlichen Wichtigkeiten der einzelnen Treiber gegeben sind, und entsprechende Störgrößen (Rauschen, fehlende Werte) kontrollierbar auf die Daten angewendet werden können. Diese „wahren“ Wichtigkeiten können dann mit den Ergebnissen, die verschiedene Verfahren liefern, verglichen werden (siehe Abbildung 3). Zudem kann mit diesen Daten abgeschätzt werden, in welchen Situationen welche Verfahren eine besonders gute oder schlechte Eignung aufweisen.

In unserer Studie haben wir verschiedene Datensätze, also sowohl Treibervariablen (X) als auch die abhängige Variable (Y) als Kombinationen dieser Faktoren ($Y=B \cdot X + e$) entsprechend üblicher Verteilungsannahmen generiert. Dabei wurden verschiedene Datenkonstellationen simuliert:



© Eberl et.al.; planung & analyse 6/2015

Abbildung 3: Logik der Validitätsprüfung.

Szenario	N	Multi-kollinearität	Faktor-struktur	R ²	OLS	CCR	JOHNS	KI	KI2.0	SV	GENIZI	RF	KIL	CORR	Mittelwert	Standardabweichung
1	gering	gering	nein	gering	9,66	9,65	7,71	7,11	6,87	7,16	7,15	6,32	6,56	6,57	7,48	1,21
2	hoch	gering	nein	gering	6,65	6,65	5,66	5,64	5,72	5,64	5,64	6,2	5,96	5,96	5,97	0,40
3	gering	hoch	nein	gering	10,32	10,32	8,65	8,18	7,67	7,32	7,23	6,73	6,72	6,74	7,99	1,38
4	hoch	hoch	nein	gering	8,71	8,7	6,45	6,42	6,43	6,42	6,4	6,68	6,8	6,82	6,98	0,92
5	gering	gering	ja	gering	9,19	9,06	8,01	6,97	6,91	7,01	6,95	6,82	6,79	6,84	7,46	0,95
6	hoch	gering	ja	gering	8,14	8,19	6,02	5,79	5,84	5,84	5,79	6,18	5,94	5,95	6,37	0,95
7	gering	hoch	ja	gering	10,37	10,8	8,95	7,64	7,57	7,58	7,42	7,16	7,25	7,27	8,20	1,36
8	hoch	hoch	ja	gering	8,86	8,97	6,47	6,43	6,52	6,44	6,47	6,67	6,66	6,66	7,02	1,01
9	gering	gering	nein	hoch	7,26	7,28	6,12	6,12	6,21	6,09	6,05	6,45	6,33	6,38	6,43	0,46
10	hoch	gering	nein	hoch	3,98	3,96	3,92	4,35	4,66	4,21	4,14	4,59	5,23	5,23	4,43	0,49
11	gering	hoch	nein	hoch	9,4	9,36	7,07	7,02	6,99	7,04	7	7,38	7,19	7,22	7,57	0,96
12	hoch	hoch	nein	hoch	5,58	5,58	5,41	5,49	5,82	5,87	5,87	5,49	6,49	6,49	5,81	0,39
13	gering	gering	ja	hoch	8,3	8,13	6,16	6,2	6,41	6,22	6,14	6,53	6,56	6,61	6,73	0,80
14	hoch	gering	ja	hoch	4,81	4,89	4,86	5,49	5,73	5,33	5,32	5,51	6,03	6,04	5,40	0,45
15	gering	hoch	ja	hoch	9,85	9,64	6,76	6,5	6,62	6,53	6,47	6,87	6,73	6,78	7,28	1,31
16	hoch	hoch	ja	hoch	6,5	6,52	5,4	6,02	6,25	6,06	6,08	5,9	6,57	6,58	6,19	0,38
Mittelwert					7,97	7,98	6,48	6,34	6,39	6,30	6,26	6,34	6,49	6,51		

© Eberl et.al.; planung & analyse 6/2015

Abbildung 4: Mittlere betragsmäßige Abweichung von den wahren Wichtigkeiten in Abhängigkeit von Datenstruktur-Variationen (kleinere Abweichungen sind zu bevorzugen).

planung & analyse

- ▶ Relation von Fallzahl zu Itemzahl: hoch (25x), gering (5x)
 - ▶ Multikollinearität: hoch, gering
 - ▶ Struktur der Treibervariablen: keine, 2 Faktoren
 - ▶ R² (Modellgüte): hoch (0.7), gering (0.4)
- Insgesamt resultieren so 16 verschiedene Typen von Datensätzen, auf die dann alle dargestellten Verfahren der Treiberanalyse angewendet werden. Um sicherzustellen, dass nicht einzelne Datenkonstellationen die Ergebnisse treiben, werden für jeden Datentyp 30 Datensätze generiert. Die mit den Verfahren ermittelten geschätzten Wichtigkeiten der Treibervariablen werden dann mit den jeweils bekannten *wahren* Wichtigkeiten, die im Rahmen der Datengenerierung erzeugt wurden, abgeglichen und jeweils die durchschnittliche betragsmäßige Abweichung ermittelt und über alle 30 Wiederholungen je Datentyp gemittelt.

Aus der beschriebenen Validitätsprüfung resultiert Abbildung 4, die für alle 16 Datentypen und untersuchten Verfahren die Simulationsergebnisse darstellt. Da mit standardisierten Variablen gerechnet wurde, können die wahren Betas – quadriert und relativiert – als prozentuale Wichtigkeiten interpretiert werden, die mittleren Abweichungen sind also als Prozentpunkte-Abweichung vom wahren Wert lesbar.

Zunächst fällt auf, dass die Unterschiede in der Anpassungsgüte zwischen den Szenarien – also in einer zeilenweisen Betrachtung – immens sind. Während bei hoher Fallzahl, ge-

ringer Multikollinearität und hohem R² die mittlere Abweichung über alle Verfahren hinweg bei 4,4 (Szenario 10, ohne Faktorstruktur) bzw. 5,4 (Szenario 14, mit Faktorstruktur) Prozentpunkten liegt, ist dieser Mittelwert im jeweils *spiegelbildlichen* Szenario deutlich höher. Ist die Fallzahl gering, korrelieren die Treiber stark und können zudem nur wenig Varianz der abhängigen Variable erklären, so liegt die mittlere Abweichung zwischen den geschätzten und vorgegebenen Wichtigkeiten bei zirka acht Prozentpunkten (Szenario 3: 8,0 Prozentpunkte, Szenario 7: 8,2 Prozentpunkte). Das schwierigste Datenszenario liegt demnach bei geringer Fallzahl, geringem R² und hoher Multikollinearität vor, wenn also die Beziehungen zwischen den Treibern im Verhältnis zur abhängigen Variable besonders stark sind. Der Blick auf die Standardabweichung zeigt, dass bei den „schwierigen“ Datenfällen außerdem große Variabilität zwischen den Verfahren in den Ergebnissen zu Tage tritt (1,38 bei Szenario 3 bzw. 1,36 bei Szenario 7). Das bedeutet, dass die Verfahren im *einfachen* Fall alle ähnlich gute Ergebnisse liefern, sich demgegenüber bei *schwieriger* Datenlage deutliche Unterschiede in der Eigenschaft der Verfahren zeigen.

OLS und CCR erweisen sich als problematisch, während die meisten partitionierenden Methoden (Shapley Value Regression, Genizi und die Verfahren der Kruskal-Klasse) sowie Random Forest in verschiedenen Datenstrukturen eine hohe Validität aufweisen, also auch in „schwierigen“ Datenkonstellationen funktio-

R ²		Multi-kollinearität		geringe Fallzahl		hohe Fallzahl	
gering (< 0,5)				<ul style="list-style-type: none"> ■ Korrelation (CORR) ^{M, I} ■ Kruskal Light (KIL) ^{M, I} 	Hohe Eignung: grundsätzlich <ul style="list-style-type: none"> ■ Kruskal Importance (KI) ^{M, I} ■ Kruskal 2.0 (KI2.0) ^{M, I} ohne Missings: <ul style="list-style-type: none"> ■ Genizi (GEN) ^I ohne Missings / geringe Itemzahl: <ul style="list-style-type: none"> ■ Random Forest (RF) ^I ■ Shapley (SV) ^I 		
				<ul style="list-style-type: none"> ■ Korrelation (CORR) ^{M, I} ■ Johnson (JOHNS) ^{M, I} ■ Kruskal Light (KIL) ^{M, I} 			
hoch (> 0,7)		gering		<ul style="list-style-type: none"> ■ Korrelation (CORR) ^{M, I} ■ Johnson (JOHNS) ^{M, I} ■ Kruskal Light (KIL) ^{M, I} 	Hohe Eignung: grundsätzlich <ul style="list-style-type: none"> ■ Kruskal Importance (KI) ^{M, I} ■ Kruskal 2.0 (KI2.0) ^{M, I} ohne Missings: <ul style="list-style-type: none"> ■ Genizi (GEN) ^I ohne Missings / geringe Itemzahl: <ul style="list-style-type: none"> ■ Random Forest (RF) ^I ■ Shapley (SV) ^I 		
		hoch		<ul style="list-style-type: none"> ■ Lineare Regression (OLS) ^{M, I} ■ CCR ^{M, I} ■ Johnson (JOHNS) ^{M, I} 			

Legende: P: Prognosefähigkeit vorhanden | M: Handhabung von Missings unproblematisch | I: Hohe Itemzahl möglich

© Eberl et.al.; planung & analyse 6/2015

Abbildung 5: Fähigkeiten und Eignung der Verfahren in Abhängigkeit der Datenstruktur.

nieren – einzige Ausnahme hier ist Johnsons Verfahren, das bei geringer Fallzahl und geringem R² deutlich höhere Fehler als die anderen Verfahren seiner Klasse produziert.

Empfehlungen

Die Gesamtschau unserer Bewertungen und Analysen erlaubt also generelle Empfehlungen, die in Abbildung 5 zusammengefasst werden:

Das OLS-Regressionsverfahren (genauso wie CCR) ist nur in spezifischen Situationen, nämlich bei hohem (erwarteten) R², hoher Fallzahl und höchstens moderater Multikollinearität geeignet, ansonsten jedoch wenig robust.

Random Forest (RF) und die partitionierenden Verfahren (JOHNS, KI, SV, GEN) weisen im Vergleich zu OLS und CORR eine deutlich höhere Eignung auf, die unabhängiger ist von der Datenlage. RF ist jedoch in der Regel weniger präferiert, da die Ergebnisse abhängig von Startlösungen sind und somit jeder Analyse-durchlauf mit denselben Daten immer (wenn auch nur graduell) unterschiedliche Ergebnisse liefert. Außerdem steigt in Abhängigkeit von Fall- und Itemzahl die Laufzeit für die Berechnungen stark an.

Genizi zeigt gegenüber den anderen Verfahren Vorteile in der Validität. Die Differenzen zu Shapley Value, Kruskal Importance und Kruskal 2.0 sind aber nur marginal. Shapley Value und Genizi haben allerdings einen gravierenden Nachteil: Aufgrund des listenweisen Ausschlusses können diese Verfahren schlecht mit Daten umgehen, die fehlende Werte aufweisen. Die Shapley Value-Analyse ist außerdem unter Effizienzkriterien äußerst problematisch: Die Laufzeit steigt hier exponentiell mit der Anzahl an Treibervariablen an, so dass das Verfahren bereits ungeeignet ist, wenn mehr als 20 bis 25 Treiber analysiert werden. Da dies jedoch in den allermeisten Anwendungen

der Marktforschungspraxis der Fall ist, ist dieses Verfahren überwiegend nicht einsetzbar. In der Praxis erweist sich also die Kruskal-Methode als das brauchbarste Verfahren zur Treiberanalyse: Sie ist valide, in unterschiedlichen Datenkonstellationen einsetzbar und effizient. Es ist abhängig von der inhaltlichen Zielsetzung, ob auf Kruskal Importance, oder Kruskal 2.0 zurückgegriffen werden sollte. Wenn diese Verfahren in schwierigen Datenstrukturen (geringe Fallzahl und geringes R²) weniger sinnvoll eingesetzt werden können, eignet sich die Kruskal Light-Variante besonders gut. Egal in welcher Datenkonstellation, eine der Kruskal-Varianten wird sich als *fitteste* Verfahren erweisen. ◀

► Literatur

Bocken, Pierre: Key Driver Analysis – comparisons of Relative Importance Measures in Linear regression, in: Research International Marketing Science, TNS 2007.

Breiman, Leo: Random Forests, in: Machine Learning, 45/2001, S. 5-32.

Genizi, Abraham: Decomposition of R² in Multiple Regression with Correlated Regressors, in: Statistica Sinica, 3/1993, S. 407-420.

Grömping, Ulrike: Relative Importance for Linear Regression in R: The Package relaimpo, in: Journal of Statistical Software, 17/1 -2006, S. 1-27.

Johnson, Jeff W.: A Heuristic Method for Estimating the Relative Weight of Predictor Variables in Multiple Regression, in: Multivariate Behavioral Research, 35/1-2006, S. 1-19.

Kruskal, William: Relative Importance by Averaging Over Orderings, in: The American Statistician, 41/1-1987, S. 6-10.

Liaw, Andy; Wiener, Matthew: Classification and Regression by Random Forest, in: R News, 2/3-2002, S. 18-22.

Lindeman, Richard H.; Merenda, Peter F.; Gold, Ruth Z.: Introduction to Bivariate and Multivariate Analysis. Glenview, IL 1980.

Magidson, Jay: Correlated Component Regression: A Prediction/Classification Methodology for Possibly Many Features. 2010: <http://statisticalinnovations.com/technicalsupport/CCR.AMSTAT.pdf>, S. 1-19.

Tonidandel, Scott; LeBreton, James M.: Relative Importance Analysis: A Useful Supplement to Regression Analysis, in: Journal of Business and Psychology, 26/1-2011, S. 1-9.



Der Artikel hat Ihnen gefallen und Sie wollen mehr lesen?

planung & analyse ist eine der führenden Fachzeitschriften für Marktforschung und Marketing und veröffentlicht **sechs Mal im Jahr** praxisorientierte, aktuelle und fundierte Fachbeiträge zu Projekten, Konzepten und Methoden der Markt- und Marketingforschung. Ergänzt wird das Spektrum durch zwei Sonderhefte.

Einzelartikel bei planung & analyse

Bestellen Sie diesen und weitere Artikel in druckfähiger Version für nur **EUR 11,50/Artikel**

Abonnement von planung & analyse

Sie haben die Wahl:

- Testen Sie planung & analyse print inkl. digital im **Probeabonnement** (Inland):
3 Hefte für **EUR 95,00** (inkl. Versand und MwSt.)
- Abonnieren Sie planung & analyse print inkl. digital direkt im **Jahresabonnement** (Inland):
8 Hefte (6x reguläre Ausgaben, 2x Sonderhefte) **EUR 219,00** (inkl. Versand und MwSt.)
- Abonnieren Sie planung & analyse nur digital direkt im **Jahresabonnement** (Inland):
8 Hefte (6x reguläre Ausgaben, 2x Sonderhefte) **EUR 75,00** (inkl. Versand und MwSt.)

Alle weiteren Informationen zu Bezugsbedingungen im In- und Ausland, Inhalte & Abstracts der bereits erschienenen Ausgaben sowie Themen und Termine 2015 finden Sie unter

www.planung-analyse.de.

Fragen oder Anregungen? Ihr direkter Kontakt zu planung & analyse:

+49 (0)69 7595 2014; info@planung-analyse.de

**planung
&analyse**