

Tobias Aubele

Hands-on: Predictive Analytics mit Excel

Aktuelle Analysesysteme ermöglichen einen sehr guten Blick in die Vergangenheit. Nahezu jeder Klick kann ermittelt, jeder Seitenaufruf dokumentiert und Umsatz centgenau gemessen werden. Eine Herausforderung stellt immer die Betrachtung der Zukunft dar. Um die Zukunft etwas genauer darstellen zu können, stehen viele Prognoseinstrumente, Modelle und Programme zur Verfügung. Ein sehr wirkungsvolles Programm ist dabei Excel, welches durch Funktionen einen validen Blick in die Zukunft ermöglicht. Die rückwärts gerichtete deskriptive Analyse wird damit zur zeitlich vorwärts gerichteten, vorhersagenden Analyse. Tobias Aubele zeigt, wie man auch mit kleineren Budgets hohen Nutzen aus dem richtigen Umgang mit Daten ziehen kann, wie das im Detail funktioniert und dass sich eine intensivere Beschäftigung mit dem Thema durchaus lohnt.

Welcher Umsatz wird nächsten Monat erzielt? Welche potenziellen Kunden lohnt es, verstärkt zu kontaktieren? Zahlt der Kunde die Rechnung oder kann er nur per Vorkasse beliefert werden? Diese und ähnliche Fragen stellen sich Händler permanent. Eine Antwort abseits vom Bauchgefühl liefert eine statistisch valide Prognose. Diese wird im Zeitverlauf immer zuverlässiger, wenn die Abweichungen anschließend analysiert werden und darauf fußend eine permanente Überprüfung bzw. Modifizierung des zugrunde liegenden Prognosemodells stattfindet. Den statistischen Modellen ist gemein, dass basierend auf historischen Daten nach Einflussfaktoren bzw. Ähnlichkeiten gesucht wird, welche die Zukunft möglichst exakt vorhersagen. Google kann bspw. eine regionale Grippewelle oder Denguefieber aufgrund bereits ermittelter Abhängigkeiten zwischen Suchanfragen und Ereignissen gut prognostizieren (siehe www.google.org/flutrends/).

Korrelationen – Wechselwirkungen als Prognosehelfer

Die Suche nach Abhängigkeiten von Ereignissen steht in einem engen Kontext mit dem Zusammenhang von Merkmalen. Ein Zusammenhang wird mit statistischen Maßen wie bspw. dem Korrelationskoeffizienten bewertet. Dieser hat einen Wert von -1 bis 1 und besagt, ob sich zwei Merk-

male positiv (1) bzw. negativ (-1) zueinander verhalten oder überhaupt in keinem Zusammenhang stehen (0). Google correlate (www.google.com/trends/correlate) zeigt bspw. grafisch an, wie stark Suchphrasen in einem Zeitraum miteinander in Beziehung stehen (siehe Abb. 1). So steht der zeitliche Verlauf der Suche nach einem Kinderwagen in einem starken Zusammenhang mit den Suchphrasen Reihenhaus ($r = 0,8208$) oder 30. Geburtstag ($r = 0,8229$). Eigene Werte können entweder über die Zeichnen-Funktion („Search by Drawing“) oder direkt als Werte („Enter your own data“) eingegeben werden. Aus diesen Eingaben berechnet Google correlate entsprechende Suchphrasen, welche einen vergleichbaren Suchverlauf haben. Dies kann ggf. für Kooperationen oder Komplementärprodukte sehr sinnvoll sein. Über die Möglichkeit, die Suchphrasen zeitlich (wochen- bzw. monatsweise) zu verschieben („Shift series“) können vor- bzw. nachgelagerte Zusammenhänge offenbart werden.

Vorsicht: Korrelation bedeutet niemals Kausalität. Hierzu ein bizarres Beispiel: Der Marktanteil des Internet Explorers in den USA geht relativ gesehen in ähnlichem Maße zurück wie die Mordrate. Zwischen den beiden Merkmalen besteht damit eine große Korrelation. Dies heißt jedoch ausdrücklich nicht: Sollte die Nutzung des Internet Explorers eingestellt werden, geschehen keine

DER AUTOR



Tobias Aubele ist wissenschaftlicher Mitarbeiter und Dozent im Studiengang E-Commerce an der Hochschule Würzburg-Schweinfurt sowie Doktorand im Bereich Konsumpsychologie an der University of Gloucestershire.

INFO

Der Autor stellt wie beim letzten Mal wieder eine XLS-Beispieldatei zum parallelen Ausprobieren während des Lesens zur Verfügung, in der das im Bericht beschriebene Makro ebenfalls schon integriert ist. Bitte beachten Sie, dass für eine reibungslose Funktion das SEO-Tool Add-In aus der letzten Ausgabe bereits in Excel integriert sein muss. Die Datei kann unter www.websiteboosting.com/xls2 kostenlos heruntergeladen werden. Bitte beachten Sie weiterhin, dass der Autor leider keinen Support für Excel-Probleme anbieten kann!

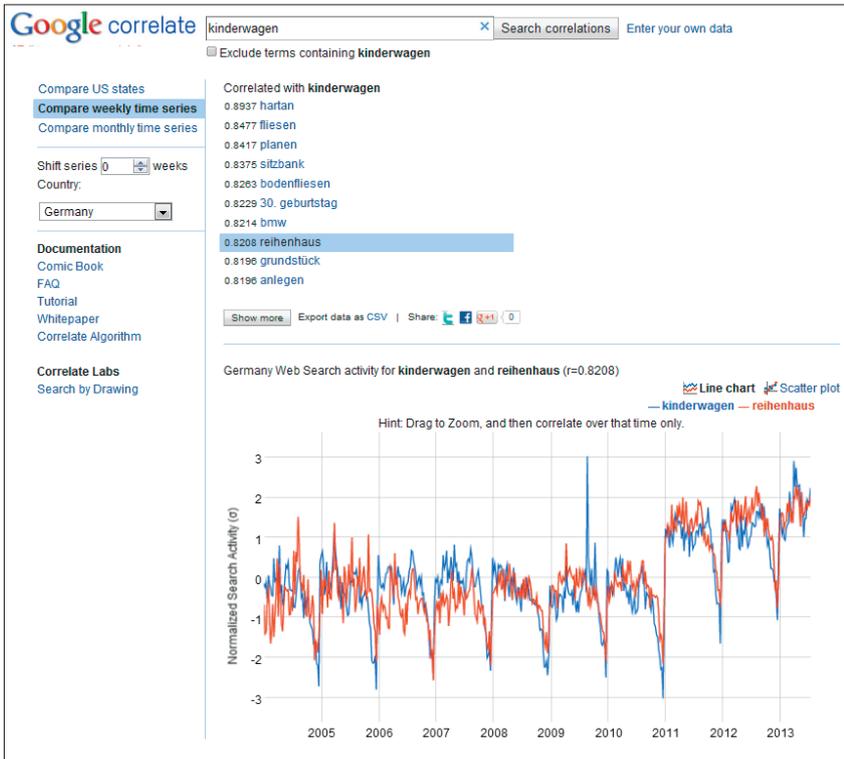


Abb. 1: Google correlate eines Suchbegriffes ohne Zeitverschiebung

Morde mehr und umgekehrt (zu den Daten dafür siehe <http://einfach.st/iemorde>).

Exploration möglicher Einflussfaktoren

Korrelationen helfen, Trends zu erkennen und Prognosen für die Zukunft zu beschreiben. Gern wird von Webseitenbetreibern der Einfluss des Wetters auf die wechselnden Besucher- und Umsatzzahlen aufgeführt. Ob dies für die eigene Webseite zutrifft, kann durch die Kombination von Wetterdaten und Besucherzahlen der jeweiligen Orte verifiziert werden. Sehr detaillierte Wetterinformationen von über 40.000 Stationen können kostenlos über die API von <http://openweathermap.org> abgefragt werden. Alternativ stellt Yahoo ebenfalls per API Wetterinformationen zur Verfügung (<http://developer.yahoo.com/weather>). Mithilfe des Add-ins SeoTools (ausführlicher Bericht in der Website Boosting Ausgabe 20) können über die Funktion

```
=XPathOnUrl
```

aktuelle Wetterinformation oder detaillierte Wettervorhersagen für die

nächsten Tage abgefragt werden (Abb. 2). Historische Daten je Station auf Stundenbasis sind ebenfalls über die API abrufbar.

Mittels der Excel-Funktion =KORREL kann festgestellt werden, ob Werte miteinander korrelieren, d. h. in einem Zusammenhang stehen. Im fiktiven Beispiel in Abb. 3 wurden die Luftfeuchtigkeit in einer Stadt sowie die Webseitenbesucher aus dieser Stadt gemessen. Je stärker die Luftfeuchtigkeit war (100 % = Regen), desto mehr Besucher verzeichnete die Webseite. Die Korrelation beträgt 0,95 und repräsentiert diesen starken Zusammenhang. Über die Summe der einzelnen Regionen kann damit die erwartete Gesamtbesucherzahl der Webseite prognostiziert werden.

	A	B	C	D	E	F	G	H	I
1	Einfluss des Wetters auf den Umsatz								
2	http://api.openweathermap.org/data/2.5/forecast/daily?q=Konstanz&mode=xml&units=metric&cnt=7&lang=de								
3	http://api.openweathermap.org/data/2.5/weather?q=Konstanz&mode=xml&units=metric&lang=de								
4	Abfrage aktueller Wetterdaten:			Konstanz					
5	Temperatur	8.68							
6		Min	6.67						
7		Max	11.11						
8	Luftdruck	923 hPa							
9	Luftfeuchtigkeit	83 %							
10	Sonnenuntergang	2013-11-01T16:05:17							
11	Wolken	24	ein paar Wolken						

Abb. 2: Aktuelle Wetterdaten - kostenlos in Excel ziehen mittels XPathOnUrl

Prognosen mit linearen Regressionen

Predictive Analytics bedeutet, dass die Zukunft basierend auf Modellen bzw. eines Modells vorhergesagt wird. Diese Modelle werden mittels Analyse historischer und aktueller Daten aufgestellt und kontinuierlich weiterentwickelt. Das heißt, das Modell korrigiert bzw. optimiert sich im Idealfall durch die kontinuierliche Berücksichtigung weiterer Daten. Eine einfache Möglichkeit, die Zukunft zu prognostizieren, sind Regressionen. Regressionen versuchen, abhängige Variable (bspw. Besucher) durch unabhängige Variablen (bspw. Luftfeuchtigkeit) abzubilden. Dies spiegelt sich in einer Gleichung der Form $y = mx + c$ wider. Über die Excel-Funktionen

```
=ACHSENABSCHNITT
```

(Y-Werte; X-Werte) sowie

```
=STEIGUNG(Y-Werte;X-Werte)
```

kann c bzw. m berechnet werden. Im Beispiel in Abb. 3 können Besucher über die Funktion $y = 42,07x + 545,64$ prognostiziert werden. Das heißt, bei einer vorhergesagten Luftfeuchtigkeit von 83 % (= x) werden 4.038 Besucher (= y) erwartet. Alternativ können mittels der Funktion

```
=TREND(historische Y-Werte; historische X-Werte; neue X-Werte)
```

die Besucher direkt prognostiziert werden. Grafisch ist die Ermittlung der Regressionsgeraden durch die Funktion „Trendlinie“ im Untermenü Layout der Diagrammtools lösbar (Ergebnis siehe Abb. 3).

Berücksichtigung mehrere Einflüsse – multiple Regressionen

Meist ist nicht nur ein Faktor vorhanden, der ein Ereignis beeinflusst, sondern es ist die Kombination unterschiedlicher Einflüsse. TREND kann ebenfalls für die Prognose unter Bezugnahme mehrerer Faktoren herangezogen werden (multiple Regression). Soll nicht nur die Luftfeuchtigkeit, sondern bspw. auch die Temperatur als unabhängiger Faktor in die Analyse einfließen, so liefert die Funktion TREND die entsprechenden Prognosewerte (siehe Abb. 4). Als Alternative können über die Matrix-Funktion RGP sowohl die Koeffizienten, die Konstante als auch deren Güte gleichzeitig berechnet werden. Bei Matrix-Funktionen muss die Formeleingabe in Excel immer mit Umschalt- + STRG + Eingabetaste abgeschlossen werden. Sie sind durch die eingabebedingt erzeugte Klammerung, bspw. {=RGP(Attribute)}, erkennbar. Matrix-Funktionen sind dadurch gekennzeichnet, dass sie mehrere Rechenoperationen simultan durchführen und deren Lösung, ein sogenanntes Array, mehrere Ergebnisse enthält. Sofern vor der Formeleingabe mehrere Zellen markiert wurden, liefert dieser Funktionstyp die entsprechenden Kennzahlen bzw. Ergebnisse unmittelbar zurück (siehe Abb. 4).

Die Abb. 4 zeigt, dass in diesem Beispieldatensatz die Temperatur stark negativ korreliert.

Das zeigt, je kälter das Wetter ist und je höher die Luftfeuchtigkeit, desto mehr Besucher sind vorrauschichtlich auf der Webseite.

Die Hinzunahme eines weiteren Faktors verbesserte die Prognose. Dadurch konnte ein weiterer Einflussfaktor entdeckt und die zukünftigen Prognosen valider gestaltet werden.

Analyse-Add-in für weitreichende statistische Tests

Excel stellt für die Datenanalyse bereits zwei umfassende Add-ins zur Ver-

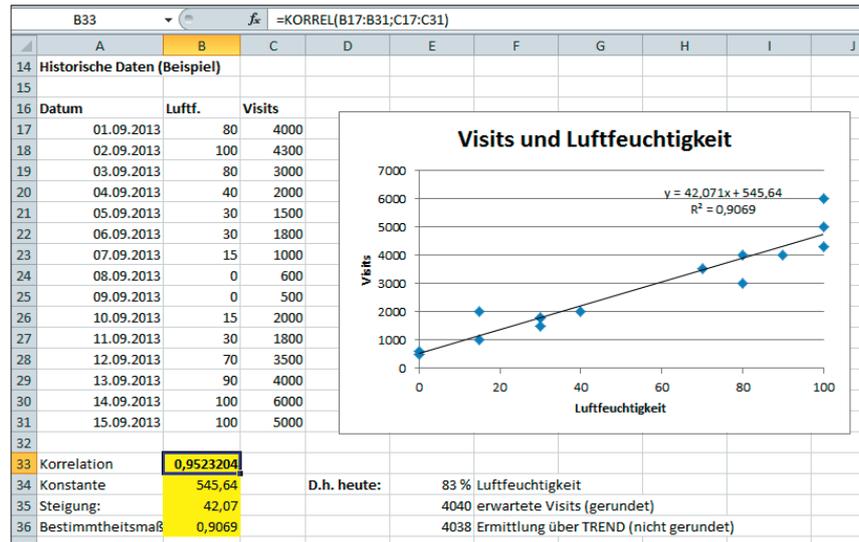


Abb. 3: Luftfeuchtigkeit und Visits: Beispiel einer Korrelation und Regressionsrechnung in Excel

L40		={RGP(N17:N31;L17:M31)}					
	K	L	M	N	O	P	Q
14	Historische Daten (Beispiel)						
15							
16	Datum	Luftf.	Temp.	Visits	Prognose	via TREND	Residual
17	01.09.2013	80	12	4000	4166	4166	-166
18	02.09.2013	100	10	4300	4679	4679	-379
19	03.09.2013	80	17	3000	3116	3116	-116
20	04.09.2013	40	22	2000	1880	1880	120
21	05.09.2013	30	23	1500	1623	1623	-123
22	06.09.2013	30	23	1800	1623	1623	177
23	07.09.2013	15	25	1000	1133	1133	-133
24	08.09.2013	0	27	600	643	643	-43
25	09.09.2013	0	26	500	853	853	-353
26	10.09.2013	15	24	2000	1343	1343	657
27	11.09.2013	30	21	1800	2043	2043	-243
28	12.09.2013	70	17	3500	3069	3069	431
29	13.09.2013	90	10	4000	4632	4632	-632
30	14.09.2013	100	7	6000	5309	5309	691
31	15.09.2013	100	9	5000	4889	4889	111
32							
33							
34	Korrelation	Visits	D.h. heute:		83 % Luftfeuchtigkeit		
35	Luftfeucht.	0,952	8,68		8,68 Temperatur		
36	Temp.	-0,974			4492 erwartete Visits		
37	Progn.	0,975					
38							
39	Koeffizient	Temp	Luft	Konstante			
40		-209,9	4,7	6309,5			

Abb. 4: Multiple Regression und Korrelation

fügung. Über den Menüpunkt „Datei – Optionen – Add-ins“ und Klick auf den Button „Gehe zu“ werden alle aktiven und inaktiven Add-ins angezeigt. Sofern „Analyse-Funktionen“ und „Solver“ noch nicht aktiviert sind, kann dies an dieser

Stelle durchgeführt werden. Ggf. müssen die entsprechenden Add-ins dadurch nachinstalliert werden. Die neuen Analysefunktionen stehen anschließend als Untermenü im Register Daten bereit. Neben Regressionen kann damit eine

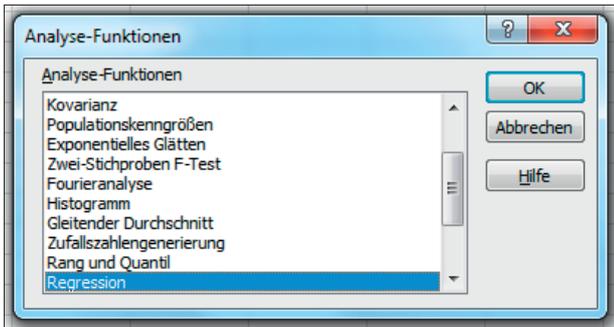


Abb. 5: Auszug der Analyse-Funktionen in Excel

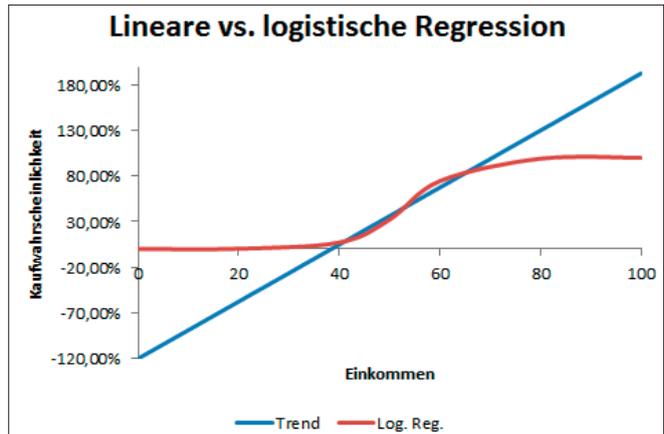


Abb. 6: Vergleich lineare und logistische Regression

Vielzahl weiterer statistischer Berechnungen (u. a. gleitende Durchschnitte zur Abdämpfung von Ausreißern in den Werten) durchgeführt werden. Ähnlich den Statistikprogrammen R und SPSS liefern die Berechnungen neben den Ergebnissen wichtige Aussagen zur Güte der Ergebnisse. Die Güte besagt, ob die berechneten Koeffizienten bzw. Konstanten statistisch valide sind oder ob es sich um Zufall handelt. Entsprechende Gütemaße werden sowohl auf Modellebene (bspw. hoher F-Wert, F krit möglichst null) oder auf Variablenebene (hoher |t|-Wert und zugehöriger p-Wert möglichst null) berechnet. Damit kann das Modell kontinuierlich optimiert und der Zufall ausgeschlossen werden.

Logistische Regression – Möglichkeit der Ja/Nein-Prognose

Fragestellungen wie die Wahrscheinlichkeit eines Kaufes oder der Bezahlung der Rechnung, d. h. Ja/Nein-Antworten, sind durch lineare Regression unzureichend abbildbar. Lineare Regressionen hätten nicht nur mögliche Wahrscheinlichkeiten von über 100 % bzw. unter 0 %, sondern unterstellen einen kontinuierlichen gleichmäßigen Zusammenhang. Damit hätte bspw. eine Änderung des Jahreseinkommens von 50.000 auf 55.000 € eine identische Auswirkung auf die Kaufwahrscheinlichkeit eines Gutes wie die Änderung des Einkommens von 95.000 auf 100.000 €. Es ist offensichtlich, dass dies voraussichtlich nicht der Realität entspricht. Diesen Nachteil können statistische Verfahren wie bspw.

eine logistische Regression lösen, da diese sich einer dichotomen Variablen (0 bzw. 1 für Nichtkauf bzw. Kauf) langsam annähert.

Bei einer logistischen Regression wird ein Modell in mehreren Iterationen optimiert. Vereinfacht gesagt werden die einzelnen Koeffizienten bzw. die Konstante geschätzt und anschließend die Passung des Modells mit den zugrunde liegenden historischen Daten ermittelt. Die Güte des Modells wird durch die Summe des Log-Likelihood (Logarithmus der Prognosewahrscheinlichkeit für Kauf/Nichtkauf) gemessen. Dies geschieht so lange, bis das Gütekriterium nicht mehr verbessert werden kann. Diese Simulation bzw. erweiterte Zielwertsuche übernimmt der Solver. Der Solver kann unter Beachtung von Bedingungen mittels diverser Lösungsalgorithmen einen Zielwert optimieren, in dem die dem Modell zugrundeliegenden Variablen permanent modifiziert werden („Was-wäre-wenn“-Analysen).

Die Besonderheit bei logistischen Regressionen ist die Nutzung von Odds (Chancen). Aus diesen Odds lassen sich die Wahrscheinlichkeiten ermitteln (Chance 1:1 entspricht bspw. 50 %). Der natürliche Logarithmus aus diesen Odds ist der Logit. In der logistischen Regression wird der Logit durch die unabhängigen Variablen sowie die Konstante bestimmt, d. h., im folgenden Beispiel bestimmen Einkommen, Alter und die Region die Kaufwahrscheinlichkeit (siehe Abb. 7).

Die Funktion
=EXP(Zahl)

ermittelt aus dem Logit die Odds. Rechnerisch verbirgt sich hinter der Funktion die Potenz der eulerschen Zahl und damit die Umkehrung des natürlichen Logarithmus. Mittels der Formel
=Odds/(1+Odds)

ergibt sich die Kaufwahrscheinlichkeit basierend auf diesem Modell. Das Modell ist präzise, sofern es einen Kauf oder einen Nichtkauf möglichst präzise vorhersagt, d. h., die Chance eines Nichtkaufs ist damit 1 – Kaufwahrscheinlichkeit (in Abb. 7: =WENN(A6=1;G6;1-G6)). Das Modell ist dann optimal, wenn diese Kauf-/Nichtkaufwahrscheinlichkeit 1 ist. Die Güte des gesamten Modells ergibt sich durch das Produkt der einzelnen Wahrscheinlichkeiten. Da dieses Produkt bei einer großen Anzahl an historischen Daten sehr klein wird (bspw. 0,2*0,2*0,2 = 0,008), empfiehlt es sich, diese Wahrscheinlichkeit mit der Funktion
= LN(Zahl)

in den natürlichen Logarithmus umzuwandeln. Dies hat den Vorteil, dass diese logarithmierte Wahrscheinlichkeit summiert werden kann.

Solver – umfassende „Was-wäre-wenn“-Analysen

Im Modell wurde die Konstante zum Start auf den Faktor 1 und die einzelnen Koeffizienten auf 0 gesetzt. Es müssen nun die Werte gefunden werden, die zum Modell mit der höchsten Güte führen. Dazu könnten manuell Werte eingesetzt und diese so lange modifiziert werden, bis die Summe des Log-Likelihood das Ideal von 0 hätte. Diese Aufgabe übernimmt der Solver, welcher im Menüpunkt

Daten aufgerufen werden kann (siehe Abb. 8). Das Ziel der Simulation durch den Solver ist das Erreichen des Maximums der Zelle H2 (Summe Log-Likelihood; in Abb. 7: =SUMME(I:I)). Dabei sollen und können die Variablen A3 bis D3 modifiziert werden. Weitere Nebenbedingungen sind nicht zu beachten. Damit die logistische Regression gelöst werden kann, muss das vorgelegte Feld „Nicht eingeschränkte Variablen ...“ deaktiviert werden (Einstellung unter „Optionen“ neben der Wahl der Lösungsmethode). Ein Klick auf „Lösen“ leitet die Simulation/Berechnung ein.

Tipp: Der Solver ist ein hervorragendes Instrument, um Datensimulationen vorzunehmen. Im Internet gibt es hierzu umfassende Tutorials und Videos.

Basierend auf den historischen Daten, die das aktuelle Modell determinieren, ergibt sich der Logit aus $0,26 \cdot \text{Einkommen} + 0,27 \cdot \text{Alter} - 2,32 \cdot \text{Region}$ (siehe Abb. 9). Über die beschriebene Umwandlung kann damit für unbekannte Daten eine Wahrscheinlichkeit prognostiziert werden. Hat ein neuer Kunde ein Einkommen von 45.500 €, ist 63 Jahre alt und wohnt in der Region 1, hat er eine Kaufwahrscheinlichkeit von 83,9 %. Diese Daten könnten herangezogen werden, wenn ein Unternehmen vor der Entscheidung steht, ob der Kunde bei knappen Ressourcen weiter bearbeitet werden soll oder nicht. Sind in der Datenbank 5.000 Kunden und es sollen nur die 1.000 Kunden per Post angeschrieben werden, welche wahrscheinlich konvertieren, so kann ein solches Modell die statistisch relevantesten Personen offenbaren.

Wie immer gilt: Je größer die zugrunde liegende Datenbasis, desto besser ist das Modell und desto genauer sind die Prognosen. Weiterhin ist empfehlenswert, dass dem Modell kontinuierlich die tatsächlichen Daten zur Verfügung gestellt werden und es damit weiterentwickelt bzw. angepasst wird.

E6 fx =+\$A\$3+B6*\$B\$3+C6*\$C\$3+D6*\$D\$3										
	A	B	C	D	E	F	G	H	I	
1		Koeffizient						Summe		
2	Konstante	Einkommen	Alter	Region				Log Likelihood:	-34,27742	
3	1,00000	0,00000	0,00000	0,00000						
4										
5	Kauf (0=N;1=J)	Einkommen (T€)	Alter	Region	Logit	Chance (Odd)	Kauf- wahrscheinl.	Kauf/Nicht- kaufwahrs.	Log Likelihood	
6	0	39,9	52	3	1,0000	2,7183	0,7311	0,2689	-1,3133	
7	0	41,4	59	3	1,0000	2,7183	0,7311	0,2689	-1,3133	
8	0	42,8	53	2	1,0000	2,7183	0,7311	0,2689	-1,3133	
9	0	42,3	49	3	1,0000	2,7183	0,7311	0,2689	-1,3133	
10	0	44,8	45	1	1,0000	2,7183	0,7311	0,2689	-1,3133	
11	0	43,4	57	1	1,0000	2,7183	0,7311	0,2689	-1,3133	
12	0	43,8	54	2	1,0000	2,7183	0,7311	0,2689	-1,3133	
13	0	42,4	57	2	1,0000	2,7183	0,7311	0,2689	-1,3133	
14	0	44,2	52	2	1,0000	2,7183	0,7311	0,2689	-1,3133	
15	0	31,9	58	3	1,0000	2,7183	0,7311	0,2689	-1,3133	
16	1	45,1	62	2	1,0000	2,7183	0,7311	0,7311	-0,3133	

Abb. 7: Logistische Regression

The screenshot shows the Excel Solver interface. At the top, the Solver Parameters dialog box is open. The 'Ziel festlegen' (Set Objective) field contains '\$H\$2'. The 'Durch Ändern von Variablenzellen' (By Changing Variable Cells) field contains '\$A\$3:\$D\$3'. The 'Lösungsmethode' (Select a GRG Nonlinear engine for Solver Problems that are smooth nonlinear) is selected. The 'Nicht eingeschränkte Variablen als nicht-negativ festlegen' (Make Unconstrained Variables Non-Negative) checkbox is unchecked. The 'Lösen' (Solve) button is highlighted in blue.

Abb. 8: Solver zur Optimierung/Simulation

VBA – freie Erweiterung von Excel durch Programmierung

Mittels VBA (Visual Basic for Applications) können sehr viele Schritte in Excel automatisiert werden. Komplett

eigene Funktionen, Prognoseverfahren mit den entsprechenden notwendigen Abläufen sind damit denkbar. Über die Tastenkombination Alt + F11 ist die Oberfläche von VBA erreichbar. Das Beispiel in Abb. 12 zeigt den Einsatz

L10											
A	B	C	D	E	F	G	H	I	J	K	L
=EXP(\$A\$3+L7*\$B\$3+L8*\$C\$3+L9*\$D\$3)/(1+EXP(\$A\$3+L7*\$B\$3+L8*\$C\$3+L9*\$D\$3))											
1	Koeffizient			Summe							
2	Konstante	Einkommen	Alter	Region	Log Likelihood: -10,25291						
3	-24,97300	0,25557	0,27485	-2,32056							
4											
5	Kauf (0=N;1=J)	Einkommen (T€)	Alter	Region	Logit	Chance (Odd)	Kauf-wahrscheinl.	Kauf/Nicht-kaufwahrs.	Log Likelihood		
6	0	39,9	52	3	-7,4450	0,0006	0,0006	0,9994	-0,0006		
7	0	41,4	59	3	-5,1377	0,0059	0,0058	0,9942	-0,0059	Einkommen	45,5
8	0	42,8	53	2	-4,1084	0,0164	0,0162	0,9838	-0,0163	Alter	63
9	0	42,3	49	3	-7,6562	0,0005	0,0005	0,9995	-0,0005	Region	1
10	0	44,8	45	1	-3,4756	0,0309	0,0300	0,9700	-0,0305	Kauf	83,90%
11	0	43,4	57	1	-0,5351	0,5856	0,3693	0,6307	-0,4610		
12	0	43,8	54	2	-3,5780	0,0279	0,0272	0,9728	-0,0275		
13	0	42,4	57	2	-3,1113	0,0445	0,0426	0,9574	-0,0436		
14	0	44,2	52	2	-4,0255	0,0179	0,0175	0,9825	-0,0177		
15	0	31,9	58	3	-7,8405	0,0004	0,0004	0,9996	-0,0004		
16	1	45,1	62	2	-1,0469	0,3510	0,2598	0,2598	-1,3478		
17	0	46,0	59	2	-1,6415	0,1937	0,1623	0,8377	-0,1771		
18	0	37,3	50	3	-8,6592	0,0002	0,0002	0,9998	-0,0002		
19	0	50,1	59	3	-2,9142	0,0542	0,0515	0,9485	-0,0528		
20	1	51,1	62	1	2,8070	16,5609	0,9431	0,9431	-0,0586		

Abb. 9: Ergebnis einer logistischen Regression und deren Prognosemöglichkeit

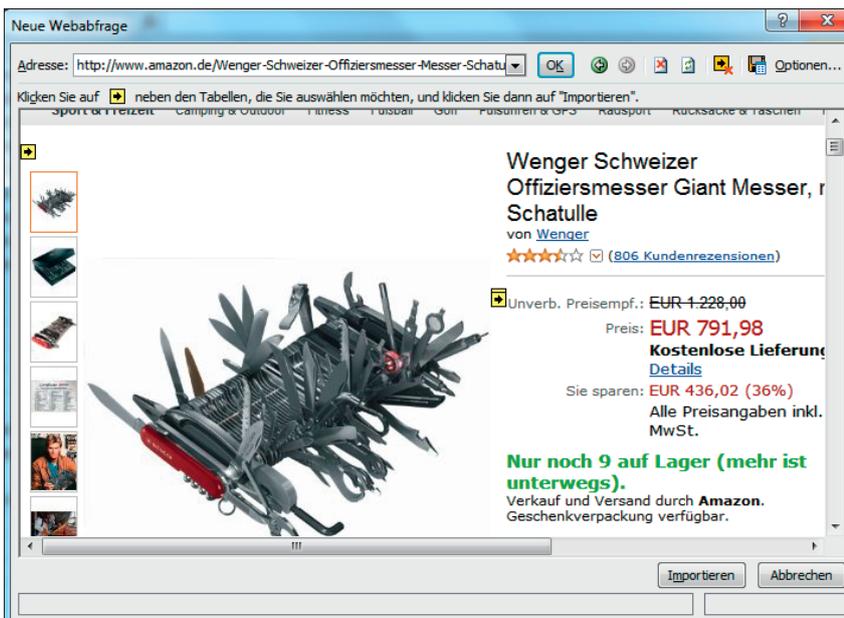


Abb. 10: Definition der zu importierenden Bereiche - gezeigt am Beispiel Amazon

Datei			Start			Einfügen			Seitenlayout			Formeln			Daten			Überprüfen			Ansicht			Entwicklertools			SeoT		
A3 = Produktgewicht inkl. Verpackung: 3,2 Kg																													
Produktinformation																													
Größe und/oder Gewicht: 30 x 30 x 20 cm ; 1,4 Kg																													
Produktgewicht inkl. Verpackung: 3,2 Kg																													
Modellnummer: 19201																													
EAN/UPC: 7611640003646																													
ASIN: B000R0JDSI																													
Im Angebot von Amazon.de seit: 4. April 2007																													
Durchschnittliche Kundenbewertung: 3.7 von 5 Sternen Alle Rezensionen anzeigen (804 Kundenrezensionen)																													
804 Rezensionen																													
5 Sterne: -296																													
4 Sterne: -239																													
3 Sterne: -127																													
2 Sterne: -46																													
1 Sterne: -96																													
Alle 804 Kundenrezensionen anzeigen...																													
Amazon Bestseller-Rang: Nr. 84.772 in Sport & Freizeit (Siehe Top 100 in Sport & Freizeit)																													
Möchten Sie Feedback zu Bildern geben oder uns über einen günstigeren Preis informieren?																													

Abb. 11: Ergebnis des Datenimports aus dem Web mit Aktualisierungsfunktion

von VBA, um automatisiert Produktinformationen von Amazon zu laden und damit Prognosen über Abverkauf von Nischenprodukten geben zu können. Sofern bspw. die Überlegung besteht, das mit über 800 Kundenrezensionen beliebte Produkt des Universal-Schweizer-Messers in das Sortiment aufzunehmen, kann dessen Verkaufsrate über den Amazon-Beststeller-Rang prognostiziert werden (siehe Abb. 10).

Damit dies nicht manuell stattfinden muss, kann eine einfache VBA-Prozedur eingesetzt werden. Über die Excel-Funktion

„Externe Daten abrufen – Aus dem Web“

im Menü Daten kann ein Bereich bzw. eine ganze Seite in Excel importiert werden. Hierzu müssen die gewünschten Bereiche mittels der gelben Pfeiltasten aktiviert werden (Abb. 10). Sofern dieser Import bereits durchgeführt wurde (bspw. mit Start des Imports in Zelle A1), können die Werte über die Schaltfläche „aktualisieren“ im Menü Daten oder über die rechte Maustaste immer wieder aktualisiert werden. Hierzu wird eine Verbindung zur Zielseite geöffnet und die entsprechenden Bereiche werden neu importiert (Abb. 11).

ACHTUNG!

Diese(r) Prozedur/Ablauf soll nur beispielhaft zeigen, wie einfach in Excel Daten von Webseiten extrahiert und analysiert werden können. Excel beansprucht für die Abfrage der Daten vergleichsweise viel Zeit, wodurch im Normalfall nicht viele Anfragen in einer kurzen Zeitspanne durchgeführt werden können. Dennoch kann eine zu hohe Beanspruchung ggf. nachteilige Auswirkungen auf den Webseitenbetreiber haben und dessen Dienste belasten. Wie immer beim unerwünschten automatisierten Abfragen von Daten, könnte das bei intensiver Nutzung z. B. als eine sog. Denial of Service Attacke (DoS) interpretiert werden.

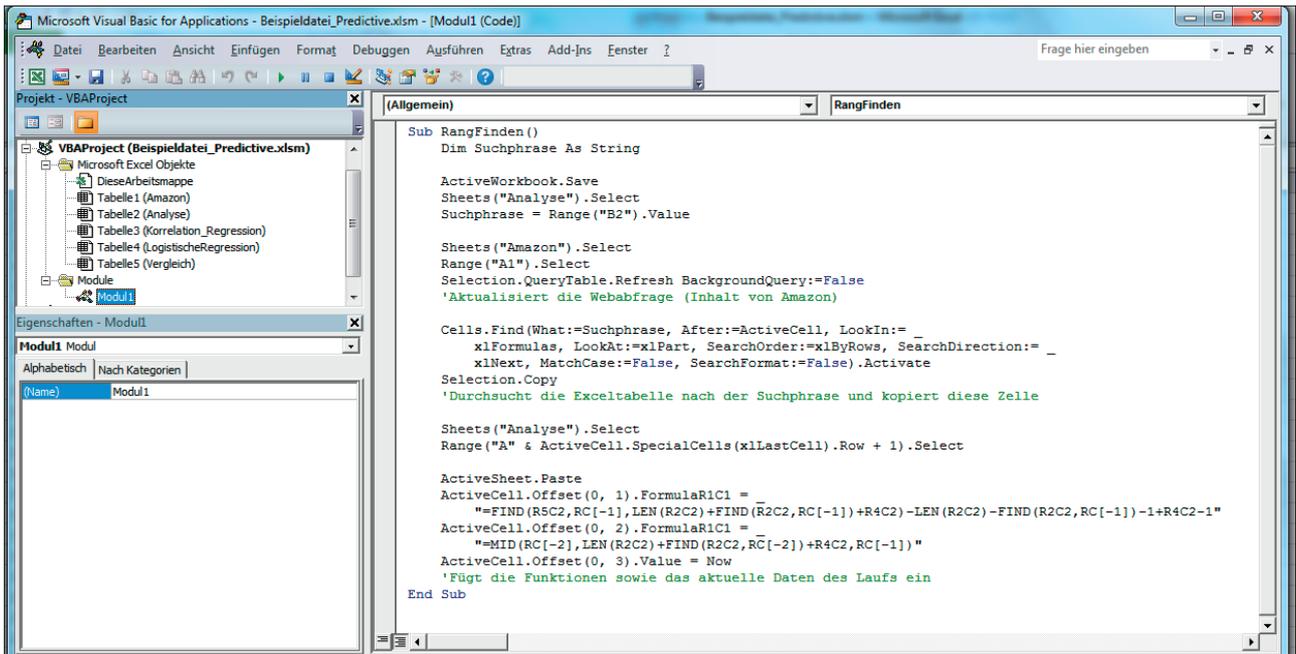


Abb. 12: Beispiel einer VBA-Prozedur, welche Daten aktualisiert und Formeln einfügt

B9 =FINDEN(\$B\$5;A9;LÄNGE(\$B\$2)+FINDEN(\$B\$2;A9)+\$B\$4)-LÄNGE(\$B\$2)-FINDEN(\$B\$2;A9)-1+\$B\$4-1					
A	B	C	D	E	
1	http://www.amazon.de/Wenger-Schweizer-Offiziersmesser-Messer-Schatulle/dp/B000R0JDSI/				
2	Suchphrase	Bestseller-Rang: Nr.	Prüfe Ranking		
3	Länge Suchphrase	20			
4	Unberücksichtigte Zeichen nach Suchphrase	1			
5	Zeichen nach dem auszulesenden Ergebnis	Eingabe Leerzeichen			
6					
7	Inhalt Suchphrase	Länge Ziel	Ziel	Datum	
8	Amazon Bestseller-Rang: Nr. 74.557 in Sport & Freizeit (Siehe Top 100 in S	6	74.557	30.10.2013 16:14	
9	Amazon Bestseller-Rang: Nr. 75.562 in Sport & Freizeit (Siehe Top 100 in S	6	75.562	30.10.2013 19:19	Nein
10	Amazon Bestseller-Rang: Nr. 75.540 in Sport & Freizeit (Siehe Top 100 in S	6	75.540	30.10.2013 19:56	JA
11	Amazon Bestseller-Rang: Nr. 84.772 in Sport & Freizeit (Siehe Top 100 in S	6	84.772	31.10.2013 19:53	Nein
12	Amazon Bestseller-Rang: Nr. 83.890 in Sport & Freizeit (Siehe Top 100 in S	6	83.890	31.10.2013 20:53	JA

Abb. 13: Ergebnis der automatischen Datenerhebung per VBA

Eine einmalig geschriebene VBA-Prozedur kann die Datenaktualisierung per Code aufrufen, die Position des aktuellen Verkaufsranges ermitteln, in eine Tabelle schreiben und bei einer positiven Veränderung mindestens einen Kauf prognostizieren. Dieser Code könnte über eine weitere Prozedur zeitgesteuert ablaufen und damit die Daten bspw. jede Stunde automatisch aktualisieren. Durch ein sog. Error-Handling kann gesteuert werden, was die Prozedur tun soll, wenn ein Fehler auftritt. Dies ist bei einer zeitgesteuerten Durchführung auf jeden Fall empfehlenswert.

Diese Prozedur liefert den aktuellen Rang und lässt bei einer Erhöhung auf einen Kauf zurückschließen.

Fazit: Webanalysesysteme sind sehr

gut in der Darstellung der Vergangenheit. Die Kombination dieser historischen Daten und die Nutzung von Prognosemodellen erlauben einen validen Blick in die Zukunft. RapidMiner, SPSS, SAS und weitere Anbieter helfen, Modelle zu entwickeln, zu trainieren und anschließend auf unbekannte Daten anzuwenden. Die Standardsoftware Excel muss sich hierbei nicht verstecken. Durch die Vielzahl der enthaltenen Funktionen sowie die Offenheit über Add-ins können einfache Zusammenhänge aus verschiedenen Datenquellen hergestellt werden und in einem Modell zusammenlaufen. Die gewohnte Excel-Umgebung kann damit eine gute Alternative sein, obwohl deren Umfang im Vergleich zu R, SPSS o. ä. reduziert ist.

In diesem Sinne: Verifizieren Sie Ihr Bauchgefühl und ermitteln Sie Zusammenhänge. Im Idealfall heißt es zukünftig: Zufall? – Nein danke!¶

ACHTUNG!



Wenn Sie weitergehendes Interesse an Predictive Analytics sowie deren Anwendung in Excel haben, finden Sie tiefer gehende und erläuternde Hinweise als Leseempfehlung im Buch „Predictive Analytics: Microsoft Excel“ von Conrad Carlberg, Que Publishing (290 S., englisch, in D als Taschenbuch 24, 95 €, als Kindle-Version 15,13 €).