



Im Daten-Dschungel

Wofür Big Data die Marktforschung braucht

Immer mehr, immer größer, immer komplexer. Dank Big Data verfügen Unternehmen über endlos viele Kunden-Informationen. Mit diesen Daten können sie erfassen, „Was“ der Einzelne macht, das wichtige „Warum“ bleibt aber ungeklärt. Welche Rolle die Marktforschung mithilfe von Machine Learning übernehmen kann, erläutert **Hannes Schettler**.

Big Data ist wie der Urwald – unendlich reich und vielfältig, aber ebenso schwer zu durchdringen. So wie ein Urwald über eine lange Zeit gewachsen ist, so haben sich auch viele Datenbanken in den Unternehmen über die Jahre gefüllt und liefern einen reichen Fundus für Datenanalysen. Viele Fragen, in denen Meinungen, Wünsche und Einstellungen eine Rolle spielen, können allerdings nicht beantwortet werden. Neue Möglichkeiten ergeben sich, wenn hochwertige Marktforschungsdaten mit einfließen. Marktforschung hat bekanntermaßen wenig mit einem Urwald zu tun. Studien werden

sorgsam geplant und vorbereitet, Befragte mit Bedacht ausgewählt, die Feldzeit ist begrenzt und die Daten sind einheitlich. Nichts wird dem Zufall überlassen.

Marktforschung ist nicht ersetzbar

Die Hoffnungen oder Befürchtungen, dass die klassische Marktforschung durch Data Mining der ohnehin vorhandenen Daten überflüssig gemacht werden könnte, sind kaum angebracht. In den Da-

tenbanken steht viel über das, was die Kunden oder Konsumenten tun. Was sie kaufen, was sie abonnieren, was sie reklamieren. Daraus lassen sich aber nur bedingt Rückschlüsse auf ihre Wünsche ziehen oder auf die Gründe für ihr Handeln. Letztere sind allerdings oft die relevanten Informationen für unternehmerische Entscheidungen. Eine intelligente Kombination von Daten, die bereits vorliegen mit solchen aus der Marktforschung kann das Unternehmen tatsächlich weiterbringen. Verhaltensdaten könnten zum Beispiel mit Erkenntnissen aus Studien angereichert werden. Die passenden Werkzeuge dafür finden sich im Bereich des Machine Learning.

Machine Learning: ein Fallbeispiel

Ein Schuhhändler möchte die Vorlieben seiner Kundinnen besser verstehen, um ihnen zielgerichtete Angebote machen zu können. Ein Marktforschungsinstitut wird beauftragt, eine Befragung durchzuführen. Anhand der Einstellungen zu verschiedenen Stilrichtungen und zu Schuhmode werden die Befragten mithilfe einer Clusteranalyse in fünf Gruppen eingeteilt. Da es dem Auftraggeber um eine konkrete Ansprache seiner Kundinnen geht (etwa durch ein Mailing), genügt es nicht, die Charakteristika der Gruppen zu kennen. Er möchte für jede einzelne Kundin wissen, welchen Stil sie bevorzugt. Die Kundendatenbank muss also um den Stiltyp angereichert werden.

Mit den Vorhersage-Methoden des Machine Learning kann in den Befragungsdaten nach Mustern gesucht werden, die den Stiltyp erklären. Ein derart trainiertes Modell kann im Anschluss auf die Kundinnen in der Datenbank angewendet werden, um deren Stiltyp abzuschätzen. Bedingung dafür ist ein Set an gemeinsamen Variablen. Diese Variablen dienen als Brückenpfeiler („Prädiktoren“), um den Bogen von den Befragungsdaten zu den Verhaltensdaten zu schlagen. Hier wurden das Alter, das Körpergewicht und die Schuhgröße als Prädiktoren verwendet.

Für die Untersuchung lag ein großer vollständiger Befragungsdatensatz von ungefähr 40.000 Fällen vor. Um ein realistisches Szenario zu simulieren, wurde allerdings nur ein Teil der Fälle als Befragungsdaten angesehen, der Rest als Kundendatenbank. Zunächst untersuchte man, inwieweit die Segmentierung selbst stabil ist. Dazu wurden wiederholt unterschiedliche Teildatensätze als Befragungsdaten betrachtet und für die Clusteranalyse verwendet. Die jeweils übrigen Fälle wurden anhand der Clustervariablen den Gruppen zugeordnet. Der Vergleich dieser unabhängigen Segmentierungen zeigt, wie eindeutig die Gruppeneinteilung ist. Mit Stichproben von 700 Fällen konnte im Schnitt eine 66-prozentige Übereinstimmung zwischen den Clusterlösungen gefunden werden. Bei 2.800 Fällen waren es schon 78 Prozent. In der Praxis ist dieser Stabilitätstest natürlich nicht möglich, da die Clustervariablen nur in den Befragungsdaten vorliegen, nicht aber in der Kundendatenbank. Hier soll der „wahre“ Stiltyp je Kundin dazu dienen, die Vorhersagen aufgrund der Prädiktoren zu testen und einzuordnen.

Prognosegüten im Vergleich

Im Machine Learning steht eine ganze Reihe an Verfahren zur Verfügung, um Vorhersagen zu generieren. Meist lässt sich nicht im Voraus sagen, welches Verfahren für eine Aufgabenstellung zum besten Ergebnis führt.

In der Tabelle werden die Prognosegüten der Methoden Diskriminanzanalyse, Boosted Tree, Random Forest und Support Vector Machine (SVM) verglichen (Abb. 1). Die Modelle wurden auf 700 Fällen

Prognosegüte mit Machine Learning

Diskriminanzanalyse	39
Boosted Tree	37
Random Forest	34
SVM	38

Abb. 1 Quelle: IfaD; in Prozent

trainiert. Alle Verfahren liefern ein ähnliches Ergebnis, lediglich der Random Forest schneidet etwas schlechter ab. Bei 2.800 Fällen erhöhen sich die Trefferquoten um jeweils ein bis zwei Prozentpunkte. Bei Vervierfachung der Fallzahl gibt es eine geringe Verbesserung und die Diskriminanzanalyse schneidet in diesem Vergleich am besten ab. Das sind Indizien, dass nichtlineare Zusammenhänge oder Wechselwirkungen zwischen den Prädiktoren kaum zur Erklärung des Stiltyps beitragen. Aber sind knapp 40 Prozent Trefferquote nun gut oder schlecht? 20 Prozent der Kundinnen wären in der richtigen Gruppe, wenn man sie einfach zufällig aufteilen würde (bei fünf gleich großen Clustern). Mit den Modellen verdoppelt sich dieser Anteil. Eine Grenze ist durch die Clusterstabilität gegeben. Wenn die Clusterzugehörigkeit an sich nur zu 66 Prozent (beziehungsweise 78 Prozent) eindeutig ist, kann auch kein Prognoseverfahren mehr erreichen.

Vorhersagen im realen Leben testen

In diesem Beispiel ging es darum, nur eine Variable (nämlich den Stiltyp der Kundin) abzuschätzen. In anderen Situationen kann es sinnvoll sein, die Kundendaten mit mehr Information aus den hochwertigen Marktforschungsdaten anzureichern und für weitere Analysen zu verwenden. Allerdings zeigen sich auch deutliche Grenzen. Beispielsweise ist es kaum vielversprechend, anhand weniger demographischer Merkmale zu versuchen, die Einstellung zu komplexen Fragestellungen auf Einzellebene vorherzusagen. Besteht kein inhaltlicher Zusammenhang zwischen den Prädiktoren und der prognostizierten Variable, wird das Ergebnis im Allgemeinen dürftig ausfallen. In manchen Situationen mag es genügen, wenn die Vorhersage nur wenig besser als der Zufall ist, in anderen kann das Ergebnis nutzlos sein, ist die Prognose nicht zu 99 Prozent sicher. Sofern möglich, ist es stets ratsam, die Prognosen im realen Leben auf die Probe zu stellen. Im genannten Beispiel könnte ein kleiner Teil der Kundinnen nicht nach dem vorhergesagten Stiltyp angeschrieben werden, sondern nach einer zufälligen Aufteilung. Nur wenn die Responserate in dieser Testgruppe tatsächlich niedriger ist, hat sich die Vorhersage gelohnt. ■

Mehr zum Thema **Big Data**⁴
und zur Schnellsuche Fachartikel

www.research-results.de/fachartikel



Dr. Hannes Schettler

verstärkt seit 2014 IfaD im Bereich Data Sciences. Dort arbeitet er an der Entwicklung und Programmierung multivariater Analysemethoden sowie von Data Mining-Modellen mit.

www.ifad.de

