

Psychologische Zielgruppensegmentierung durch Verknüpfung von Bestandsdaten und Marktforschungsergebnissen

Ein Projekt der psychonomics AG mit *STATISTICA Data Miner*

Einleitung

Die psychonomics AG betreibt im Kundenauftrag nicht nur Marktforschung oder branchenübliche Kundensegmentierung, sondern beherrscht darüber hinaus auch psychologische Zielgruppensegmentierung. Herkömmliche analytische Werkzeuge der Marktforschung reichen für derartige Aufgabenstellungen nicht aus, deshalb verwendet psychonomics seit einigen Jahren *STATISTICA Data Miner* zur Bewältigung komplexer Aufgaben. Der folgende Bericht beschreibt eines von vielen Projekten.

Projektüberblick

Gesamtziel des vorgestellten Projektes war die Optimierung von Dialogmarketing-Maßnahmen für den Gesamtkundenbestand eines deutschlandweit tätigen Versicherungsunternehmens. Drei Gesichtspunkte wurden berücksichtigt: Kommunikationstyp, Produktaffinität und Ansprachekanal. Als Ergebnis wurden numerische Scorewerte zur Qualifikation des Kundenbestandes im CRM-System des Versicherers hinterlegt, auf die folgende Business-Entscheidungen fußen.

Die Sonderstellung des Projektes bestand darin, dass eine Vielzahl der vorherzusagenden Größen nicht unmittelbar aus einer reinen Bestandsdatenanalyse auf der Basis historischer Transaktionsdaten abgeleitet werden konnte. Die notwendigen Informationen wurden daher auf Basis einer repräsentativen Kundenbefragung generiert. Die Antworten der Befragten wurden anhand der vorgegebenen Ziele voranalysiert und erst danach erfolgte eine Verknüpfung der Befragungsergebnisse mit den zugehörigen Bestandsdaten.

Nach der Bestandsdatenaufbereitung für die eigentlichen Analysen erfolgte die Auswahl geeigneter Algorithmen zur Prognose der drei Zielgrößen. Die Herausforderung bestand dabei darin, unter Ausschöpfung möglichst vieler Informationen bzw.

Informationskombinationen sowohl treffsichere als auch stabile Prognosemodelle für die einzelnen Zielgrößen zu generieren. Gleichzeitig musste aus Kostengründen ein begrenzter Stichprobenumfang (n=4.000) eingehalten werden. Dies gelang mit *STATISTICA Data Miner*, die Funktionen Feature-Selection mit Prüfung von Interaktionseffekten und das Verfahren Boosted Trees in Kombination mit verschiedenen multivariaten Techniken lieferten die gewünschten Prognoseergebnisse.

„Die Analytik von *STATISTICA Data Miner* beherrscht große Variablenanzahlen auch mit Interaktionseffekten. So werden hochkomplexe, multivariate Aufgaben zu handhabbaren Scorewerten für den Anwender.

Robuste und trennscharfe Prognosen werden schon bei kostengünstigen Stichproben erreicht.“

Guido Kiell, Projektleiter und Analytiker

Projektbericht

Gesamtziel des vorgestellten Projektes war eine Optimierung von Dialogmarketing-Maßnahmen für den Gesamtkundenbestand eines deutschlandweit tätigen Versicherungsunternehmens. Im Gesamtprojekt wurden verschiedene Teilziele verfolgt, u.a.

- a) eine psychologische Segmentierung des Kundenbestandes in verschiedene Kommunikationstypen,
- b) die Qualifikation der Bestandskunden nach zukünftig geplanten Produktaffinitäten und
- c) die Qualifikation des Bestandes nach Präferenzen für bestimmte Ansprachekanäle.

Zu a) Die psychologische Qualifikation des Kundenbestandes nach Kommunikationstypen gibt darüber Auskunft, mit welchen Schlüsselreizen je Kundensegment (textlich, bildlich) Informationsmaterial wie Printwerbung, Mailings, Beileger etc. bevorzugt gestaltet werden sollen.

Zu b) Die Qualifikation des Bestandes nach geplanten Produktabschlüssen informiert darüber, bei welchen Kunden oder Kundengruppen mit den höchsten Abschlusswahrscheinlichkeiten zu rechnen ist, wenn diese kontaktiert werden.

Zu c) Die Qualifikation nach Präferenzen gibt Auskunft darüber, bei welchen Kunden höhere oder niedrigere Affinitäten für bestimmte Informationskanäle vorliegen (schriftlich, telefonisch, elektronisch, Print, Radio etc.).

Reine Bestandsdatenanalysen (z.B. Abschlusswege bei Produktabschluss, bisherige Produktnutzung, Response-Reaktionen nach Mailings usw.) geben häufig nur Auskunft über das *tatsächliche* Verhalten der Kunden. Informationen über das *geplante* (z.B. geplante Produktabschlüsse) oder *gewünschte* Verhalten des Anbieters (z.B. kundenseitig gewünschte Kontaktwege, Gestaltung von Informationsmaterial) bis hin zu Informationen über die Produktnutzung beim Wettbewerb können aus einer reinen Bestandsdatenanalyse nur begrenzt gewonnen werden.

Eine umfassende Qualifikation des Kundenbestandes zur Optimierung von CRM-Aktivitäten steht dabei vor zwei Hürden:

- Notwendigkeit der Generierung umfassender Informationen zum Kundenbestand aus externen Quellen, d.h. Marktforschungsdaten und
- einem begrenzten Budget, das nur eine eingeschränkte Anzahl von Befragungen erlaubt. Als Hintergrundinformation: Eine Befragung von etwa 20 Minuten Länge bei einem Stichprobenumfang von ca. 4.000 Personen kann schnell zu Kosten zwischen 50.000€ bis 100.000€ führen. Aus dieser begrenzten Menge an Befragungen sollten allerdings gleichzeitig treffsichere Prognosen auf der einen Seite und möglichst stabile Prognosemodelle auf der anderen Seite entwickelt werden.

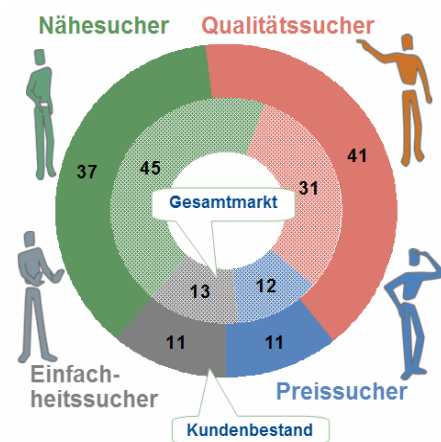
Projektstufe 1: Marktforschung

Die psychologische Qualifikation des Kundenbestandes erfolgte auf Basis einer repräsentativen Befragung von 4.000 Kunden des Auftraggebers, die u.a. Fragen zu den oben skizzierten Zielen umfasste. Die Befragungsdaten dienten zunächst als Basis zur Ermittlung der Kommunikations- und Affinitätstypen. Für die Ermittlung der Kommunikationstypen wurde auf eine bewährte Kundentypologie (Typologie privater Versicherungsnehmer) der Firma psychonomics zurückgegriffen. Die Befragten durchlaufen während der Gesamtbefragung u.a. eine standardisierte Fragenbatterie und werden im Anschluss an die Erhebung mittels eines auf einem diskriminanzanalytischen Modells basierenden Verfahren bereits existierenden Kundensegmenten zugeordnet. Es existieren vier Basistypen:

- Qualitätssucher,
- Nähesucher,
- Einfachheitssucher,
- Preissucher.

Diese vier Typen unterscheiden sich u.a. hinsichtlich ihrer

- Betreuungs-,
- Service-,
- Leistungs- und
- Preisansprüche.



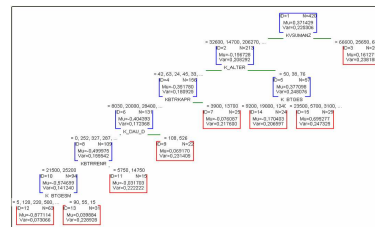
Für jeden der befragten Kunden wurden zudem Affinitätscores für geplante Produktabschlüsse (Leben- und Sachprodukte) sowie bevorzugte Ansprachekanäle des Kunden (telefonisch, schriftlich, via email) ermittelt.

Projektstufe 2: Data-Mining

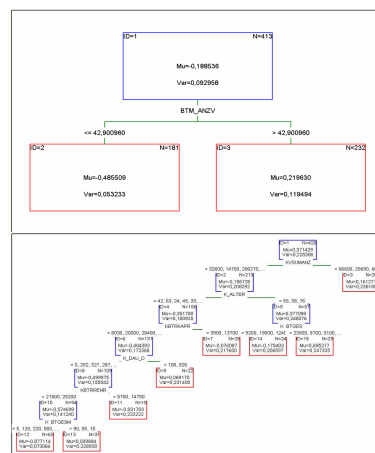
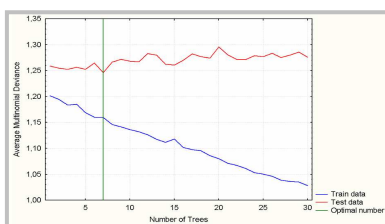
Algorithmen von STATISTICA Data Miner: Im Anschluss an die Marktforschung wurden die Befragungsdaten mit sämtlichen verfügbaren Kundeninformationen aus dem Datawarehouse des Auftraggebers verknüpft. Nach der Bestandsdatenaufbereitung für die eigentlichen Analysen erfolgte mit *STATISTICA Data Miner* die Entwicklung geeigneter Algorithmen zur Prognose der drei Zielgrößen.

Prognose: Eine besondere Herausforderung stellte dabei die Prognose der vier Kommunikationstypen dar. Die vier Segmente waren unterschiedlich stark im Bestand verteilt: Während die beiden kleinen Segmente jeweils knapp 11% des Gesamtbestandes ausmachten, hatten die großen Segmente einen Umfang von jeweils etwa 40%. Zur Prognose der Kommunikationstypen wurden verschiedene Verfahren (u.a. einfache Chaid- und C&RT-Bäume, Random Forest, logistische Regression) getestet. Befriedigende Ergebnisse ließen sich allerdings nur durch Variationen diskriminanzanalytischer Modelle und Variationen verschiedener Modelle basierend auf dem Boosted Trees-Algorithmus erzielen.

Gewöhnliche Entscheidungsbaumverfahren: Die Hauptproblematik bei dem Einsatz gewöhnlicher Entscheidungsbaumverfahren als auch dem Einsatz neuerer Baumverfahren wie Random Forest lag in der Größe der Stichprobe. Die Tiefe der Bäume ist begrenzt durch den Stichprobenumfang. Insbesondere kleine Segmente können aufgrund der geringen Fallzahlen daher häufig nur sehr schlecht prognostiziert werden, da bei Kombination verschiedener Merkmale die Fallzahlen schnell erschöpft sind. Zudem bietet der begrenzte Stichprobenumfang nur sehr eingeschränkte Möglichkeiten, die generierten Ergebnisse in hinreichend großen Trainings- und Teststichproben abzusichern.



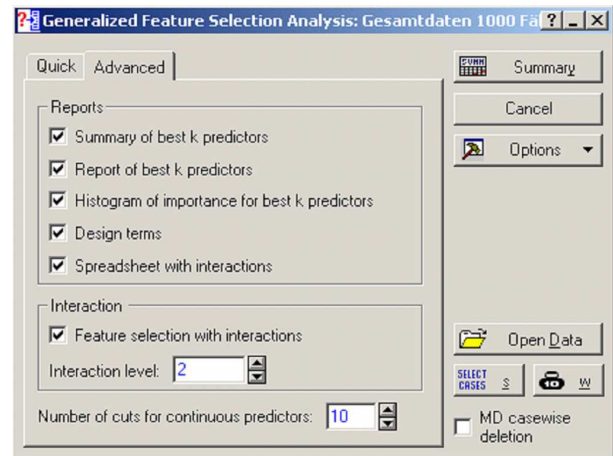
Boosted Trees: Das Verfahren Boosted Trees bietet Lösungsansätze der geschilderten Problematik und kam aus diesem Grund zum Einsatz. Wesentliche praktische Vorteile in der Anwendung des Verfahrens liegen in der geringen Knotenanzahl (drei bis fünfzehn) bzw. der damit einhergehenden Baumtiefe die dem Vorgehen zu Grunde liegen. Aus der (gewichteten) Kombination der Prognosen einer Vielzahl sehr einfacher Bäume wird die finale Gesamtprognose erstellt. Gleichzeitig wird durch ein stochastisches Gradientenboosting das Overfitting-Problem anwenderfreundlich gelöst. Das Verfahren stoppt die Modellierung weiterer Bäume, bevor es zum Overfitting kommt.



Multivariate Verfahren: Parallel wurde neben verschiedenen Entscheidungsbaumverfahren auch die Prognosemöglichkeit mittels multivariater Verfahren geprüft. Der Vorteil: Die Verfahren benötigen für zufrieden stellende Prognosen in der Regel kleinere Stichprobenumfänge als einige Entscheidungsbaumverfahren. Multivariate Verfahren wie logistische Regressionen sind diskriminanzanalytischen Verfahren in der Prognosegüte häufig überlegen, in der Regel wird dies jedoch durch höhere Rechenzeiten und einer eher unkomfortablen Anwendung der Prognosemodelle auf neu zu klassifizierende Objekte bezahlt. „Einfache“ multivariate Verfahren, wie die oben erwähnte Diskriminanzanalyse oder auch logistische Regressionen, haben aber im Gegensatz zu Entscheidungsbaumverfahren den Nachteil, dass sie per se zunächst keinerlei Modellierungen von Interaktionen (also z.B. Dauer der Kundenzugehörigkeit*Anzahl genutzter Produkte) erlauben.

Interaktionseffekte: Interaktionen zwischen mehreren unabhängigen Variablen „prognostizieren“ häufig sehr viel besser einzelne Kundensegmente, als einfache lineare Beziehungen. Beispiel: Unter älteren Kunden, die zusätzlich bereits sehr lange Kunden sind, findet man überdurchschnittlich viele „Nähesucher“ und zwar deutlich mehr, als wenn man nur ältere Kunden oder nur Kunden mit einer langjährigen Beziehung zum Unternehmen betrachtet. Genau diesen „prognostischen Mehrwert“ von Merkmalskombinationen machen sich auch Entscheidungsbaumverfahren zu nutze. Mit jeder zusätzlichen Stufe eines Baumes wird ein bereits in den Baum aufgenommenes Merkmal mit einem neuen Merkmal kombiniert. Die Merkmalskombinationen sind also nichts anderes als Interaktionseffekte bezogen auf die Zielgröße. Zwar sind diese Möglichkeiten der Modellierung von Interaktionen in STATISTICA auch bei multivariaten Verfahren im Rahmen des Allgemeinen Linearen Modells, hier

speziell der General Discriminant Analysis (GDA) möglich, die theoretische Anzahl von zweifachen oder dreifachen Interaktionen bei einer Ausgangsdatenbasis von ca. 400 Kundenmerkmalen übersteigt jedoch schnell die rechnerische Durchführbarkeit da die zu prüfenden Kombinationen leicht in die 10.000 bis 100.000 Interaktionsterme geht. Um diesen Nachteil entgegenzuwirken, wurden zunächst sämtliche theoretisch möglichen zweifachen Interaktionen und weite Teile theoretisch plausibler dreifacher Interaktionen zwischen den vorliegenden Prädiktoren und den Zielvariablen mittels sogenannter *Feature Selection* geprüft. Auch hier wurden zur Überprüfung der Stabilität der entdeckten Beziehungen



verschiedene Test- und Trainingsdatensätze konstruiert, bevor die prognosestärksten Interaktionsterme an multivariate diskriminanzanalytische Modelle weitergegeben wurden. Der große Vorteil der *STATISTICA Feature Selection* besteht vor allem darin, zunächst eine relativ große Anzahl an Interaktionstermen vorselektieren zu können, die gebildeten Interaktionsterme gleichzeitig zu berechnen und in einem Spreadsheet zur (multivariaten) Weiterverarbeitung bereitzustellen. Als Ergebnis dieser Funktionalität von *STATISTICA* kann das folgende Modell alle signifikanten Interaktionseffekte bis zur dritten Ordnung berücksichtigen und so robuste und trennscharfe Modellierungsqualität erreichen.

Prognosegüte für kleine Segmente: Neben der Aufteilung des Datensatzes in verschiedene Trainings- und Teststichproben wurden Zufallsstichproben mit unterschiedlichen Umfängen der zu prognostizierenden Kommunikationstypen zusammengestellt, um so die Prognosegüte der kleinen Segmente zu verbessern. Die verschiedenen Einzelprognosen wurden schließlich mittels eines einfachen diskriminanzanalytischen Ansatzes zu einer Gesamtprognose verrechnet, da ein einfaches Voting-Verfahren einer solchen Gesamtverrechnung unterlegen war. Die durchschnittliche Treffergenauigkeit für die vier Kommunikationstypen lag am Ende des Projektes bei rund 70 Prozent. Für die Kundentypen variierte die Treffergenauigkeit zwischen 55 und 74 Prozent.

Projektstufe 3: Implementierung und Anwendung

Nach Abschluss des Projektes wurden für jeden einzelnen Kunden einfache Scorewerte im Datawarehouse des Versicherers hinterlegt. Die Implementierung der Datenaufbereitungsschritte und Algorithmen für zukünftige Klassifikationen des Kundenbestandes wurde softwareunabhängig durch die IT-Abteilung vorgenommen, nachdem diverse Testfelder für unterschiedliche Dialogmarketingmaßnahmen die Überlegenheit der systematischen, segmentspezifischen Kundenbestandsbearbeitung nachgewiesen hatten.

Über psychonomics: Die psychonomics AG ist ein international tätiges Marktforschungs- und Beratungsinstitut mit Hauptsitz in Köln und einer Niederlassung in Wien. psychonomics [psychology - economics] ist Name und Programm: Die psychologisch fundierte Markt- und Organisationsforschung liefert den Auftraggebern hochwertiges Entscheidungswissen für Marketing, Vertriebssteuerung und Organisationsentwicklung. In der Umsetzung berät psychonomics Consulting als ausgewiesener Spezialist für Kundenbeziehungs- und Qualitätsmanagement.