Peergroup-Vergleich von kategorisierten Transaktionsdaten

Studiengang: MAS Data Science

Im E-Banking dieses Finanzinstitutes wird dem Kunden bereits heute eine Übersicht seiner kategorisierten Transaktionen angeboten. Ein Mehrwert würde ein Vergleich dieser Transaktionen mit jenen einer Gruppe ähnlicher Kunden (einer Peergroup) bieten. Ein solcher Peergroup-Vergleich wurde vor einigen Jahren umgesetzt, aber nach wenigen Monaten aufgrund von Mängeln (bezüglich Datenschutz und Qualität) wieder deaktiviert.

Ausgangslage

Die Probleme bezüglich Datenschutz wurden behoben, nicht aber jene bezüglich Qualität. Die vorliegende Master Thesis erarbeitet eine alternative Peergroup-Einteilung mit dem Ziel, die qualitativen Mängel zu beheben.
In der ersten Version des Peervergleiches konnten die Kunden ihre Peergroup aufgrund von Werten in den Kategorien Alter, Geschlecht und Einkommen selber bestimmen. Eine Peergroup war also sehr ähnlich in Bezug auf diese Kategorien. Dieses Vorgehen betrachtet nicht die Gesamtähnlichkeit der Kunden: sollten sich Kunden ähnlich sein, nur weil sie ein ähnliches Alter, Geschlecht und Einkommen besitzen?

Vorgehen

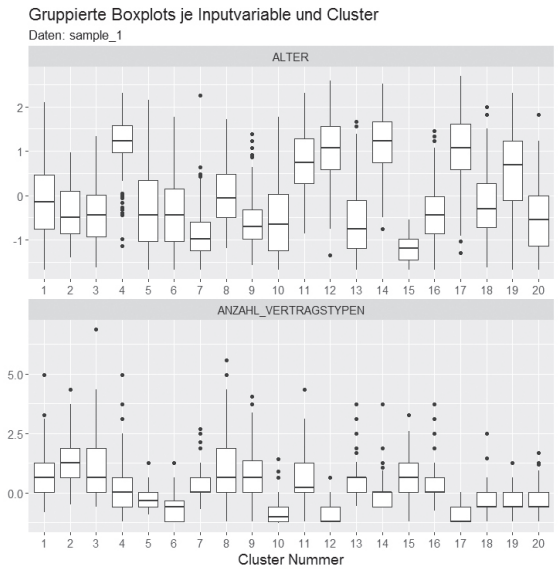
Um die Gesamtähnlichkeit einer Peergroup zu maximieren, werden die Kundengruppen anhand einer Clusteranalyse (k-Means Algorithmus) gebildet. Es werden alternative Clusterings (unterschiedliche Anzahl Cluster, unterschiedliche Inputvariablen, unterschiedliche Samples) quantitativ gegenübergestellt.

Das beste Clustering wird dann qualitativ in Bezug auf die Inputvariablen und die Transaktionen beurteilt. Die Clusteranalyse zeigt: bezüglich Inputvariablen lassen sich einige Cluster sehr klar definieren. Bezüglich Transaktionen ist das Bild weit weniger scharf.

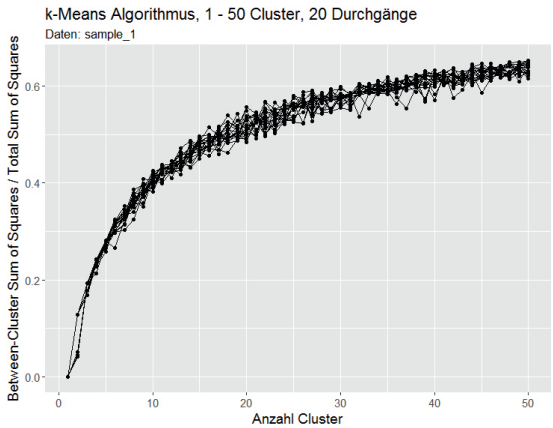
Fazit

Der Peervergleich hängt neben dem Kundenclustering auch von der Transaktionsklassifizierung ab. Das unscharfe Bild in Bezug auf die kategorisierten Transaktionen legt nahe, dass auch die Qualität der Transaktionsklassifizierung verbessert werden könnte. Zudem stellen die unterschiedlichen Personenkonstrukte in den verschiedenen Datenbanken ein Problem dar: es handelt sich um technische Konstrukte, welche nicht der Realität entsprechen und für Kunden- und Transaktionsdaten nicht übereinstimmen. Dies erschwert die Analyse des Verhaltens realer Personen.

Nicole Keller



Gruppierte Boxplots pro Cluster für die Inputvariablen Alter und Anzahl Vertragstypen (standardisierte Daten)



Qualität des Clusterings mit 1–50 Clustern und 20 Durchgängen