Segmentierung von Nutzern der autoscout24.de-Plattform unter Verwendung clusteranalytischer Methoden

BACHELORARBEIT
ZUR ERLANGUNG DES AKADEMISCHEN GRADES
BACHELOR OF SCIENCE (B. Sc.)

Autor: Yvonne Barth

Matrikelnummer: XXXXXXXX

Gutachter: Prof. Dr. Christian Heumann

Abgabedatum: 15. September 2015
Eidesstattliche Erklärung


München, 15. September 2015 ....................................................

(Unterschrift)
Inhaltsverzeichnis

1 Einleitung 1
  1.1 Nutzersegmentierung und Intention der Arbeit ......................... 1
  1.2 Zu AutoScout24 .............................................. 2

2 Datengrundlage 3
  2.1 Datenbeschaffung ............................................... 3
    2.1.1 HDFS .................................................... 4
    2.1.2 Map-Reduce ............................................... 5
  2.2 Datenaufbereitung .............................................. 6
    2.2.1 Basisdaten ............................................... 6
    2.2.2 Transformation der Logfiles ............................... 8

3 Deskriptive Datenanalyse 12

4 Clusteranalyse 20
  4.1 Die Funktionsweise des K-Means Algorithmus ......................... 21
  4.2 K-Means in der Anwendung ....................................... 23
  4.3 Die Funktionsweise des Two-Step-Cluster Algorithmus ............... 31
  4.4 Two-Step-Clustering in der Anwendung ................................ 36

5 Ex-Post Analyse 39
  5.1 K-Means ...................................................... 39
  5.2 Two-Step-Clustering ........................................... 45
  5.3 Vergleich K-Means und Two-Step-Clustering .......................... 46
  5.4 Random Forest ................................................ 48
    5.4.1 Vorhersagekraft auf Basis der K-Means Ergebnisse ............ 51
    5.4.2 Vorhersagekraft auf Basis der Two-Step-Clustering Ergebnisse 53

6 Fazit und Handlungsempfehlungen 53

Literatur 56

Tabellenverzeichnis 59

Abbildungsverzeichnis 59

A Hive-Befehl zum Abfragen der Basisdaten aus dem Hadoop Clus-
ter 61
B Filterhäufigkeit einzelner Marken

C Variablenunterschiede einzelner Cluster des Two-Step-Cluster Algorithmus

D Koordinatenplots der vier Nutzergruppen des Two-Step-Cluster Algorithmus

E Klassifikationsvergleich K-Means - Random Forest mit veränderten Parametern
  E.1 Anzahl Bäume: 500, Split-Variablen: 8 . . . . . . . . . . . . . . . . . 75
  E.2 Anzahl Bäume: 800, Split-Variablen: 5 . . . . . . . . . . . . . . . . . 75

F Klassifikationsvergleich Two-Step-Clustering - Random Forest
  F.1 Anzahl Bäume: 500, Split-Variablen: 5 . . . . . . . . . . . . . . . . . 75

G Variablenwichtigkeit / Two-Step-Clustering

76
1 Einleitung

1.1 Nutzersegmentierung und Intention der Arbeit


Die Hauptmerkmale verschiedener Nutzer sollen innerhalb eines Segments möglichst einheitlich und zwischen den Segmenten so verschieden wie möglich sein. Um einzelne Personen individuell erreichen zu können muss klar sein, in welches Cluster man sie einordnen kann.


1.2 Zu AutoScout24


Gegründet wurde die Tochter der Scout24-Gruppe im Jahr 1998 als MasterCarAG. Das Onlineportal ist mittlerweile in 17 europäischen Ländern vertreten und insgesamt gibt es mehr als zwei Millionen Angebote.


Ruft man die Webseite auf, sieht man zunächst eine Box, in welche man Filterkriterien für das gewünschte Fahrzeug oder Motorrad eingeben kann. Klickt man weiterführend auf den Button zur Anzeige der entsprechenden Fortbewegungsmittel, erscheinen auf der Suchergebnisseite 20 Übersichtsfenster. Wählt man ein Fahrzeug der Suchergebnisseite aus, kann man auf der Detailseite nähere Informationen und Bilder betrachten.


Überdies besteht eine weitere Verdienstmöglichkeit durch die im Werkstattportal vertretenen Werkstätten, von welchen, ähnlich zu den Händlern, auch eine Gebühr erhoben wird. [vorliegende Informationen entstammen dem Unternehmensportrait von AutoScout24 (2015)]
2 Datengrundlage

2.1 Datenbeschaffung


Die zwei wichtigsten zusammenarbeitenden Komponenten im Umfeld von Hive sind HDFS und die „Map-Reduce-Engine“, weshalb im Folgenden auf diese Komponenten eingegangen wird.
2.1.1 HDFS


Folgende Grafik skizziert die eben beschriebene Architektur grundlegend:

![Grafische Darstellung eines Hadoop-Clusters und dessen Komponenten](image)

Abbildung 2: Grafische Darstellung eines Hadoop-Clusters und dessen Komponenten

2.1.2 Map-Reduce


Abbildung 3: Map-Reduce Schritte [van Groningen (2009)]

Im Schritt „Aufteilung“ werden die Daten auf Blöcke verteilt. Die Map-Funktion greift auf die Blöcke zu und erzeugt distinkte Schlüsselwert-Paare (engl.: key-value pairs). In der Zwischenverarbeitung wird nach den einzelnen Schlüsseln, wie „Hund“ oder „Katze“ sortiert. Die Reduce-Funktion komprimiert die Daten zu einer Datei, verarbeitet die Ausgabe und gibt sie anschließend aggregiert in einer Ergebnisliste zurück. In diesem Beispiel werden die Werte summiert, jedoch kann auch beispielsweise das arithmetische Mittel, das Minimum oder Maximum berechnet werden [White (2010)].

Eine weitere Komponente der Hadoop Datenbank ist die Abfragesoftware Hive. Hive erweitert Hadoop um eine Datenbank-Infrastruktur und ermöglicht das Abfragen von Daten im Hadoop System mit einer SQL-ähnlichen Sprache [Rutherglen et al. (2012)]. Die SQL-Abfragen werden im Hintergrund wieder in die Map-Reduce-
Funktion übersetzt, um somit die Daten zur Verfügung stellen zu können [Rutherglen et al. (2012)].

In dieser Arbeit werden die Variablen für die spätere Segmentierung unter Verwendung von Hive extrahiert. Der verwendete Befehl ist Anhang A zu entnehmen. Die relevanten URLs, welche Detailseiten-, Suchergebnisseiten- und Filterseitenauftrufe beinhalten, werden im Vorfeld schon durch die Hive Abfrage gefiltert. Gespeichert werden die Basisdaten mit circa 12 Millionen verschiedenen URLs von 89.821 Nutzern im csv-Format, um das spätere Einlesen mit R zu ermöglichen.

2.2 Datenaufbereitung

2.2.1 Basisdaten


Während des Ladeprozesses einer Seite werden für csuriquery, csuristem und csreferer sekundär Informationen weggescrieben. Daher werden für jede Aktion, die der Nutzer tätig, viele verschiedene Teilinformationen erfasst. Vorrangiges Ziel ist es, wiederkehrende Bestandteile relevanter Aktionen, wie das anfängliche Suchen der gewünschten Marke, herauszufiltern und als eigenständige Aktion zu erfassen. Tabelle 1 beschreibt den Datensatz ausschnittsweise:
<table>
<thead>
<tr>
<th>sdate</th>
<th>stime</th>
<th>as24visitorID</th>
<th>csuristem</th>
<th>csuriquery</th>
<th>csreferer</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-05-10 16:49:01</td>
<td>000588f4-9758-ffe5-4cf2-b9d0-2162</td>
<td>/ArticleList</td>
<td>atype=C &amp;make=60 &amp;mmvmk0=1 &amp;fregto=1985 &amp;cy=D &amp;uystate=N,U &amp;sort=threetier, price&amp;resu lts=20&amp;pa ge=1&amp;tabg =guidedfull</td>
<td>csreferer <a href="http://ww4.ffe5-4cf2-9758-4cf2-">http://ww4.ffe5-4cf2-9758-4cf2-</a> &amp;make=60 &amp;mmvmk0=60 &amp;mmvco=1 &amp;fregto=1985 &amp;cy=D &amp;uystate=N,U &amp;sort=price &amp;results=20 &amp;page=1 &amp;event=sort &amp;dtr=s autoscout24.de/fahrzeuge</td>
<td></td>
</tr>
<tr>
<td>2015-05-10 16:49:10</td>
<td>000588f4-9758-ffe5-4cf2-b9d0-2162</td>
<td>/classified</td>
<td>asrc=st</td>
<td>as &amp;testvariant=list3tiers &amp;tierlayer=st</td>
<td><a href="http://ww4.ffe5-4cf2-9758-4cf2-">http://ww4.ffe5-4cf2-9758-4cf2-</a> autoscout24.de/fahrzeuge? atype=C &amp;mmvmk0=21 &amp;mmvmd0=185 45&amp;mmvco=1 &amp;make=21 &amp;model=18545 &amp;fuel=D &amp;fregfrom=200 00&amp;pricefrom=1000&amp;priceto=2500 &amp;cy=D &amp;uystate=N,U &amp;fromhome=1 &amp;intcidm=HP Searchmask Button&amp;dtr=s</td>
</tr>
<tr>
<td>2015-05-10 16:52:47</td>
<td>000588f4-9758-ffe5-4cf2-b9d0-2162</td>
<td>/Parkdeck</td>
<td>/Add</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Tabelle 1: Rohdatensatz mit mehreren Zeilen pro Nutzer

1 aus Gründen des Datenschutzes anonymisiert
Der in dieser Tabelle beispielhaft ausgewählte Nutzer hat am 10. Mai 2015 zwischen 16:49 Uhr und 16:52 Uhr drei relevante Aktionen auf der Webseite ausgeführt. Diese sind zunächst ein Aufruf der Suchergebnisseite (\ArticleList\GetCounters in csuristem) mit zuvor eingestellten Filtern, welche durch das „&“-Zeichen im csreferer und csuriquery getrennt sind. Wie aus der zweiten Zeile hervorgeht wurde ein Fahrzeug mit Artikelnummer „268648722“ im Detail angesehen (\classified\268648722 in csuristem) und die letzte Aktion war das Anlegen eines Merkzettels mit betreffendem Automobil (\Parkdeck\Add\268648722 in csuristem).

2.2.2 Transformation der Logfiles

Um das Nutzungsverhalten der Besucher besser zu beschreiben, werden sogenannte „Key Performance Indicators“, übersetzt Messwerte oder Kennzahlen, ermittelt. So wird der Rohdatensatz, welcher pro Nutzer mehrere Zeilen mit unterschiedlichen URLs enthält, in einen Datensatz mit jeweils einer Zeile pro Nutzer transformiert. Diese Zeile enthält vor allem individuelle Informationen über das Surf-Verhalten. Für alle 89.821 Besucher werden sowohl Kennzahlen zum Nutzungsverhalten auf den verschiedenen Seitentypen als auch fahrzeugspezifische Messwerte ermittelt.

Für die Berechnung aller Kennzahlen wird die open-source Software R [R Core Team (2015)] verwendet.

Angefangen mit der Messung der Intensität eines Besuchs, wird das arithmetische Mittel aller relevanten Aktionen, die der Nutzer pro Tag auf der Webseite tätigte, berechnet.


Auf Basis dieses Datensatzes kann die durchschnittliche Anzahl der relevanten Aktionen pro Tag für jeden Nutzer berechnet werden. Anschließend wird jeweils der prozentuale Anteil an den eben genannten Seiten im Verhältnis zu allen Aktionen ermittelt.

Des Weiteren wird berechnet, wieviele Filter durchschnittlich in jeder Initialsuche gesetzt werden, um die Phase der Kauffentscheidung eines Nutzers einordnen zu können. Wenn der Nutzer viele Filter setzt, liegt die Vermutung nahe, dass er schon eine klare Vorstellung vom gewünschten Auto hat.
Eine Initialsuche zeichnet sich dadurch aus, dass sich der Nutzer nach der Filtereinschätzung die erste Suchergebnisseite aller von ihm gefilterten Fahrzeuge ansieht.


Der Score berechnet sich aus:

\[
\frac{(\text{Anzahl aktiver Tage})^2}{\text{Zeitspanne}}, \text{ Wertebereich } \left[\frac{4}{15}, 15\right]
\]


Der Score ist somit \(\frac{2^2}{5} = 0,8\).

Ein Nutzer, der in der gleichen Zeitspanne an vier Tagen aktiv war, erhält den Score \(\frac{4^2}{5} = 3,2\). Demnach besitzt der zweite Nutzer im Vergleich zum ersten einen höheren Wert, da er im selben Zeitraum häufiger auf der Webseite war. Die auf dem Score basierenden Daten entstammen der Spalte „Datum“. Für alle Nutzer wurde ein dataframe folgender Form erstellt:

<table>
<thead>
<tr>
<th>as24visitorID</th>
<th>03.05.</th>
<th>04.05.</th>
<th>05.05.</th>
<th>06.05.</th>
<th>07.05.</th>
<th>08.05.</th>
<th>09.05.</th>
<th>10.05.</th>
</tr>
</thead>
<tbody>
<tr>
<td>00005274-0290-4843-b7e1-e5d6209XXXXX</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11.05.</th>
<th>12.05.</th>
<th>13.05.</th>
<th>14.05.</th>
<th>15.05.</th>
<th>16.05.</th>
<th>17.05.</th>
<th>aktive_Tage</th>
<th>Zeitspanne</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>10</td>
<td>6,4</td>
</tr>
</tbody>
</table>

Tabelle 2: Berechnung des Maßes für aktive Tage

Ist das betreffende Datum mit einer Eins vermerkt, war der Nutzer an dem Tag
online. Eine Null bedeutet, dass er die Plattform an dem Tag nicht besucht hat. Die aktiven Tage ergeben sich aus der Summe der Einträge, die Zeitspanne ist die Differenz aus dem letzten und ersten aktiven Tag plus Eins. Der Score berechnet sich aus der oben genannten Formel. Die letzte Kennzahl gibt Auskunft darüber, ob ein Nutzer mindestens ein Fahrzeug inseriert hat. Trifft dies zu, erhält der betreffende Nutzer eine Eins für dieses Variable, falls nicht, wird eine Null notiert.

Als nächstes werden fahrzeugspezifische Kennzahlen ermittelt. Hier wird zuerst berechnet, wie hoch der Durchschnittspreis aller Fahrzeuge ist, die sich der Nutzer im Detail angesehen hat. Um die Preisbereitschaft herauszufinden, wird sowohl das absolute Maximum aller Minimumpreise, als auch das absolute Minimum aller Maximumpreise jedes Nutzers notiert. Zusätzlich wird noch die Information aufgenommen, ob ein Besucher spezifisch nach Fahrzeugen von Händlern oder privaten Anbietern gesucht hat, um die Präferenzen verschiedener Nutzer zu messen.


Durch die Kennzahlen für jeden Nutzer erhält man nun einen Datensatz, der die Nutzer mit verschiedenen Merkmalen und nicht mehr durch URLs beschreibt. Die Merkmale sind zusammengefasst Messgrößen zu den unterschiedlichen Seitentypen und zu den Eigenschaften der betrachteten Detailseiten. Außerdem enthält der
Transformierter Datensatz auch Informationen darüber, wie viele und welche Filtereinstellungen verwendet wurden und wie aktiv der Nutzer ist.

Tabelle 3 enthält die Kennzahlen ausschnittsweise für einen Nutzer.

<table>
<thead>
<tr>
<th>as24visitorID</th>
<th>alle_Aktionen</th>
<th>rel_suchen</th>
<th>rel_detail</th>
<th>rel-mails</th>
<th>rel_bookmarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>00002cb7-9758-44b2-95de-47dee09XXXXX</td>
<td>59,75</td>
<td>0,75</td>
<td>0,25</td>
<td>0,00</td>
<td>0,00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>active_score</th>
<th>mind_1_Inserat</th>
<th>Haendler</th>
<th>Privat</th>
<th>Filter_Inital</th>
<th>Durchschnittspreis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,77</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2,54</td>
<td>7.663,16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Minimumpreis</th>
<th>Maximumpreis</th>
<th>keine_Marke</th>
<th>VW</th>
<th>Mercedes</th>
<th>BMW</th>
<th>Audi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>12.000</td>
<td>0,30</td>
<td>0,00</td>
<td>0,00</td>
<td>0,24</td>
<td>0,08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ford</th>
<th>Opel</th>
<th>Skoda</th>
<th>Toyota</th>
<th>Renault</th>
<th>Peugeot</th>
<th>uebrige_Marken</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,05</td>
<td>0,00</td>
<td>0,00</td>
<td>0,30</td>
<td>0,00</td>
<td>0,00</td>
<td>0,03</td>
</tr>
</tbody>
</table>

Tabelle 3: Transformierter Datensatz mit einer Zeile pro Nutzer

Diese Person hat durchschnittlich 59,75 Aktionen pro Tag, an dem er auf der Plattform war, getätigt.

Anteilig hat dieser Nutzer zu 75% Suchergebnis- und Filterseiten aufgerufen und zu 25% Detailseiten betrachtet. Es wurden weder Emails versendet, noch Lesezeichen angelegt.

Der Aktiv-Score liegt bei 1,77, was bedeutet, dass dieser Nutzer in dem Beobachtungszeitraum nicht oft auf der Plattform tätig war.

Er hat kein Fahrzeug inseriert und mindestens einmal explizit nach Händlerfahrzeugen gesucht. Am Anfang jeder Suche wurden im Mittel 2,54 Filter gesetzt.

Im Schnitt waren alle Fahrzeuge, die er sich im Detail angesehen hat, 7.663,16€ wert.

Für Fahrzeuge, die billiger als 1.000€ sind, hat er sich nicht interessiert und seine minimale Obergrenze lag bei 12.000€.

Betrachtet man alle seine Markenfilter hat er in 30% seiner Filtereinstellungen keine Marke angegeben. Zu 24% wurde die Marke BMW, zu 8% Audi, zu 5% Ford und zu 30% Toyota gesucht. 3% aller seiner Sucheinstellungen bestanden aus mindestens einer der 353 übrigen Marken.

Durch diese Kennzahlen wird der wichtigste Informationsbedarf, welcher für die Nutzersegmentierung interessant ist, abgedeckt.
3 Deskriptive Datenanalyse

Die deskriptive Analyse soll dazu dienen, Daten grafisch darzustellen und zu beschreiben [Fahrmeir et al. (2004)]. So kann ein Überblick über Strukturen und Besonderheiten gewonnen werden, was eine sehr wichtige Grundvoraussetzung für die weitere Arbeit mit den Daten ist.
Im Folgenden werden die Segmentierungsvariablen einzeln deskriptiv betrachtet. Alle Grafiken wurden unter Verwendung des ggplot2-Pakets für R [Wickham (2009)] erstellt.

Zunächst wird die Verteilung aller getätigten Aktionen betrachtet.

Abbildung 4: Histogramm der Variable „alle Aktionen“


Insgesamt 103 Nutzer führen im Schnitt weit mehr als 300 Aktionen pro Tag durch, diese sind jedoch nicht in der Grafik enthalten.

Nutzer, die mehr als 300 Aktionen pro Tag tätigen, fallen vermutlich in die Gruppe „grabber“ oder „crawler“. Für diese Gruppe wird davon ausgegangen, dass automatisiert oder computergestützt gezielt Informationen der Webseite gesammelt werden.

Abbildung 5: Boxplots der Variablen „Detailseiten“ und „Suchseiten“

Der Median der Detailseiten liegt bei circa 25% während sich das 50%-Quantil für Suchseiten knapp unter 75% befindet. Die Interquartilsabstände beider Boxplots sind nahezu gleich breit. Die Verteilung der Detailseiten ist linkssteil, die der Suchseiten rechtssteil. Für die Hälfte der Grundgesamtheit machen Detailseiten zwischen 15% und 40% aller Aktionen aus, wohingegen Suchseiten mit 60% bis 80% öfter aufgerufen werden. Seiten, auf welchen Emails versendet oder Merkzettel angelegt werden, machen einen wesentlich kleineren prozentualen Anteil aus, wie man Abbildung sechs und sieben entnehmen kann. Diese Seitentypen wurden anhand zweier Histogramme dargestellt, da ein Boxplot vergleichsweise keinen anschaulichen Interquartilsabstand liefert und primär Ausreißer zu erkennen sind.

Abbildung 6: Histogramm der Variable „Merkzettel“

Abbildung 7: Histogramm der Variable „Mails“

In der Grafik rechts wird die Verteilung der durchschnittlich gesetzten Initialfilter dargestellt\(^1\). Hierbei handelt es sich um eine bimodale Verteilung. Die erste Gruppe charakterisiert sich dadurch, wenige Filter in der Initialsuche zu setzen. Die zweite Gruppe stellt im Schnitt sehr viele Filter ein und scheint daher schon im Vergleich zur ersten Gruppe genauere Vorstellungen vom Fahrzeug zu besitzen.

\(^1\) Wählt man im Filterfenster zu Beginn keine Filter aus und sucht alle Fahrzeuge, sind automatisch drei Filter eingestellt: Ein Minimumpreis von 1.000€, Fahrzeuge des Landes Deutschland und Neu- & Gebrauchtwagen
Zu den Durchschnittspreisen aller angesehenen Fahrzeuge ist zu bemerken, dass der Wert, welcher bei der Datenaufbereitung zu Beginn durch den Mittelwert aller Preise ersetzt wurde, ausgeschlossen wird, da nur die tatsächlich angesehenen Durchschnittspreise aller Nutzer betrachtet werden sollen. Damit die Grafik anschaulich wird, werden die 4.000 Nutzer mit einem Durchschnittspreis über 50.000€ ebenfalls entfernt.

Abbildung 10: Histogramm der Variable „Durchschnittspreis“

Aus der Grafik kann man ablesen, dass um die 4% aller Nutzer günstige Fahrzeuge mit einem Durchschnittspreis zwischen 2.000€ und 4.000€ betrachtet haben. Die Verteilung fällt mit steigendem Preis konstant, was bedeutet, dass die Nachfrage nach billigeren Fahrzeugen höher ist.

Ferner werden die Preisfilter betrachtet. Beim Minimumpreis ist standardmäßig 1.000€ eingestellt, um Unfallfahrzeuge und nicht fahrbereite Automobile auszuschließen. Beim Maximumpreis ist keine obere Grenze voreingestellt, weshalb die fehlenden Werte durch 100.000€ ersetzt wurden. Sowohl 1.000€ als auch 100.000€ werden für folgende Grafiken ausgeschlossen, um besser erkennen zu können, wonach explizit gefiltert wurde. Des Weiteren sind 984 Nutzer, welche einen Minimumpreis gesetzt haben, der 25.000€ überschreitet, zur besseren Anschaulichkeit nicht in der Grafik enthalten.

Abbildung 11: Histogramm der Variable „Minimumpreis“
Abbildung 12: Histogramm der Variable „Maximumpreis“

Der Standard-Minimumpreis wurde von über 30% der Nutzer auf 500€ reduziert. An zweiter Stelle beschränken circa 20% aller Personen den Filter auf ein Minimum von 0€. Weniger als jeweils 5% setzen den Mindestpreis höher als 7.500€ und ein kleiner Teil der beobachteten Grundgesamtheit sucht nach Fahrzeugen ab 10.000€. Zusammenfassend liegt das 50%-Quantil des Minimumpreises bei 500€, was bedeutet, dass die Hälfte der in der Grafik betrachteten Nutzer den Minimumpreis auf einen Preis unter oder gleich 500€ setzt, während die andere Hälfte höhere Minimumpreise zwischen [500, 25.000]€ eingibt.

Als Obergrenze für den Fahrzeugpreis wurde am häufigsten von 20% aller Personen ein Wert um die 9.000€ gesetzt. An zweiter Stelle befindet sich ein kleiner Maximalpreis zwischen 2.000€ und 4.000€, welcher von circa 10% der Nutzer eingestellt wurde. Die meisten Personen filtern nach verschiedenen Maximalpreisen im Inter-
vall bis 30.000€. Insgesamt suchen die unteren 50% aller Besucher nach Fahrzeugen mit einem Preis unter 10.000€, wohingegen die oberen 50% nach Automobilen im Intervall zwischen [10.000, 100.000]€ filtern.
Nachfolgend wird der explizite Filter nach Händlern und privaten Verkäufern betrachtet.

Abbildung 13: Balkendiagramm der Verkaufstypen

Circa Neun Prozent aller Nutzer haben im Laufe ihrer Suche mindestens einmal spezifisch nach Automobilen von Händlern gesucht, während im Vergleich ungefähr Fünf Prozent den Filter „private Fahrzeuge“ gesetzt haben. Insgesamt nutzen 14% aller Nutzer einen der beiden Filter, was bedeutet, dass 86% der Grundgesamtheit nicht explizit nach Verkaufstypen unterscheidet.
In Abbildung 14 wird visualisiert, welchen Anteil die unterschiedlichen Personen für die Suche nach der Marke VW aufweisen.
Es ist deutlich zu erkennen, dass bei einer Mehrheit von 75% VW nie in den Filterkriterien nach Marken vorkommt.
Lediglich 5% der Nutzer sind markentreu und suchen ausschließlich nach VW.
Bei den übrigen 20% wird nicht vollständig nach VW gefiltert, was auf eine geringe Markenpräferenz und -treue hinweist.
Die Verteilung für die Top 2-10, als auch für keine und übrige Marken ist sehr ähnlich zu der von VW, was in Anhang B nachgesehen werden kann.
Ergänzend zu der deskriptiven Analyse kann gesagt werden, dass 0,05% aller Nutzer mindestens ein Fahrzeug inserieren. Da dies ein sehr kleiner prozentualer Anteil der Grundgesamtheit ist, wird diese Variable nicht mit in die Clusteranalyse aufgenommen.

Bei der Verteilung der durchschnittlichen Aktionen pro Tag, des Minimumpreises, des Durchschnittspreises und der relativen Häufigkeit von Mails und Merkzetteln treten im oberen Bereich Ausreißer auf. Da ohne Ausreißerbehandlung die Clusterschätzung verzerrt wird, werden im folgenden alle Werte der Variablen, welche größer als das jeweilige 99%-Quantil sind, entfernt. Wie Tabelle 4 zu entnehmen ist, werden beispielsweise für die Variable „Minimumpreis“ alle Beobachtungen, welche für diese Variable Werte größer als 25.000 enthalten, beseitigt.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Wert</th>
</tr>
</thead>
<tbody>
<tr>
<td>durchschnittliche Aktionen pro Tag</td>
<td>168,86</td>
</tr>
<tr>
<td>Minimumpreis</td>
<td>25.000€</td>
</tr>
<tr>
<td>Durchschnittspreis</td>
<td>161.036, 24€</td>
</tr>
<tr>
<td>relativer Anteil Mails</td>
<td>2,90%</td>
</tr>
<tr>
<td>relativer Anteil Merkzettel</td>
<td>14,96%</td>
</tr>
</tbody>
</table>

Tabelle 4: Werte des 99%-Quantils ausgewählter Variablen

Durch die Ausreißerbehandlung fallen ungefähr 5.000 der ursprünglich circa 90.000 Nutzer weg und rund 85.000 Personen bilden somit die Basis für eine Clusteranalyse.
4 Clusteranalyse

Die Clusteranalyse ist ein multivariates statistisches Verfahren. Ziel einer Clusteranalyse beziehungsweise Datensegmentierung ist es, Objekte aufgrund vorgegebener MerkmalsvARIABLEN in disjunkte Gruppen $\{C_1, ..., C_k\}$ zu unterteilen [Steinhausen & Langer (1977)]. Die Clusterbeobachtungen sollen innerhalb einer Gruppe möglichst homogen und untereinander so heterogen wie möglich sein. Darüberhinaus dient das Clustering der deskriptiven Analyse, also der Beantwortung der Frage, ob in den Daten überhaupt verschiedene Untergruppen, welche sich im Wesentlichen voneinander unterscheiden, zu finden sind [Handl (2010)].

Bei der Clusteranalyse handelt es sich um „unsupervised learning“ [Hastie et al. (2009)]. Hier ist die Gruppenzugehörigkeit, im Gegensatz zu „supervised learning“, einzelner Beobachtungen und die Anzahl der verschiedenen Cluster a priori nicht bekannt.


4.1 Die Funktionsweise des K-Means Algorithmus

K-Means ist ein partitionierendes Clusterverfahren und gehört zu den iterativen Algorithmen. Im Gegensatz zu hierarchischen Verfahren kann sich bei dieser Methode die Klassenzugehörigkeit einzelner Beobachtungen im Iterationsverlauf ändern [Handl (2010)]. Grundlegend für diesen Algorithmus sind metrische Variablen und als Distanzmaß wird die quadrierte euklidische Distanz verwendet, welche sich nach Hastie et al. (2009) wie folgt berechnet:

\[ d(x_i, x'_i) = \sum_{j=1}^{p} (x_{ij} - x'_{ij})^2 = ||x_i - x'_i||^2 \]  

(1)

Da bei K-Means zu Beginn die Anzahl der gewünschten Cluster \( k \) festgelegt werden sollen, müssen diese zunächst durch ein geeignetes Gütekriterium bestimmt werden. Abhängig von dieser Anzahl werden zu Beginn \( k \) zufällige Startpartitionen festgelegt.

Jeder Unterteilung \( k \) wird ein Mittelwert beziehungsweise Zentroid \( \bar{x}_k \) zugeordnet. Alle Individuen werden iterativ dem Clusterzentroid, zu dem die kleinste Distanz besteht, zugewiesen.

Ziel ist es hier unter Verwendung der quadratischen euklidischen Distanz die Streuung zwischen den Clustern zu minimieren.

Die Streuung zwischen den Clustern ergibt sich aus folgender Formel:

\[
W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x_i, x'_{i'}) \\
= \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} ||x_i - x'_{i'}||^2 \\
= \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \bar{x}_k||^2
\]  

(2)

wobei \( \bar{x}_k = (x_{1k}, ..., x_{pk}) \) der Mittelwertsvektor des \( k \)-ten Cluster ist und \( N_k = \sum_{i=1}^{N} I(C(i) = k) \)
Somit wird im ersten Schritt versucht, die Clustervarianz zu minimieren:

$$\min_{C, m_k} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - m_k||^2$$  \hspace{1cm} (3)

wobei \{m_1, ..., m_K\} die Menge der Zentroide darstellt. Die Streuung wird kleiner, je ähnlicher die Werte der Beobachtungen den Zentroiden sind. Nach diesem Vorgang werden im zweiten Schritt auf Basis der neu angeordneten Datenpunkte neue Zentroide wie folgt berechnet:

$$\bar{x}_S = \arg\min_m \sum_{i \in S} ||x_i - m||^2$$  \hspace{1cm} (4)

wobei \(m\) der derzeitige Mittelwert des zugeteilten Cluster ist. Anschließend werden die Beobachtungen wieder neuen Clustern C(i) unter Anwendung von (5) zugeordnet.

$$C(i) = \arg\min_{1 \leq k \leq K} ||x_i - m_k||^2$$  \hspace{1cm} (5)


Abbildung 15: sukzessive Iteration des K-Means Clusteralgorithmus für simulierte Daten [Hastie et al. (2009)]

\[22\]

4.2 K-Means in der Anwendung


Zusammenfassend lässt sich sagen, dass in drei große Segmente, Cluster fünf, sechs und sieben, drei mittelgroße, Cluster eins, vier und acht, und drei kleine Gruppen, Cluster zwei, drei und neun, aufgeteilt wurde.
Tabelle 5: Resultat Clustering: Verteilung der Nutzer auf alle neun Cluster / K-Means

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Cluster 5</th>
<th>Cluster 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>8,89%</td>
<td>4,64%</td>
<td>3,71%</td>
<td>12,63%</td>
<td>16,14%</td>
<td>21,48%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 7</th>
<th>Cluster 8</th>
<th>Cluster 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>17,11%</td>
<td>12,74%</td>
<td>2,66%</td>
</tr>
</tbody>
</table>

Abbildung 16: Bestimmung der Clusteranzahl anhand des Ellbogenkriteriums
Nachdem die Daten segmentiert wurden, wird nun auf die Unterschiede zwischen den einzelnen Clustern eingegangen. Hierfür werden Boxplots verwendet, welche die z-standardisierten Werte der einzelnen Variablen in jedem Cluster veranschaulichen. Die Boxplots werden ohne Ausreißer dargestellt, damit der Wertebereich der y-Achse übersichtlicher wird und die einzelnen Interquartilsabstände somit besser interpretierbar sind.

Abbildung 17: Boxplots der Variablen „alle_Aktionen“, „aktiv_Score“ & „Filter_Initialsuche“ / K-Means

Anknüpfend werden die relativen Anteile der Seitentypen betrachtet.

Übergreifend ist zur Interpretation der Boxplots für die prozentualen Anteile der einzelnen Seitentypen zu sagen, dass es keine großen Unterschiede zwischen den Clustern gibt und lediglich Gruppe zwei heraussticht. Dieses Segment besitzt eine verhältnismäßig kleine Quote an Filter- und Suchergebnisseiten und hebt sich durch einen hohen Anteil an Detailseiten von den restlichen Nutzergruppen ab. Des Weiteren legen Nutzer des zweiten Clusters viele Lesezeichen an und versenden mehr Mails als die übrigen Nutzer, was darauf hinweist, dass diese schon sehr fortgeschritten sind in der Suche nach einem Fahrzeug. Bei den übrigen Gruppen ist das Verhältnis zwischen Such- und Detailseiten ausgeglichen. Der Emailversand ist im Vergleich zum zweiten Segment sehr gering und Lesezeichen werden kaum angelegt.

Abbildung 18: Boxplots der vier Seitentypen / K-Means
Im weiteren Verlauf wird auf die Unterschiede der Variablen „Durchschnittspreis“, „Minimumpreis“ und „Maximumpreis“ eingegangen.

Abbildung 19: Boxplots der Variablen „Durchschnitts-“, „Minimum-“ und „Maximumpreis“ / K-Means

Bei der Betrachtung des Durchschnittspreises aller angesuchten Fahrzeuge fallen die Gruppen eins, sechs und sieben auf. Diese Segmente suchen im Vergleich zu den anderen Gruppen eher Fahrzeuge im höheren Preisbereich. Die niedrigsten Durchschnittspreise sind in Gruppe vier zu finden.

Wirft man einen Blick auf den Minimumpreis, fällt auf, dass der Median in allen Segmenten identisch ist, was darauf zurückzuführen ist, dass der voreingestellte Standardpreis von 1.000€ in allen Segmenten oft vorkommt. In Cluster acht ist der Abstand zwischen dem 75% und 25% Quantil am größten, in Gruppe vier überschneidet sich das dritte Quantil mit dem Median.

Beim Vergleich der Maximumpreise differenzieren sich die Gruppen vier und acht von den übrigen, da der Interquartilsabstand hier wesentlich kleiner ist. In Segment
eins, sechs und sieben ist das 50% Quantil gleich dem 75% Quantil. Diese Gruppen filtern nach Fahrzeugen mit hohem Maximalpreis. In den Segmenten zwei, drei, vier, fünf, acht und neun liegt der Median eher im unteren Bereich, was bedeutet, dass Nutzer dieser Gruppe einen niedrigen Maximalpreis setzen. Insgesamt kann man über Gruppe eins, sechs und sieben sagen, dass sich Nutzer dieser Cluster für Fahrzeuge im durchschnittlich oberen Preisbereich interessieren, einen zu anderen Segmenten ähnlichen Minimumpreis angeben und nach einem hohen Maximumpreis filtern. Gruppe vier charakterisiert sich durch die Betrachtung von Fahrzeugen im unteren Preisniveau und sucht im unteren Preisbereich, da nach einem niedrigen Minimumpreis sowie einem niedrigen Maximumpreis gefiltert wird. Nutzer des Clusters acht sehen sich Fahrzeuge mit einem niedrigen Durchschnittswert an, setzen eine hohe Grenze für den Minimumpreis und einen niedrigen Maximumpreis fest.


Abbildung 20 zeigt, aufgeteilt nach Segmenten, den prozentualen Anteil der Nutzer, welche im Laufe ihrer Suche mindestens einmal explizit nach Händlerfahrzeugen gesucht haben (Wert 1) im Vergleich zu denjenigen, welche nicht danach gefiltert haben (Wert 0). Am häufigsten suchen über 25% der Nutzer aus Gruppe acht nach Händlerfahrzeugen. Im Gegensatz dazu steht die übrigen Cluster, in denen nur zwischen
6% und 10% aller Personen mindestens einmal ausschließlich Interesse an Händlerfahrzeugen zeigen.
Abbildung 21 zeigt eine ähnliche Verteilung der „Privatfahrzeuge“ wie die der „Händlerfahrzeuge“.
Hier suchen auch wieder um die 25% der Nutzer, welche sich in Cluster acht befinden, nach Privatfahrzeugen, während verglichen mit Segment sechs am wenigsten Nutzer nach privaten Automobilen filtern.
Insgesamt wird also nicht nach den Filterkriterien „Händler“ und „Privat“ unterschieden, sondern es werden diejenigen Nutzer, welche generell den Filter setzen, in ähnliche Cluster eingeteilt.

Abbildung 22 visualisiert, analog zu den beiden vorherigen Grafiken, den prozentualen Anteil der Nutzer, welche sich im Laufe ihrer Suche mindestens einmal eingeloggt haben.

Abbildung 22: gestapeltes Balkendiagramm - Logins / K-Means

Hier sind kaum Unterschiede zwischen den einzelnen Clustern zu erkennen. Das zweite Segment hat mit circa 5% am meisten Personen, die sich mindestens einmal eingeloggt haben. In den anderen Segmenten sind vergleichweise wenige Nutzer vertreten.

Als Resultat der Clusterunterschiede kann man sagen, dass kaum Unterschiede für die drei quasi-stetigen Variablen zu finden sind. Ursache hierfür ist, dass generell wenige Nutzer nach Händler- oder Privatfahrzeugen filtern und sich auch nur ein
kleiner Anteil mindestens einmal einloggt, was bereits in der deskriptiven Analyse der Daten gezeigt wurde. Diese Variablen wurden dennoch mit in die Segmentierung aufgenommen, da sie in der Analyse der Entscheidungsbäume (Abschnitt 5.4.1 und 5.4.2) als ausschlaggebend erachtet werden.

Abschließend wird auf Unterschiede zwischen der Markensuche in den Segmenten Bezug genommen.

Abbildung 23: Boxplots der Markenunterschiede / K-Means


Abbildung 25 macht deutlich, dass Personen aus Gruppe zwei und acht verstärkt nach den übrigen Marken filtern, jedoch auch nach keinen Marken. Gruppe vier wird charakterisiert dadurch, dass generell nach keiner Marke gefiltert wird, was bedeutet, dass die Marke für diese Nutzer von geringer Bedeutung ist. Cluster sechs sucht vorwiegend nach den übrigen Marken und interessiert sich auch nicht für die Top 10 Marken.

Insgesamt lässt sich sagen, dass die Cluster sich in der Markenpräferenz und -treue stark voneinander unterscheiden.

### 4.3 Die Funktionsweise des Two-Step-Cluster Algorithmus


Die zweistufige Clusteranalyse ist nach SPSS (2006) für große Datenmengen konzipiert und kann im Vergleich zu BIRCH sowohl mit metrischen als auch mit katego-
riellen Variablen umgehen. Wie der Name bereits verrät, wird der Algorithmus in zwei Schritten durchgeführt.


Im ersten Schritt werden alle Objekte sukzessive bis zu den Subclustern in einen „Cluster-Feature“ - Baum (CF-Baum) eingeordnet. SPSS verwendet für die CF-Bäume voreingestellt eine Tiefe von drei und die Knotenanzahl acht. Wie Abbildung 26 zu entnehmen ist, ermöglicht diese Kombination insgesamt maximal $8 \times 8 = 64$ Blätter und jeweils $8 \times 64 = 512$ Sub-Cluster. Dieses grobe Vorgehen ist notwendig, um große Datenmengen bearbeiten zu können. „Mögliche Fehler in der Clusterzuordnung scheinen dabei vertretbar, weil das Ergebnis der ersten Stufe noch nicht abschließend ist und Fehler der ersten Stufe in der zweiten durchaus wieder korrigiert werden können“ [Brosius (2013)].

Abbildung 26: Cluster-Feature Baum der ersten Stufe [Brosius (2013)]

In diesen Subcluster sind die „Cluster Features“ (CF) beziehungsweise Kennzah-
len „Anzahl der Dateneinträge zu jedem Cluster j“: $N_j$, „absolute Häufigkeiten der stetigen Merkmale jeder Kennzahl $N_j$“: $s_j$, „absolute quadrierte Häufigkeiten der stetigen Merkmale jeder Kennzahl $N_j$“: $s_j^2$ und „Anzahl der Dateneinträge, bei denen das k-te kategorielle Merkmal der l-ten Kategorie zu jedem Cluster j angehört“: $N_{jkl}$ enthalten. Der Vektor der Cluster Features hat folgende Form:

$$CF_j = (N_j, s_j, s_j^2, N_{jkl})$$ (6)

Diese Informationen dienen als Grundlage zur Berechnung der Distanzen, welche man zur Einteilung in die Subcluster benötigt. Wie bereits erwähnt, kann dieser Algorithmus stetige und kategoriale Variablen behandeln. Liegen ausschließlich stetige Variablen vor, wird die Distanz zweier Objekte $X$ und $Y$ mit der euklidische Distanz berechnet:

$$d_E(X, Y) = \sqrt{\sum_{i=1}^{v} (X_i - Y_i)^2}$$ (7)

wobei

$v$: Anzahl der in die Analyse mit einbezogenen Variablen

und

$X_i$ beziehungsweise $Y_i$: die Ausprägungen der Variablen $i$ der beiden Objekte $X$ und $Y$

Die euklidische Distanz entspricht der Wurzel aus der Summe der quadrierten Abweichungen zwischen zwei Clustern $X$ und $Y$.

Falls jedoch zusätzlich kategoriale Variablen geclustert werden sollen, wird das Log-Likelihood Distanzkriterium, welches auf der BIRCH-Methode [Zhang et al. (1999)] aufbaut, verwendet.

„Steht man nun vor der Frage, ob ein Fall einem Cluster A oder Cluster B zugeordnet werden sollte, lässt sich für beide Varianten die Log-Likelihood-Distanz berechnen und es wird die Zuordnung vorgenommen, die mit der höchsten Wahrscheinlichkeit (Likelihood) verbunden ist“ [Brosius (2013)]. BIRCH kann ausschließlich stetige Variablen behandeln und verwendet die „Anzahl der Dateneinträge“, die „absoluten Häufigkeiten der stetigen Merkmale“ und die „absoluten quadrierten Häufigkeiten der stetigen Merkmale“, was analog die ersten drei CF für die Two-Step-Clusteranalyse sind.

Der SPSS Algorithmus erweitert BIRCH’s CF, indem die „Anzahl der Dateneinträge, bei denen das k-te kategoriale Merkmal der l-ten Kategorie zu jedem Cluster j..."
angehört“ hinzugefügt wird.
Die genaue Berechnung der Distanz zwischen zwei Beobachtungen, welche kategorielle Ausprägungen enthalten, ist an dieser Stelle nicht näher erläutert und unter Zhang et al. (1999) genauer ausgeführt.
Im weiteren Verlauf wird genauer darauf eingegangen wie der CF-Baum aufgebaut wird. Jede zu klassifizierende Beobachtung wird chronologisch in den Baum eingeordnet und dem Knoten zugewiesen, zu dem die Distanz am kleinsten ist. Der Abstand wird durch das zuvor beschriebene Distanzmaß berechnet.
Um Heterogenität zwischen den Unterclustern zu sichern, wird ein Schwellenwert T, welcher den Distanzwert berücksichtigt, vorgegeben. Die einzuordnenden Beobachtungen dürfen diesen Wert nicht überschreiten.
Ist dies jedoch der Fall, wird der Baum abhängig von den unterschiedlichen Beobachtungen und deren zugehörigen Schwellenwerten T reorganisiert. Grundsätzlich gilt: Je größer T, desto kleiner ist der Entscheidungsbaum, da die Regeln der Clusterzuweisung umso flexibler sind.
Des Weiteren hängt der Aufbau des CF-Baums stark von der Reihenfolge der eingeordneten Punkte ab. Um diesen Effekt zu umgehen, bietet sich eine randomisierte Einordnung an.
Das Ablaufschema zur Aufnahme eines Falles/Objekts in den CF-Baum wird in Abbildung 27 dargelegt.

Nachdem nun der CF-Baum erstellt wurde, werden die Untercluster im zweiten Schritt unter Verwendung der hierarchischen Clusteranalyse zu der gewünschten Anzahl an Gruppen zusammengefasst.
Die Fusion erfolgt durch die schrittweise Zusammenfassung zweier Untercluster, welche gemäß dem Distanzmaß als ähnlich erachtet werden. Der Algorithmus kann zu einer vorgegebenen Clusterzahl zusammenfassen oder eine Anzahl vorschlagen.
Abbildung 27: Ablaufschema der Aufnahme eines neuen Falles in den CF-Baum [Brosius (2013)]
4.4 Two-Step-Clustering in der Anwendung


<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Cluster 5</th>
<th>Cluster 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>7,62%</td>
<td>5,41%</td>
<td>7,04%</td>
<td>15,76%</td>
<td>9,35%</td>
<td>8,79%</td>
</tr>
<tr>
<td>Cluster 7</td>
<td>Cluster 8</td>
<td>Cluster 9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10,41%</td>
<td>15,49%</td>
<td>20,13%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tabelle 6: Resultat Clustering: Verteilung der Nutzer auf alle neun Cluster / Two-Step-Cluster-Component

Man kann, wie bei K-Means auch, drei größere Segmente (Cluster vier, acht und neun), drei mittelgroße (Cluster fünf, sechs und sieben), und drei kleine Gruppen (Cluster eins, zwei und drei), erkennen.

Um die beiden Algorithmen gegenüberstellen zu können, werden analog zu K-Means erneut die Boxplots für die Variablen der unterschiedlichen Cluster verglichen. Alle dazu produzierten Grafiken finden sich im Anhang B.


Im nächsten Schritt werden analog zur Analyse der K-Means Ergebnisse aufs Neue


Beim Vergleich der Markenpräferenzen werden, wie bei K-Means auch, deutliche Unterschiede zwischen den Gruppen klar. Cluster zwei und vier interessiert sich primär für die Top Vier Marken VW, Mercedes, BMW und Audi, sucht jedoch auch nach Fahrzeugen ohne Markeneinschränkung.


Zur besseren Übersicht kann man neben Anhang B auch Tabelle 7 entnehmen, wie sich verschiedene Variablenwerte innerhalb der Cluster unterscheiden.
<table>
<thead>
<tr>
<th>Aktivität</th>
<th>Initialfilter</th>
<th>Anteil Detailseiten</th>
<th>Anteil Suchseiten</th>
<th>Anteil Emails</th>
<th>Anteil Merkzettel</th>
<th>Durchschnittspreis</th>
<th>Marken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>-</td>
<td>-</td>
<td>o</td>
<td>o</td>
<td>-</td>
<td>-</td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ford, Opel, übrige und restliche Marken</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VW, BMW, Mercedes, Audi, übrige und restliche Marken</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>-</td>
<td>-</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Skoda, Toyota, Renault, Peugeot, übrige Marken</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>+</td>
<td>-</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VW, BMW, Mercedes, Audi, Ford, Opel, übrige und restliche Marken</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>-</td>
<td>-</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VW</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>-</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mercedes</td>
</tr>
<tr>
<td>Cluster 7</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>keine Marken</td>
</tr>
<tr>
<td>Cluster 8</td>
<td>-</td>
<td>-</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>BMW, Audi</td>
</tr>
<tr>
<td>Cluster 9</td>
<td>-</td>
<td>+</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>übrige Marken</td>
</tr>
</tbody>
</table>

Tabelle 7: Ausprägungen ausgesuchter Variablenwerte innerhalb der Cluster, 
+: hoher Wert, o: mittel, -: niedrig
5 Ex-Post Analyse

5.1 K-Means

Im Folgenden werden die neun Cluster auf vier in der Praxis relevante Nutzergruppen reduziert, da sich diese in den wichtigsten Merkmalen, wie der relative Anteil an Seitentypen oder dem Aktivitätsmaß nicht wesentlich voneinander unterscheiden. Ergänzend wird die Interpretation von neun Nutzergruppen, welche sich nur durch marginale Feinheiten von den übrigen distanzieren, als zu komplex erachtet und aus diesem Grund erscheint die Einschränkung auf vier Segmente, welche sich im Wesentlichen unterscheiden, sinnvoller.


Der Anteil der Nutzer in den Gruppen ist in folgender Tabelle dargestellt.

<table>
<thead>
<tr>
<th>Orientierungsphase</th>
<th>aktive Suche</th>
<th>Entscheidungsphase</th>
<th>Fahrzeugliebhaber</th>
</tr>
</thead>
<tbody>
<tr>
<td>35,15%</td>
<td>12,74%</td>
<td>4,30%</td>
<td>47,81%</td>
</tr>
</tbody>
</table>

Tabelle 8: Clusterzuweisung des K-Means Algorithmus

Die Cluster sind farblich getrennt, wie es in der jeweiligen Legende dargestellt ist.

\[\text{unternehmensinterne Auswertung (basierend auf qualitativen Interviews)}\]
Abbildung 28: Koordinatenplot der „Orientierungsphase“/ K-Means

In Abbildung 28 werden Cluster drei, vier, fünf und neun charakterisiert durch Nutzer, die in der Suchphase noch am Anfang zu stehen und sich orientieren. Der relative Anteil an Suchseiten ist im Vergleich zu Detailseiten, Mailversand und dem Anlegen von Merkzetteln etwas höher als der Durchschnitt, was vermuten lässt, dass sich diese Nutzer zunächst einen groben Überblick verschaffen möchten.


Es werden außerdem wenige Initialfilter festgelegt, was darauf schließen lässt, dass diese Personengruppe noch keine konkrete Vorstellung vom gewünschten Fahrzeug hat, sondern sich in dieser Phase noch „durchklickt“. Diese Personen filtern auch noch nicht unter- oder überdurchschnittlich oft nach Händler- oder Privatfahrzeugen. Des Weiteren wird mit Ausnahme einzelner Cluster und Marken unterdurchschnittlich selten nach den Top 10 Marken, aber auch nach keinen und übrigen Marken gesucht.

1 Wird aus rechtlichen Gründen nicht veröffentlicht und kann beim Autor angefragt werden
Abbildung 29 veranschaulicht Cluster acht. Dieses Segment spiegelt Nutzer wider, welche aktiv auf der Fahrzeugsuche sind.

Personen dieses Clusters differenzieren sich von der vorherigen Gruppe durch eine hohe Präsenz auf der Plattform, da sie sowohl überdurchschnittlich viele Aktionen pro Tag tätigen, als auch im Verhältnis zu ihrer Zeitspanne an vielen Tagen die Webseite besuchen.

Der prozentuale Anteil an Suchseiten ist wie in der Anfangsphase leicht überdurchschnittlich, während der Anteil an Detailseiten, Emails und Merkzetteln leicht unterdurchschnittlich ist.

Des Weiteren werden auch wenige Initialfilter gesetzt, was darauf schließen lässt, dass diese Nutzer immer noch keine klare Vorstellung vom gewünschten Automobil haben.

Außerdem loggen sich diese Nutzer im Vergleich zum Durchschnitt oft ein, was ein Hinweis darauf ist, dass sie Kontakt zum Verkäufer aufnehmen möchten. Dies ist nur möglich, sofern man eingeloggt ist.

Zusätzlich wird im Vergleich zu der vorherigen Phase ein hoher Minimumpreis und ein niedriger Maximumpreis festgelegt, was bedeutet, dass diese Nutzer in einem konkreteren Preisintervall suchen.

Überblickend zeichnen sich Nutzer dieser Phase dadurch aus, im Vergleich zu Personen, welche sich in der Anfangsphase befinden, öfter und intensiver die Plattform
zu nutzen, ein kleineres Preisintervall festzulegen und sich überdurchschnittlich oft einzuloggen.

Im weiteren Verlauf werden die Nutzer beschrieben, die in der Entscheidungsphase weit fortgeschritten sind und vermutlich kurz vor dem Fahrzeugkauf stehen.

Abbildung 30: Koordinatenplot der „Entscheidungsphase“ / K-Means

Wie Abbildung 30 zu entnehmen ist, sind diese Nutzer im Vergleich zu denjenigen der aktiven Suche nicht mehr so stark, jedoch trotzdem noch überdurchschnittlich oft, auf der Webseite aufzufinden, da nur noch gezielt nach einer kleinen Grundgesamtheit an Fahrzeugen gesucht wird.


Es wird, wie in der aktiven Suche, ein überdurchschnittlich hoher Minimum- und unterdurchschnittlich tiefer Maximumpreis eingestellt, was auch wieder auf eine sehr klare Preisvorstellung schließen lässt.

Insgesamt wird diese Phase also, im Vergleich zu der aktiven Suche, durch den Versand überdurchschnittlich vieler Emails und Kontaktanfragen sowie durch das Anlegen vieler Merkzettel charakterisiert. Analog zur aktiven Suche besitzen diese
Nutzer einen hohen Aktiv-Score und tätigen vergleichsweise viele Aktionen.

Der vierte Nutzertyp besteht aus „Fahrzeugliebhabern“ beziehungsweise Automobilenthusiasten.


Zusammenfassend kann man sagen, dass dieser Nutzertyp nicht daran interessiert ist, die angesehenen Fahrzeuge zu kaufen, sondern diese „Wunsch-/Traumfahrzeuge“ nur betrachten möchte.

Abbildung 32: Heatmap / K-Means


Segment acht, welches Nutzer der aktiven Suche enthält, und zwei, welches Personen in der Entscheidungsphase repräsentiert, bilden jeweils eigene Nutzergruppen. Diese heben sich durch die Messzahlen der Aktivität, der verschiedenen Seitentypen und der Filtereinstellungen von den übrigen Clustern ab.

Die vierte Gruppe, die Fahrzeugenthusiasten, besteht aus Cluster eins, sechs und

### 5.2 Two-Step-Clustering


<table>
<thead>
<tr>
<th>Orientierungsphase</th>
<th>aktive Suche</th>
<th>Entscheidungsphase</th>
<th>Fahrzeugliebhaber</th>
</tr>
</thead>
<tbody>
<tr>
<td>34,30%</td>
<td>14,76%</td>
<td>6,40%</td>
<td>44,54%</td>
</tr>
</tbody>
</table>

Tabelle 9: Clusterzuweisung des Two-Step-Cluster Algorithmus

Die durchschnittlichen Variablenwerte für jede Phase unterscheiden sich nicht wesentlich zwischen den beiden zu vergleichenden Algorithmen. Aus diesem Grund wird an dieser Stelle auf Anhang D verwiesen, in dem die Koordinatenplots zu finden sind und lediglich überblickend im Folgenden auf die Heatmap eingegangen, welche die Ähnlichkeiten innerhalb einzelner Cluster und die Unterschiede zwischen den Gruppen verdeutlicht.

5.3 Vergleich K-Means und Two-Step-Clustering


<table>
<thead>
<tr>
<th>k/t</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>∑</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>87,52%</td>
<td>9,40%</td>
<td>1,19%</td>
<td>1,89%</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>17,99%</td>
<td>73,77%</td>
<td>2,91%</td>
<td>5,33%</td>
<td>100%</td>
</tr>
<tr>
<td>3</td>
<td>2,23%</td>
<td>0,35%</td>
<td>97,23%</td>
<td>0,19%</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>2,40%</td>
<td>6,30%</td>
<td>0,93%</td>
<td>90,37%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Tabelle 10: Klassifikationssvergleich beider Clusteralgorithmen, k:K-Means, t:Two-Step-Cluster Algorithmus
Die grüne Diagonale stellt eine übereinstimmende Clusterzuweisung beider Algorithmen dar. 87,52% aller Nutzer aus Cluster Eins des K-Means Algorithmus werden auch durch den zweistufigen Clusteralgorithmus als Käufer in der frühen Phase klassifiziert. 8059 Personen, welche 73,77% der durch K-Means dem zweiten Cluster zugeordneten Personen entsprechen, werden auch vom Two-Step-Cluster-Verfahren als Nutzer in der aktiven Suche befunden. 97,23% der durch K-Means als Nutzer kurz vor dem Fahrzeugkauf beurteilten Personen werden auch durch die Two-Step-Clustermethode als solche eingestuft. 90,37% aller sich durch K-Means in Gruppe vier befindlichen Personen, welche 41.000 entsprechen, klassifiziert der Two-Step-Cluster Algorithmus auch als Fahrzeugliebhaber.


Des Weiteren ist noch eine mangelnde Trennschärfe zwischen Cluster zwei und vier festzustellen. 6,3% der Nutzer, die von K-Means als Fahrzeugliebhaber befunden werden, weist Two-Step-Clustering dem zweiten Segment, welches die aktive Suche darstellt, zu. Andererseits werden von denjenigen Personen, welche K-Means in die aktive Suche einteilt, circa 5% vom zu vergleichenden Algorithmus als Fahrzeugliebhaber eingestuft.

Insgesamt werden circa 87% aller Personen von beiden Algorithmen in die gleiche Klasse eingeordnet, während um die 13% nicht eindeutig von beiden Clusteralgorithmen segmentiert werden.

In Allgemeinen lässt sich zu den übereinstimmenden Ergebnissen sagen, dass sich anhand der beiden Clusteralgorithmen circa 30% aller Nutzer in der frühen Kaufphase, 9% in einer Zwischenentwicklung und um die 4% in einem fortgeschrittenen Entscheidungsabschnitt befinden. Bemerkenswert ist, dass fast die Hälfte aller Beobachtungen als Fahrzeugenthusiasten gesehen werden, welche nicht am Kauf interessiert sind.

Übergreifend kann man sagen, dass die Klassifikationen der beiden voneinander unabhängigen Algorithmen sehr ähnlich sind. Die vier Segmente unterscheiden sich im Wesentlichen deutlich und ein Großteil der Nutzer kann klar in einzelne Gruppen eingeteilt werden.
5.4 Random Forest


Grundsätzlich besteht der Random Forest aus mehreren unkorrelierten Entscheidungsbäumen. Für jeden Entscheidungsbaum werden Beobachtungen aus einer Bootstrap-Lernstichprobe verwendet und für jeden Knotenpunkt stehen üblicherweise nur \( m \approx \sqrt{p} \) Einflussvariablen zur Verfügung, wobei \( p \) die Summe aller Einflussvariablen ist [Hastie et al. (2013)]. Die Stichprobe des Bruchteils an Prediktoren wird an jedem Knoten neu bestimmt. Die zufällige Auswahl an Entscheidungsknoten stellt sicher, dass die einzelnen Bäume nicht wiederholt aus den gleichen und wichtigsten Startvariablen aufgebaut werden, wie es bei „bagged-trees“ der Fall ist [Hastie et al. (2013)].

Durchschnittlich werden nach Hastie et al. (2013) \( \frac{(p - m)}{p} \) Knoten die wichtigsten Zielvariablen nicht enthalten. Folglich schließt diese Vorgehensweise die Korrelation einzelner Bäume aus.

In Abbildung 34 wird noch einmal verdeutlicht wie sich die Korrelation zwischen den Bäumen in Abhängigkeit der Split-Variablen \( m \) verändert.

Abbildung 34: Korrelation zwischen Baumpaaren abhängig von den zufällig ausgewählten Split-Variablen \( m \) [Hastie et al. (2009)]
Wie man Abbildung 35 entnehmen kann besteht der Vorteil von Random Forest daraus, dass die Varianz mit weniger Variablen $m$, welche an jedem Knotenpunkt zur Verfügung stehen, reduziert wird.

Zu beachten ist bei dieser Grafik, dass die Werte der Varianz rechts abgetragen sind, während Werte des MSE und Bias links zu finden sind.

Zumal der Erwartungswert eines Baums gleich dem Erwartungswertes des Durchschnitts aller Bäume ist, verringert sich der Bias mit absteigender Split-Variablenanzahl nicht, was ebenfalls aus Abbildung 36 ersichtlich werden soll.

Da sich die mittlere quadratische Abweichung (MSE) aus Varianz plus $Bias^2$ berechnet, verläuft diese auch im oberen Bereich für wenige Split-Variablen.

Abbildung 35: Varianz-Bias-MSE-Vergleich des Random-Forest Ensemble [Hastie et al. (2009)]

In Abbildung 36 wird exemplarisch ein Baum des Ensembles zur Klassifikation eines Nutzers anhand von drei Variablen verdeutlicht.

Die wichtigste Variable, welche den ersten Knotenpunkt darstellt, ist „relativer Anteil Bookmarks“. Falls mehr als 4% aller Aktionen für einen Nutzer das Anlegen von Merkzetteln ausmacht, wird dieser als Person in einer fortgeschrittenen Kaufphase klassifiziert.
Abbildung 36: exemplarischer Entscheidungsbaum zur Nutzereinteilung

Ist dies nicht der Fall wird zusätzlich die zweite Variable „Aktiv-Score“ betrachtet. Handelt es sich um einen Nutzer, der einen hohen Score besitzt, welcher den Wert 3,5 übersteigt, wird dieser in die aktive Suche eingeteilt. Hat diese Beobachtung jedoch einen prozentualen Anteil an Merkzetteln, welcher 4% nicht übersteigt und einen Aktiv-Score unter oder gleich 3,5, bezieht man eine dritte Variable, das Filterkriterium Maximumpreis, mit ein. Hat die zu klassifizierende Beobachtung im Laufe ihrer Suche eine minimale Obergrenze von 30.000€ überstiegen, wird sie den „Fahrzeugenthusiasten“ zugeordnet. Falls dies nicht zutrifft, wird der Nutzer in die frühe Kaufphase eingeordnet.

Um nun einen Nutzertyp zu klassifizieren, werden alle Entscheidungsbäume mit den Eigenschaften dieses Nutzers einzeln analysiert. Eingeordnet wird nach dem Mehrheitsprinzip, folglich wird die Klasse, welche am häufigsten aus den einzelnen Bäumen resultiert, derjenigen Person zugewiesen.

In den nachfolgenden Abschnitten 5.4.1 und 5.4.2 wird versucht, die Klassifikation des K-Means beziehungsweise Two-Step-Cluster Algorithmus durch ein Random Forest Modell vorherzusagen.
5.4.1 Vorhersagekraft auf Basis der K-Means Ergebnisse


<table>
<thead>
<tr>
<th>k/rf</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97,30%</td>
<td>0,96%</td>
<td>0,30%</td>
<td>1,44%</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>8,65%</td>
<td>80,88%</td>
<td>0,43%</td>
<td>10,04%</td>
<td>100%</td>
</tr>
<tr>
<td>3</td>
<td>3,15%</td>
<td>2,55%</td>
<td>90,74%</td>
<td>3,56%</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>0,94%</td>
<td>0,73%</td>
<td>0,15%</td>
<td>98,18%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Tabelle 11: Klassifikationsvergleich K-Means - Random Forest, k:K-Means, rf: Random Forest

Der Tabelle ist zu entnehmen, dass Random Forest die Beobachtungen im Großteil richtig klassifiziert. Wie beim Vergleich zwischen den beiden Clusteralgorithmen K-Means und Two-Step-Clustering, existieren die meisten Fehlklassifikationen sowohl zwischen Cluster eins und zwei als auch zwischen Segment zwei und vier. Durch Random Forest werden insgesamt 95,33% aller Beobachtungen richtig klassifiziert, wohingegen 1.602 Personen, welche circa 4% ausmachen, falsch segmentiert werden.


Um diese Frage zu beantworten wird nun Abbildung 37 betrachtet, welche die Wichtigkeit der Variablen im Random Forest Ensemble darstellt.
Abbildung 37: Variablenwichtigkeit / K-Means

Die Grafik zeigt jede Variable auf der y-Achse und die „mean decrease accuracy“ auf der x-Achse. Je höher der Wert der x-Achse für die jeweiligen Variablen, desto wichtiger sind diese für die Genauigkeit des Modells und desto mehr Einfluss hat diese darauf, die Fehlerrate zu erhöhen, falls die Variable nicht mit eingeht. Die wichtigsten Variablen, welche weit oben im Baum vorkommen, sind die relative Häufigkeit an Merkzetteln und Mails, ob nach Privat- oder Händlerfahrzeugen gefiltert wird, sowie die Anzahl der Aktionen pro Tag, als auch der Aktiv-Score.
5.4.2 Vorhersagekraft auf Basis der Two-Step-Clustering Ergebnisse


<table>
<thead>
<tr>
<th>K-Means</th>
<th>Two-Step-Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relativer Anteil der Merkzettel</td>
<td>Filter: Händlerfahrzeuge</td>
</tr>
<tr>
<td>Relativer Anteil der Mails</td>
<td>Relativer Anteil der Merkzettel</td>
</tr>
<tr>
<td>Aktiv-Score</td>
<td>Minimumpreis</td>
</tr>
<tr>
<td>Filter: Privatfahrzeuge</td>
<td>Aktiv-Score</td>
</tr>
<tr>
<td>Alle Aktionen</td>
<td>Filter: Privatfahrzeuge</td>
</tr>
</tbody>
</table>


6 Fazit und Handlungsempfehlungen

Zusammenfassend lässt sich die Stichprobe der 85.743 beobachteten Nutzer durch vier Segmente beschreiben. Beide Algorithmen resultieren in annähernd identische Nutzergruppen, wovon drei die „consumer journey“, die jede Person vor dem Fahrzeugkauf durchläuft, und das vierte Segment die Fahrzeugenthousiasten beschreibt. Das erste Segment, welches circa 35% der Grundgesamtheit ausmacht, wird charakterisiert durch Personen, die sich noch ganz am Anfang ihrer Suche befinden und sich einen groben Überblick über die Auswahl verschaffen möchten. Zweite Gruppe hat die erste Phase schon durchlaufen und differenziert sich von den Übrigen durch genauere Vorstellungen vom gewünschten Fahrzeug. Diese Nutzer befinden sich bereits in einer aktiven Suche. Insgesamt weist diese Gruppe einen Anteil um die 15% aller
untersuchten Nutzer auf. Das dritte Cluster enthält mit einem prozentualen Anteil von 5% am wenigsten Nutzer, welche jedoch als am Wichtigsten erachtet werden: Diese Personen haben die ersten beiden Phasen bereits überwunden und befinden sich in einer Entscheidungsphase kurz vor dem Kauf eines Fahrzeugs. Theoretisch muss jeder kaufinteressierte Nutzer im Laufe seiner Suche die ersten drei Phasen mindestens einmal durchlaufen, kann aber jederzeit wieder in eine der ersten beiden Phasen zurückfallen, falls sich seine Vorstellungen ändern.

Personen des vierten und größten Segments, anteilig 45%, können als Fahrzeugliebhaber beschrieben werden, welche ausschließlich daran interessiert sind, Automobile anzusehen und (noch) nicht vorhaben, diese auch zu erwerben. Die Ergebnisse der Segmentierung werden in Abbildung 38 verdeutlicht.

Resultate der Nutzersegmentierung

Abbildung 38: Grafische Darstellung der vier Nutzergruppen

In Abschnitt 5.4.1 und 5.4.2 wurden die wichtigsten Variablen zur Bestimmung neuer Nutzer ermittelt, woraus resultierte, dass die Variablen „relativer Anteil an Merkzetteln“, „Filter: Privatfahrzeuge“ und „Aktiv-Score“ bei beiden zu vergleichenden Algorithmen unter den fünf wichtigsten Variablen vertreten sind und somit einen großen Einfluss besitzen. Aufbauend darauf könnten beispielsweise für die zehn wich-
Literatur


Tabellenverzeichnis

1  Rohdatensatz mit mehreren Zeilen pro Nutzer .......................... 7
2  Berechnung des Maßes für aktive Tage .................................. 9
3  Transformierter Datensatz mit einer Zeile pro Nutzer .................. 11
4  Werte des 99%-Quantils ausgewählter Variablen ....................... 19
5  Resultat Clustering: Verteilung der Nutzer auf alle neun Cluster / K-Means ................................................................. 24
6  Resultat Clustering: Verteilung der Nutzer auf alle neun Cluster / Two-Step-Cluster-Component ................................. 36
7  Ausprägungen ausgesuchter Variablenwerte innerhalb der Cluster, +: hoher Wert, o: mittel, -: niedrig ........................................ 38
8  Clusterzuweisung des K-Means Algorithmus ............................. 39
9  Clusterzuweisung des Two-Step-Cluster Algorithmus ................... 45
10 Klassifikationssvergleich beider Clusteralgorithmen, k:K-Means, t:Two-Step-Cluster Algorithmus ........................................... 46
11 Klassifikationsvergleich K-Means - Random Forest, k:K-Means, rf: Random Forest ................................................................. 51
12 Auflistung der fünf wichtigsten Variablen basierend auf Ergebnissen des K-Means- und Two-Step-Cluster Algorithmus ................... 53

Abbildungsverzeichnis

1  Grafische Darstellung der Nutzersegmentierung .......................... 1
2  Grafische Darstellung eines Hadoop-Clusters und dessen Komponenten 4
3  Map-Reduce Schritte [van Groningen (2009)] ............................ 5
4  Histogramm der Variable „alle Aktionen“ .......................... 13
5  Boxplots der Variablen „Detailseiten“ und „Suchseiten“ .............. 14
6  Histogramm der Variable „Merkzettel“ .......................... 14
7  Histogramm der Variable „Mails“ .................................. 14
8  Histogramm der Variable „Aktiv-Score“ .......................... 15
9  Histogramm der Variable „Initialfilter“ .................................. 15
10 Histogramm der Variable „Durchschnittspreis“ .......................... 16
11 Histogramm der Variable „Minimumpreis“ .......................... 17
12 Histogramm der Variable „Maximumpreis“ .......................... 17
13 Balkendiagramm der Verkaufstypen .................................. 18
14 Balkendiagramm der Variable „VW“ .................................. 19

59
sukzessive Iteration des K-Means Clusteralgorithmus für simulierte Daten [Hastie et al. (2009)] .................................................. 22
Bestimmung der Clusteranzahl anhand des Ellbogenkriteriums .... 24
Boxplots der Variablen „alle_Aktionen“, „aktiv_Score“ & „Filter_Initialsuche“ / K-Means .......................................................... 25
Boxplots der vier Seitentypen / K-Means .................................... 26
Boxplots der Variablen „Durchschnitts-“, „Minimum-“ und „Maximumpreis“ / K-Means ....................................................... 27
gestapeltes Balkendiagramm der Variable „Händler“ / K-Means ... 28
gestapeltes Balkendiagramm der Variable „Privat“ / K-Means ..... 28
gestapeltes Balkendiagramm - Logins / K-Means ....................... 29
Boxplots der Markenunterschiede / K-Means ............................ 30
Boxplots der Top 6 bis 10 Marken / K-Means ............................. 31
Boxplots der übrigen Marken & keine Marken / K-Means ........... 31
Cluster-Feature Baum der ersten Stufe [Brosius (2013)] ............. 32
Ablaufschema der Aufnahme eines neuen Falles in den CF-Baum [Brosius (2013)] ......................................................... 35
Koordinatenplot der „Orientierungsphase“ / K-Means ............... 40
Koordinatenplot der „aktiven Suche“ / K-Means ....................... 41
Koordinatenplot der „Entscheidungsphase“ / K-Means ............... 42
Koordinatenplot der „Fahrzeugliebhaber“ / K-Means ................. 43
Heatmap / K-Means ................................................................ 44
Heatmap / Two-Step-Cluster Algorithmus ................................. 46
Korrelation zwischen Baumpaaren abhängig von den zufällig ausge-wählten Split-Variablen $m$ [Hastie et al. (2009)] ....................... 48
Varianz-Bias-MSE-Vergleich des Random-Forest Ensemble [Hastie et al. (2009)] ................................................................. 49
exemplarischer Entscheidungsbaum zur Nutzererreichung ........... 50
Variablenwichtigkeit / K-Means .............................................. 52
Grafische Darstellung der vier Nutzergruppen ......................... 54
Appendix

A  Hive-Befehl zum Abfragen der Basisdaten aus dem Hadoop Cluster

```r
username <- 'ybarth'
system(paste('kinit ', username, '@AS24.LOCAL -k -t ~/. ', username,
               '.keytab && echo 'kinit done' ||
               echo 'kinit failed', sep = ' '))
library(RODBC)
library(sqldf)
# ODBC-Verbindung herstellen
Hive <- odbcConnect('Hive', readOnly = TRUE)
sqlQuery(Hive, "use temp")
sqlQuery(Hive, "show tables")
# Daten in eine Tabelle schreiben
userData <- sqlQuery(Hive,"CREATE TABLE
hadoop_user_data_yvonne
STORED AS PARQUET
AS
SELECT
sdate, stime, csuristem, csuriquery, csreferer, cshost, as24visitorguid
FROM weblogs.iislog_archive
WHERE time >= 20150503
AND time <= 20150517
AND as24visitorguid IN (SELECT as24visitorguid
FROM weblogs.iislog_archive
WHERE time = 20150510
AND cshost LIKE '%autoscout24.de'
AND as24visitorguid != ''
LIMIT 300000)
")
```

61
# relevante URL's aus der Tabelle extrahieren

```r
sqlQuery(Hive, "
SELECT sdate, stime, as24visitorguid, csuristem, 
csuriquery, csreferer, cshost
FROM temp.hadoop_user_data_yvonne
WHERE (csuristem = '/GN/CountV1.ashx' OR 
   csuristem LIKE '/classified/%' OR 
   csuristem = '/sendcontactmail/contact' OR 
   csuristem LIKE '/Parkdeck/Add/%' OR 
   csuristem = '/ArticleList/GetCounters' OR 
   csreferer LIKE '%login%' OR 
   csuristem = '/offerb2c/data/NewDecision/Taxonomy/GetVehicleIdentificationData' OR 
   csuriquery LIKE '%tabg=guidedfull')
ORDER BY as24visitorguid"
)
```

# Testen, wieviele unterschiedliche Nutzer in Datensatz vorhanden

test <- sqldf('SELECT COUNT(DISTINCT as24visitorguid) FROM userData')

# Dataframe im csv-Format speichern

```r
write.csv2(userData, file = '/data/R/export/yvonne_BA.csv')
```

62
B Filterhäufigkeit einzelner Marken
C  Variablenunterschiede einzelner Cluster des Two-Step-Cluster Algorithmus
einzelne Cluster
z-standardisierter Wert
variable
rel_Suchen
rel_Detail
rel_Bookmarks
rel_Mails
rel_Suchen rel_Detail rel_Bookmarks rel_Mails
0
5
10
1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9

66
<table>
<thead>
<tr>
<th>einzelne Cluster</th>
<th>z-standardisierter Wert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Durchschnittspreis</td>
<td>Minimumpreis</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

variable
- Durchschnittspreis
- Minimumpreis
- Maximumpreis
Clusterprozentuale Verteilung

Filter:Privat

Wert
0
1

Cluster

prozentuale Verteilung

0%
25%
50%
75%
100%

D Koordinatenplots der vier Nutzergruppen des Two-Step-Cluster Algorithmus
E  Klassifikationsvergleich K-Means - Random Forest mit veränderten Parametern

E.1 Anzahl Bäume: 500, Split-Variablen: 8

<table>
<thead>
<tr>
<th>k/rf</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97,12%</td>
<td>1,18%</td>
<td>0,32%</td>
<td>1,38%</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>7,38%</td>
<td>83,72%</td>
<td>0,48%</td>
<td>8,42%</td>
<td>100%</td>
</tr>
<tr>
<td>3</td>
<td>3,23%</td>
<td>2,14%</td>
<td>91,41%</td>
<td>3,22%</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>0,93%</td>
<td>0,87%</td>
<td>0,15%</td>
<td>98,05%</td>
<td>100%</td>
</tr>
</tbody>
</table>

E.2 Anzahl Bäume: 800, Split-Variablen: 5

<table>
<thead>
<tr>
<th>k/rf</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97,28%</td>
<td>0,97%</td>
<td>0,30%</td>
<td>1,45%</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>8,4%</td>
<td>91,20%</td>
<td>0,43%</td>
<td>9,97%</td>
<td>100%</td>
</tr>
<tr>
<td>3</td>
<td>3,15%</td>
<td>2,55%</td>
<td>90,81%</td>
<td>3,49%</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>0,93%</td>
<td>0,73%</td>
<td>0,14%</td>
<td>98,20%</td>
<td>100%</td>
</tr>
</tbody>
</table>

F  Klassifikationsvergleich Two-Step-Clustering - Random Forest

F.1 Anzahl Bäume: 500, Split-Variablen: 5

<table>
<thead>
<tr>
<th>t/rf</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97,49%</td>
<td>1,04%</td>
<td>0,42%</td>
<td>1,05%</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>4,23%</td>
<td>90,05%</td>
<td>0,65%</td>
<td>5,07%</td>
<td>100%</td>
</tr>
<tr>
<td>3</td>
<td>2,73%</td>
<td>3,90%</td>
<td>90,43%</td>
<td>2,94%</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>0,80%</td>
<td>1,03%</td>
<td>0,30%</td>
<td>97,87%</td>
<td>100%</td>
</tr>
</tbody>
</table>
G  Variablenwichtigkeit / Two-Step-Clustering