Predictive Analytics Models and Collaborative Filtering

Matthieu Heinzelmann
Prof. Dr. Peter Bug

Reutlingen University, Reutlingen 2016

Abstract

In times of e-commerce and digitalization, new markets are opening, young companies have the possibility to grow and new perspectives arise in terms of **customer relationship**. Customers require more possibilities of personalization. In the same time, companies have access to new and especially more information about the customer. Seems like it was a correlation that could evolve greatly if there weren't **privacy issues**. Vast amount of data about consumers are collected in **Big Data** warehouses. These shall be analyzed via **predictive analytics** and customers shall be classified by algorithms like **clustering models**, **propensity models** or **collaborative filtering**. All these subjects are growing in importance, as they are shaping the global **marketing landscape**. Marketers develop together with IT scientists new ways of analyzing customer databases and benefit from more accurate **segmentation methods** as that have been used until now. The following paper shall provide a literature review on new methods of consumer segmentation regarding the high inflow of new information via e-commerce. It will introduce readers in the subject of predictive analytics and will discuss several predictive models. The writing of the paper is not based on own empirical researches, but shall serve as a reference text for further researches. A conclusion will complete the paper.

List of Contents

List of Tables

List of Figures

Introduction

As the global market is getting more and more competitive and relentless, companies are searching desperately for ways to get more information about their (potential) customers. Whether through their habits, needs or identity, they try hard to approximate to each individual. This is why consumer segmentation is a core subject for every company that has to classify their clients. In Addition, there is an increasing pressure in companies to make marketing accountable. Around three decades ago, consumer segmentation was a subject of simple data ascertainment and based on the decision of few marketers. In the early 90s e-commerce spread and a new dimension of customer data got available. It is the talk of the development of predictive analytics. A process of classifying your customers, that is undertaken by high performance programs and key technologies, which are created by humans, but work more efficient than humans. Due to new segmentation landscapes like digitalization or social media and modern technology, new consumer segmentation methods are being developed and will have a significant impact on global industry and retailing issues. However present technologies are already pushing the boundaries of information accumulation. The following paper shall provide a literature review on **new methods of consumer segmentation**. It is focusing on the **high inflow of new consumer data** via e-commerce and **how companies can make use of this data via predictive analytics and new marketing models**. It will first introduce the reader in general knowledge about predictive analytics. This will be followed by a the core chapter about predictive models. Then, a discussion returns to the title subject and the review is concluded.

1 Predictive Analytics and New Technology

In the last twenty years, a large "... growth in data about ..." consumers is being observed, "...making previously invisible consumer behavior pattern visible..." (Falcon, Gorbis, Jeffery, & Spalding, 2003, p. 9). This phenomenon has started with the birth of E-Commerce and the spread of collaborative filtering. In the early 1990s, this technology based on computing algorithms "...[arose] as a solution for dealing with overload in online information spaces" (Ekstrand, Riedl, & Konstan, 2011, p. 84). It is one of three predictive analytics models that were made a subject of discussion in chapter 2. Collaborative filtering is commonly used for recommendations in online stores. Especially online retailer Amazon is known for making this model popular to enhance its marketing strategy (Levin, 2015). The other predictive analytics models you will find in chapter 2 are clustering models and propensity models. All three have in common, to deal with data warehouses and to bring order into a high inflow of information.

As new segmentation methods are not only developed by inventing and shaping new Models, but also as an answer to new developments in technology, we have to start researching with where you get new information from and how you get new information about your customers. Therefore it is important, to catch up with new methods of data ascertainment and how this data is organized. The first of the following chapters shall give you a short introduction into modern technology of data collection. The second part will inform you about Data organization.

1.1 New Technologies Provide More Information

There are technologies, which were especially developed to collect information. In retail business, this information is mainly given by the customer - whether directly or indirectly. We distinguish between technology that captures "... data [...] from people (for example, from on-line transactions and social networks) and ..." technology that captures data through "... sensors (for example, from GPS mobile devices) ..." (Guazzelli, 2012a). There are many technologies already existing and used. As this paper shall not be focusing on technology, Table 1 only shows some technologies and their definitions, that have a high impact on the process of consumer segmentation and that are relevant for collecting data. They are part of a listing, the Institute for the Future has designated as "... key emerging technologies ...", (Falcon et al., 2003, pp. 12–13):

Table 1: Key Emerging Technologies that are relevant for collecting data

Biometrics	"Biometrics allow companies to capture biological data about consumers ... that

	could be used to either identify individual consumers, or ..." to measure the emotional state of particular customers.
Collaborative filtering	"Collaborative filtering technologies (also known as recommender systems) use product evaluations, stated tastes, clickstreams, transaction histories, or other data about specific individuals to create a user profile that can be compared to others' profiles to generate customized cross-selling or upselling opportunities."
Peer-to-peer architectures	"Peer-to-peer (P2P) architectures demonstrate a computing model in which there is truly no center: storage and processing resources are distributed over the Internet ... However, P2P architectures will be important for capturing consumer data from distributed sensors and for managing the massive amounts of data that sensors will generate in the future."
Physical tagging	Physical tagging informs companies about an objects movement and allows "... to monitor products and processes more closely. Radio frequency identification (RFID) tags which communicate to a receiver through short-range radio waves will allow products to efficiently contain large amounts of updateable information.
Positioning technologies	"Positioning technologies will allow companies to obtain more information on the geographical position of people or things.

	Networked-based and satellite-based systems ... can be used to help understand shopping patterns as well as communicate location-relevant messages."
Sensors	"New [smaller and cheaper] sensors will allow companies to cheaply and easily identify and track consumers, and continuously evaluate the quality of products."
Soft tags	Soft tags will generate electronic identities for pieces of data that are standardized and sharable across systems and companies. Technologies such as XML will provide a method to track and access specific data, thereby facilitating data exchange across formats and languages.
Wireless technologies	"Pervasive wireless technologies [(like cellulsr phone networks, Bluetooth or Wi-Fi)] will allow companies and consumers alike to move information faster, easier, and into new places and spaces unhindered by the need for physical connections."

Adapted from: (Falcon et al., 2003, pp. 12–13)

1.2 Big Data and Data Mining

When talking about data collection, it is not only interesting to know "how" but also to know "what" information is collected. Therefore we must know, that predictive analyses are customized. Each company has different segmentation methods. Hence it is firstly not important to list specific attributes, as each company has a different target group and mostly different products.

> The building of a predictive solution starts by a clear definition of the problem it is trying to solve. With a well-defined goal, data analytics scientists can then embark on the building of a predictive model that will be accurate, whose benefits will be readily explainable to all the parties involved. (Guazzelli, 2012c)

The criteria that is used for the analysis is hence depending on the definition of the goal. This is why it is not possible to give a definite list of all criteria, data collections are based on. In addition, assessment criteria of most new segmentation methods (for example, clustering models) are chosen automatically by a mathematical algorithm (Fraley & Raftery, 1998). This is due to the high amount of data and the point where it comes to **Data Mining**. The dictionary of economics from the German publishing house Springer Gabler, describes Data Mining as an application of methods and algorithms that extract empirical coherences between planning objectives, whose data are supplied in a therefore created database (Lackes, 2015). To be brief, Data Mining is about using computing systems, that organize an inscrutable accumulation of data and make it clear for human understanding.

The Process of data collection demands the creation of a data warehouse. It is a space, where all collected data is saved. If these warehouses have accumulated a high amount of "... data that can be captured, communicated, aggregated, stored, and analyzed ...", the space is named **Big Data** (Manyika et al., 2011, p. iv). In a report of 2011 by the McKinsey Global Institute (MGI) it is also mentioned, that "**Big Data** refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Manyika et al., 2011, p. 1). How big is Big Data?

It is an interesting fact, that in most definitions of the term there is no precise size of Big Data itself. The MGI justifies this phenomenon by "... [assuming] that, as technology advances over time, the size of datasets that qualify as big data will also increase" (Manyika et al., 2011, p. 1). To understand this growth and to get an idea of the amount of stored information, it is worth having a look onto an earlier study of Hal Varian and Peter Lyman from the University of California Berkeley. As part of their project "How much information?", they have estimated that "[print], film, magnetic, and optical storage media produced about 5 exabytes of new information in 2002. Ninety-two percent of the new information was stored on magnetic media ..." (Lyman & Varian, 2003, p. 1). Due to their estimation, "... that new stored information grew about 30% a year between 1999 and 2002 ..." they estimated "... that the amount of new information stored ...has about doubled ..." (Lyman & Varian, 2003, p. 2).

Additional researches, like a white paper from the International Data Corporation (IDC) from 2012 and sponsored by the EMC Corporation estimated, that "from 2005 to 2020, the digital universe will grow by a factor of 300, from 130 exabytes to 40,000 exabytes, or 40 trillion gigabytes ..." and "[from] now until 2020, the digital universe will about double every two years" (Gantz & Reinsel, 2012, p. 1).

Even if the volume of data may not be comprehensible, it is the growth rate that proves the growing significance of Big Data. Data collection and the development of technologies that facilitate it is one part of building a profound base of predictive analytics. But it is also important not to neglect a well organized data warehouse.

1.3 Privacy Issues

To supply customers with the best service or product, marketers need a detailed view into their lifestyle, including preferences and behavior. This can be a way, how products and services can become more and more personalized (Bohnert, 2004). According to Kauffelt (2002), studies have verified that 56% of consumers prefer shopping on personalized websites and 90% of consumers are willing to give their private e-mail address for personalization purposes. Meanwhile, in a report from the McKinsey Global Institute, Manyika et al. (2011) predict an aligned increasing importance of Big Data with the increase of policy issues "... including, but not limited to, privacy, security, intellectual property, and liability" (Manyika et al., 2011, p. 11). When predictive analytics is used (especially when using collaborative filtering (see Chapter 2.3)), every web activity can be recorded and assigned (Chen & McLeod, 2006). There are different approaches to maintain the consumers' trust. One way could be to improve the usability of anonymous user information. Another way to maintain consumer trust could be to increase transparency in web marketing (Chen & McLeod, 2006). There are applications like *Platform for Privacy Preferences* (P3P), the WWW Consortium is working on, that shall give consumers transparency and power over their web activity (Bohnert, 2004; WWW Consortium, 2016). In Addition, there are governmental institution and laws like the Bundesdatenschutzgesetz

(BDSG) in Germany, that shall provide a legal framework (Bohnert, 2004). However, it as to be mentioned, that by searching for personalization, consumer loose anonymity.

2 Predictive Analytics Models for Marketing

"Today we live with an ever-expanding sea of data. To navigate it safely, we use analytics" (Guazzelli, 2012a). Therefore we need to distinguish between descriptive and predictive analytics.

The purpose of descriptive analytics is to describe raw data from the past after putting them into patterns. Then, descriptive analytics is about "...[summarizing what happened ..." in the past (Wu, 2013). In an Interview by InformationWeek "... Dr. Michael Wu, chief scientist of San Francisco-based Lithium Technologies, which develops social customer experience management software for businesses ... estimated, that more than 80% of business analytics ... are descriptive" (Wu, 2013). "Descriptive analytics are useful because they allow us to learn from past behaviors, and understand how they might influence future outcomes" (Halo Business Intelligence, 2015). "Common examples of descriptive analytics are reports that provide historical insights regarding the company's production, financials, operations, sales, finance, inventory and customers" (Halo Business Intelligence, 2015).

> Our need to understand past events has led to a discipline that we now call *business intelligence*. It allows us to make decisions based on statistics obtained from historical data. For example: How many customers have churned or defected due to attrition in the last six weeks? ... Where are customers located (maybe shown using Google maps)?...". (Guazzelli, 2012a)

This decisions are consequently profound as based on facts. It is still criticized, that descriptive analytics is not going to be enough in future, as humans will not be able to keep an overview over their collected data. In the first of his series of four articles Guazzelli emphasizes, that "... descriptive analytics [wasn't] enough ..." and that "... in the society we live in today, it is imperative that decisions be highly accurate and repeatable" (Guazzelli, 2012a). Therefore predictive analytics shall provide information that makes it possible to generate highly accurate and repeatable decisions.

While descriptive analytics gives us analyses about the past, predictive analytics focuses on the future. Even if the term sounds like telling the future, "...the purpose of predictive analytics is NOT to tell you what will happen in the future ... as no analytics can do that" (Wu, 2013). Wu wrote, that predictive analytics could only forecast what might happen in the future, because all predictive analytics were probabilistic in nature (Wu, 2013). "Predictive analytics provide estimates about the likelihood of a future outcome" (Halo Business Intelligence, 2015).

The relevance of predictive analytics "... in industry increased together with the amount of data being captured [(as mentioned in Chapter 1)] from people ... and sensors ... as well as the availability of cost-effective processing power, be it Cloud or Hadoop-based" (Guazzelli, 2012a). Hence, there is no issue of time and technological process, but of controlling "the problem of ... quality and ... reliability of the information that is found on the web ..." (Majó & Révész, 2011, p. 62). This is why it is important to capture high quality data, as it "... will directly reflect the quality of the model" (Guazzelli, 2012a). Majó and Révész call this conflict the coping of "... the abundance and the reliability of information" (Majó & Révész, 2011, p. 62).

It seems as if algorithmic takes over from human knowledge the responsibility to make decisions about the future of e.g., a company. Guazzelli does not agree. In fact,

> ... [expert] knowledge is based on experience and is used everyday by all companies to influence day-today operations. Given how we can translate expert knowledge into a set of business rules, we've built decision-based systems to automatically apply the knowledge elicited from human experts". (Guazzelli, 2012a)

As a consequence, "... we may use a series of rules to trigger business decisions depending upon the output obtained by a predictive model" (Guazzelli, 2012a). The following chapters will inform you about three models of predictive analytics: clustering models, propensity models and collaborative filtering. This information is based on a customer relationship management (CRM) perspective.

2.1 Clustering Models

From the perspective of the client, clustering models are probably the most classic way of segmentation. Clustering in general describes the process of dividing a total in several groups (Oxford University Press, 2015). Thereby members of the same group, have related qualities. In predictive analytics regarding consumer segmentation, the term is used to describe the process of classifying customers.

Through modern technology "... you let the algorithms, rather than the marketers, create customer segments" (Levin, 2015). In information technology (IT), there is the "... problem of determining the structure of clustered data, without prior knowledge of the number of clusters or any other information about their composition" (Fraley & Raftery, 1998).

> ... [i.d.] cluster analysis is a set of data-driven partitioning techniques designed to group a collection of objects into clusters, such that the number of groups (clusters) as well as their forms are unknown, the degree of association or similarity is strong between members of the same cluster [and] is weak between members of different clusters. (Brusilovsky, 2015)

Therefore, before being able to penetrate day-today operations, these Models have to learn how to analyze the database. "A predictive model is simply a mathematical function that is able to learn the mapping between a set of input data variables, usually bundled into a record, and a response or target variable" (Guazzelli, 2012a). Concise, in the end, with a bit of help by putting expert knowledge in training and developing the program, it is able to find the best way of clustering data itself. This paper will not deepen the problem of IT, but how clustering models are used in reality.

Especially "[retail] marketers are constantly looking for ways to improve the effectiveness of their campaigns" not only to improve sales by itself, but also to create customer loyalty (Tirosh, 2012). Therefore clustering can help to "... target customers with the particular offers most likely to attract them back to the store and to spend more on their next visit" (Tirosh, 2012).

The main advantage in clustering models is the fact, that "[algorithms] are able to segment customers based on many more variables than a human being ever could" (Levin, 2015). With clustering models, it is possible to create completely individual groups. The predictive marketing cloud AgilOne calls this individual dimensions the cluster DNA. Figure 1 shows possible components of a cluster DNA (Levin, 2015).

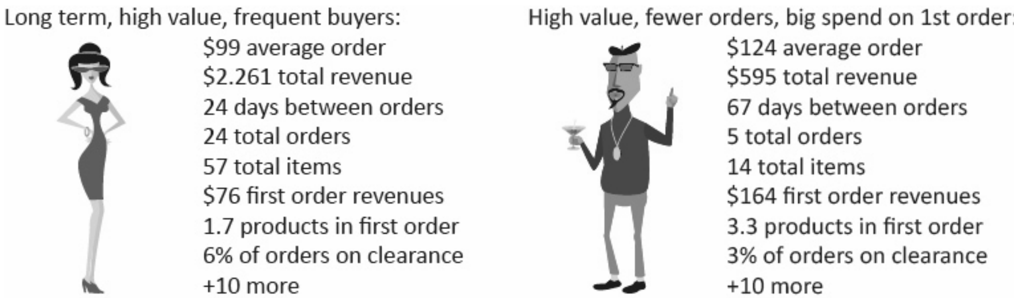

Figure 1: Example: cluster DNA
Adapted from: (AgilOne, 2015)

There are different types of clustering models. "The most used clustering algorithms are behavioral clustering, product based clustering ... and brand based clustering" (Levin, 2015)

2.1.1 Behavioral Clustering

To reach efficiency marketing campaigns "... should be directed at the customers most likely to respond to it" (Tirosh, 2012). Therefore a company (especially retailers) need to understand what characteristics their customers have. With behavioral clustering you learn how they behave. Their behavior can manifest themselves in factors like loyalty, information behavior, the reaction towards discounts, shopping frequency or customer wallet (and more). "By using customer segmentation to determine actionable “customer prototypes,” the marketer is able to test different campaigns on particularly-relevant target groups of customers ... [to] find the most effective repeatable offers to each customer segment" (Tirosh, 2012). "For instance, customers that buy frequently but with low sized orders might react well to offers like ‘Earn double rewards points when you spend $100 or more" (Levin, 2015).

En example of clustering customers on their behavior: *An auto insurance is planning to improve their offerings for their customer. The company "... can ask them to rate how important the following two attributes are to them ..." (Darden Business Publishing, n.d., p. 1):*

- *"Saving on premium*
- *Existence of a neighborhood agent*

Importance of the attributes are measured using a seven-point ... scale, where a rating of one represents not important and seven represents very important" (Darden Business Publishing, n.d., p. 1). The analysis might look like the Figure below:

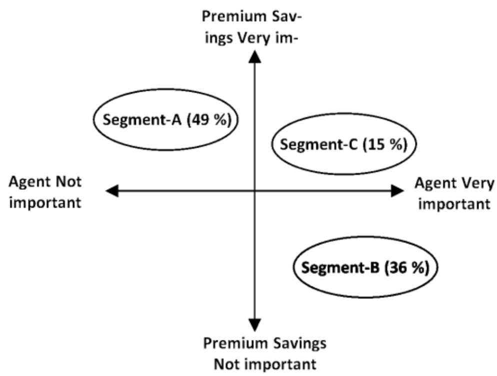

Figure 2: Segmentation from Insurance customers
Adapted from: (Darden Business Publishing, n.d., p. 2)

"[The] analysis shows three segments. The majority of ... [the] customers [(A)]... prefer savings on their premium and do not prefer having a neighborhood agent" (Darden Business Publishing, n.d., p. 2). With customers of segment B it happens to be the other way round. Segment C, preferring both attributes, is represented as a minority. "The analysis shows that [the insurance company] can benefit ..." by concentrating on both offerings separately, as segment C contains not as much customers as the other two segments (Darden Business Publishing, n.d., p. 2).

Although the amount of variables includes in predictive analytics is mostly much higher due to data mining technology than in the given example, it shows how a company can use behavioral clustering for consumer segmentation.

2.1.2 Product Based Clustering

Another "... [way] to improve the effectiveness of a campaign ..." (Tirosh, 2012) and to reach the right target group, is to cluster your customers by the offered groupings of products (Levin, 2015).

The first step would be, that a company "... [prepares] the customer spend data for each product category. Grocery stores and supermarkets would typically look at categories such as packaged foods, meat, dairy, produce, seafood and bakery" (Tirosh, 2012).

A data mining program will then "... identify distinct homogeneous groups of customers with minimal variance between their purchasing behaviors. This identifies unique “customer prototypes” (such as meat lovers, produce lovers and gourmet lovers) to which specific marketing offers can be targeted" (Tirosh, 2012).

Table 2: Product Based Clustering on Supermarket Customers

Discovered Customer Patterns	% of customer	Departments						
		Fresh Meat	Packaged Foods	Dairy	Fish & Seafood	Gourmet	Fresh Produce	Bakery
Basic Shoppers	39%	3%	75%	2%	6%	1%	10%	3%
Meat Lovers	15%	59%	15%	4%	5%	3%	9%	5%
Produce Lovers	8%	9%	21%	5%	6%	2%	49%	7%
Gourmet Lovers	3%	1%	12%	0%	3%	73%	6%	4%
Variety Shoppers	35%	14%	39%	8%	12%	6%	19%	2%

Adapted from: (Tirosh, 2012)

The chart shows that the analysis resulted in 5 different product based clusters. Thereby more than two thirds of customers are allocated in only two clusters: Basic Shoppers (39 %), who buy mainly packaged food and Variety Shoppers (35 %), "... whose shopping behavior is widely spread among departments" (Tirosh, 2012).

Now as marketers "... [have] a clear view of the various customer prototypes, it makes sense to target relevant marketing campaigns to the most interesting segments" (Tirosh, 2012).

2.1.3 Brand Based Clustering

Brand based clustering is more or less an advanced perspective of product based clustering. Though instead of focusing on the product category, brand based clustering models are concentrating on the brand preferences of customers. "When a brand releases new products ..." a company can find out "... who is likely to be interested" (Levin, 2015). This way, marketers have and overview which brands are preferred by which cluster.

2.2 Propensity Models

"The end goal [of propensity models] is to automate the offer selection and placement based on analysis and predictive models for your particular customer base" (Miller, 2012). With propensity models a marketer gets to know what the customers are tending to do, whether talking about customer churn rate, fraud detection, transaction repeat etc. (Bradley & Stewart, 2015; Guazzelli, 2012a; SAS Institute, 2015; Tirosh, 2015). With propensity models marketers can assume how consumers will behave in the future. This is important on the one hand to anticipate the cost to acquire a customer (CAC), on the other hand to monetize those customers, as it is important to "... show a return on marketing investment" (Gupta et al., 2006, p. 140; Skok, 2015; Walker, 2015). As there are many output variables and hence many way of using propensity models, the literature review concentrates on three models: predicted lifetime value, predicted share of wallet and propensity to churn.

2.2.1 Predicted Lifetime Value

Predicted lifetime value, in other words predicted customer lifetime value (CLV), describes the value of the projected revenue or profit a customer generates over his or her life of relationship with a company (Custora, 2015; Fairclough, 2015; Gupta et al., 2006). When a customer makes its first purchase, the marketer can collect more information beside the initial transaction record. He may also get "... email and web engagement data for example, as well as demographic and geographic information" (Levin, 2015). "By comparing a customer to ..." other customers from the past it is possible to "... predict with a high degree of accuracy their future lifetime value" (Levin, 2015).

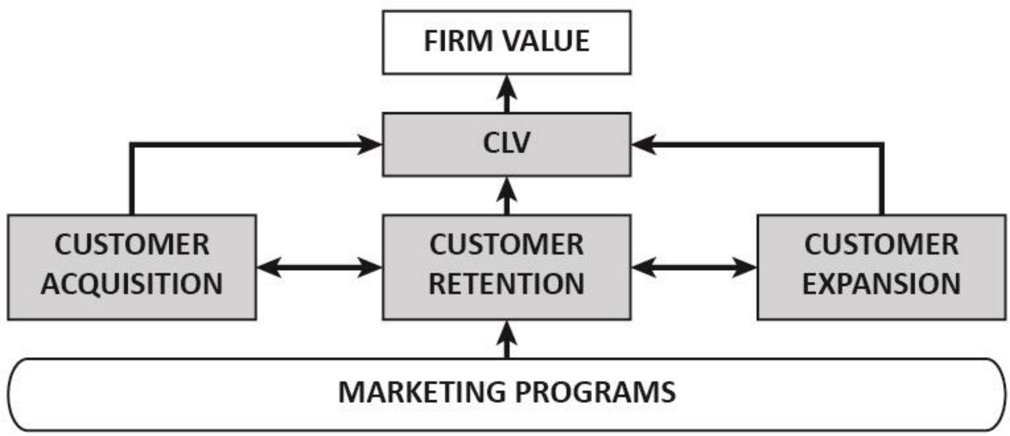

Figure 3: Conceptual Framework for Modeling Customer Lifetime Value
Adapted from: (Gupta et al., 2006, p. 140)

The framework shown in Figure 5 visualizes after Gupta et al. (2006, p. 141) the basic functions of customer lifetime value. It shows "... the impact of marketing programs on customer acquisition, retention and/or expansion ..." (Gupta et al., 2006, p. 140).

The first function, customer acquisition, is told to be a heavy factor when it comes to costs. Thomas, Reinartz, and Kumar (2003) "... found that customers should be acquired based on their profitability rather than on the basis of the cost to acquire and retain them" (Gupta et al., 2006, p. 144). Tommy Walker writes on ConversationXL, that "... many businesses focus on transactional customer value, and forget to invest in the experience that happens after the conversion" (2015). This is why CLV can be used as a tool when it comes to longevity of companies and especially start-ups. In an article of venture capitalist David Skok, he puts two variables to balance (2015):

- "Cost to Acquire Customers (CAC)
- The ability to monetize those customers ..." (CLV)

"To compute the ... CAC, you ... take your entire cost of sales and marketing over a given period, including salaries and other headcount related expenses, and divide it by the number of customers that you acquired in that period" (Skok, 2015). To generate the CTV, you "... look at the Gross Margin that you would expect to make from that customer over the lifetime of your relationship" (Skok, 2015). There are different ways of calculating CLV. In a case study about Starbucks, the Blog Kissmetrics explains that the company would "... typically use several different equations to calculate the CLV" (Kissmetrics, n.d.). When used in combination you take the average CLV. If the CLV is higher than the CAC it signifies, that customer retention is given (Skok, 2015). The fact that customers are satisfied after being acquired will lead to their repeat purchase (Gupta et al., 2006).

How is this model going to help the marketers of a company to classify customers? As the calculation is based on average customer data, there are different types of customers existing. Kissmetrics gives the example for creating a segment of "average" and "good" customers. The difference between those segements is that, "... good customers might cost more to acquire, but they'll likely be more profitable as well" (Kissmetrics, n.d.).

2.2.2 Propensity to Churn

Models that predict the propensity that a customer churns are somehow the counterpart to the predicted customer lifetime value (see chapter 2.2.1). Rouse defines churn rate as "... a measure of customer ... attrition [that] is defined as the number of customers who discontinue a service ... during a specified time period divided by the average total number of customers ... over that same time period" (2015).

The website "churn-rate.com" offers a churn rate calculating service and provides advice in terms of predictive analytics. It is emphasized, that especially for e-commerce the calculation of the churn rate was "... worth the effort" (RJMetrics, 2015). To clarify this propensity model, Guazzelli gives the following example:

Customer 1	Customer 2
No complaints in last 6 month	3 complaints in last 6 month
Opened 1 support tickets in the last 4 weeks	Opened 1 support tickets in the last 4 weeks
Spent a total of $9,876 buying merchandise	Spent a total of $1,234 buying merchandise
Spent a total of $987 in services	Spent a total of $123 in services
Purchased 12 items in last 4 weeks	Purchased 2 items in last 4 weeks
Is 54 years old	Is 34 years old
Is a male	Is a male
Lives in Chicago	Lives in Los Angeles
...	...

Figure 4: Two customers and their input features
Adopted from: (Guazzelli, 2012a)

You as a company need "... a predictive model that will be able to tell who among your customers is most likely to churn" (Guazzelli, 2012a). The first step would be to search in your "... historical data ... [for] attrition-related features for both existing and past customers that churned" (Guazzelli, 2012a). This list may include the number of complaints in the last 6 months, the number of support tickets opened in the last 4 weeks, how often and how much money the customer spent buying merchandise or services (on-line or in-store) ..." (Guazzelli, 2012a). In addition it may include "... generic information" (Guazzelli, 2012a).

Figure 4 shows two customers: "Customer 1 is an existing customer and seems to be satisfied. Customer 2, however, has churned" (Guazzelli, 2012a). Following this "... you present all your customer data to a predictive technique ..." (see Figure 5). In this time the data mining program will be trained. After this the predictive model has to be validated and therefore you must answer the question if it works.

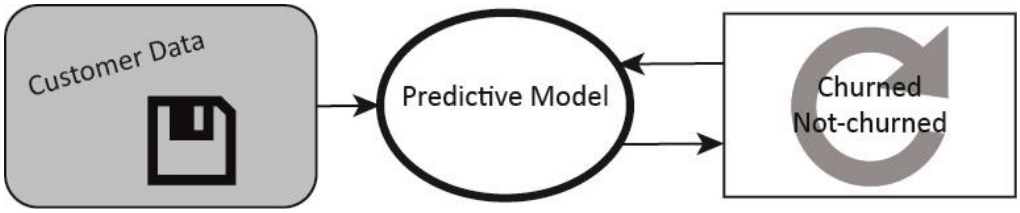

Figure 5: Predictive model during training
Adopted from: (Guazzelli, 2012a)

After the predictive model was validated to use in day-today operations, it has to be deployed. Therefore "... a standard called PMML (Predictive Model Markup Language) ... allows predictive models to ... move between different systems" (Guazzelli, 2012a). "After deployment, [you] can use the churn model to monitor all existing customer activity [, as a] good predictive model is able to generalize its knowledge to compute the churn risk even for customers it has never encountered before" (Guazzelli, 2012a) (see Figure 6). "If a high churn risk is detected, procedures may be put in place to mitigate it" (Guazzelli, 2012a).

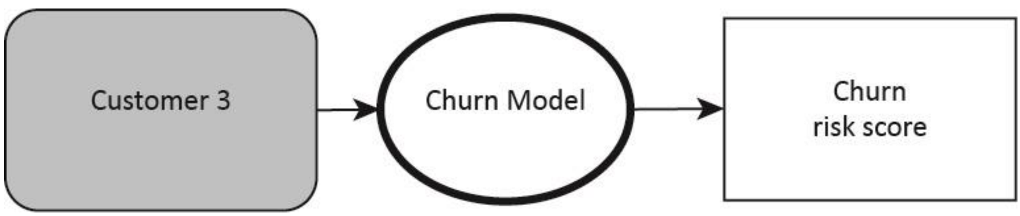

Figure 6: Churn model
Adopted from: (Guazzelli, 2012a)

This is how customers can be segmented by calculating the churn rate. Another research about propensity to churn was done by Optimove.

Optimove is a market leading company that sells the service of"... [enabling] marketers to grow their business through their existing customers, by giving them the ability to maximize customer spend, engagement, loyalty, retention and lifetime value" (Optimove, 2015). In an analyze of 2015 of over one million e-commerce transactions "... over a two-year period across a variety of categories including fashion, gadgets, fast food and others ...", they found out, that "... the number of items in a shopper's first order indicates future behavior...", especially when it comes to churn rate (Tirosh, 2015). Tirosh comments, that it is a big challenge for online retailer not only to acquire customers, but also to retain them for more than one purchase.

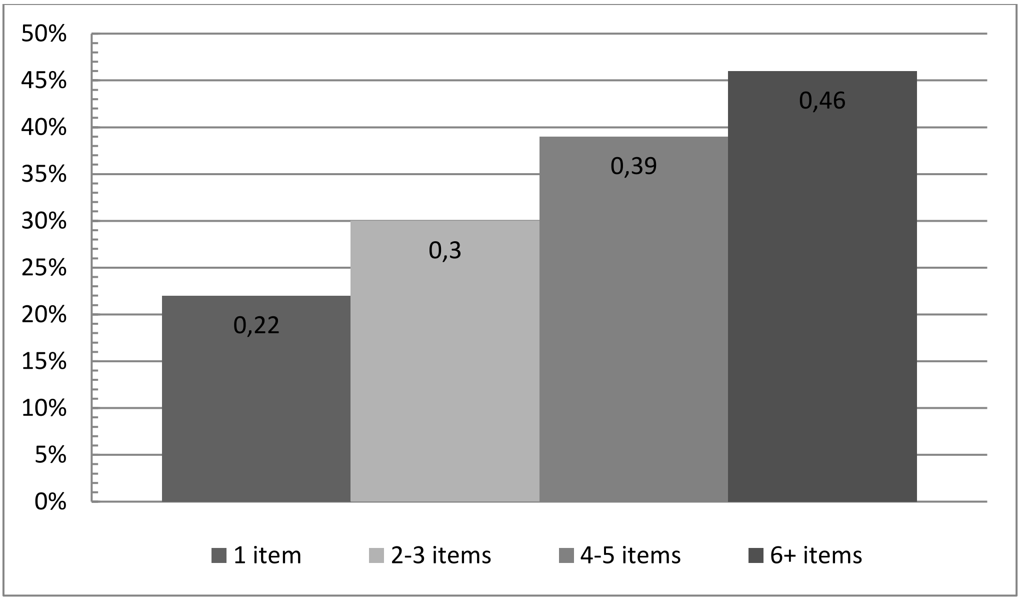

Figure 7: Likelihood of a Second Order per Number of Items in First Order
Adapted from: (Tirosh, 2015)

Through the analysis, the company found out that "... while 50% of shoppers buy exactly one item during their first order, [shoppers] who purchases two or three items [are] 36% more likely to place a second order than [shoppers] who purchased just one item in the first order" (Tirosh, 2015) (see Figure 9).

Another result of this analysis was that the number of future orders increased per number of items in first order and hence has influence on the customer lifetime value (CLV) - the time before the customer churns (Tirosh, 2015). The number of items in first order is consequently another indicator for customer churn. This information will then lead to create customer segments regarding churn rates.

With that knowledge about the customers churn likability marketers can react on the churn rate of particular groups of clients. One reaction could be to make discounts depending on the likability to churn. This way you can retain customers longer (Guazzelli, 2012a). Even so marketers can put more effort in increasing the number of items of first orders (Tirosh, 2015).

2.2.3 Predicted Share of Wallet

Customers may be satisfied with a particular brand of products or services and "... recommend it to others—but if they like your competitors just as much (or more), ..." the brand loses sales (Keiningham, Aksoy, Buoye, & Cooil, 2011). This is where it comes to share of wallet. In an article from the Harvard Business Review, the importance of this subject is emphasized, as the share of wallet provides companies with an additional tool of customer segmentation (Keiningham et al., 2011). Levin defines share of wallet as a model that "... can estimate what percentage of a person's category spend you currently have achieved" (2015). The following chapter is not concentrating on how share of wallet is calculated, but how it is used by marketers.

In an article from the Journal of Marketing, Du, Kamakura and Mela describe the share of wallet as another useful metric of customer relationship management (CRM) guidance. The main conflict in predictive analytics is said to be the fact, that "... firms [were] often compelled to manage customer relationship using a view of their customers that [was] based mostly on internal records" (Du, Kamakura, & Mela, 2007). Bell and colleagues indicated "... that this lack of individual-level, industry-wide consumer data is a primary barrier to CRM" (Bell, Deighton, Reinartz, Rust & Swartz, 2002 as cited in Du et al., 2007, p. 94). As a solution, marketers combine survey results, where customers can out themselves about other brands and companies, with "... information already stored in the internal database (e.g., transaction history, demographics)" (Du et al., 2007, p. 94)(Keiningham et al., 2011; Rosset et al., n.d.).

In most literature, the share of wallet is constructed with three variables. The first variable describes the relation of your customer between other companies (e.g., expenditure, requirements). The second variable describes the relation between your company and your customer and mostly the income, the third variable, is implied somehow as well (Du et al., 2007; Keiningham et al., 2011; Reinartz & Kumar, 2001; Rosset et al., n.d.). However, there are different ways of calculating the share of wallet. The main difference between all of them, is the application of the variables. In researches about the share of wallet, Du et al. talk about several key findings (2007).

The first finding and probably the most important one, as it emphasizes the importance of customers share of wallet, is the refusal of a correlation between customer loyalty and share of wallet (Du et al., 2007). Keiningham et al. (2011), as well as Reinartz and Kumar (2001) confirm that both, customer loyalty and the share of wallet may be used in CRM, but are not linked directly together. The confusion comes from their linkage to customer lifetime value (CLV) (Reinartz & Kumar, 2001). Ratner explains, that share of wallet my only help to monitor customer loyalty (2015).

The second finding is the positive correlation of customers' decisions. "Customers with high share in one category also tend to have high share in another category" (Du et al., 2007, p. 95). This fact is interesting when it comes to cross- and up-selling applications (Du et al., 2007).

Finding three is relating to the negative correlation between customers' share of wallet and total purchase decisions. It means, that for "... some categories, customers with small shares within the focal firm tend to transact a large volume outside it. These customers might represent significant opportunities for volume growth to the extent that the focal firm can induce them to switch" (Du et al., 2007, p. 95).

In finding 4, the third variable, the customers income is implied. Du et al. explain that "[customers] with higher incomes tend to balance share of requirements across firms" (2007, p. 95). This is probably due to insufficient customer satisfaction or for some reason "... customers with higher incomes have incentives to allocate business across firms" (Du et al., 2007, p. 95).

In Addition, Keiningham et al. explain that a low share of wallet has not only to do with customer satisfaction, but also with the position of your company in the customers' rank (2011). An often done mistake was to "... measure customer satisfaction or ... other metrics that are based on customers' perceptions of ..." the own company alone (Keiningham et al., 2011). The goal was actually to compare the company with others and to improve the company's rank (Keiningham et al., 2011). This can be done, by calculating the share of wallet. Keiningham et al. suggest a calculation called Wallet Allocation Rule (see figure 10). Thereby this model is comparably simple to others, as it implies only two variables: the rank of a company or brand and the number of companies or brands. The following example of a wallet allocation rule is reduced to 3 customers and shall only provide an idea of how share of wallet can be calculated. Keiningham et al. structure this calculation in 3 steps:

The first step is to collect all the needed data to fill in the variables. In our case we need the number and names of the companies or brands that shall be compared (e.g., Acme, Mega and Brand X).

The second step would be to "[survey] customers and obtain satisfaction or other loyalty scores for each brand [and to] convert the scores into ranks" (Keiningham et al., 2011). In the example the customers Stuart, Mary and Joe were surveyed.

Table 3: Ranks of the brands according to the customers satisfaction scores

	Acne	Mega	Brand X
Stuart	3	1	2
Mary	3	2	1
Joe	3	1	2

Adapted from: (Keiningham et al., 2011)

Thirdly, you insert the number of companies or brands and their ranks into the following formula:

$$\text{Share of Wallet}=\left(1-\frac{\text{Rank}}{\text{Number of Brands+1}}\right)*\left(\frac{2}{\text{Number of Brands}}\right)$$

$$\text{Share of Wallet}_{\text{Stuart \& Acme}}=\left(1-\frac{3}{3+1}\right)*\left(\frac{2}{3}\right)=0.1675\text{, Rounded to 17\%}$$

When this calculation is repeated for all variables, the following results (see Table 3) are being retained.

Table 4: Brand Share of Wallet

	Acme	Mega	Brand X
Stuart	17%	50%	33%
Mary	17%	33%	50%
Joe	17%	50%	33%
Brand Share of Wallet	17%	44%	39%

Adopted from: (Keiningham et al., 2011)

These results show two significant values. One value is the position of your company or brand compared to others (Keiningham et al., 2011). The other value is the visualization of different customer segments. In a publication of the Journal of Marketing, Du et. al list some other methods of share of wallet estimation (Du et al., 2007). Rosset et al. found therefore again other calculation models, due to the researches with the IBM T. J. Watson Research Center (Rosset et al., n.d.). There is a vast amount of calculation models found in literature. But the main subject of this paper is the use in customer segmentation.

Table 5: Strategy Based on Share of Wallet vs. Profitable Lifetime Duration

Share of Wallet		Profitable Lifetime Duration: Low	Profitable Lifetime Duration: High
	High	◘Use selective or optimal mailings/strategy to reduce cost ◘Attempt to cross-sell and up-sell	◘Invest to nurture, defend, retain ◘Reward with loyalty program
	Low	◘Lower marketing expenses ◘Consider divestment strategy ◘Possibly customer outsourcing	◘Use conversion strategy to hire from competition ◘Encourage cross-buying and up-buying ◘Offer lucrative loyalty programs
		Low	High
		Profitable Lifetime Duration	

Adapted from: (Reinartz & Kumar, 2001, p. 27)

The survey results can now be combined with information from the internal database, by using a predictive model that discovers patterns and creates customer segments (Du et al., 2007). This way, share of value can be applied in big data based marketing. By using a 2x2 matrix, Reinartz and Kumar have built up a model that shows how share of wallet and customer lifetime value can be combined in planning marketing strategies (see table 4). Depending, to which of the 4 segments a customer belongs to, marketers may be able to react with a different strategy.

2.3 Collaborative Filtering

Clustering models as well as propensity models, are models its importance has risen with the spread of e-commerce, mainly because the data base capacity has increased with predictive analytics. Collaborative filtering (CF) though is a result of web marketing and reaction to a growing e-commerce environment (Lee, Jun, Lee, & Kim, 2005). Turban et. al mention collaborative filtering to the same effect as behavioral targeting. Both terms are subsidiary subjects of personalization and behavioral marketing (Turban, King, Lee, Liang, & Turban, 2015).

In *Electronic Commerce* they also mention related filtering methods, that are not to be confused with CF (Turban et al., 2015). **Rule-based filtering** is the word for the method of surveying "... consumers about their preferences via multiple choice questions and [using] the collected information to build patterns for predicting customers' needs" (Turban et al., 2015, p. 417). Hence the data is collected from the general consumer and not from existing customers. **Content-based filtering** "... allows vendors to identify customer preferences by the attributes of the product(s) they buy or intend to buy ... [and] ... recommend products with similar attributes to the user" (Turban et al., 2015, p. 417). In this case the filtering system is focusing a unique customer and his purchasing behavior, without searching for relations to other customers. **Activity-based filtering** is call the method of "... analyzing the website's visiting level and activities"(Turban et al., 2015, p. 418). Thereby marketers may find the highest potential where to find new customers or possibilities of product placement.

All of these methods are about one task marketers have in customer relationship management (CRM), namely to "... predict what products or services ..." the customer is interested in (Turban et al., 2015, p. 417). This can be done by analyzing the customers profile or his past buying behavior. But in the case of a new customer you have no knowledge about those features. To solve this problem Goldberg, Nichols, Oki and Terry (1992) named and introduced the term CF in a special issue of Communications of the ACM on information filtering in 1992. At that time, a solution was searched to cope with e-mail overload. In an article of Expert Systems with Applications, Lee et al. say of CF it "... has been known to be the most successful recommendation technique ..." (Lee et al., 2005, p. 700). Of course as a reaction also other branches had a growing interest in using this method, as it was not only useful in e-mail filtering, but also to classify web page users. In terms of customer segmentation, CF is now used to connect "... preferences and activities of many customers that have similar characteristics to predict preferences of new customers and recommend products to them" (Turban et al., 2015, p. 417). An early successful technical solution of this was developed by GroupLens in 1994 (Bohnert, 2004). According to Majó and Révesz, as well as Bohnert, GroupLens was one of the first to introduce automated CF (Bohnert, 2004; Majó & Révész, 2011). "GroupLens is a research lab in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities specializing in recommender systems, online communities, mobile and ubiquitous technologies, digital libraries, and local geographic information systems" (GroupLens, 2015). The most reviewed project of GourpLens and a convenient example for CF is called MovieLens. It is a website that provides movie recommendation (Bohnert, 2004; Chen & McLeod, 2006; Kobsa, 2007; Majó & Révész, 2011). But before fixing CF as a method of recommendation, it is important to understand that CF is a method of predictive analytics, developed to quantify and systemize a vast amount of consumer preferences data (Kwak, 2001; Majó & Révész, 2011); then, put the data "... into an easy-to-communicate form" (Majó & Révész, 2011, p. 59). Only then marketers are able to recommend, i.d. "... influencing customers in their choice" (Majó & Révész, 2011, p. 59).

At MovieLens, new users have to register first with personal information. Following this, the user is asked to distribute three assessment points on a given list of genre. After the registration process, MovieLens is able to recommend a collection of some movies you might like. In order of rating some films you, as user have previously seen, on a scale from 0.5 to 5 stars, recommendations become more personalized. According to Bohnert (2004), the recommendation engine has access to multiple millions of ratings. The ratings of a particular user are being compared with ratings of other users. This way "... the system looks for the closest neighbour who is most similar to that particular user's profile" (Majó & Révész, 2011, p. 60). The accuracy increases with the amount of films, that the user has rated (Bohnert, 2004). Additional features, that enhance the accuracy of recommendation are a favorite list and a personalized search engine.

The example shows, that whatever the action of a users might be, it is leading to collect data and improve the product or service. Consequently (on other websites using CF), this data is obtained as soon as "... a visitor asks questions, places orders, fills in a registration form or a questionnaire, or writes to the company or a discussion group about his or her opinion of a product or a particular situation" (Majó & Révész, 2011, p. 62).

Breese, Heckerman, and Kadie (1998) classify CF into memory-based and model-based. Memory-based CF algorithms predict the votes of a new active user based on information given by him (e.g., through a registration form) and based on data from existing users - a reason, for calling it a also user-based approach (Breese et al., 1998; Lee et al., 2005). The first model of Goldberg et al. was such a CF system (Goldberg et al., 1992; Lee et al., 2005). Model-based CF algorithms calculate the expected value of a vote from a particular user.

As mentioned above, MovieLens is an "Automated CF" system, i.d. it collects, stores and analyzes customer preference data, compares it with other customer data and sends a recommendation to the user (Majó & Révész, 2011). In his seminar paper, Bohnert refers to a second CF system called "Active CF". This term describes a recommendation process, where users, not the website, recommend objects to each other (Bohnert, 2004). Majó and Révész subdivide this term into pull-active CF and push-active CF. Thereby pull-active CF describes the process when users search for recommendations, whereas push-active CF is the process when a "... user can simply recommend (push) the information to others if they find it interesting or useful" (Majó & Révész, 2011, p. 61).

2.3.1 Collaborative Filtering in Marketing

There are different aspects, how collaborative filtering can be useful for marketers. Beside an automated improvement of the product or the service for a particular user, recommendations are a way of strengthening a cross- or up-selling strategy, as customers find other products, they didn't intent to buy (Winoto & Tang, 2008). The information, that users or customers leave, are an imprint of their behavior and show their individual preferences and expectations. This creates a chance for marketers, to target easily existing customers, but also new customers directly with recommendations, only by comparing them to profiles of similar customers. CF becomes also interesting when it comes to the cost factor. The development of a CF system creates mainly fix costs, whereas new users entail practically no extra costs (Majó & Révész, 2011).

2.3.2 Issues in CF and Problem-Solving Approaches

Kwak (2001) as well as Chen and McLeod (2006) mention a limit in CF and so risks in marketing. The model relies on information from past users, so results are inaccurate when recommending new products, as the program needs information about object orientated customer experiences. In Addition, it happens that misunderstandings are created as well as it is difficult to differ manipulated recommendations from honest ones (Kwak, 2001). Another main issue is the difficulty to obtain an appropriate amount of scores, as customers seldom give ratings on used products (Chen & McLeod, 2006; Lee et al., 2005). Any try of generating scores artificially result mostly in a poor recommendation accuracy. As user behavior is thereby dynamic, CF can not be applied for short time periods (Lee et al., 2005). Aside from that, CF "... may not cover the extreme case ..." (Chen & McLeod, 2006, p. 122); i.d. users that have unique tastes and do not fit in average patterns. To revise errors, Ariely suggests to combine different approaches when using a predictive model (Aparicio & Ariely, n.d. as cited in Kwak, 2001). Balabanovic and Shoham (1997) introduced therefore hybrid models. These combine CF with other filtering models like content-based filtering (for definition see above). By doing so, problems like extreme cases can be covered, as on the one hand, the profile of a particular customer is compared to others, on the other hand, objects that this customer has already rated or bought, are included in the calculation (Balabanovic & Shoham, 1997; Chen & McLeod, 2006; Ekstrand et al., 2011; Liu, Lai, & Lee, 2009).

3 Discussion

The paper writing was temporal limited and took place between October 2015 and January 2015. It shall only present a literature review that provides a collection of literature on new methods of consumer segmentation regarding the high inflow of new information via e-commerce. Themes like Big Data, predictive analytics, data mining and consumer classification were made subject of discussion. Selected literature was used to selected literature to give an overview on the given issue. The preparation of the paper was limited by free web sources, the library of Reutlingen University and the academic research engine EDDI, which has access to EBSCO Discovery Service. Source citations were done in APA style, using the citation program Zotero. In the following discussion chapter the used literature will be reflected and evaluated, starting with the first literature mentioned in the review.

"Beyond Consumer Segmentation: New Technologies, New Market Lenses" of Falcon et al. (2003) provides an up to date report about the most relevant technological developments that have influenced the marketing environment. Furthermore, it focuses on changes in the consumer segmentation landscape. Falcon et al. give advice in how the market can be viewed by using six so called "market lenses". The message of the report to marketers, is to widen the scale of perspective when segmenting consumers, as new technology is influencing the trading business and hence, new consumer requirements have to be covered. A wide range of technological developments is given, that was worth to study (see Table 1), whereas profound information about particular segmentation methods is not included in the report or does not fit into contemporary science. The Institute for the Future, that published the report, has deepened the subject in another report from July 2003 called "Reinventing Customization: New Technologies, New Markets, and New Strategies " brought together by Spalding (2003). The report contributed to the literature review in introducing new ways of information ascertainment, regarding a focus on technology and giving a different perspective on the consumer market.

Another reference, introducing in the paper's subject, is given by Levin (2015). She runs marketing at the AgilOne blog and summarized a broad collection on how predictive analytics is used in marketing. The author appraised the post as the probably best reference when searching a reliable overall view on how predictive analytics may be useful for marketers and used in reality. As the importance of the subject increases, such references will grow in importance as well, for that one may stay able to have a sufficient overview on basic knowhow. Following this, readers have to make further researches to obtain profound literature on the particular subjects.

Even if not new, Big Data and Data Mining is a contemporary subject of many discussions in literature and part of the first chapter of this paper. The Mckinsey Global Institute (MGI) published a report in 2011 about the ever growing influence of Big Data. It discusses how Big Data transforms the global economy and how the leadership of companies may react on this development. The report of the MGI, that was written by Manyika (2011) et al. contributed to the literature review by giving a definition on Big Data, respectively explaining also, why Big Data can hardly be defined. This information has been proven to be an acceptable alternative to many different approaches of defining Big Data found in other literature. Still, the online encyclopedia of Springer Gabler is giving another acceptable definition and refers also to the application of Big Data in reality. Whereas the MGI focuses on a volume based definition of Big Data, Lackes (2015) (an author from Springer Gabler) focuses on an analytics based approach and is based on Data Mining. Both are highly recommendable sources, although the report from the MGI is very extensive, general and has some lacks considering using Big Data as a tool for consumer segmentation. Furthermore, the study by Lyman and Varian (2003) from the University of California Berkeley, was useful to grow understanding of the world wide amount of transmitted data. As this study was set up in 2003, newer studies are necessary to retain scientific reliability. Another study, from the International Data Corporation (IDC) by Gantz and Reinsel (2012) from 2012, confirms the predicted growth of data volume by Lyman and Varian.

When talking about privacy issues, the last chapter of the seminar paper by Bohnert (2004) gives an overview on the influence of collaborative filtering and web mining on the consumers privacy. Additionally, he provides information about laws and institutions like the WWW Consortium that amongst other things protects the internet users' private sphere. Another approach comes from Chen and McLeod (2006), who emphasize that an increased transparency would strengthen the trust of consumers. Both approaches can be fortified: a backbone in legacy and the spread of transparency. The report from the MGI predicts an increase of policy issues (Manyika et al., 2011). Still, in the authors opinion the privacy issue is mostly not emphasized enough and in most literature it is either not made a subject of discussion or the subject is only mentioned as a theme that may be subject of further researches. One outstanding work was written by Kobsa (2007). He was a contributing author to the book "The Adaptive Web" and author of its 21st chapter "Privacy-Enhanced Web Personalization". The chapter focused on the privacy issue and acts as literature review on this subject.

When researching predictive analytics, researchers are mostly confronted with IT literature, whereas in his post, Guazzelli (2012a, 2012b, 2012c, 2012d) gives a decent mix of IT and application methods. The IBM scientist works as an expert in predictive technology and provides in his post laypersons with a deeper understanding of how predictive analytics works. Furthermore, his work deepens the subject of propensity models. Especially an example about the propensity to churn visualizes the role of a predictive model. Also, Guazzelli lists the steps of introducing a predictive model in a company and what preconditions have to be covered. Wu (2013) emphasizes the important fact, that predictive analytics does not predict the future, but probable outcomes. These are in the authors opinion important information especially for young founders, as this is the way, how limitations can be found out. Predictive models are not the best way of segmenting customers for every type and size of company. Guazzelli as well as Majó and Révész (2011) write about the requirement of quality and reliability of information. Especially when it comes to collaborative filtering, vast amounts of high quality data are needed and not every company might be able to afford such a data warehouse. In Addition it is important to distinguish between predictive and descriptive analytics. Therefore, Halo Business Intelligence (2015) provides clear definitions. Predictive analytics may be a more accurate way of shaping a marketing strategy, but for some companies descriptive analytics might stay a more affordable way, by learning from the past.

There are several models of predictive analytics. Levin (2015) described clustering models, propensity models and collaborative filtering as the most significant models in marketing, as these provide tools to enhance the customer relationship management (CRM). There is a universally accepted definition of clustering. All enlisted authors describe cluster analysis as a model, where segmentation clusters are automatically created by algorithms and not by marketers. On the one hand, Fraley and Raftery's (1998) report is a reliable reference regarding the technological background of clustering models. On the other hand, it shows some weaknesses in explaining how clustering models are applied in reality. Therefore Levin (2015) and Tirosh (2012) give relevant examples of real life application of clustering models. Whereas Tirosh focuses rather on product based clustering, Levin adds examples of behavioral and brand based clustering. Darden Business Publishing (n.d.) provides therefore a vivid paper on behavioral clustering. In the authors opinion brand based clustering can be seen as an alternative perspective to product based clustering, as the predictive models behind it works practically the same way. Marketers do only have to find out, if their customers shop rather product orientated or brand orientated. This might be an important issue in fashion business, e.g. when buying a suit, customers value the brand differently as when buying luxury casual wear. All in all clustering models are a useful opportunity to obtain an overview over segments of your customer base.

Whereas literature on clustering models lays weight on life praxis, literature on propensity models is often heavily based on IT. The working paper of Reinartz and Kumar (2001) is thereby a significant reference in the business, especially when it comes to customer lifetime value (CLV). The article by Gupta et al. (2006) is a profound rather theoretical work on the overall process of predicting CLV. Levin as well as Skok (2015) and Fairclough (2015) (from Kissmetrics.com) provide rather practically based literature. Thereby Kissmetrics provides researchers with information on CLV from a marketer point of view and hence from a perspective of its application in customer segmentation. CLV is thereby probably on these grounds important, as in growing branches the number of suppliers grows and companies have to retain their customers as long as possible. As mentioned above, Guazzelli (2012a) refers in his post about predictive analytics to the propensity to churn. On the one hand, such classical approaches are significant to understand the process when calculating a churn rate. On the other hand, there are newer life praxis orientated approaches like an article from Tirosh (2015) on the Optimove blog proves. In the authors opinion, such approaches are reactions on the life praxis and are released from theoretical models.

In their working paper, Reinartz and Kumar (2001) refer when explaining CLV to the prediction of the share of wallet. According to them, these are two segmentation tools that can be combined, respectively are dependent (see Table 5), even if they are not directly linked together. The authors agree, that this approach is very useful when it comes to customer relationship management, as Du et al. (2007) describe, as well as it provides another tool of customer segmentation, like Keiningham et al. explains. In addition Du et al. emphasizes, that it is one of few predictive models, that includes the position of a particular company in a group and is hence particularly worthy. In the authors opinion, this is an significant fact, because doing something good as a company, does logically not exclude the fact that other companies do it as good or even better. Models that predict the share of wallet are, due to the surveying, very statistical and my relent prediction speed and accuracy, as human intervention is higher than at other models.

Regarding overall works in E-Commerce, "Electronic Commerce: A Managerial and Social Networks Perspective (8th ed.)" by Turban et al. (2015) belongs probably to reliable basic literature, regarding its strong connection to a commercial view. In terms of collaborative filtering (CF) it gives a simple overview and even refers to the privacy issue. Furthermore it provides information about further filtering methods and is hence a significant item in terms of consumer classification. The article from Goldberg et al. (1992) has a high significance in terms of CF, as it introduced the term in business. It also provides researches with information about the original idea of CF. Furthermore, this researches acted as a base for further ones considering the application of CF in marketing business. The author found out, that most reliable literature refers to recommender systems, when talking about CF. This is a consequence of the fact that CF is nowadays the most spread way of recommender techniques and has hence already achieved high acceptance in commercial business. Amazon is thereby a common example of using CF in CRM. There are different approaches of classifying CF. However, the biggest issue in this area is on one hand the accuracy of recommending, as these are mostly based on an average database, and on the other hand the privacy issue, as of all presented predictive models, CF is the most inquisitive one.

4 Conclusion

In order to complete the paper, the most important facts are summarized and significant results of the literature review are concluded. It includes all insights, readers are receiving when reading the paper. Thereby the set up of the conclusion is aligned to the structure of the paper, starting with a general view on predictive analytics.

Predictive Analytics is a result of searching a way to use vast amounts of collected data through e-commerce and data ascertainment technology. The technological development is attributed to the increase of marketing interests in customer identity. The accumulation of data is called **Big Data**. The term includes beside the high accumulation of data, that the intention of collecting this data was to use **Data Mining** methods to extract particular valuable information. The organization of Big Data is called **data warehousing**. The process of using data to predict future behavior of particular objects is called predictive analytics. This way of analyzing data shall prevent, that statistical bases for strategic decisions are influenced by humans. The decision by itself has still to be done by humans, as expert knowledge is highly relevant when it comes to the amount of experience. Predictive analytics is consequently no detachment from marketers, but a backup system that relies on data instead of feelings.

The paper was focusing on how predictive analytics can be useful in marketing in terms of customer segmentation. Clustering models, propensity models and collaborative filtering are methods that can be used for segmenting customers and putting marketing efforts at the right place. **Clustering models** group customers into patterns. These patterns can be based i.a. on customer behavior (**behavioral clustering**), preferences for product categories (**product based clustering**) or preferences for brands (**brand based clustering**). When customers are clustered by using behavioral clustering, the patterns are build up basing on the intentions why they buy a particular object. Furthermore, similarities are found between product based clustering and brand based clustering, as it only depends on the customer base and their shopping behavior.

Propensity models put customers into patterns, that make predictions about possibly developing tendencies of the behavior of customer groups, that belong to a particular segment. This way, i.a. **CLV**, the **propensity to churn** or the **share of wallet** can be predicted. Marketers have then the possibility, to target marketing activities to particular customers to acquire and retain them and probably expand the customer base. This way, marketers can allocate their budget more precise. They may also react quick to customer activities due to models like proposed by Reinartz and Kumar.

One main goal of marketers is to influence customers in their choice. A way to do so is making recommendations. In terms of predictive analytics, recommender systems are being developed under a predictive model called **collaborative filtering (CF)**. The authors found out, that CF is one of the most spread filtering methods in web marketing. The algorithm connects profiles of users (received through registration forms) with their preferences and activities on a particular web site and compares it to other users. Depending on the outcome of the algorithm and the allocation of the user, marketers can decide what reaction rule may be used on such user types. CF models are mostly automated and handle these processes on their own. This autonomy is based on input rules and the information given by new users and the given database. This method is interesting in cost factors, as only the installation needs high investments in fix costs, whereas new users entail practically no extra costs. Hence, this method is unprofitable for small companies, as a vast amount of high quality data (so a powerful data warehouse) is required to provide a high recommendation accuracy. Such cases should base their strategy rather on **descriptive analytics**, instead of predictive analytics. Additionally, the model somehow excludes extreme cases, as it is based on average data. These weaknesses can be covered., by combining CF with other filtering methods.

The choice for a predictive model depends on the size of the company and the expected output, but also on the quality of the database. All models (especially CF) have difficulties with handling **privacy issues**. During the whole process of researching, there could not easily be found a definite solution for the conflict of coping with customers requiring personalization, but in the same time keeping privacy saved. As the environment of privacy issues in marketing was not largely worked out in this paper, it would be worth further researches, as it is predicted to grow in importance.

Appendix

The following table shows the number of hits, when searching for a keyword on Google, Google Scholar and EDDI. The number of results is neither limited on a language, nor on a country. It was set up on January 5th 2016, Germany.

Table 6: Number of keyword results in different search engines

Keywords	Hits on research engine		
	Google	Google Scholar	EDDI
Predictive Analytics	13,200.000	51,500	6,999
Big Data	703,000,000	4,120,000	107,075
Data Mining	123,000,000	2,670,000	311,338
Clustering Model	22,500,000	2,810,000	88,992
Propensity Model	11,500,000	741,000	30,454
Collaborative Filtering	1,140,000	368,000	10,043

List of References

AgilOne. (2015). The Definitive Guide to Predictive Analytics Models for Marketing: Figures. Retrieved from http://www.agilone.com/academy/the-definitive-guide-to-predictive-analytics-models-for-marketing/

Balabanovic, M., & Shoham, Y. (1997). Content-based, collaborative recommendation: EDDI - Discovery Service der Hochschule Reutlingen. *Communications of the ACM*, *40*(3), 66–72.

Bohnert, F. (2004, January). Einsatz von Collaborative Filtering zur Datenprognose. Universität Ulm. Retrieved from http://www.mathematik.uni-ulm.de/sai/ws03/dm/arbeit/bohnert.pdf

Bradley, S., & Stewart, D. (2015). Five trends in fraud solutions. Retrieved 22 December 2015, from http://www.sas.com/en_us/insights/articles/risk-fraud/five-trends-in-fraud-solutions.html

Breese, J. S., Heckerman, D., & Kadie, C. (1998). *Empirical Analysis of Predictive Algorithms for Collaborative Filtering* (Technical Report No. MSR-TR-98-12).

Redmont: Microsoft Corporation. Retrieved from http://research.microsoft.com/apps/pubs/default.aspx?id=69656

Brusilovsky, P. (2015). Cluster Analysis vs. Market Segmentation. Retrieved 21 December 2015, from http://www.bisolutions.us/Cluster-Analysis-vs.-Market-Segmentation.php

Chen, A. Y.-A., & McLeod, D. (2006). Collaborative Filtering for Information Recommendation Systems. In M. Khosrow-Pour (Ed.), *Encyclopedia of E-Commerce, E-Government, and Mobile Commerce:* (pp. 118–123). Web: IGI Global. Retrieved from http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-59140-799-7

Custora. (2015). Predictive Customer Lifetime Value (CLV) Analysis for Retail | Custora. Retrieved 23 December 2015, from https://www.custora.com/tour/feature_predictive_customer_lifetime_value_clv_retail

Darden Business Publishing. (n.d.). Cluster Analysis for Segmentation. Retrieved from http://faculty.darden.virginia.edu/GBUS8630/doc/M-0748.pdf

Du, R. Y., Kamakura, W. A., & Mela, C. F. (2007). Size and Share of Customer Wallet. *Journal of Marketing*, (71), 94–113.

Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative Filtering Recommender Systems. *Found. Trends Hum.-Comput. Interact.*, *4*(2), 81–173. http://doi.org/10.1561/1100000009

Fairclough, J. (2015, December 22). How To Calculate Lifetime Value - The Infographic. Retrieved from https://blog.kissmetrics.com/how-to-calculate-lifetime-value/

Falcon, R., Gorbis, M., Jeffery, L., & Spalding, L. (2003). *Beyond Consumer Segmentation: New Technologies, New Market Lenses* (No. SR-784A). USA: Institute for the Future. Retrieved from http://www.iftf.org/our-work/people-technology/technology-horizons/beyond-consumer-segmentation/

Fraley, C., & Raftery, A. E. (1998). *How Many Clusters? Which Clustering Method?: Answers Via Model-Based Cluster Analysis* (Technical Report No. 329). Washington: Department of Statistics University of Washington. Retrieved from https://www.stat.washington.edu/research/reports/1998/tr329.pdf

Gantz, J., & Reinsel, D. (2012, December). THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadow s, and Biggest Grow th in the Far East. Retrieved from https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61–70. http://doi.org/10.1145/138859.138867

GroupLens. (2015). What is GroupLens? Retrieved 31 December 2015, from http://grouplens.org/about/what-is-grouplens/

Guazzelli, A. (2012a, May 29). Predicting the future, Part 1: What is predictive analytics? [CT316]. Retrieved 20 December 2015, from http://www.ibm.com/developerworks/analytics/library/ba-predictive-analytics1/index.html

Guazzelli, A. (2012b, June 19). Predicting the future, Part 2: Predictive modeling techniques [CT316]. Retrieved 20 December 2015, from http://www.ibm.com/developerworks/library/ba-predictive-analytics2/

Guazzelli, A. (2012c, July 3). Predicting the future, Part 3: Create a predictive solution [CT316]. Retrieved 20 December 2015, from http://www.ibm.com/developerworks/analytics/library/ba-predictive-analytics3/index.html

Guazzelli, A. (2012d, July 10). Predicting the future, Part 4: Put a predictive solution to work [CT316]. Retrieved 20 December 2015, from http://www.ibm.com/developerworks/analytics/library/ba-predictive-analytics4/index.html

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Nathaniel, L., & Sriram, S. (2006). Modeling Customer Lifetime Value. *Journal of Service Research*, *9*(2), 139–155.

Halo Business Intelligence. (2015). Descriptive, Predictive, and Prescriptive Analytics Explained. Retrieved 21 December 2015, from https://marketing.halobi.com/acton/form/9007/000b:d-0001/1/index.htm?id=000b

Kauffelt, S. (2002). Personalisierung im Web Schlüssel zum Stammkunden - Print-Archiv - absatzwirtschaft.de. *Absatzwirtschaft*, (03), 98.

Keiningham, T. L., Aksoy, L., Buoye, A., & Cooil, B. (2011). Customer Loyalty Isn't Enough. Grow Your Share of Wallet. *Harvard Business Review*. Retrieved from https://hbr.org/2011/10/customer-loyalty-isnt-enough-grow-your-share-of-wallet

Kissmetrics. (n.d.). How To Calculate Lifetime Value - The Infographic. Retrieved 23 December 2015, from https://blog.kissmetrics.com/how-to-calculate-lifetime-value/

Kobsa, A. (2007). Privacy-Enhanced Web Personalization. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web* (pp. 628–670). Berlin, Heidelberg, New York: Springer. Retrieved from https://www.ics.uci.edu/~kobsa/papers/2007-AWBS-privacy-kobsa.pdf

Kwak, M. (2001, July 15). Web Sites Learn To Make Smarter Suggestions. Retrieved from http://sloanreview.mit.edu/article/ebusiness-web-sites-learn-to-make-smarter-suggestions/

Lackes, R. (2015). Definition » Data Mining « | Gabler Wirtschaftslexikon. Retrieved 21 December 2015, from http://wirtschaftslexikon.gabler.de/Archiv/57691/data-mining-v8.html

Lee, J.-S., Jun, C.-H., Lee, J., & Kim, S. (2005). Classification-based collaborative filtering using market basket data. *Expert Systems with Applications*, (29), 700–704.

Levin, D. (2015). The Definitive Guide to Predictive Analytics Models for Marketing. Retrieved from http://www.agilone.com/academy/the-definitive-guide-to-predictive-analytics-models-for-marketing/

Liu, D.-R., Lai, C.-H., & Lee, W.-J. (2009). A hybrid of sequential rules and collaborative filtering for product recommendation. *Information Sciences*, *179*(20), 3505–3519. http://doi.org/10.1016/j.ins.2009.06.004

Lyman, P., & Varian, H. R. (2003, October 27). How Much Information?: 2003. Retrieved from http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

Majó, Z., & Révész, B. (2011). The role of collaborative filtering in marketing. *The Diversity of Research at the Szeged Institute of Business Studies*, 59–71.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Miller, S. (2012, February 6). 6 Tips for Using Propensity Models to Improve Response and Revenue. Retrieved from https://www.clickz.com/clickz/column/2142788/tips-propensity-models-improve-response-revenue

Optimove. (2015). About Us. Retrieved 23 December 2015, from http://www.optimove.com/about-us

Oxford University Press. (2015). cluster - definition of cluster in English from the Oxford dictionary. Retrieved 21 December 2015, from http://www.oxforddictionaries.com/definition/english/cluster?q=clustering#cluster__12

Ratner, B. (2015, December 22). Predicting Share of Wallet without Survey Data. Retrieved 28 December 2015, from http://www.geniqmodel.com/res/predicting-share-of-wallet.html

Reinartz, W. J., & Kumar, V. (2001, November). The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. INSEAD R&D. Retrieved from https://flora.insead.edu/fichiersti_wp/inseadwp2001/2001-108.pdf

RJMetrics. (2015). Churn Rate 101. Retrieved 23 December 2015, from http://churn-rate.com

Rosset, S., Perlich, C., Zadrozny, B., Merugu, S., Weiss, S., & Lawrence, R. (n.d.). Wallet estimation models. IBM T.J. Watson Research Center. Retrieved from http://www.tau.ac.il/~saharon/papers/wallet-final.pdf

Rouse, M. (2015). What is churn rate (Predictive Churn Modeling)?: Definition. Retrieved 23 December 2015, from http://searchcrm.techtarget.com/definition/churn-rate

SAS Institute. (2015). Predictive Analytics: What it is and why it matters. Retrieved 21 December 2015, from http://www.sas.com/en_us/insights/analytics/predictive-analytics.html

Skok, D. (2015). Startup Killer: the Cost of Customer Acquisition. Retrieved 23 December 2015, from http://www.forentrepreneurs.com/startup-killer/

Spalding, L. (2003). *Reinventing Customization: New Technologies, New Markets, and New Strategies* (No. SR-807B). USA: Institute for the Future. Retrieved from http://www.iftf.org/our-work/people-technology/technology-horizons/reinventing-customization/

Thomas, J. S., Reinartz, W., & Kumar, V. (2003). Getting the most out of all your customers. *Harvard Business Review*, *82*(7-8), 116–23.

Tirosh, V. (2012, December 10). Behavior-based Customer Segmentation for More Effective Retail Marketing. Retrieved from http://www.optimove.com/blog/customer-segmentation-for-more-effective-marketing

Tirosh, V. (2015, October 14). The Number of Items in a Shopper's First Order Indicates Future Behavior. Retrieved from http://www.optimove.com/blog/first-order-item-quantity-is-a-good-indicator-of-future-behavior

Turban, E., King, D., Lee, J. K., Liang, T.-P., & Turban, D. C. (2015). *Electronic Commerce: A Managerial and Social Networks Perspective* (8th ed.). Cham, Heidelberg, New York, Dordrecht, London: Springer. Retrieved from http://swbplus.bsz-bw.de/bsz425835103cov.htm

Walker, T. (2015). How To Calculate & Increase Customer Lifetime Value. Retrieved from http://conversionxl.com/customer-lifetime-value/

Winoto, P., & Tang, T. (2008). If You Like the Devil Wears Prada the Book, Will You also Enjoy the Devil Wears Prada the Movie? A Study of Cross-Domain Recommendations. *New Generation Computing*, *26*(3), 209–225.

Wu, M. (2013, December 31). Big Data Analytics: Descriptive Vs. Predictive Vs. Prescriptive. Retrieved from http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-descriptive-vs-predictive-vs-prescriptive/d/d-id/1113279

WWW Consortium. (2016). P3P: The Platform for Privacy Preferences. Retrieved 2 January 2016, from http://www.w3.org/P3P/