MODELING CUSTOMERS' ONLINE PURCHASING BEHAVIOR USING CLICKSTREAM DATA

by

Bahar Yeşiladalı

B.S., Chemical Engineering & Food Engineering, İstanbul Technical University,

2013

Submitted to the Institute for Graduate Studies in

Science and Engineering in partial fulfilment of

the requirements for the degree of

Master of Science

Graduate Program in Industrial Engineering

Boğaziçi University

2018

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation and sincere gratitude to Assist. Prof. Mustafa Gökçe Baydoğan, my thesis advisor for all his encouragement, motivations, most importantly his guidance throughout this study. He provided great research opportunities to explore my area of interest and enable to improve my knowledge. I feel indeedly grateful and honoured to have the opportunity to study with him.

I am also most grateful to Prof. Gürhan Kök, for his generous support and priceless advices. Working with him, I have had a great experience.

I would like to thank to Prof. Gürhan Kök and Prof. Taner Bilgiç, the members of my thesis committee, for their valuable comments and suggestions.

I am also thankful to Prof. Victor Martinez de Albeniz and his research assistant, Arnau Planas Bahi for their contribution and their help for obtaining the clickstream data.

I would like to offer my special thanks to my family for their invaluable encouragement, understanding and endless love. I feel always their precious support in every step of my life. Without them, this study could not have been completed.

I would like to extend my thanks to my closest friends for their help and confidence in me.

ABSTRACT

MODELING CUSTOMERS' ONLINE PURCHASING BEHAVIOR USING CLICKSTREAM DATA

In the digitalizing world, the rapid development of the Internet has reshaped the customers' expectations, attitudes and has also changed shopping habits. Visiting online stores for different reasons such as easier product-price comparison, effortless searching and browsing has been becoming much more preferable than the traditional shopping. For this reason, estimating the future behavior of customers is becoming important day by day in order to take advantage of the competitive market. With this motivation, this research focuses on building different behavioral models using clickstream data, which contains the factors that have an impact on the purchasing probability, such as customer past transactions, behavioral frequencies, season and channel. Since the objective of this study is to estimate the likelihood of whether a customer makes a purchase or not, alternative classification approaches such as logistic regression, random forest and boosting are considered. It is determined that models constructed with logistic regression and boosting methods have better predictive accuracy than that of built with other method. According to the results, customers' past behavior, its frequencies, seasonality and conversion rate related factors are found as significant on the purchasing probability. Moreover, when the computation time of logistic regression and boosting methods are benchmarked, it is investigated that logistic regression requires less time to train a model.

ÖZET

TIKLAMA VERİLERİ KULLANILARAK MÜŞTERİLERİN ONLINE SATINALMA DAVRANIŞININ MODELLENMESİ

Dijitalleşen dünyada, İnternetin hızla gelişmesi müşterilerin beklentilerini, tutumlarını yeniden şekillendirdi ve de alışveriş alışkanlıklarını değiştirdi. Daha kolay ürün-fiyat karşılaştırması, zahmetsiz şekilde arama ve tarama yapabilmek gibi farklı nedenlerle çevrimiçi mağazaları ziyaret etmek, geleneksel alışverişten çok daha tercih edilir hale geldi. Bu sebeple de, müşterilerin gelecek davranışlarının tahmini rekabetçi pazarda avantaj sağlamak için günden güne önem kazanmaktadır. Bu motivasyonla, bu çalışma müşteri geçmiş hareketleri, davranışsal frekansları, mevsim ve kanal gibi satınalma olasılığına etkisi olan faktörleri içeren tıklama verisiyle farklı davranış modellerinin oluşturulmasına odaklanmaktadır. Çalışmanın amacı, bir müşterinin satınalma işlemi yapıp yapmayacağının olasılığını tahmin etmek olduğundan, lojistik regresyon, rastgele orman ve boosting yöntemi gibi alternatif sınıflandırma yaklaşımları incelenmiştir. Lojistik regresyon ve boosting methodları ile kurulan modellerin tahmin performanslarının diğer method ile kurulan modellerinkinden daha iyi olduğu saptanmıştır. Sonuçlara göre, müşterilerin geçmiş davranışları, frekansları, mevsimsellik ve dönüşüm oranı ile ilgili faktörlerin satınalma olasılığı üzerinde etkili olduğu bulunmuştur. Ayrıca, lojistik regresyon ve boosting yöntemlerinin hesaplama zamanı kıyaslandığında, lojistik regresyonun bir modelin geliştirilmesi için daha az zaman gerektirdiği tespit edilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

B	Number of trees
d	Interaction depth
F	Predicted negative class
H_0	Null hypothesis
k	Number of parameters
L(.)	Likelihood function
m	Number of predictors at each split
n	True negative class
N	Total negatives
p	True positive class
P	Total positives
t	Total number of predictors
T	Predicted positive class
β_0	Intercept
β_1	Coefficient of first variable
$\widehat{\beta 0}$	Estimated value of intercept
$\widehat{\beta 1}$	Estimated value of coefficient of first variable
$\hat{\beta}$	Maximized value of beta function
λ	The shrinkage parameter

LIST OF ACRONYMS/ABBREVIATIONS

AIC	Akaike's Information Criterion
AUC	Area under the Receiver Operating Curve
B2B	Business to Business
B2C	Business to Consumer
CF	Consumers Consulting and Following a Product Recommendation
CNF	Consumers Consulting but not Following a Product Recommendation
C2B	Consumer to Business
C2C	Consumer to Consumer
EC	Electronic Commerce
FP	False positives
GLM	Generalized Linear Models
MLE	Maximum Likelihood Estimation
NC	Consumers not Consulting an Online Product Recommendation
OOB	Out of Bag Error
ROC	Receiver Operating Curve
TN	True negatives
TP	True positives

1. INTRODUCTION

Internet allows us to do many things that would not have ever imagined few years ago. Therefore, it is almost impossible to think the world without the Internet, nowadays. So much so that it takes up a lot of space in people's lives. Besides, the way people live is changing day after day due to the convenience enabled with digitalization. The evolution of people's lifestyles causes the differentiation of traditional way of communication, entertainment and even shopping.

Moreover, technology helps build a platform that enables the business transactions available over the Internet for huge population. The rapid developments on the Internet change the way in which products and services are sold and bought, thus causing e-commerce to become more important. The more people engage in Internet for various purposes, the more opportunities arise to develop for e-commerce.

According to the Statista ecommerce market results (Statista, 2017), it is predicted that the number of e-commerce active paying customers all over the world will increase over time. By 2022, the number of users is expected to be 2.5 billion, seen in Figure 1.1.

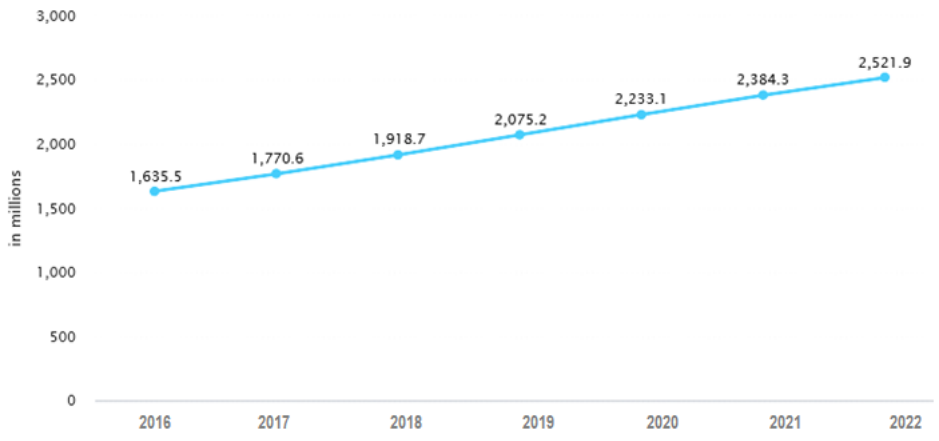

Figure 1.1. The Number of Users in E-commerce Market Worldwide (Statista, 2017).

Looking at the change in worldwide e-commerce revenue from the Figure 1.2, it is estimated that in 2020 the total revenue in different markets will reach 2 trillion US dollar by increasing every year (Statista, 2017).

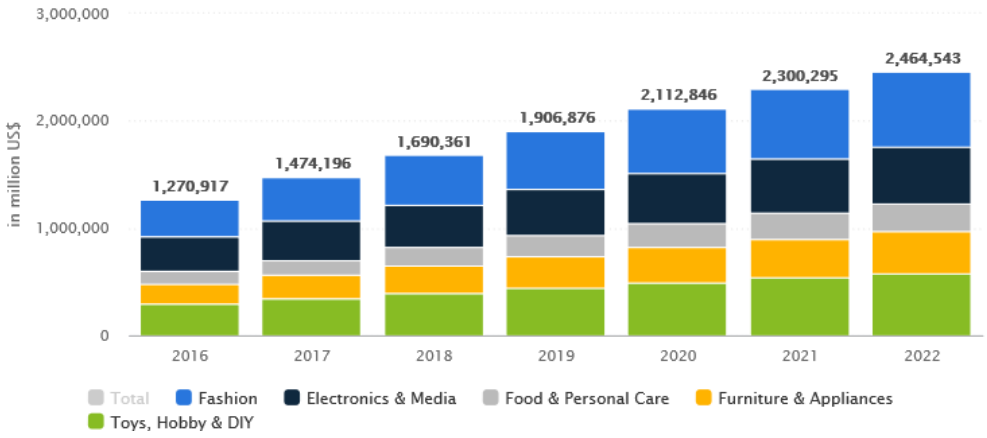

Figure 1.2. Revenue in Million US Dollar (Statista, 2017).

In addition, eMarketer states that retail e-commerce sales will rise to 2,860 trillion dollars in 2018 and compose 11.5% of total retail spending worldwide in Figure 1.3. While the growth in total retail sales is slowing, the digital portion of sales continues to grow swiftly with a 21.6% growth rate. Also, it is expected that retail e-commerce sales will reach 4.058 trillion dollars by 2020 and will account for approximately 15% of total retail spending in the same year (eMarketer, 2016). For this reason, e-commerce has a huge market potential with growing sales volume from day to day. The success of e-commerce is the indication of the evolution of business type from brick-and-mortar model to brick-and-click model (Lim *et al.*, 2016).

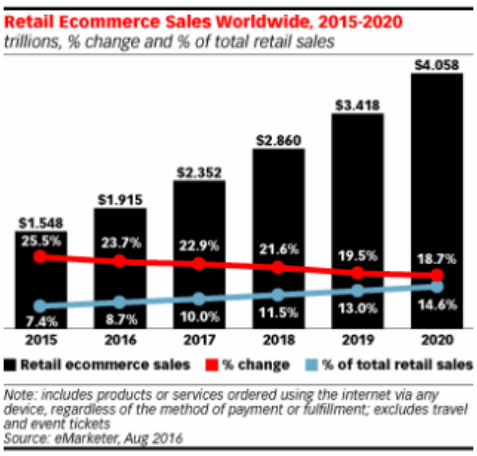

Figure 1.3. Retail E-commerce Sales Worldwide (eMarketer, 2016).

Due to the around-the-clock availability, rapid and easy accessibility, offered various product and service range, consumers today no longer prefer to go to the nearest store; they instead grab their computers or smart phones and use their digital device along the entire path to purchase. They complete their tasks like browsing, information search, and product - price comparison in more easier and convenient way in an online medium. The convenience of online shopping brings about an emerging trend among consumers as they can buy anything wherever and whatever they want. As can be seen in Figure 1.4, in 2017, 60.2% of Internet users prefer to end up their purchases online. In 2021, with the digital buyer penetration it is estimated to exceed 65% of Internet users worldwide (Statista, 2017). Therefore, the gap between online retail and its brick-and-mortar counterpart is enlarging day by day, and customer experience journey is evolving in turn.

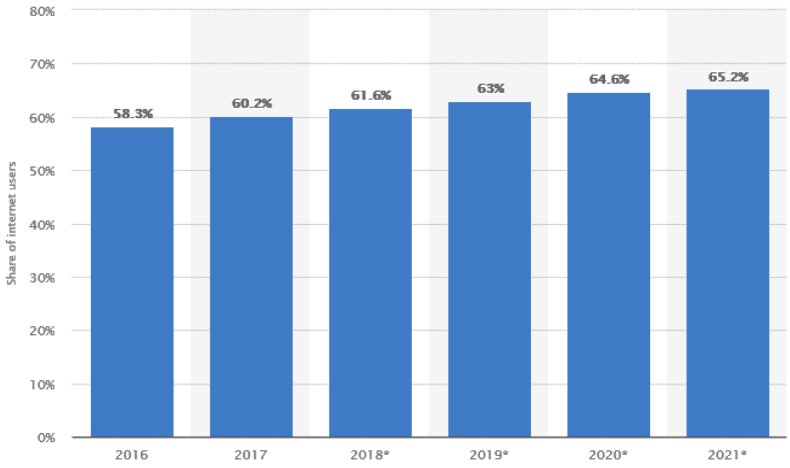

Figure 1.4. Digital Buyer Penetration Worldwide from 2016 to 2021 (Statista, 2017).

The expeditious evolution of customer journey induces a dynamic market place and challenge among all shopping channels. In order to survive in a competitive environment, companies should adapt their marketing strategies and business models to the dynamic conversion. Therefore, the business world has shifted its focus from product or service oriented to customer driven, recently.

Retailers are now seeking to know who their customers are and how their customer respond to the service they offer. At any given time, there are millions of potential customers of online shopping mediums. For retail marketers, the problem is to know what to do with the customer data rather than collecting it. According to survey results of a marketing firm in Figure 1.4, 34% of UK and US retailers identify creating an insight from

past data as a challenge. Also, same survey shows that demographic and geographic data are commonly used as criteria for personalization; while psychographic data and behavioral data are less used for targeting, at 53% and 34%, respectively (eMarketer, 2017). Since shopping is about more than just buying things, there are a complicated relationship between buyer and seller who affects the shopper not only with the product itself but also with psychological and behavioral influences. Therefore, understanding what and how customers shop with its mental processes behind has become prominent recently.

Figure 1.5. Challenging Marketing Initiatives (eMarketer, 2017).

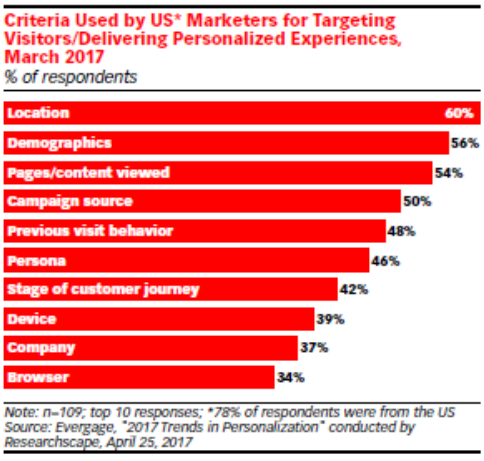

Figure 1.6. Criteria Used by US Marketers for Targeting Visitors (eMarketer, 2017).

Although a lot of studies have focused on understanding the customer; the rationale behind their chosen behaviors is left somewhat underserved. However, there is an invaluable resource in an online medium for understanding customer and optimizing user experience, which is clickstream data. Since clickstream data tracks all historical

transactions, it is possible to identify each customer as a unique individual by creating a data driven customer profile. Clickstream data gives information about which products is desirable for customers (browsing /adding to cart/ building a wish list), which products categories are most relevant to that customer, the price sensitivity of users and all other related information on the customers' buying decision (Tulsyan, 2017). Through clickstream data, it is possible to find various conversion rates - which customers visits turn into an action- in terms of transactions (buy, add to cart, remove from the cart), search (natural, paid search) or traffic (site navigation within websites or across websites, repeated visits). Different conversion rates can be used as different metrics to segment various users, to improve the quality of traffic or strength of offers (Chaffey, 2017). As the visit to purchase conversion rate is low, there is much concern among retail marketers (Bucklin *et al*., 2002). Despite of big data available, few researches have been made to find out predicting customers' intention to buy online.

If the tendency to buy of customers can be predictable, it is possible to decide who to target due to the knowledge which customers are ready to make a purchase. Also, "the propensity to buy models" provides information to optimize the strategies from the email send frequency, to sales representative time, to money, containing discounts. Hereby, retailer provide right action in their offer by having knowledge who to aim. Marketing strategies can be reshaped accordingly for example, those who are most likely to buy will not need high discounts; while others that are probably abandon their cart may need a more attractive incentive as a sweetener. Therefore, it is possible to choose the level of discounts for different segments through the learnings from purchasing likelihood models (Artun and Levin, 2015).

Due to the importance of understanding customer needs and expectations, making true predictions about a customers' future behavior having propensity to buy is evolving into an interesting area. Unlike the other researches proposing predictive models with determinants of number of visits, purchases or view time for product detail (Moe and Fader, 2004; Olbrich and Holsing, 2011), in this study, it is aimed to improve our knowledge about customer behavior by analyzing click stream data and quantify the effects of past behaviors, time duration, channels by considering other aspects as well, such as seasonality, total amount spent, behavioral frequencies etc. Moreover, it is tried to

understand and give an insight about how purchasing behavior of online customers' is affected.

The goal of this research is to propose a predictive model of purchasing behavior using clickstream data. Questions wanted to be responded in this research are namely:

- How do past transactions (visits, purchases etc.) affect customer's buying decision in the near future?
- What are the effects of time duration of transactions, channel, season and times of day on purchasing behavior?
- Which factors are more significant for online purchasing stage?
- How do these factors influence purchasing behavior?
- Which statistical techniques can be used to model customers' purchasing behavior and how?
- How precise can be estimated the customer' purchasing behavior model?

Finding reasonable answers to these questions may bring about understanding customers' rationale in a much better way. In order to fulfill this purpose, customer sales data obtained from an online website is analyzed and different prediction models with various statistical approaches are built to assess the significance of factors influencing purchasing decision.

The thesis consists of three literal and two analytical chapters. In the analytical chapters all calculations are performed using R program. In Chapter 2, related studies in this field with different customer behavior models are explained. In Chapter 3, conceptual framework including a review of clickstream data, electronic commerce and types of online shoppers are given as a background. In Chapter 4, as a methodological framework, general information about statistical models used in this study is covered. More specifically, it involves logistic regression, classification tree, random forest and boosting model. Chapter 5 focuses on properties of the clickstream data and its analysis with a descriptive statistical approach. In Chapter 6, various customer behavior models built with different factors are evaluated by comparing different statistical approaches. Lastly, Chapter 7 brings to a conclusion with an interpretation and critic of findings, the contribution and restraints of the research together with inferences for future studies.

2. BACKGROUND

2.1. Clickstream Data

Clickstream data is known as the comprehensive electronic footprint of users which is followed and recorded about web usage and online actions. It is an electronic history of a user's transaction on the web including the route a user follow, like the pages viewed and time expended on every page. The recorded track presents a series of decisions taken both within a website (for instance, which web pages to review, how much time spent to visit page, and whether to buy something) and across websites (e.g. which sites to visit) (Bucklin *et al*., 2002; Moe and Fader, 2004).

Offline data captures only purchasing activities, while non-purchasing activities such as visit characteristics are neglected or need to be gathered through self-reports which might be inaccurate. However, online data enable to capture both the purchasing and non-purchasing activities (e.g. product viewing) in a prompt and proper manner (Moe and Fader, 2004; Senecal *et al.*, 2005). Since internet choice behavior includes a number of interdependent decisions; both customer and retailer have an essential role in structuring the consecutive choice actions. Therefore, the clickstream data is a modern day treasure that can be excavated in terms of products and services. The collected data is used to keep track of operations, forecast consumer needs and understand the customer expectation in order to offer products to the market successfully. Through modeling methods to define users' behavior in as much as their past transactions and choices, deep insights are offered for marketing strategies (Parise *et al*., 2012). By using clickstream data, several types of goods and services are examined and inquiries into decision making are permitted in dynamic medium in which marketers can customize the stimuli according to the customer expectations. Therefore, clickstream data presents a notable opportunity to improve the understanding and predicting the consumer online behavior.

Clickstream data are seized in several ways. First way, server log files store all information transmitted from the user's computer to the server during a web usage. Second, a panel data record web addresses of all pages by transferring the electronic pathway from the consumer's computer to the data provider (e.g. MediaMetrix). Third, data is captured through a user's Internet service provider (Bucklin *et al*., 2002).

Due to the prevalence of the Internet and its impact on source of information, there is a growing interest on research about the prediction of online customer behavior. Several different approaches to understand and model online purchase behavior of visitors have emerged, by tracking clickstream data variously (Bucklin and Sismerio, 2009). Some of the researches analyze the clickstream data by applying choice models. According to Table 2.1, there are four fields of choice model applications where researches have started to investigate (Bucklin *et al*., 2002):

- Within site navigation choice
- Choice among websites
- Purchase choices involving electronic agents, especially shopbots
- E-commerce purchase decisions

Table 2.1. Internet choice applications (Bucklin *et al*., 2002).

	Within Website	**Across Websites**
Search	Site navigation	Site Choice
Purchase	E-commerce and recommendation systems	Shopbots

Site navigation choice include the decision whether stay on a given page, which pages to view, the number of viewed pages, duration of viewing a page. These decisions may affect the stickiness of a website. The second research area is about user's decision to choose one website versus another. These decisions are related with the site loyalty. Thirdly, researchers have begun to investigate choice process that users make at shopbots which presents a group of alternative goods and various prices from competitive firms.

These decisions are relevant to online purchases across multiple sites. Lastly, other application area involving within site behavior focuses explicitly on e-commerce. Since the visit to purchase conversion rate is low, there is much concern among practitioners. To overcome this problem, understanding factors affecting the purchase decision and building models representing the online customer behavior are necessary. The first essential step for choice models is assessing the objective of individual Internet user (e.g. a person searching information, or a person having intention to make a purchase?) Another step is modeling the probability of an online-purchase by considering a series conditional user's activity (e.g. visit to site, items added to the cart, order submitted) (Bucklin *et al*., 2002).

2.2. Electronic Commerce (EC)

Electronic commerce (E-commerce) stands for the process of purchasing and selling of goods and services, or relaying information via electronic channels, mainly the Internet. The revolution of e-commerce industry has started in the 1990s in the world. E-commerce business model have four types, which are Business to Business (B2B), Business to Consumer (B2C), Consumer to Consumer (C2C) and Consumer to Business (C2B). Electronic commerce is performed through using various applications, such as electronic mail, online shopping lists, carts and web services. Some of the advantages of e-commerce comprise its around-the-clock availability, rapid and easy accessibility, offered variety of goods and services for the consumer, and international reach. A few of its perceived drawbacks involve not having the opportunity of touch and feel a product before buying, limited customer service and required waiting period for delivery (TechTarget, 2016).

E-commerce sales have shown a remarkable growth with its diffusive power in the past decade. Especially, internet has emerged as a beneficial marketing tool for e-commerce to offer service as a platform for up country and overseas transactions (Lim *et al.,* 2016). The penetration of Internet buyers in Figure 2.1 has prevailed in the worldwide and continues to increase in the future (eMarketer, 2014).

Digital Buyer Penetration Worldwide, by Region, 2013-2018
% of internet users

	2013	2014	2015	2016	2017	2018
North America	72.0%	73.6%	74.9%	76.3%	77.7%	78.8%
Western Europe	64.0%	65.2%	66.3%	76.3%	68.2%	69.0%
Asia-Pacific	42.1%	44.1%	46.8%	48.9%	50.4%	50.9%
Central & Eastern Europe	41.6%	43.4%	44.3%	44.4%	44.6%	44.6%
Middle East & Africa	31.3%	33.1%	34.0%	35.0%	36.0%	37.0%
Latin America	28.2%	29.9%	30.9%	31.8%	32.7%	33.7%
Worldwide	**41.3%**	**42.7%**	**44.3%**	**45.4%**	**46.4%**	**47.3%**

Note: ages 14+; internet users who have made at least one purchase via any digital channel during the calendar year, including online, mobile and tablet purchases
Source: eMarketer, July 2014

Figure 2.1. Digital Buyer Penetration in the World (eMarketer, 2014).

According to eMarketer forecasts in Figure 2.2, retail ecommerce sales will reach $1.915 trillion with a 23.7% growth rate in 2016, composing 8.7% of worldwide total retail consumption. eMarketer estimates ecommerce retail sales in 2020 will show increase to $4.058 trillion, accounting for approximately 15% of whole sales in retail sector in the same year (eMarketer, 2016). Therefore, e-commerce has enormous market potential due to the continuous sales increase. The success of e-commerce converts the existent business type that is brick-and-mortar model to brick-and-click model (Lim *et al.*, 2016).

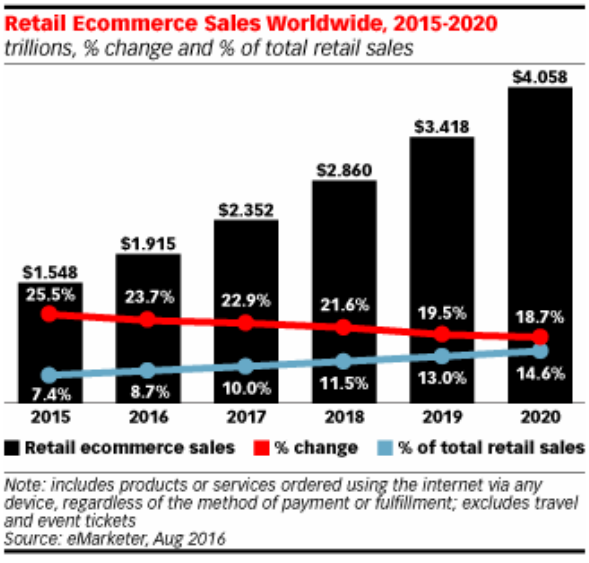

Figure 2.2. E-commerce Sales Worldwide (eMarketer, 2016).

An essential business model in e-commerce is online shopping. E-tail is sometimes used as a term relating to transactional processes for online shopping (TechTarget, 2016). Online retailing is a model of B2C e-commerce type where businesses sell goods or

services to consumers. A customer's online buying experience involves information inquire, product assessment, making a decision, purchasing, delivery, returns and after sales service (Rewatkar, 2014).

Consumers' shopping behavior is changing rapidly with the fast evolution in the online retailer system. Shoppers today no longer prefer to go to the nearest store; they instead grab their computers or smart phones and use their digital device along the entire path to purchase. Since online shopping offers some conveniences like low prices, broad alternative products, around-the-clock availability, deliver services to the even customer's home that brick and mortar stores cannot, online shopping stores attract people more than offline stores (Nielsen, 2014; TechTarget, 2016). eMarketer foresees that retail sales in online medium will become over twice as large as sales in 2015 within four years (Nielsen, 2016). The online stores differ in a lots of ways such as time & travel cost and information availability from the traditional market in which the direct experience of the transactions and face to face contact with sales person are necessary (Chiang and Dholakia, 2003; Degeratu *et al*., 2000; Moe and Fader, 2004; Yeh *et al*., 2007).

One of the most outstanding differences of customer behavior among in online and traditional channels is the low shipping cost which is necessary to visit an online store (Moe and Fader, 2004). In addition to location constraint, time constraint is also key factor influencing the preference of online shopping and traditional shopping (Chiang and Dholakia, 2003). Offline shopping behavior studies show that customer's decision about which stores to choice for shopping and buying behavior is dependent on the costs – both financial and emotional- (Moe and Fader, 2004). Customers are able to make a purchase from wherever and whenever they want with just a few clicks since time and travel cost are eliminated virtually in online environment. Thus, consumers can get further information online and make comparison of the product price amongst online retail websites (Chiang and Dholakia, 2003). Nevertheless, being costless to visit a virtual store influences customers' behavior in a various way. Initially, online shoppers may tend to visit a store without having a buying intention due to the lower costs; while in the offline world wasting a trip - not buying- is less likely to be since the consumer incurs costs through spending the time and making an effort to perform a store visit. Therefore, online shopping has lower conversion rates. Moreover, online shoppers are more probably to postpone their

buying decision and finalize purchasing stage later. For these reasons, online shoppers are inclined to make a couple of visits to the same store (Moe and Fader, 2004). Also, it is observed that concern about bank card security was utmost significant disincentive of online shopping, due to the lack of ability to see and touch the product (56% of consumers), having distrust to smooth online ordering (43%), having worries about personal data (43%), and the cost of shipping (43%) (Swinyard and Smith, 2003).

Furthermore, product attributes drive differences in online and offline shopping. Several studies indicate that high touch products that shoppers are in the need of touching, smelling or trying on are those that slightly necessitate physical existence at the purchasing phase. Nonetheless, the low touch products like airplane tickets and electronic devices are preferred to be bought online because of importance of shopping them quickly. According to survey conducted by Nielsen, non-consumable products have higher online browse/buy intention rate than consumable ones in Figure 2.3 (Nielsen, 2014).

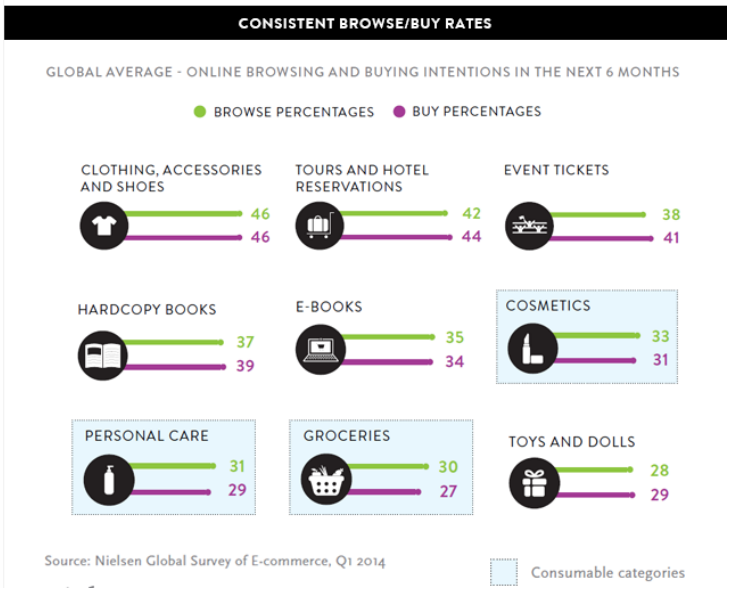

Figure 2.3. Online Browsing and Buying Rates (Nielsen, 2014).

In brief, when characteristics like broad selection availability and shopping fast are predominant, online shopping is preferable. If face to face service and ability for seeing and touching attributes are predominant, offline shopping is more desirable than another (Levin *et al*., 2005).

In addition, information availability influences customer decision making process and this is a distinctive factor for online shopping and offline shopping. The offline

environment offers more total information available for the product categories which have numerous sensory features (for instance; fruits – visual cues) to ease customer choices; whereas the online environment presents more information for product categories having numerous non-sensory attributes (for example; margarine – fat content) than offline.

Also, in an online environment, information availability is easier to search by sorting on price or on the promotion indicator. The possibility of customers to get more knowledge about both price and other specifications is pointed out as an important difference between online and offline stores (Degeratu *et al*., 2000). Therefore, in comparison to offline shopping, the online medium enables shopping in an interactive and customized way that has an influence on customer satisfaction and loyalty. With more associated additional information, customers decide better and improve their choices, which in turn, lead to higher satisfaction when online environments are preferred. Since the satisfaction has a reciprocal correlation with the loyalty such that each supports the other in a positive way (Shankar *et al*., 2003).

Briefly, the discriminations between online and traditional business types are summed up in Table 2.2.

Table 2.2. Comparison of online shopping and offline shopping (Chiang and Dholakia, 2003; Degeratu *et al*., 2000; Moe and Fader, 2004; Nielsen, 2014; TechTarget, 2016; Yeh *et al*., 2007).

Factors	**Offline Shopping**	**Online Shopping**
Shopping Medium	Physical store	Website
Time Cost	High	Low
Transportation Cost	High	Low
Availability	Limited	24/7
Price	Promotion	Price comparison / Promotion indicator
Variety of Products	Narrow	Broad
Product Type	High touch / Experienced products	Low touch / Search products
Information Availability	Sales person	Online search / Product comparison
Service	Service quality offered	Payment security / Privacy

2.3. Factors Influencing Online Shopping

According to literature research on factors affecting online shopping behavior, in general sense the main factors are summarized as three main categories in Figure 2.4, particularly: consumer concerning factors, web vendor concerning factors, and web/website concerning factors (Turan, 2011).

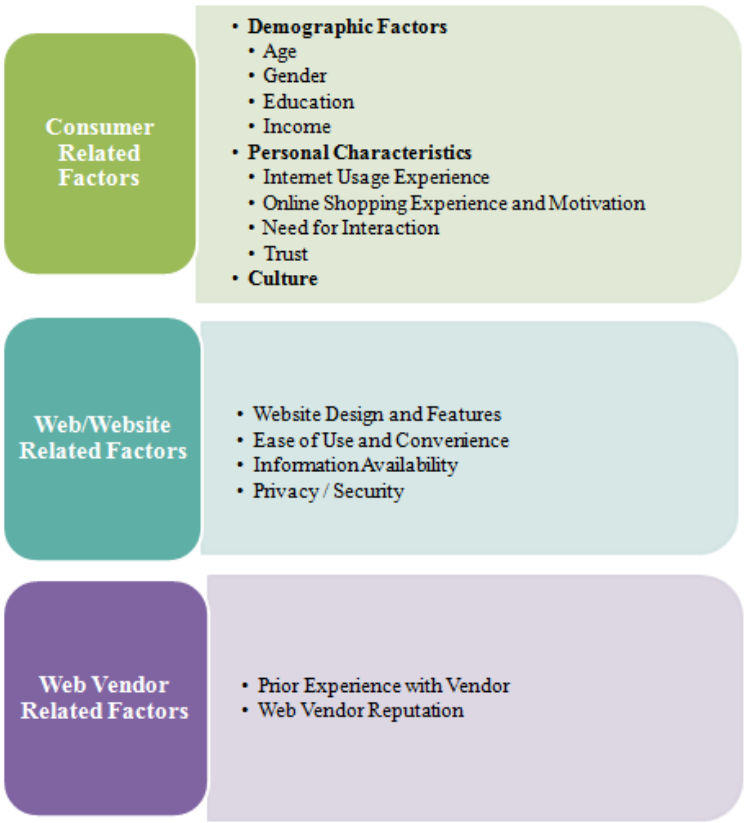

Figure 2.4. Online Shopping Factors (Turan, 2011).

2.3.1. Consumer Related Factors

Consumer related factors are divided into three categories, which are demographic factors, personal traits and culture. These factors play an essential role in customer decision making model.

Firstly, the primary category of customer related factors are demographic factors, which are key variables influencing online shopping behavior. The main demographic determinants are age, gender, education and income (Bednarowska and Jedruszek, 2012; Hou, 2015; Punj, 2012; Rodgers and Harris, 2003; Turan, 2011; Zhou *et al.,* 2007).

- Age is one of the first key factors having effect on buying decision. Several researches demonstrate that young people are more inclined to use technology and to make online purchases (Turan, 2011). Additionally, it is found that people aged below 34 had purchased online in the last half year with the percentage amounts to 68%, whereas people aged up to 55 the percentage comes to 42%, on the oldest people aged above 55 the percentage was merely 14% (Bednarowska and Jedruszek, 2012).

- The second demographic factor, about gender difference studies show that women are more likely to be shoppers preferring to shop from physical stores, while men have more tendencies to be online customers (Hou, 2015; Rodgers and Harris, 2003). Also, men are more inclined to shop online since they perceive the online environment as a more favorable shopping medium than do women (Rodgers and Harris, 2003). It is also stated that the interest in products differs among the genders. Men are more interested in computer and electronic products, while women are mostly interested in food, clothes and cosmetics (Zhou *et al.,* 2007).

- Another demographic factor influence shopping behavior is education. It is stated that the level of education of online shoppers is higher than that of online non-shoppers since education is relevant to the individuals' info search and info usage attitude (Hou, 2015). Therefore, highly educated people are more positive to online shopping (Turan, 2011).

- Last demographic factor affecting purchasing behavior is income. Lower income consumers do not have information about the advantages of electronic commerce when compared with their counterparts, since less educated and elderly people are slow to adapt and utilize the Internet. The consumers having higher income have a greater inclination on saving time, whereas their counterparts are more likely to save money (Punj, 2012). Generally, it is argued in the literature that the income level of the consumer affects online shopping behavior positively (Turan, 2011).

Secondly, personal characteristics are considered as factors related to customer related factors. Consumer decision making process in the online medium is influenced by the personal characteristics of the consumer, namely the level of expertise with the Internet, online shopping experience, motivation, the need for interaction, and trust on ecommerce (Close and Kukar-Kinney, 2010; Hou, 2015; Sultan and Uddin, 2011; Turan, 2011; Wang *et al.*, 2008).

- People with more closely related to Internet have a positive attitude toward online shopping (Turan, 2011). If people use the Internet longer and more frequently, they are more probably to make online purchase (Hou, 2015).

- Furthermore, if the online shopping experience of the consumer is satisfactory and positive, this encourages the person to engage in online shopping (Turan, 2011). The motivations of making secure online price promotions, gathering more product details, acquiring desired items and having a pleasure promote the willingness of the consumer to shop online (Close and Kukar-Kinney, 2010).

- Besides, people high need for interaction is not willing to shop online since there is not any possibility to have a face to face communication with the sales representative in online shopping (Turan, 2011).

- Also, trust is a vital determinant of online shopping behavior. The most outstanding problem encountered in online environment stems from the lack of trust on safeguarding the individual information during online transactions (Turan, 2011). Online transition of personal information is easily intercepted and illegally utilized (Wang et al., 2008). Therefore, many consumers refrain from online shopping due to different reasons like non-delivery risk, privacy factors, and so on (Sultan and Uddin, 2011).

Thirdly, the culture of country to which customer pertains shapes the online shopping decisions of customer since culture reflects a common set of values that affect social believes, attitudes, choices, and reactions (Turan, 2011; Zhou *et al.*, 2007).

In brief, demographic, attitudinal and cultural impacts related to customer side like education, age, degree of Internet usage, security and the enjoyment of shopping, play an outstanding role on online purchase behavior (Punj, 2012).

2.3.2. Web/Website Related Factors

Web and website related factors influencing the online customer behavior are categorized into four groups, which are website design, ease of use, information availability and security. These are external factors having strong influence on online decision making process.

Firstly, website design and characteristics are key elements for online buying behavior. The animated pictures, scenario images, and colors produce emotional effects on people, therefore; the structure of web design is essential for attracting and retaining online customers (Yeh *et al*., 2007). Moreover, the graphic style, improved product list, feedback section and the amount of updates to the website in the earliest time and links to other sites are influential to encourage consumers to shop online, as well (Turan, 2011).

Secondly, ease of use is the factor defined as uncluttered screens, explicit organization, logical flow, and navigational patterns facilitating the efficient use of website. It is stated that the more the ease of use on online medium, the more the motivation to engage in e-commerce (Turan, 2011). Moreover, customers can easily browse or search information in the product catalog through online, but it is difficult and time consuming to find the same product or item manually in a physical store. Thereby, online consumers can more easily compare the price of products by saving time with a low physical effort (Sultan and Uddin, 2011).

Moreover, informativeness is defined as a website's potential to provide information to the customers about products for ensuring utmost pleasure. The consumers examine the products depending on the online information including text, tables, graphs, photos and video in order to assess and choose the product they need. If the website enables high quality product information with ads, consumers make better decisions with high satisfaction level, in turn this leads to improve online shopping attitude (Turan, 2011).

Lastly, privacy is another one of the most vital factors having influence on online shopping since information security is a major barrier which limits buying on the web (Hou, 2015; Sultan and Uddin, 2011; Turan, 2011). As it is stated that there is a large segment of internet users who do not prefer online shopping due to their worries about the safety of their personal information. The more risk they face, the less conduct of online shopping they attend since the benefit of online shopping may not compensate the loss originating from the privacy issue (Turan, 2011).

In short, website attributes, ease of use, availability of information and security are the web/website related determinants having significant effects on customer decision making.

2.3.3. Web Vendor Related Factors

There are two major elements related to web vendor shaping the customer behavior on the web, which are prior experience with vendor and the reputation of web vendor.

Firstly, online shopping experience with a web vendor has a strong influence on the future actions of the customer towards online shopping. The more positive and satisfied the consumer's past experience with web vendors, the more the user have willingness to make a purchase on the web.

Secondly, web-vendor reputation is a critical enabler of online actions, because it diminishes the perceived risk and worry of the consumer. Web vendor's perceived reputation and size affect the consumers' trust on the web vendor in a positive way; in turn this gives rise to positive attitude and willingness of customer to shop online (Turan, 2011).

Briefly, prior experience with web vendor and the reputation of web vendor are essential influencers of online customer behavior.

2.4. Types of Online Shoppers

Internet users are classified into two groups as Internet shoppers and Internet browsers. While Internet shoppers are those who make online purchases, Internet browsers are those who just surf in the internet out of shopping purpose (Sultan and Uddin, 2011).

Also, there is a differentiation between online shopper and their counterparts; online shoppers get more nervous about convenience and time saving while non online shoppers are concerned about security, confidentiality and on time delivery (Turan, 2011). Several studies have shown that online shoppers are younger, better educated, wealthier, more probably to expend their time on computer and consider online shopping to be practical, less nervous about the monetary loss arising from online actions, and to have more time constraint than non-Internet shoppers (Swinyard and Smith, 2003). Also, it is stated that online shoppers are more convenience searchers than their counterparts (Hou, 2015).

Moe constitutes a typology of store visits where visits differ from each other in terms of shoppers' goals. According to the research, visits are classified in Table 2.3 as buying, browsing, information seeking, or knowledge building visit depending on in-store navigational patterns and general contents of viewed pages. The content of pages viewed is necessary for both defining the type of shopper involved and predicting purchases.

Table 2.3. Typology of shopping behavior (Moe, 2003).

Purchasing Horizon	**Search Behavior**	
	Directed	**Exploratory**
Immediate	Directed buying	Hedonic browsing
Future	Search/deliberation	Knowledge building

- Directed buyers are persons having an intention to make a purchase and targeted toward a specific and immediate purchase. These shoppers have very focused search patterns, that is more product level pages rather than category level pages are viewed by them. These consumers view the targeted products repeatedly.

- Search and deliberation visits are goal directed like directed buying visits; however, they are motivated by a future purchase not an immediate purchase. These buyers are not still sure about which specific product within a product category to purchase. They have a focused search within product category.

- Hedonic browsers are less focused in store behavior. Their shopping goal is dominated by exploratory search behavior. Therefore, they spent most of their time to view the category level pages than the product level pages. Since hedonic browsers explore new stimuli during visits, they exhibit a lot more variety in both products and categories viewed than directed buyers and buyers having search/deliberation strategy.

- Knowledge building shoppers' objective is to increase their expertise on product and marketplace. They focus more on informational pages such as advice columns, community discussion areas. These consumers do not consider any specific purchase, but acquired information may affect the future purchase (Moe, 2003).

According to another research customers are categorized into four archetypes by considering decision making type (maximizers/satisficers) and the information of products (high/low);

- Satisficers with low level of information
- Satisficers with high level of information
- Maximizers with low level of information
- Maximizers with high level of information

Whereas maximizers are defined as persons seeking the best possible result, searching for more information, as a result browsing more intensively; sacrificers are identified as persons who prefer a well enough choice that satisfies several criteria, spend less time to search, consider smaller consideration set (Karimi *et al.*, 2015).

Another study shows that price motivated buyers more likely to place products in their online shopping carts than those not stimulated by discount. Hence, promotion programs are a stimulus for the usage of online cart and as a result, buying (Close and

Kukar-Kinney, 2010). Additionally, long term customers have more tendencies to buy more and less costly to deliver a service to them; while gaining new customers or replacing present ones is more expensive and risky since those are more sensitive to maintain their churning attitude in the future (Lariviere and Van den Poel, 2005).

2.5. Customer Decision Making Processes

Customer decision process involves five stages seen in Figure 2.5, beginning with problem definition and information inquiry and following with assessment of alternatives, buying decision and eventually post purchase behavior. The problem definition begins with the perception of need and description of expectations and then customers collect and utilize the internal and external sources for information search. In the next step, customers evaluate the alternatives with the assessment of the products by giving weights. Then, they move toward purchase decision where they face some probabilities like from whom to buy and when to buy. Once they have made a purchase, post purchase behavior stage starts, which includes whether they are satisfied or not with their choices (Sultan and Uddin, 2011).

Figure 2.5. The Stages of Customer Decision Process (Sultan and Uddin, 2011).

The current purchase intent is found as the strongest factor of online purchasing, as well as cart placement. Many shoppers using their virtual cart have an intention to buy at that time since the purchase process (e.g. one click buying option) is simple, once the item is placed into the cart. In addition, a sizeable segment of shoppers utilize their online shopping carts with the aim of storing desired item list for a possible future purchase or compiling a wish list. Therefore, e-tailers need to provide their buyers with special shopping carts that hold the items after the customers log off without purchasing so that they do not require any effort of finding products again (Close and Kukar-Kinney, 2010).

Consumers' decision making strategies are different from each other's. Many researches show that the product feature influences kind of information searched and in turn the decision process to make a choice. When encountered with a product choice, consumers perform an internal search in accordance with the personal preferences (e.g. relying on the previous knowledge of brands) and an external search including actions like collecting more information about brands and searching product advices. Due to the unique characteristics of Internet, the way consumers search for information is modified and facilitated with the presence of new information sources like recommendation tools and intelligent-agent based systems.

Senecal *et al.* (2005) try to find out the differences between online choices of consumers (for example, time spent during decision making, pages viewed etc.) and their influences on the complexity of online purchasing attitude. They conduct an experiment, in which subjects are demanded to answer an online questionnaire measuring subjective knowledge of product category and to complete an online shopping task providing product recommendations. According to the research results, the complexity of consumers' online purchasing attitude does not be influenced by the attendants' subjective information of the product category. It is also observed that the shopping behavior of subjects not consulting a product recommendation is less complicated than that of shoppers consulting the product recommendation, as they have less information to process. Besides, participants who decide not to take a product advice (NC) view fewer pages, view fewer product detail pages and review the pages they visited before. In the study, one of the hypothesize is about that subjects consulting but not following (CNF) a product recommendation have a more broad and complex shopping behavior (e.g. spending more time) than the subjects consulting and following (CF) a product suggestion, since the recommendation and the preferred alternative do not match with each other and leads to increase in decision difficulty and more deliberation. However, online shopping behavior of CNF and CF is surprisingly found as identical. Another result shows that there is not any difference in time spent to select a product between NC and CF.

Karimi *et al.* (2015) try to find out how the individuals' choice behavior and their information about products affect the online purchase process. In the study, consumers are categorized into four types by considering decision making style (maximizers/satisficers) and the knowledge of products (high/low). Regarding decision making style, maximizers

are defined as consumers seeking best possible result, while satisficers are defined as consumers selecting a well enough choice that matches with several criteria. The dependent variables of decision-making style are: (1) the number of cycles, (2) the duration, (3) the number of alternatives, and (4) the number of criteria. The findings show that maximisers with low knowledge have the highest number of cycles, expend more time, find more alternatives and assess more criteria than other three type consumers. Unlike maximisers (with low and high knowledge), satisficers (with low and high knowledge) have a lower number of consideration sets with a same path for other outcomes. In the study, Karimi states that: “People with low level of knowledge have decision making processes with a higher number of cycles, a higher number of alternatives, a higher number of criteria and longer duration compared with their counterparts”. In short, maximizers carry out more complex decision making processes than satisficers.

3. LITERATURE REVIEW

Marketing researches show that customers' behavior when purchasing products differs in an online environment with regard to in a traditional shopping environment (Lee *et al*., 2007). Therefore, several studies have recently focused on modeling customer preferences by taking different aspects into consideration, which are summarized in Table 3.1 by comparing their similarities and differences with the proposed model in this thesis.

In order to understand (1) browsing behavior and web usage, (2) Internet advertising and (3) online purchasing behavior (electronic commerce), key developments from clickstream data analysis are pointed out by Bucklin and Sismerio. In the article, studies based on clickstream data are discussed, their limitations for predicting the behavior of Internet users are explained and future opportunities for new researches and emerging areas are defined. Three broad research themes are described in detail. Firstly, researches into how people choice a website, navigate the new medium, search across websites are discussed. According to result of a study, it is found that visitors spend less time if they visit the same website; that is users' behavior change consistently with the learning effects. Also, the empirical results about the navigational behavior indicate that dynamic model of browsing behavior is better predictive than a static model. Secondly, researches about advertising, email, paid search are examined. Understanding the consumer response to advertisements is helpful to aim and personalize advertising vehicles to boost their efficiency. One study shows that repeated banner exposures stimulate the click-through rate of less click-prone users. Another one states that banner advertising expedites the timing of purchases and influences positively the repeat purchasing. Lastly, researches into how people shop online are investigated and models to predict online purchase conversion are compared. Stochastic models and forms of binary choice models are developed not only to forecast, but also to understand the factors on purchase conversion (Bucklin and Sismerio, 2009).

Johnson *et al.* (2002) develops a probabilistic model that defines the search behavior with respect to depth, dynamic and activity of search. Clearly, depth of search is related with the decision to visit more than one store at a given time. Also, whereas dynamic of

search is about the number of sites visited, the activity of search is the total amount of category-level shopping activity. According to the results of proposed model, it is found that the more active shoppers are inclined to visit more sites. Also, experience brings about a small decline in the number of visited sites. The experience and browsing activities changed by product category and activity level do not affect each other positively.

Moe and Fader (2004) build a stochastic model of customer behavior that predicts each shopper's purchasing probability depending on the click stream data of visits and purchases. They predict purchase as a function of previous visits to website. The developed model considers various forms of customer diversity by dividing individual's conversion behavior into two parts: one for visit effects and other for purchasing threshold effects. Visit effects examine the role of store visits in the buying stage, whereas purchasing threshold effects deal with the information about customers' bias to online buying. With respect to the proposed conversion model, three main dynamics are evaluated as a result, which are: the influence of visits, visiting and purchasing threshold evolving effects over time. It is stated that if the number of visits on a website increases, the probability of purchasing also increases. Also, it is found that, subsequent visits have a decreasing effect on buying behavior and purchasing threshold increase based on previous purchasing experiences, since the novelty of buying online decreases. Different versions of the model by altering these two components are tested and compared with the alternative statistical methods such as logistic regression. As a result, it is found that the dynamic conversion model performs better than other benchmark models.

Fan *et al.* (2012) propose a model on the conversion rate determinants by examining factors about vendor like seller's marketing strategies, reputation scores, product quality ratings, and service responsiveness. They analyze the state dependence among conversion rate and seller's covariates with the Hidden Markov model. In the study, two states (low and high) are considered and high state is indicated a more appropriate baseline conversion rate. According to the results, it is found that seller-level covariates have a significant effect on purchasing. The findings show that the relation between the conversion rate and the vendor associated variables varies among the states. For example, enlarging the product variety at low state has an adverse effect on the conversion rate because of not being able to persuade visitors to purchase a product. On the other hand, when a vendor in the high

state, enlarging product variety increase the purchasing likelihood since the consumer finds what they desire and high reliability of seller convince them to buy.

According to article of Olbrich and Holsing (2011), social shopping features are examined by analyzing clickstream data to predict purchasing behavior by developing logistic regression model. Since social shopping communities have gained popularity with high interest recently, the research paper aims to find the significant factors for predicting conversion rate within social shopping communities. Social shopping is an emerging type of e-commerce, which is created by a linkage between online shopping and social networking. Recommendation lists, ratings, assortments arranged by users and tags are user self-created social shopping attributes. According to the findings of the article, social shopping attributes like recommendation lists and ratings are highly influential factors. Tags and high ratings affect the participation to online shopping positively (i.e. click out) due to reducing the perceived risk, whereas the lists and styles influence the user negatively. Still, lists and assortments arranged by users enlarge site stickiness and browsing. Besides, it is stated that members are more prone to make a click-out than ordinary users. In addition, it is observed that the longer the view time and the longer average view time per product result in higher probability of a click-out. Also, the frequent use of home page indicates the lower probability of click-out. Lastly, the price of products correlates positively with click-out; so it is a principal key of decision making process.

Another article of Moe (2006) empirically builds a two stage model that describes unstable decision rules and interdependences between choices within each stage. The model evaluates the click-stream data by analyzing the customer choices for two stages: products viewed and products purchased. When two stages are compared, simpler and less cognitively effortful decisions are made in earlier stage. In the study, customer heterogeneity is also examined in preferences and in decision rules. The observed choices are assessed with the proposed model according to attribute preference ratings and criterion characteristics. More clearly, criterion attributes are factors affecting the costumer's evaluation of a product when choosing a product. The effect of various product attributes in the viewing stage to the purchasing stage is investigated. The findings indicate that fewer product features are observed in product viewing phase than those observed in purchasing phase. Also, consumers are more inclined to make simple decisions with regard to a subset of attribute information in earlier stages. Furthermore, it is found that attributes

like price and size are prone to be considered in purchase stage, whilst ingredient attributes tend to be considered in both stages. As a result, it is found that the predictive performance of two stage model is than that of a single stage model.

Sato and Asahi (2012) propose a binomial logit model for users' two decision phases: viewing and buying. In the research, the likelihood of buying is divided into conditional likelihood of buying given a visit and likelihood of visiting. According to the findings, the total of money spent in a month has a negative influence on purchasing; whereas the total number of product pages viewed in a week affects visiting positively.

Chen and Su (2013) develop an interest oriented model that is applicable not only for within category but also for across category. In the study, three main factors reflecting users' interest and behavior are proposed, which are visiting path of category, the frequency of browsing and relative time spent on a web page. The clustering algorithm is used to gather the user with similar interest. As a result of experiments, it is observed that users have various interests with respect to commodity categories. According to the article, it is suggested that across-category recommendations can be offered such as mixed promotions in order to improve cross-selling ability and to transform online shopping to buying.

Kurmiawan (2000) tries to model online customers' preference and stickiness with a structural modeling method from the standpoint of five aspects, which are site's appeal, website environment, the convenience the site, customer satisfaction and entertainment. In the study, the customer preference and retention are investigated separately as a dependent variable. The results show that only customer satisfaction and entertainment affect directly customer preference and stickiness; whereas others' impact on customer retention and preference is mediated through customer satisfaction and entertainment. Also, customer preference is predicted more powerfully with customer satisfaction than entertainment, whereas the situation is opposite for customer retention. That is, ensuring customer satisfaction is essential to make the customers choose a specific website; however, it does not guarantee to make the customers stick to that site.

Iwanaga *et al.* (2016) explore the effect of customers' page views on the likelihood of the product choices. They build a nonparametric optimization model for the estimation of product choice likelihoods based on the property of being recent and frequent of

customers' past purchases. Many other studies emphasis on the prediction of conversion rates, whereas the study investigates purchase probability to each product. Since the predicted product-choice probabilities describe the customer desires for the products, the research helps collaborative filtering technique. When the predictive performance of the model is compared with the generally known methods of binary classification such as logistic regression model and kernel-based support vector machine, the nonparametric approach outperforms other methods.

In the article of Lee *et al.* (2007), a decision tree model is built to predict e-commerce success by defining an independent variable as customers' preference to offline medium. Decision tree model is preferred since relationship between variables can be assessable. Data is collected from a survey conducted in the website. Response variables obtained from studies on both online and offline transactions are selected. The findings point out that buying frequency and price are main decision criteria in the beginning, while relative importance of sales representatives, the degree of urgency in processing service about to purchase and contact time spent by a customer per a transaction are factors at second level. When the decision tree model is compared with logistic regression and discrimination analysis, the prediction accuracy of proposed model is stated as superior to that of others.

Lariviere and Van den Poel (2005) try to explore three essential metrics of customer response: "next buy", "partial-defection" and "profitability evolution". The customer profitability evolution in the way of the profitability is represented by a binary (profit drop) and a linear (profit evolution) response variable. The selection of a new product, the choice of canceling an uncompleted status product (partial-defection) and profit drop are binary measures involving a classification problem; whereas the profit evolution representing the change in the customers' profitability is analyzed with regression forest. When the random and regression forests are benchmarked with linear and logistic regression methods, significant improvements are found in terms of prediction accuracy. The results show that past customer behavior has a significant role in repeated purchase activities and favorable profitability evolutions; whereas salespersons' role affect customers' defection tendency.

Van Wezel and Potharst (2007) build a choice model by analyzing sales data for two product categories: ketchup and peanut butter. In the study, demographic variables (household size, household income) and situational variables (day of the week, loyalty for

brands, price, and ad of brands) are used as input variable; whereas brand name is defined as target variable. Ensemble methods (Bagging, boosting and multi-boosting) outperform the model when compared with the individual decision trees and logistic regression.

Table 3.1. Summary of literature framework (Chen and Su, 2013; Fan *et al.*, 2012; Iwanaga *et al.*, 2016; Johnson *et al.*, 2002; Kurmiawan, 2000; Lariviere and Van den Poel, 2005; Lee *et al.*, 2007; Moe, 2006; Moe and Fader, 2004; Olbrich and Holsing, 2011; Sato and Asahi, 2012; Van Wezel and Potharst, 2007).

Author, Year	Model Type	Similarities	Differences
Johnson et al., 2002	Search behavior probabilistic model	Clickstream data analyse The results show that more active online shoppers tend to search across more sites.	They model an individual's tendency to search as a logarithmic process. Only browsing probabilities are examined. Purchasing decisions of users are not considered seperately since their clickstream data does not identify purchases. Time periods of searching and product classes (books, compact disks, air travel services) are taken into consideration. It is found that browsing attitude is influenced by product category and level of activity but does not increase with experience.
Moe and Fader, 2004	Stochastic probabilistic model of dynamic conversion	Clickstream data analyse Customers' purchasing probability model based on past visits and past purchases is built. The predictors of their research are as follows: 1.the number of past visits 2.the number of past purchases 3. the number of visits since the last purchase 4. time elapsed since the last visit 5. time elapsed since the last purchase The findings show that if the number of visits on a website increases, the probability of purchasing also increases.	They build a dynamic conversion model which includes visit effects and purchasing threshold effects. Visit effects examine the role of store visits in the buying stage, whereas purchasing threshold effects deal with the information about customers' bias to online buying. The number of units purchased and the total amont spent are not analyzed in this study.
Fan et al., 2012	Hidden markov model of buying behavior	They propose a model of customers' purchasing probability.	They investigate the effects of seller-covariates on the conversion rate. They analyze the state dependence among conversion rate and seller's covariates. The results indicate that the relation between the conversion rate and the vendor associated variables varies among the states.
Olbrich and Holsing, 2011	Customer purchasing behavior model	Clickstream data analyse Binary coded dependent variable Logistic regression model	It is tried to find the factors affecting consumer purchasing behavior within social shopping communities. The longer the view time and the longer average view time per product detail site result in greater likelihood of a click-out. The more product detail sites viewed, the lower likelihood of click-out. Tags and high ratings have a positive effect on the participating online shop (i.e. click-out) due to reducing the perceived risk.
Moe, 2006	Two stage choice model	Online retailer clickstream data analyse	In this research, a model that defines varying decision rules between product viewing and purchasing stages is built.

Table 3.1. Summary of literature framework (Chen and Su, 2013; Fan *et al.*, 2012; Iwanaga *et al.*, 2016; Johnson *et al.*, 2002; Kurmiawan, 2000; Lariviere and Van den Poel, 2005; Lee *et al.*, 2007; Moe, 2006; Moe and Fader, 2004; Olbrich and Holsing, 2011; Sato and Asahi, 2012; Van Wezel and Potharst, 2007). (cont.)

Author, Year	Model Type	Similarities	Differences
Sato and Asahi, 2012	Binomial logit model of visiting and purchasing probabilities	Clickstream data analyse Both purchasing and visiting probabilities are estimated. Total amount spent by each customer is used as a factor in this research. The model performances are measure with AUC method.	In the model, they defined a heterogenity parameter to capture each customer's different behavior. The number of product pages is also taken into account as an independent variable.
Chen and Su, 2013	Interest oriented model	Clickstream data analyses	Category visiting path, browsing frequency and access time are considered as key attributes in the model. Clustering algorithm is used to segment users' similar interest. Only browsing of category and item page is considered. Operations like searching, putting into cart not taken into account.
Kurmiawan, 2000	Structrual equation model of customer preference and stickiness		Customer preference and stickiness are investigated in terms of five aspects: 1. site's appeal, 2. the community atmosphere the site created, 3. the convenience the site offered, 4. customer satisfaction and 5. entertainment.
Iwagana et al., 2016	Product-choice probabilities model	Clickstream data analyses	The recency and frequency of page views are defined as independent variables. * Recency : The time of last visit of customer to a product * Frequency : The number of visits of customer to a product In this study, they model purchase probability to each product.
Lee et al., 2007	Prediction model of choosing online medium for shopping	Decision tree method and logistic regression is used.	They try to identify identify service characteristics encouraging customers to make a purchase in an online channel.
Lariviere and Van den Poel, 2005	Prediction model of customer retention and profitability	Random forest and regression tree methods are used. The performance of the models are evaluated with AUC method. The number of products owned is used as a past behavior predictor in this study.	Demographic features of customers are used to segment them according to their similarities.
Van Wezel and Potharst, 2007	Choice modeling	Ensemble methods (Bagging, boosting and multi-boosting), decision tree and logistic regression methods are used. Twofold cross-validation method is applied to have an estimate of the prediction error of the models.	They include demographic variables (household size, household income) and situational variables (day of the week, loyalty for brands, price, and ad of brands) are to their model and they use the brand name as target variable.

4. METHODOLOGY

4.1. Classification Methods

In linear regression model, quantitative dependent variable is predicted; however, in many cases the dependent variable is qualitative instead. Linear regression model is not applicable when the response variable is qualitative. Normally, it is not possible to transform a qualitative dependent variable with more than two levels into a variable having equally reasonable coding that is appropriate for linear regression method. For a categorical dependent variable with 0/1 coding, regression analysis by minimizing the sum of squares of the residuals is reasonable; however, some of estimates might be outside [0,1] interval.

The process to predict the qualitative (categorical) response variable is referred to as classification. Classification approaches include dividing the observation into categories or classes; nevertheless, these methods predict first the probability of any categories of a qualitative variable. There are a lot of classification methods; one of widely used classifiers is logistic regression. Moreover, there are more computer-intensive classifiers such as trees, random forest and boosting methods (James *et al.*, 2013).

4.1.1. Logistic Regression

Generalized linear models (GLMs) are extended type of linear models which allows for the distributions other than normal such as Poisson, Binomial and Gamma distributions. The generalized linear model used for Binomial data is called Logistic Regression (Hormann and Guler, 2015).

The limitation of linear regression is that it is not appropriate for categorical dependent variable. Many variables in the real data are qualitative: for instance, consumers

decide to purchase or not, a product has a good quality or poor, an employee may be promoted or not. Therefore, logistic regression method is desirable for data with categorical dependent variable.

Logistic regression constitutes a best fitting equation using the maximum likelihood estimation method (MLE), which maximizes the likelihood of classifying the observed data into the appropriate class (Logistic Regression, n.d.).

For logistic regression method, the following function (4.1) is used,

$$p(X) = \frac{e^{\beta 0 + \beta 1 X}}{1 + e^{\beta 0 + \beta 1 X}} \tag{4.1}$$

When a linear line is fit to a binary response coded as 0 or 1, p(X) <0 for certain X values and p(X) >1 for other values can be predicted as a principle. On the other hand, the logistic function produce S-shaped curve and it enables to obtain a sensible prediction, irrespective of the value of X. For low values of X, the likelihood of Y is predicted close to zero, yet never below zero. Similarly, for high values of X, the probability of Y is predicted close to, but never above, one. The coefficients β_0 and β_1 are estimated through MLE method such that the estimated $\widehat{\beta 0}$ and $\widehat{\beta 1}$ are chosen to maximize the likelihood function. The standard deviations of the coefficient β_0 and β_1 estimates are a measure of their accuracy.

The following equation (4.2) is found after a bit of manipulation,

$$logit(p) = \frac{p(X)}{1 - p(X)} = e^{\beta 0 + \beta 1 X} \tag{4.2}$$

The proportions of probabilities are described as the odds. It expresses the ratio of the probability of a success to the probability of failure; its range is between 0 and ∞. The odd values close to zero and ∞ point out very low and very high likelihoods of response, in order.

After taking the logarithm of left and right side of (4.2),

$$logit(p) = log\left(\frac{p(X)}{1-p(X)}\right) = \beta 0 + \beta 1X \tag{4.3}$$

The left hand side of equation (4.3) is called the "log-odds" or "logit". The logit function is the inverse CDF of the logistic distribution. In the logistic regression equation (4.2) has a linear relationship with independent variable (X).

In logit equation (4.3), every one unit increase in X corresponds to the change of log odds by β1. Equally the odds are multiplied by $e^{\beta 1}$. Since the relation of p(X) and X in (4.1) is non-linear, β1 is not equal to the change in the probability of X when X increases by one unit. The ratio of change in the p(X) because of increasing X by one unit is based on the current value of X.

With respect to logistic regression output, the estimated intercept is not main interest; its aim is to fit the average predicted values to the ratio of ones in the data. The intercept is the value of response variable when all independent variables set to 0. Standard errors of the estimated coefficient measure the accuracy of estimated coefficient. If p value is small for factors, the null hypothesis H_0: $\beta_1 = 0$ is rejected. That is, it is concluded that there is an association between independent variable and the probability of response variable.

If a binary response is predicted using multiple predictors where $X = (X_1,\ldots,X_p)$, the multiple logistic regression can be generalized (4.4) and (4.5) as follows (James *et al.*, 2013):

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta 0 + \beta 1X1 + \cdots + \beta pXp \tag{4.4}$$

$$p(X) = \frac{e^{\beta 0 + \beta 1 X 1 + \cdots + \beta p X p}}{1 + e^{\beta 0 + \beta 1 X + \cdots + \beta p X p}} \tag{4.5}$$

As a general comment on modeling, in a building a regression "the correct model" is not the thing searched for. Instead, a simple model is looked for that is in compliance with the data and serves the aim of interest. If the total numbers of variables is not too high, a stepwise regression procedure can be applied directly. That is, a regression model is built either by starting from one variable and trying to add more variables (known as forward stepwise regression), or by starting with a model having many variables and then removing the useless ones (called backward stepwise regression). The backward stepwise regression starts with all factors of in the model, then the method tries to answer what would have if the corresponding variable dropped. This process repeats until no further variables can be eliminated without a statistically significant loss of fit.

In order to interpret the GLM model as a good fit, the deviance is evaluated, which is called "residual deviance" in R program. Deviance is used as a "quality of fit" statistic for generalized linear model. For GLMs having a high residual deviance indicates the significance of deviance and the model does not fit well (Hormann and Guler, 2015).

4.1.1.1. Akaike's Information Criterion: AIC. AIC is a metric for finding the best regression model, which is defined as in the following equations (4.6) and (4.7):

$$AIC = -2 \log\left(L\left(\hat{\beta}\right)\right) + 2k \tag{4.6}$$

$$AIC = Deviance + 2k \tag{4.7}$$

where $\hat{\beta}$ denotes the MLE, L(.) is the likelihood function and k is the number of parameters. The formula tries to find the best simple model such that models are penalized with the lots of parameters. The best model is the model having the smallest AIC value (Hormann and Guler, 2015; Toronto University, n.d.).

Unlike logistic regression model, decision tree models are easy and advantageous for interpretation because of its graphical representation (James *et al.,* 2013).

4.1.2. Classification Tree Model

Tree based approaches include dividing the predictor space into a number of subsets. Due to the series of splitting rules applied to stratify the predictor space, these type of approaches are named as decision tree methods (James *et al.,* 2013).

Decision tree method is the form of top-down tree structure. Each data set flows through the tree where decisions are made at each node till the record hits the terminal node. In the tree structure, a series of tests are applied and final result is determined according to the value of independent variable which can be either discrete or continuous. If the response has a discrete value, a regression tree is built, yet if it has a continuous value, a classification tree is built (Lee *et al.,* 2007).

A classification tree is applied to estimate a qualitative outcome. The method builds a tree by splitting on the values of independent variables (Bertsimas, n.d.). For classification method, we predict the probability of the most frequently occurring class since each outcome belongs to the majority class of training outcomes in the subset which it falls into. Both the class prediction related to a certain terminal node region and the class proportions between the training outcomes in the region are of interest to interpret the results of a classification tree.

In decision tree methods, the predictor space is divided into high dimensional rectangles in order to interpret the results easily and simply. The goal is to find rectangles R_1,..., R_j that minimizes classification error rate. Classification error rate states the proportion of misclassified observations.

Recursive partitioning approach is the step-by-step process to construct a decision tree. These approach starts at the top of the tree (where all outcomes correspond to a single region) and then segments the predictor space in a successive manner. Each segment is represented through two new branches and it furthers down on the tree. Also, the best

segment is selected at that certain step instead of looking a better split in some future step (James *et al.*, 2013). In short, the splits in the tree are followed and the most frequent outcome in the training set that followed the same path is used to predict the outcome for an observation (Bertsimas, n.d.).

Tree based models try to split the data into subsets so that each subset becomes homogenous and pure as far as possible, seen in Figure 4.1. The standard prediction is made in accordance with the majority in each subset (Bertsimas, n.d.).

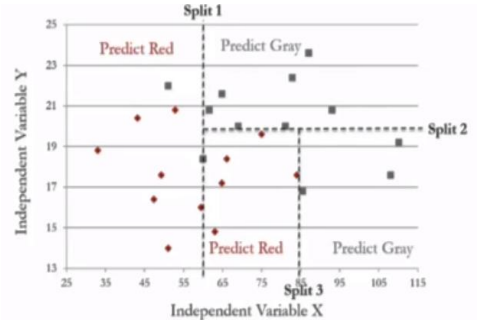

Figure 4.1. Classification Approach Splitting Rule (Bertsimas, n.d.).

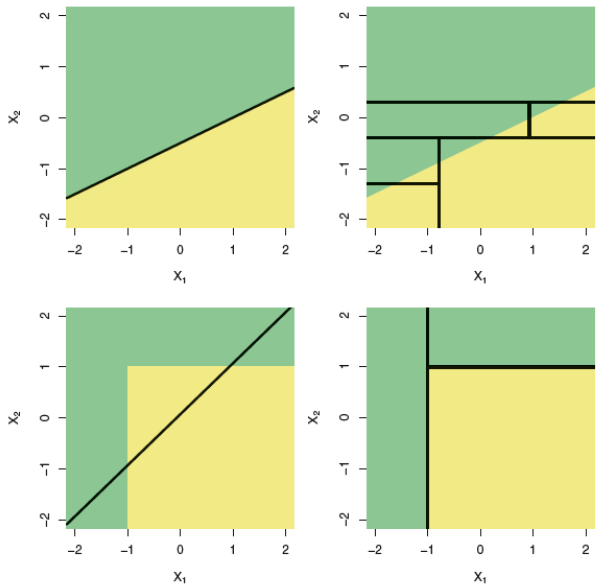

Figure 4.2. Graphics of Trees versus Linear Models (James *et al.*, 2013).

In Figure 4.2, the upper part of graph represents an example in which the decision boundary is linear and shown by shaded areas. Left graphic in the top row, linear methods

will outperform a decision tree which splits several regions (right). The decision boundary is nonlinear in bottom row. Right graphic in the bottom row, a linear model is not able to express the true decision boundary, while a decision tree successes.

There are several advantages of decision tree methods over the classical approaches:

- Trees are easily explainable.
- Trees can be shown as a graph, and are interpreted with ease.
- Trees can simply overcome qualitative independent variables without the need of dummy variables.

However, there are some disadvantages of decision tree models, as well.

- Trees can be very non-robust, means that the final predicted tree can be affected largely by a small change in the data.
- Trees can differ in the level of predictive accuracy when it is compared with some of other regression and classification methods.

The predictive accuracy of decision trees can be substantially improved by producing multiple decision trees applying methods like bagging, random forests and boosting. These decision tree based methods lead to a single consensus prediction. By combining many decision trees bring about substantial improvements in predictive performance, at the cost of some loss in interpretation (James *et al.*, 2013).

4.1.3. Random Forests

When there is one strong explanatory variable in data, together with other moderate ones, whole trees use this strong factor in the top split and trees will be quite similar one another. Thereby the predictions will be highly correlated and will not cause a considerable reduction in variance over a single tree.

Random forest algorithm overcomes the problem by forcing each segment to regard only a subset of the explanatory variables. The rationale behind this logic is related to decorrelation of the decision trees, hence making the average of the final trees lead to reduce the variance and increase the reliability (James *et al.,* 2013).

In order to optimize the decision tree method, an ensemble of trees is created with a vote for most common class, known as random forest method. This method applies the random sampling of m predictors to build each tree according to bootstrap sampling. The number of predictors, m is used to divide the nodes and is smaller than the available number of variables (Lariviere and Van den Poel, 2005).

More specifically, when building random forest tree, a random selection of a subset of m predictors is selected as split candidate from the whole set of t predictors; the split is allowed to consider one of those m predictors. Therefore, averagely (t-m)/t of the splits will not even use the strong predictor, thereby other predictors chance will be more (James *et al.,* 2013).

When a random forest is used for classification, it gets a class vote from each tree, and then stratifies each split by using majority vote. For classification, the default value for the number of predictors at each split (m) is the square root of the total number of predictors ($\sqrt{t}$) and the minimum size of terminal nodes is one (Hastie *et al.,* 2008).

The number of trees B is not critical parameter for random forest method since the presence of more trees will not bring about over-fitting. Therefore, in practice the number of trees can be selected sufficiently large for the error rate to have settled down (James *et al.,* 2013).

4.1.4. Boosting

Boosting is another approach for improving the predictions from a decision tree. This approach involves sequentially grown trees by creating multiple copies of training data set, building a separate decision tree for every copy and then aggregating whole trees to build a single predictive model. The sequential growth means that each tree is constructed by

using information from the trees that have already been grown. The boosting approach learns slowly, instead of building a single large decision tree to the data, which potentially turns into over-fitting.

There are three tuning parameters of boosting method:

- The number of trees (B): On the contrary to random forest, boosting can overfit if the value of number of trees is selected too large. This over-fitting problem arises slowly. In order to choose the number of trees, cross validation method is used.
- The shrinkage parameter (λ): This parameter is a small positive number which adjusts the learning rate of boosting. The choice of this parameter depends on the problem. Typical values are 0.01 or 0.001. If it is very small, it requires a very large number of trees to obtain good performance.
- The depth of each tree (d): This parameter defines the number of splits in each tree and adjusts the complexity of boosting. In general, d is interaction depth, which controls the interaction order of boosted ensemble, as d splits can include at most d variables.

In boosting method, since the particular tree considers the previously grown other trees, smaller trees are generally sufficient. Using small trees facilitate interpretability, also (James *et al.,* 2013).

4.1.4.1. Out of Bag Error (OOB). Out of bag error (OOB) is a method of estimating the prediction test error of a bagged model in which bootstrap aggregating is utilized to subset the training data (James et al., 2013). Each tree is grown from a different bootstrapped subset of the original data for ensemble methods. The samples left out of a bootstrap sample and not used in building of a single tree are known as out-of-bag samples (Baydogan and Runger, 2014). Around two thirds of observations are used to fit a bagged tree, the rest one-third of the observations are out-of-bag samples.

The response can be estimated for the ith observation by using only the trees that does not have ith sample in their bootstrapped sample. This yields one-third of predictions for ith observation. The predicted outcomes can be computed with the majority vote on

that set for a classification problem in order to have a single prediction. OOB error, which is a classification error, is a prediction of the test error as the response for each observation is estimated utilizing only the trees that are OOB (James *et al.*, 2013).

4.2. The Area under the Receiver Operating Curve (AUC)

A receiver operating characteristics (ROC) graph is used to visualize and choose classifiers depending on their predictive accuracy. This graphical method is advantageous for skewed class distribution and unequal classification error costs.

In a classification problem, each event is linked to one elements of set (p, n) of positive and negative class. A two by two confusion matrix (contingency table) shown in Table 4.1 is created by a classifier and a set of instances. There are four probable outcomes for given classifier and event as follows (Fawcett, 2006);

Table 4.1. Contingency table (Fawcett, 2006).

	The True Class		
The Predicted Class		**P**	**n**
	T	**T**rue **P**ositives	**F**alse **P**ositives
	F	**F**alse **N**egatives	**T**rue **N**egatives
		P	**N**

$$tp\ rate = \frac{TP\ (Positives\ correctly\ classified)}{P\ (Total\ positives)} \tag{4.8}$$

$$fp\ rate = \frac{FP\ (Negatives\ incorrectly\ classified)}{N\ (Total\ negatives)} \tag{4.9}$$

$$precision = \frac{TP}{TP + FP} \tag{4.10}$$

$$recall = \frac{TP}{P} \tag{4.11}$$

$$accuracy = \frac{TP + TN}{P + N} \tag{4.12}$$

The AUC metric is a comparison of the predicted value of the observation and the actual value of that observation. In more detail, the number of the number of correctly classified events to the total number of events (sensitivity) and the number of correctly classified non-events to the total number of non-events (specificity) are considered in a confusion matrix and expressed in a bidimensional graph yielding ROC curve (Lariviere and Van den Poel, 2005).

In ROC graphic, true positive rate is drawn on the Y axis, while false positive rate is drawn on the X axis. It shows the relative relationships between benefits (TP) and costs (FP) (Fawcett, 2006). A point (FP, TP) on the ROC curve represents each classifier with a given class distribution and cost matrix. This curve evaluates the performance of classifiers across the entire range of class distributions and error costs by comparing. According to the statistical meaning of AUC, the likelihood of randomly chosen negative event will have a smaller predicted likelihood of belonging to the positive class than a randomly chosen positive event.

In general, predictive ability of classification methods is evaluated with their accuracy or error rate on testing data. So long as the class with the largest likelihood prediction is same with the actual outcome, it is counted as correct. The accuracy measure does not take into account the likelihood of estimation. However, in many data mining applications, the overall classification error rate is not an appropriate metric if the criterion is for ordering or ranking. For instance, in marketing, it is necessary to know the top X% of customers is interested in our offer in order to deploy different promotion strategies to customers with the different probability of buying. Therefore, a ranking of customers having different purchasing probabilities is more critical than a mere classification of buyers and non-buyers. Therefore, it is stated that the ROC curve is better method than accuracy to assess classifiers that also produce rankings by representing the tradeoff between hit rates and false alarm rates. Moreover, for balanced or imbalanced, binary or

multiclass datasets, AUC is found as statistically consistent with accuracy and statistically more discriminant than accuracy (Huang and Ling, 2005).

The AUC measure is used to interpret the predictive performance of models (Lariviere and Van den Poel, 2005). In order to find better classifiers, the larger AUC is selected (James *et al.*, 2013).

5. DATA ANALYSES

In this section, we tried to identify and summarize basic features of the data with its descriptive statistics. Therefore, we analyzed the data to understand which factors might be essential and influential for purchasing decision. By sampling, we inferred whether there is any unusual behavior or not. Then, we determined the drivers affecting customers' purchasing decision available in the clickstream data set in order to use them for modeling.

The clickstream data is collected from an online retailer Privalia, operating across Spain, Italy, Germany, Brazil and Mexico. The electronic commerce website Privalia offer a service for multifarious apparel products such as shirts, dresses, and shoes for every customer segment.

In the dataset, Internet behavior of 1000 customers who made at least one visit was recorded from January 1, 2013, through December 31, 2015. Initially, data covers collectively a total of 291511 visits and 3218 purchases as a result of merged raw data. Each transaction in the data corresponding to view, add to cart and purchase comprises information on date & time, member ID, channel and etc. View data includes which member visited the website for which product and campaign at which time, whereas add to cart data involves whether the member added the product to the cart with an order number or removed it from his/her cart. Also, order data contains which member purchased which products with their amounts and prices.

Before obtaining last form of merged data with the inclusion of all three data explained above, we realized that the set of customers in each dataset are slightly different one from the other. Therefore, some customers, for example, those who are in add to cart data, but not in view data, were excluded properly from relevant data set. Likewise, some order numbers which are found in add to cart data, but not in order data, were also removed. Furthermore, ten customers who are only view homepage or product over approximately average of 1700 times but never add any product to their cart and end their shopping with a purchase are identified as outlier and they are eliminated from view data before merging. Since the data involves so many view actions compared to purchase

actions, this unproportioned distribution leads to class imbalance problem for modelling. Due to this reason, top ten customers are excluded to slightly balance the data distribution in terms of purchase and view actions. After data cleaning, further applied steps were advanced through the data available which is the combination of view, add to cart and order data. The final form of data consists of overall 271749 visits and 2636 purchases. Three-year Internet behavior of 976 customers was analyzed.

In Table 5.1, some transactions of a member can be seen as an example taken from the merged data. From this figure, when the member comes to visit online store, whether with an aim of buying or not, which product is seen or bought, how many items are placed to the shopping cart, how much is paid for an order can be seen.

Table 5.1. Sample data set of a member.

Date_Time	Member_PK	Channel	Action	Product_PK	Campaign_PK	Stockp_PK	Order_PK	Ordline_PK	Carttr_PK	Carttr_quantity	Order_Type	Qty_ordered	Order_FK_mcorder_PK	Ordline_price
17.03.2014 14:46	394124	web	removed from cart	NA	NA	41635690	78026830	NA	161932210	1	NA	0	NA	0
17.03.2014 14:46	394124	web	add to cart	NA	NA	41635990	78026830	214959650	161932230	1	NA	0	NA	0
17.03.2014 14:47	394124	web	add to cart	NA	NA	41635120	78026830	214959630	161932240	1	NA	0	NA	0
17.03.2014 14:49	394124	web	add to cart	NA	NA	41636080	78026830	214959660	161932250	1	NA	0	NA	0
17.03.2014 14:51	394124	web	add to cart	NA	NA	41631400	78026830	214959610	161932260	1	NA	0	NA	0
17.03.2014 14:53	394124	web	add to cart	NA	NA	41636330	78026830	214959670	161932270	1	NA	0	NA	0
17.03.2014 14:56	394124	web	add to cart	NA	NA	41661470	78026830	214959690	161932280	1	NA	0	NA	0
17.03.2014 15:00	394124	web	view	11185760	107860	NA	NA	NA	NA	0	NA	0	NA	0
17.03.2014 15:00	394124	web	view	11186380	107860	NA	NA	NA	NA	0	NA	0	NA	0
17.03.2014 15:00	394124	web	view	11187280	107860	NA	NA	NA	NA	0	NA	0	NA	0
17.03.2014 15:00	394124	web	view	11189240	107860	NA	NA	NA	NA	0	NA	0	NA	0
17.03.2014 15:07	394124	web	add to cart	NA	NA	41640190	78026830	214959680	161932290	1	NA	0	NA	0
17.03.2014 15:09	394124	web	add to cart	NA	NA	41640190	78026830	214959680	161932320	1	NA	0	NA	0
17.03.2014 15:11	394124	web	add to cart	NA	NA	41635720	78026830	214959640	161932330	1	NA	0	NA	0
17.03.2014 15:15	394124	web	add to cart	NA	NA	41631780	78026830	214959620	161932340	1	NA	0	NA	0
17.03.2014 15:15	394124	web	add to cart	NA	NA	41667590	78026830	214959700	161932350	1	NA	0	NA	0
17.03.2014 15:17	394124	web	purchase	11184860	107860	41631400	78026830	214959610	NA	0	S	1	78026830	24.8
17.03.2014 15:17	394124	web	purchase	11184920	107860	41631780	78026830	214959620	NA	0	S	1	78026830	24.8
17.03.2014 15:17	394124	web	purchase	11185460	107860	41635120	78026830	214959630	NA	0	S	1	78026830	10.7
17.03.2014 15:17	394124	web	purchase	11185640	107860	41635720	78026830	214959640	NA	0	S	1	78026830	12.4
17.03.2014 15:17	394124	web	purchase	11185730	107860	41635990	78026830	214959650	NA	0	S	1	78026830	8.3
17.03.2014 15:17	394124	web	purchase	11185760	107860	41636080	78026830	214959660	NA	0	S	1	78026830	14

Table 5.2 summarizes the main features of the data quantitatively. As shown in this table, the number of products purchased changes simultaneously with the number of total add to cart actions and in parallel with the increase in the number of unique campaigns throughout 3 years. Even though the main part of the number of products purchased in total is performed via web, mobile shopping is gaining importance recently year after year.

Table 5.2. Descriptive statistics of data.

Description	2013	2014	2015	Total	Mean	Std Dev.	Min	Max	Median
Number of visits	93416	82526	95807	271749	90583,00	7079,24	82526	95807	93416
Number of views	91745	77119	90299	259163	86387,67	8059,40	77119	91745	90299
Number of products viewed	44076	34450	43677	122203	40734,33	5446,05	34450	44076	43677
Number of products added to the cart	1028	4113	4129	9270	3090,00	1785,76	1028	4129	4113
then removed from cart	131	235	314	680	226,67	91,78	131	314	235
then purchased	512	1059	1065	2636	878,67	317,56	512	1065	1059
via web	311	603	401	1315	438,33	149,54	311	603	401
via mobile	177	393	597	1167	389,00	210,03	177	597	393
via webmobile	18	48	67	133	44,33	24,70	18	67	48
via fanshop	6	15	0	21	7,00	7,55	0	15	6
Number of visits in campaign time	92257	78178	91364	261799	87266,33	7883,38	78178	92257	91364
Number of unique campaign	2443	3219	4035	9697	3232,33	796,08	2443	4035	3219
Number of purchased products with campaign	512	1059	1065	2636	878,67	317,56	512	1065	1059
from single campaign (S)	305	720	743	1768	589,33	246,51	305	743	720
from multiple campaign (M)	207	339	322	868	289,33	71,81	207	339	322

When the data is analyzed according to monthly purchases as in Figure 5.1, the number of views and purchases are increasing in a similar way. It also shows that website performance is improving with overall three-year visits and orders at the first three-month period, whereas it is decreasing between 6^{th} and 7^{th}; 11^{th} and 12^{th} monthly periods, as well. The number of products purchased is increasing from August to November by showing a similar pattern with the number of products viewed; also conversion rate is increasing from July to September continuously. According to the Table 5.3, the highest conversion rate which is the measure of converting site visitors into paying customers is on September, even if the majority of views and purchases are made on November, can be seen in Figure 5.2 and Figure 5.3. Therefore, the aggregate measure does not account for true underlying dynamics, since the efficiency of marketing is measured by the total number of customers who have ended a transaction as an order divided by the number of visitors.

Table 5.3. The number of visits and purchases monthly.

Number of products viewed				
	Years			
Months	2013	2014	2015	3 YEAR
January	5350	4876	6406	16632
February	8381	5741	5833	19955
March	9870	7009	7628	24507
April	8312	7066	8274	23652
May	8708	6194	9981	24883
June	9383	4923	8885	23191
July	7024	2713	5397	15134
August	5573	4514	4877	14964
September	6395	8137	7985	22517
October	6750	9170	10010	25930
November	9583	11814	11335	32732
December	6416	4962	3688	15066
Total	91745	77119	90299	259163
Mean	7645,42	6426,58	7524,92	21596,92
Std Dev	1585,83	2430,99	2330,31	5438,98
Min	5350	2713	3688	14964
Max	9870	11814	11335	32732
Prob.of visiting	0,9787	0,9380	0,9427	0,9543

Number of products purchased					
	Years				
Months	2013	2014	2015	3 YEAR	Conversion Rate
January	0	32	38	70	0,41%
February	0	118	75	193	0,93%
March	0	119	102	221	0,87%
April	0	98	93	191	0,78%
May	0	81	143	224	0,86%
June	0	84	147	231	0,95%
July	26	41	49	116	0,74%
August	48	56	31	135	0,87%
September	92	153	83	328	1,37%
October	123	86	158	367	1,33%
November	124	105	142	371	1,07%
December	99	86	4	189	1,17%
Total	512	1059	1065	2636	
Mean	42,67	88,25	88,75	219,67	
Std Dev	52,14	34,27	51,40	95,00	
Min	0	32	4	70	
Max	124	153	158	371	
Prob. of purchasing	0,01018	0,0121	0,01101	0,01104	

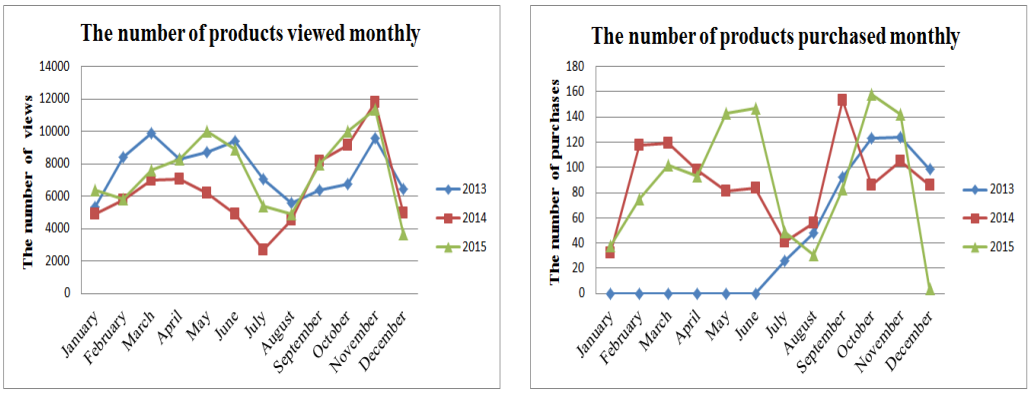

Figure 5.1. Time Series of Products Viewed and Purchased Monthly.

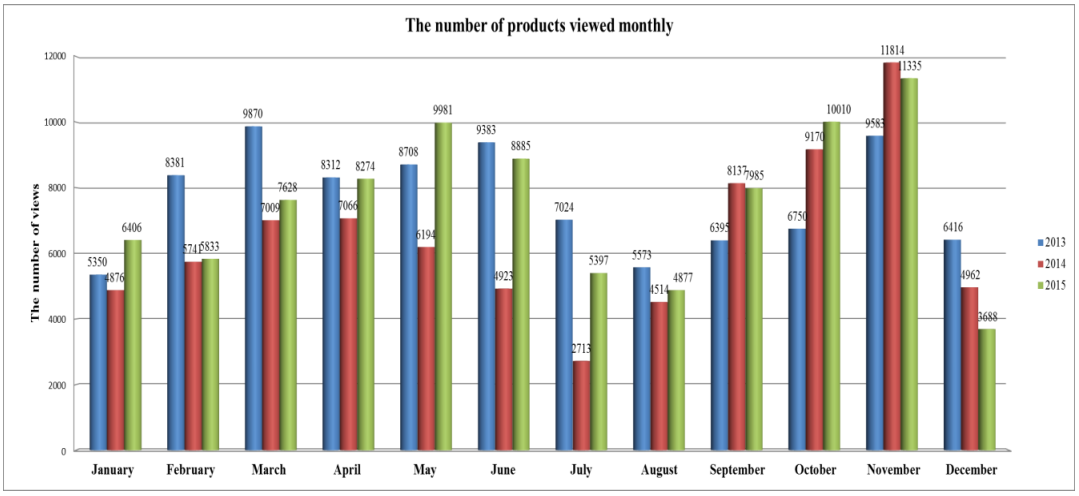

Figure 5.2. The Number of Products Viewed Monthly.

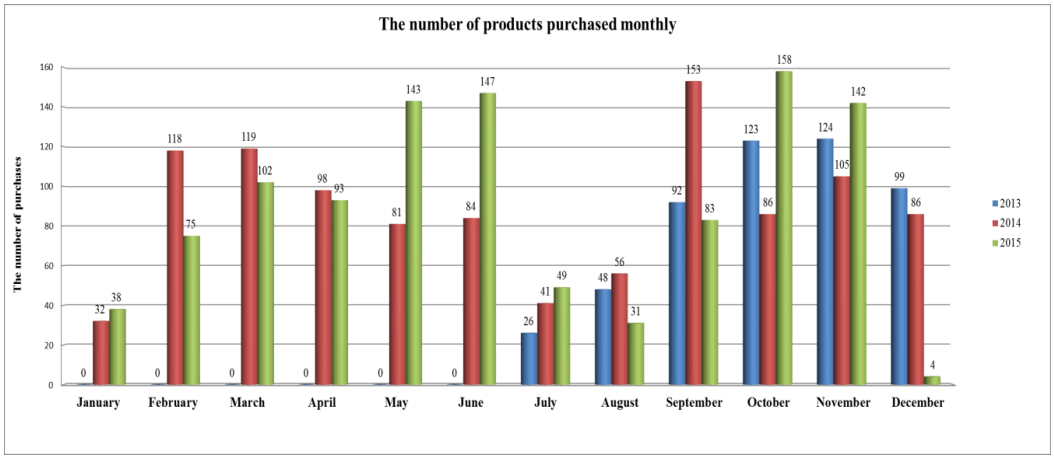

Figure 5.3. The Number of Products Purchased Monthly.

When the views and purchases are examined on a daily basis, Table 5.4 shows that people are more likely to visit a shopping website on Mondays overall 3 years-round. However, they tend to make a purchase on Wednesdays as seen in Table 5.4, Figure 5.5 and 5.7. Moreover, the conversion rate is highest on Wednesday for 3 years period. If actions are evaluated separately for each year, a majority of purchases are performed on Wednesday for first two years; whereas the visits and purchases are made mainly on Sundays in 2015, seen in Figure 5.4 and 5.5. Also, according to Figure 5.6 and 5.7, the number of visits and purchases show a decrease from Wednesdays to Fridays for each three year.

Table 5.4. The number of visits and purchases daily.

Number of products viewed

	Years			
Days	**2013**	**2014**	**2015**	**3 YEAR**
Monday	15094	11850	13914	40858
Tuesday	12968	10736	14043	37747
Wednesday	13955	12161	13470	39586
Thursday	13943	12126	11853	37922
Friday	12022	9220	11379	32621
Saturday	10936	9035	11382	31353
Sunday	12827	11991	14258	39076
Total	91745	77119	90299	259163
Mean	13106,43	11017,00	12899,86	37023,29
Std Dev	1375,64	1379,47	1304,97	3613,63
Min	10936	9035	11379	31353
Max	15094	12161	14258	40858

Number of products purchased

	Years				
Days	**2013**	**2014**	**2015**	**3 YEAR**	**Conversion Rate**
Monday	72	167	145	384	0,90%
Tuesday	92	119	154	365	0,92%
Wednesday	115	195	154	464	1,11%
Thursday	74	161	149	384	0,96%
Friday	47	154	135	336	0,98%
Saturday	45	115	126	286	0,87%
Sunday	67	148	202	417	1,02%
Total	512	1059	1065	2636	
Mean	73,14	151,29	152,14	376,57	
Std Dev	24,57	27,78	24,26	56,82	
Min	45	115	126	286	
Max	115	195	202	464	

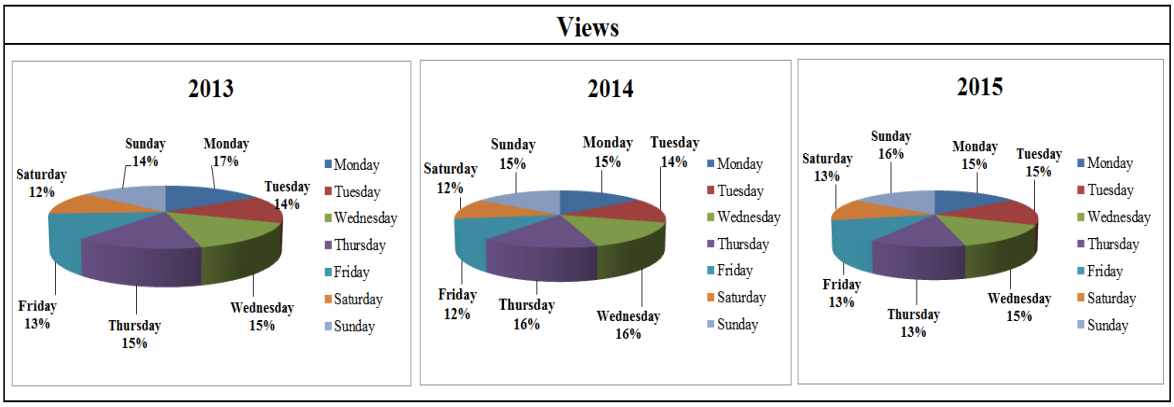

Figure 5.4. Daily Percentage Distribution of Products Viewed.

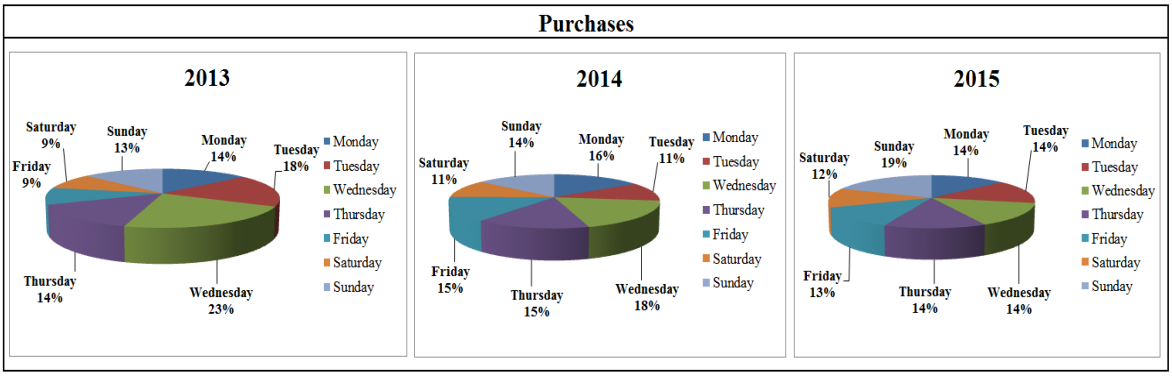

Figure 5.5. Daily Percentage Distribution of Products Purchased.

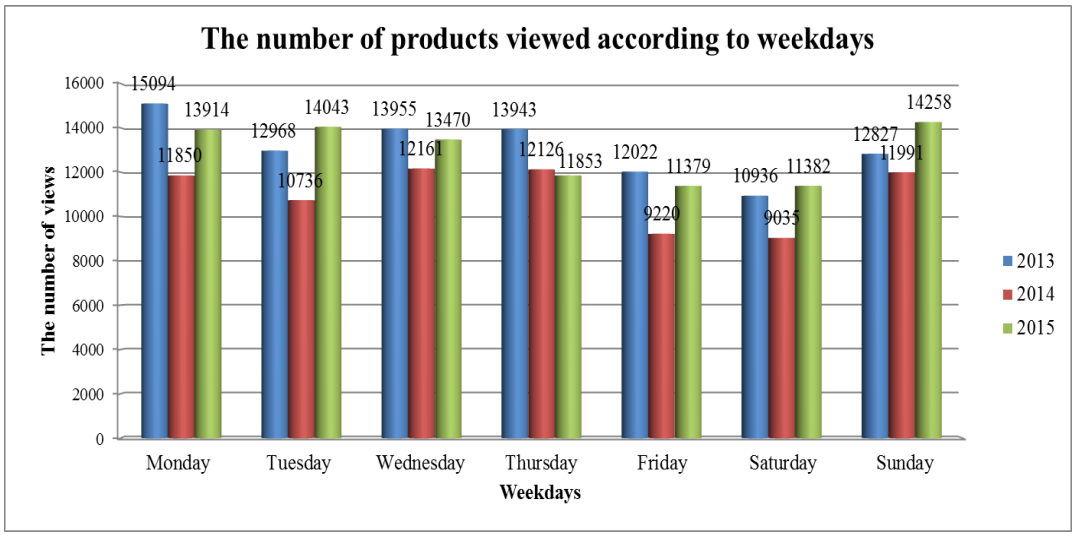

Figure 5.6. The Number of Products Viewed Daily.

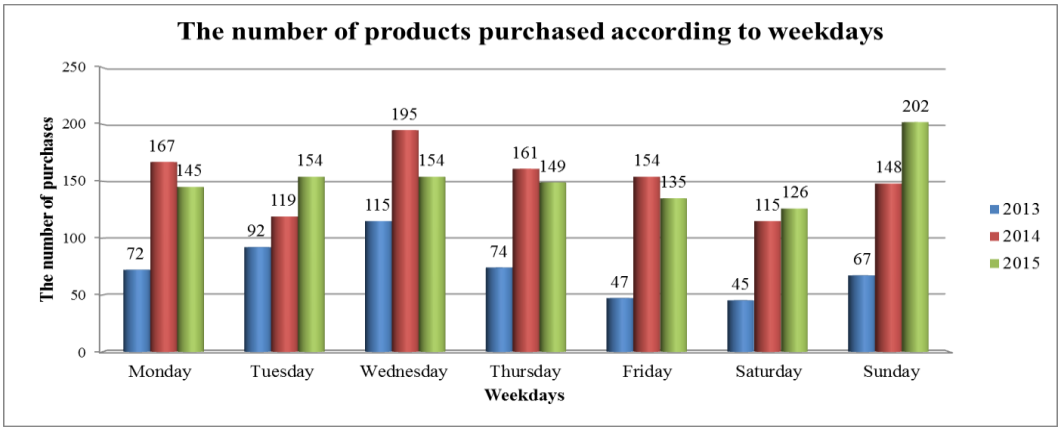

Figure 5.7. The Number of Products Purchased Daily.

Table 5.5 summarizes visiting and purchasing activities with the number of visitors and buyers at Privalia. Conversion rate is increasing and decreasing from the first year to the third year with the similar change in the number of visitors and buyers. However, the number of visits and purchases indicate a particular pattern. The number of visits shows an inversely proportional change with the conversion rate. Since these entire measures do not demonstrate the inflow of new shoppers and outflow of existing shoppers, they are not accountable for the underlying dynamics.

Table 5.5. Summary of visiting and purchasing activities.

	2013	2014	2015	3 year
Number of visitors	515	608	585	976
Number of buyers	85	150	143	226
Number of visits	93416	82529	95807	271749
Number of purchases	512	1059	1065	2636
Conversion rate (%)	0,55	1,28	1,11	0,97
Visits / Visitor	181,39	135,73	163,77	278,43
Purchases / Buyer	6,02	7,06	7,45	11,66
Purchases / Visitor	0,99	1,74	1,82	2,70

Table 5.6 accounting for the entering and existing shoppers represents conversion rate statistics for active buyers along whole data period with at least one or more visits to the store. As can be seen in Table 5.6, the number of active visitors and active buyers increase gradually for two years periods with the increasing number of visits and purchases, as well. Moreover, even if the number of new shoppers and visitors decreases and the number of outflow of existing customers increases later, the rate of conversion shows a boost due to the active buyers and visitors' activities and also the increase in the number of purchases relative to the number of visits.

Table 5.6. Summary of active visitors and buyers.

	2013-2014	2014-2015
Number of active visitors	341	368
Number of active buyers	118	121
Number of new visitors	267	217
Number of new buyers	32	22
Number of dropout existing visitors	174	239
Number of dropout existing buyers	4	10
Number of visits	175942	178333
Number of purchases	1571	2124
Conversion rate (%)	0,89	1,19

When the influence of times of day on the number of visits and purchases is investigated, it can be seen in Table 5.7 that people both visit the website and make a purchase mostly at nights according to overall visits and purchases. However, the conversion rate is highest at late in the day and the orders are performed mainly via web except fan-shop channel. The actions which are ended with an order in different times of a day are evaluated separately by channels; people mostly tend to convert their visits on the website into place an order via fan-shop and web. Moreover, the conversion rate of web channel does not differ from that of midday and late-day. According to the Figure 5.8, consumers visit more the website at nights and with their mobile phones, whereas they prefer to use web at middays in order to make a purchase for their desired item. Furthermore, the use of web channel is more common for purchasing actions at middays and late-days than other channels; while mobile phones are used extensively on purchasing decision at mornings and nights.

Table 5.7. The number of actions according to times of day.

Time of Day	Channel	Number of visits	Number of purchases	Conversion Rate %	Total visits	Total purchases	Total Conversion Rate %
Morning	Web	22420	252	1,12	66195	588	0,89
Morning	Mobile	36570	303	0,83			
Morning	webmobile	7205	33	0,46			
Midday	Web	26665	449	1,68	65826	726	1,10
Midday	Mobile	32363	238	0,74			
Midday	webmobile	6798	39	0,57			
Lateday	Web	18473	311	1,68	51150	588	1,15
Lateday	Mobile	29062	243	0,84			
Lateday	webmobile	3591	23	0,64			
Lateday	Fanshop	24	11	45,83			
Night	Web	25529	303	1,19	88578	734	0,83
Night	Mobile	57130	383	0,67			
Night	webmobile	5883	38	0,65			
Night	Fanshop	36	10	27,78			

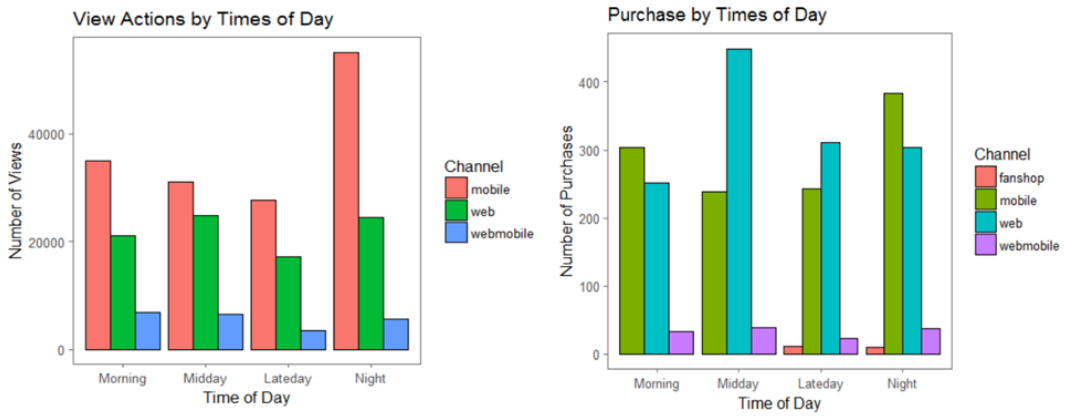

Figure 5.8. The Number of Visits and Purchases according to Times of Day.

If the seasonality effect on visits and purchases is analyzed as seen in Table 5.8, it is concluded that consumers visit online websites and buy some items mostly in the fall. Overall visits and purchases become intense in the fall and spring. Moreover, the main channel to place an order with the highest conversion rate is in the fall by using web channel except fan-shop. According to Figure 5.9, people prefer the use of mobile phones for viewing action; while they opt for web as a channel to complete their transactions throughout nearly all seasons.

Table 5.8. The number of actions according to seasons.

Season	Channel	Number of visits	Number of purchases	Conversion Rate %	Total visits	Total purchases	Total Conversion Rate %
Winter	Web	21047	255	1,21	54103	452	0,84
Winter	Mobile	28835	179	0,62			
Winter	webmobile	4207	14	0,33			
Winter	Fanshop	14	4	28,57			
Fall	Web	24938	529	2,12	86106	1066	1,24
Fall	Mobile	53473	469	0,88			
Fall	webmobile	7654	53	0,69			
Fall	Fanshop	41	15	36,59			
Spring	Web	28439	321	1,13	76018	636	0,84
Spring	Mobile	40703	285	0,70			
Spring	webmobile	6871	28	0,41			
Spring	Fanshop	5	2	40,00			
Summer	Web	18663	210	1,13	55522	482	0,87
Summer	Mobile	32114	234	0,73			
Summer	webmobile	4745	38	0,80			

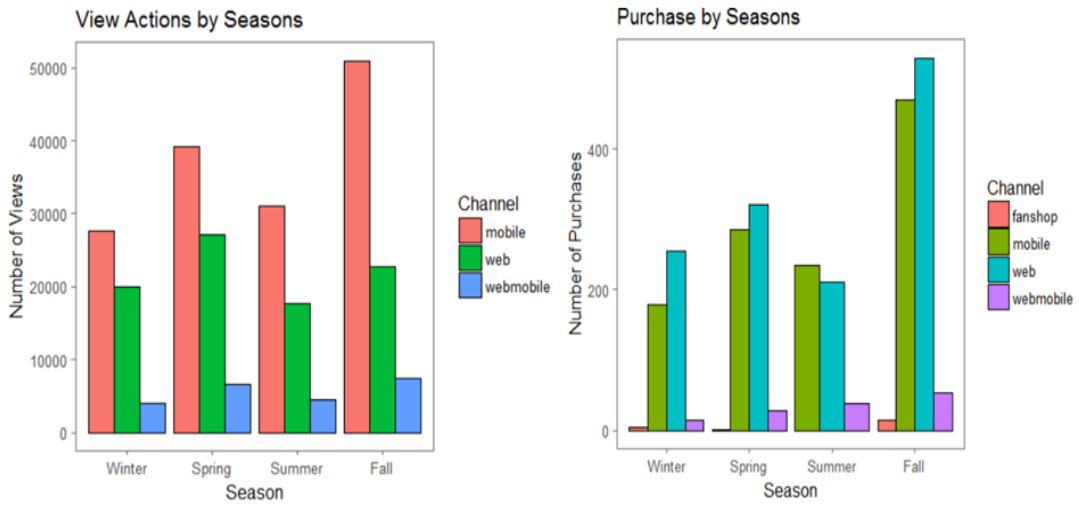

Figure 5.9. The Number of Visits and Purchases according to Seasons.

6. MODELING CUSTOMER BEHAVIOR

6.1. Model Parameters

After the analyses of data dynamics, determinants of customer decision making are defined in five main categories seen in Table 6.1:

Table 6.1. Types of model parameters.

Customer behavioral history
Channel and seasonality factors
Customer behavior frequencies
Month, channel, day and season specific sales factors
Conversion rate

Unlike factors stated in Moe and Fader (2004) article, we determined customer behavioral history factors by taking into account their averages and frequencies. In order to analyze behavioral history in a detailed way, we incorporated other key parameters of seasonality and channel utilities into our model. Moreover, we tried to understand the effect of sales history and conversion rate by analyzing it both customer and seasonal based.

To clarify, brief descriptions of factors with their above mentioned factor types are as follows in Table 6.2:

Table 6.2. Factor table with their definitions.

Type of Factors	Factors	Definition
Customer behavioral history	Lag2_View	Average of last two view actions in past
Customer behavioral history	Lag4_View	Average of last four view actions in past
Customer behavioral history	Lag6_View	Average of last six view actions in past
Customer behavioral history	Lag2_Purchase	Average of last two purchase actions in past
Customer behavioral history	Lag4_Purchase	Average of last four purchase actions in past
Customer behavioral history	Lag6_Purchase	Average of last six purchase actions in past
Customer behavioral history	Time elapsed since the last visit	The number of days between that day and last day member visited website
Customer behavioral history	Time elapsed since the last purchase	The number of days between that day and last day member purchased an item
Customer behavioral history	Average views per day	Average number of view actions made by a member per day
Customer behavioral history	Average purchases per day	Average number of purchases made by a member per day
Customer behavioral history	Average add to cart per day	Average number of add to cart actions made by a member per day
Customer behavioral history	How long customer	How many days of being customer
Customer behavioral history	Total amount spent	Total amount of an order
Channel and seasonality factors	Channel	Types of channels utilized (web, mobile, webmobile, fanshop)
Channel and seasonality factors	Weekdays	Days of a week when members being active
Channel and seasonality factors	Month	Months when members having website activity
Channel and seasonality factors	Season	Seasons when members creating a website traffic (winter, fall, spring, summer)
Customer behavior frequencies	Mean arrival frequency	Average visiting frequency of a member
Customer behavior frequencies	Mean view frequency	Average frequency of member view actions
Customer behavior frequencies	Mean purchase frequency	Average frequency of member purchase actions
Customer behavior frequencies	Mean add to cart frequency	Average frequency of member add to cart actions
Month, channel, day and season specific sales factors	Average monthly revenue	Average amount of spending per month
Month, channel, day and season specific sales factors	Average revenue by channel	Average amount of spending per channel
Month, channel, day and season specific sales factors	Average daily revenue	Average amount of spending per weekday
Month, channel, day and season specific sales factors	Average seasonally revenue	Average amount of spending per season
Month, channel, day and season specific sales factors	Average monthly units sold	Average units ordered per month
Month, channel, day and season specific sales factors	Average units sold by channel	Average units ordered per channel
Month, channel, day and season specific sales factors	Average daily units sold	Average units ordered per weekday
Month, channel, day and season specific sales factors	Average seasonally units sold	Average units ordered per season
Conversion rate	Conversion rate of member	Ratio of number of purchases to number of visits by a member
Conversion rate	Average monthly conversion rate	Ratio of number of purchases to number of visits / made in each month
Conversion rate	Average conversion rate of channels	Ratio of number of purchases to number of visits / made by each channel
Conversion rate	Average daily conversion rate	Ratio of number of purchases to number of visits / made in each day
Conversion rate	Average seasonally conversion rate	Ratio of number of purchases to number of visits / made in each season

In summary, factors corresponding to customer behavioral history are mainly about the average of past transactions of members, how long have they been a member, inter arrival time of their actions and how much they spent. Channel and seasonality factors are used to investigate the effects of channel, weekday, month and season on website activity. The customer behavioral frequencies factors refer to mean of how frequent customers visit the website and their transactions by considering time interval. The other factors which are summarized as channel and seasonality specific sales determinant are related to average of how much turnover have the website on daily, monthly, seasonally and channel based. Last factor type is about to averaging of how many visits turn into purchase actions.

6.2. Model Development

Customer behavior modeling defines the model of customer behavior to make a prediction about how customers will behave similarly under identical conditions. Mathematical expressions of customer behavior are built in order to represent the response of a particular group of shopper.

In this research, different models are constructed to predict the future purchasing behavior of customers on the basis of historical clickstream data analyses by using a statistical tool, R program. The models examine transactional data to forecast how likely a customer to make a purchase online in the near future. In order to analyze the effects of factors properly, different models comprising various factors are established initially in Table 6.3 as follows:

Table 6.3. Models with their factors.

Type of Factors	Factors	Model 1	Model 2	Model 3	Model 4	Model 5
Customer behavioral history	Lag2_View	X	X	X	X	X
	Lag4_View	X	X	X	X	X
	Lag6_View	X	X	X	X	X
	Lag2_Purchase	X	X	X	X	X
	Lag4_Purchase	X	X	X	X	X
	Lag6_Purchase	X	X	X	X	X
	Time elapsed since the last visit	X	X	X	X	X
	Time elapsed since the last purchase	X	X	X	X	X
	Average views per day	X	X	X	X	X
	Average purchases per day	X	X	X	X	X
	Average add to cart per day	X	X	X	X	X
	How long customer	X	X	X	X	X
	Total amount spent	X	X	X	X	X
Channel and seasonality factors	Channel		X	X	X	X
	Weekdays		X	X	X	X
	Month		X	X	X	X
	Season		X	X	X	X
Customer behavior frequencies	Mean arrival frequency			X	X	X
	Mean view frequency			X	X	X
	Mean purchase frequency			X	X	X
	Mean add to cart frequency			X	X	X
Month, channel, day and season specific sales factors	Average monthly revenue				X	X
	Average revenue by channel				X	X
	Average daily revenue				X	X
	Average seasonally revenue				X	X
	Average monthly units sold				X	X
	Average units sold by channel				X	X
	Average daily units sold				X	X
	Average seasonally units sold				X	X
Conversion rate	Average conversion rate of member					X
	Average monthly conversion rate					X
	Average conversion rate by channel					X
	Average daily conversion rate					X
	Average seasonally conversion rate					X

Table 6.4. Summary of models.

Model	**Explanation**
Model 1	Customer Behavioral History
Model 2	Customer Behavioral History + Channel & Seasonality
Model 3	Customer Behavioral History + Channel & Seasonality + Customer Behavioral Frequencies
Model 4	Customer Behavioral History + Channel & Seasonality + Customer Behavioral Frequencies + Sales Specific Factors
Model 5	Customer Behavioral History + Channel & Seasonality + Customer Behavioral Frequencies + Sales Specific Factors + Conversion Rate

In this study, five models are built firstly, as seen in Table 6.3 and 6.4. After training our models with logistic regression method, we checked whether there are any useless factors in the models by applying backward stepwise regression. Then totally, we obtained 10 different models. In the first model, we assess the influence of past transaction averages which are mean of views, add to carts and purchases per day for each customer, average of last two-four and six transactions in the past, elapsed time between each view actions and each purchase actions, how long the customer has been a member and total amount spent by each customer. Our further models are established with some other factor additions to this base model. Model 2 has channel and seasonality effects additionally; whereas Model 3 involves the frequency of arrival, viewing or purchasing in addition to the second model. Then, in the Model 4 the averages of revenue and units sold on the basis of month, channel, weekday and season are investigated additively. Moreover, in the last model the influences of conversion rates on the purchasing probability are examined.

In order to obtain eligible variables for all models, five models are built with logistic regression method and then backward stepwise regression is applied. All model types with their simple forms and independent variables used in these models are summarized in Table 6.5.

Table 6.5. Models with their simple forms.

Factors	Model 1	Model 2	Model 3	Model 4	Model 5	Simple Model 1	Simple Model 2	Simple Model 3	Simple Model 4	Simple Model 5
Lag2_View	X	X	X	X	X	X	X	X	X	X
Lag4_View	X	X	X	X	X					
Lag6_View	X	X	X	X	X					
Lag2_Purchase	X	X	X	X	X					
Lag4_Purchase	X	X	X	X	X			X	X	X
Lag6_Purchase	X	X	X	X	X					
Time elapsed since the last visit	X	X	X	X	X	X	X	X	X	X
Time elapsed since the last purchase	X	X	X	X	X	X	X	X	X	X
Average views per day	X	X	X	X	X	X	X	X	X	X
Average purchases per day	X	X	X	X	X	X	X	X	X	X
Average add to cart per day	X	X	X	X	X					
How long customer	X	X	X	X	X					
Total amount spent	X	X	X	X	X	X	X	X	X	X
Channel		X	X	X	X		X	X	X	X
Weekdays		X	X	X	X					
Month		X	X	X	X		X	X	X	X
Season		X	X	X	X					
Mean arrival frequency			X	X	X			X	X	
Mean view frequency			X	X	X					X
Mean purchase frequency			X	X	X			X	X	
Mean add to cart frequency			X	X	X					
Average monthly revenue				X	X					
Average revenue by channel				X	X					
Average daily revenue				X	X					
Average seasonally revenue				X	X					
Average monthly units sold				X	X					
Average units sold by channel				X	X					
Average daily units sold				X	X					
Average seasonally units sold				X	X					
Average conversion rate of member					X					X
Average monthly conversion rate					X					
Average conversion rate by channel					X					
Average daily conversion rate					X					
Average seasonally conversion rate					X					

Since we try to estimate whether a member makes a purchase or not, our response factor is identified as qualitative variable. Classification methods are selected to construct a model estimating the likelihood of purchasing because a binary independent variable exists. Three types of classification approach are used to build mathematical expressions; which are logistic regression, random forest and boosting methods. In the models, the probability of purchasing activity of customers is predicted with the assumption that there is no interaction between covariates.

In order to measure relative closeness of predictions to actual outcomes, dataset consisting of three years transactional history is separated randomly into two parts: train set and test set. The former one comprises of %80 of all data; whereas the latter includes the rest. During the division of the raw dataset into two parts, we checked the closeness of

the distribution of target probability in train and test set. We set these two dataset when we obtained roughly same conversion probability, seen from Table 6.6. While the first data partition is used to develop a model; the other is used for predictive model assessment. Due to not having access to future literally, some of present available data (test set) are treated as if it were future data.

Table 6.6. Number of observations for each class in the training and test data set.

Predictor class	**0**	**1**	**Conversion rate probability**
Train set	19405	636	0.03173
Test set	4817	159	0.03195

Since in the overall dataset the number of instances non-purchasing class far more than the total number of purchasing class data, class imbalance problem arise. As can be seen in Table 6.6, since there are so many view actions in our train dataset (0 class), we come across with a severe class imbalance problem. That is, our prediction models are more prone to generate a classifier that stratifies every sample as the majority class with the probability of % 97 approximately. To compare results of models having the problem of disproportionate number of classes, alternative measure AUC is used despite of the accuracy counting number of errors. Since defining the percentage of consumers having tendency to buy is more essential for our study than a mere classification of buyers and non-buyers, AUC is preferred as a performance assessment method instead of accuracy. If the rationale behind is explained mathematically, in this ROC curve performance method, accuracy is partitioned to sensitivity and specificity. In this way, models can be selected by considering the balance thresholds of these values. To overcome the class imbalance problem, after creating different models with different variables, we evaluated the predictive performance of all models with the help of the AUC method.

While developing our models, we also used cross validation technique to assess the predictive ability of our models to the data has not already seen. Since the models are fitted to training data set, our model might perform well on that data set. However, the prediction performance may decrease when performed to the testing set. Therefore, the small set of training data is seperated and used as a test of the trained model.

Cross validation method is used to measure how our models perform the set of data not used in estimation and to optimize several parameters of boosting method. For boosting method we tested model performances by evaluating different values of depth, shrinkage and number of tree seen in Table 6.7. In the study, we tried to compare all possible alternatives of models by the help of this method. Therefore, we applied ten fold cross validation method with five replications to our models analyzed with three different statistical approach.

Table 6.7. Parameter values of boosting method.

Parameter	**Values**
Depth	1, 3, 5
Shrinkage	0.001, 0.005, 0.01, 0.05
Number of Trees	300, 500, 1000

6.3. Model Results

In this part, we interpreted the model results by comparing them. After building and testing the predictive performance of above models, the model results are summarized in this section.

When we evaluate sensitivity and specificity separately, their relationships are ignored. In order to assess the prediction performance, ROC curve is better method since it represents the tradeoff between hit rates and false alarm rates. Therefore, we used AUC method to compare all models' predictive performance.

The AUC values of different models are summed up in Table 6.8. When AUC values of cross validation (train data AUC) and test data after cross validation of boosting method are compared, the results are found close to each other, that indicates the consistency of our results for boosting method. That is, our model AUC results of train data show similarity with AUC of test set.

Table 6.8. Model AUC results of statistical methods.

	LOGISTIC REGRESSION	**RANDOM FOREST**	**BOOSTING**	
	Test Data AUC	**Test Data AUC**	**Cross Validation AUC**	**Test Data AUC**
Model 1	0,6863	0,6966	0,7157	0,7103
Simple Model 1	0,6807	0,6585	0,7176	0,7022
Model 2	0,6832	0,7077	0,7156	0,7169
Simple Model 2	0,6912	0,6999	0,7180	0,7211
Model 3	0,6929	0,7187	0,7317	0,7075
Simple Model 3	0,7016	0,7126	0,7336	0,7218
Model 4	0,6929	0,7046	0,7316	0,7221
Simple Model 4	0,7016	0,7126	0,7336	0,7218
Model 5	0,7043	0,7205	0,6757	0,6757
Simple Model 5	0,7094	0,6997	0,7309	0,7268

If the performances of models in Table 6.8 are compared, since AUC values of Simple Model 5 are higher than that of other methods, this model outperforms to all other models. In order to interpret the results by benchmarking statistical methods, we chose Simple Model 5 as best model.

The best estimating methods of this model are boosting and logistic regression methods due to having the best AUC values. According to AUC assessment method, the balance thresholds of false positive and true positive values are considered instead of assessing them one by one.

The selected model, Simple Model 5 is obtained after the elimination of the least significant variables from Model 5, seen in Figure 6.9. When the useless factors are removed from the model, AIC value decreases. Therefore, the Simple Model 5 is better model than Model 5 since it has the smallest AIC value, seen in Table 6.10.

Table 6.9. Factor Table of Model 5 and Simple Model 5.

Factors	**Model 5**	**Simple Model 5**
Lag2_View	X	X
Lag4_View	X	
Lag6_View	X	
Lag2_Purchase	X	
Lag4_Purchase	X	X
Lag6_Purchase	X	
Time elapsed since the last visit	X	X
Time elapsed since the last purchase	X	X
Average views per day	X	X
Average purchases per day	X	X
Average add to cart per day	X	
How long customer	X	
Total amount spent	X	X
Channel	X	X
Weekdays	X	
Month	X	X
Season	X	
Mean arrival frequency	X	
Mean view frequency	X	X
Mean purchase frequency	X	
Mean add to cart frequency	X	
Average monthly revenue	X	
Average revenue by channel	X	
Average daily revenue	X	
Average seasonally revenue	X	
Average monthly units sold	X	
Average units sold by channel	X	
Average daily units sold	X	
Average seasonally units sold	X	
Average conversion rate of member	X	X
Average monthly conversion rate	X	
Average conversion rate by channel	X	
Average daily conversion rate	X	
Average seasonally conversion rate	X	

Table 6.10. Model AIC results of Model 5 and Simple Model 5.

AIC Values	
Model 5	5244
Simple Model 5	5226

6.3.1. Logistic Regression Method Results

Logistic regression and boosting methods outperform to random forest approach in terms of the predictive performance of Simple Model 5. The logistic regression performance results of Simple Model 5 are given in Table 6.11, as follows:

Table 6.11. Summary table of Simple Model 5.

Coefficients	**Estimate**	**Standard Error**	**z value**	**Pr(>|z|)**	
Intercept	-1.836e+00	1.242e+00	-1.478	0.13937	
Lag2_View	2.949e-02	5.311e-03	5.551	2.83e-08	***
Lag4_Purchase	-1.960e-01	9.337e-02	-2.099	0.03578	*
Time elapsed since the last visit	5.385e-03	2.034e-03	2.647	0.00812	**
Time elapsed since the last purchase	-7.772e-04	4.400e-04	-1.766	0.07732	.
Average views per day	-1.292e-01	4.576e-02	-2.824	0.00474	**
Average purchases per day	3.771e+00	2.582e+00	1.460	0.14415	
Total amount spent	3.327e-04	8.513e-05	3.908	9.29e-05	***
Channel_mobile	-2.645e+00	1.237e+00	-2.139	0.03247	*
Channel_web	-2.495e+00	1.236e+00	-2.018	0.04359	*
Channel_webmobile	-3.448e+00	1.249e+00	-2.760	0.00578	**
Month_August	-1.139e-01	2.473e-01	-0.460	0.64517	
Month_December	6.770e-02	2.156e-01	0.314	0.75351	
Month_February	2.366e-01	2.190e-01	1.080	0.27992	
Month_January	-3.660e-01	2.574e-01	-1.422	0.15502	
Month_July	-2.709e-01	2.597e-01	-1.043	0.29695	
Month_June	2.955e-01	2.107e-01	1.402	0.16083	
Month_March	4.127e-02	2.176e-01	0.190	0.84958	
Month_May	2.541e-01	2.062e-01	1.232	0.21794	
Month_November	2.013e-01	1.865e-01	1.079	0.28039	
Month_October	2.582e-01	1.924e-01	1.342	0.17964	
Month_September	4.134e-01	1.980e-01	2.178	0.02937	*
Mean view frequency	5.947e-02	2.108e-02	2.821	0.00479	**
Conversion rate of member	2.414e+01	3.702e+00	6.520	7.02e-11	***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
Null deviance: 5640.4 on 20040 degrees of freedom					
Residual deviance: 5177.7 on 20017 degrees of freedom AIC: 5225.7					

The summary table of selected model built with logistic regression shows that:

- The p-value for all factors is compared with the significance level to test the null hypothesis. It is assumed that in the null hypothesis that the factor's coefficient is equal to zero, that means there is no relationship between the factor and the response. Since the p-value is less than the significance level, it is concluded that there is a statistically significant association between the response variable and the independent variables. Most of factors are seen as significant since the p-values are small, which are the average of last two views in the past, average of last four purchases in the past, time elapsed since the last visit and last purchase, average views per day, total amount spent, channel (mobile, web, web-mobile), month, mean view frequency and conversion rate of member.

- For the independent variables having small p value which are continuous predictors, it can be concluded that the coefficients do not equal to zero.

- The mean of last four purchases in the past, time elapsed since last purchases, average views per day and channel factors have negative influence on the purchasing probability.

- However, the average of last two views, the time elapsed since the last visit of users, total amount spent, month (September), the average frequency of view actions and conversion rate factors affect the purchasing probability of customers positively. The mean of last two view actions, overall amount spent by members, the ratio of purchases to visits are more significant than other factors affecting probability in a positive manner.

- For every one unit change in the factor of average last four purchases, the time elapsed after the last purchasing action and the mean of daily view actions, the log odds of buying declines by -0.196, -0.0007 and -0.125 respectively. Also, various channels affect purchasing probability at different rates, such that each additional one unit change in the drivers of channel mobile, channel web and channel web-mobile result in a decrease of the log odds by -2.645, -2.495 and -3.448 in turn.

- On the other hand, for every unit change in the average of last two view actions, time elapsed since the last visit and total amount spent; the log odds increases by 0.029,

0.005 and 0.0003 correspondingly. Moreover, if the factors of month September, the average frequency of views and conversion rate of member changes one unit, the log odds of buying probability shows a boost by 0.413, 0.06 and 20.4 in order.

To sum up, the logistic equation of Simple Model 5 can be seen as stated below:

$$
\begin{aligned}
ln\frac{P(Purchase)}{1-P(Purchase)} &= -1.836 + 0.029 * Average\ last\ two\ views - 0.196 \\
&* Average\ last\ two\ purchases + 0.005 \\
&* Time\ elapsed\ since\ the\ last\ visit - 0.0007 \\
&* Time\ elapsed\ since\ the\ last\ purchase - 0.125 \\
&* Average\ views\ per\ day + 0.0003 * Total\ amount\ spent - 2.645 \\
&* Channel(mobile) - 2.495 * Channel(web) - 3.448 \\
&* Channel(webmobile) + 0.413 * Month(September) + 0.06 \\
&* Mean\ view\ frequency + 20.4 * Conversion\ rate\ of\ member
\end{aligned}
\tag{5.1}
$$

The binary classifier performance of Simple Model 5 built with logistic regression equals to 0.71. As can be seen in Figure 6.1, ROC curve is close to the top left corner of the plot, which shows the good discrimination of our model.

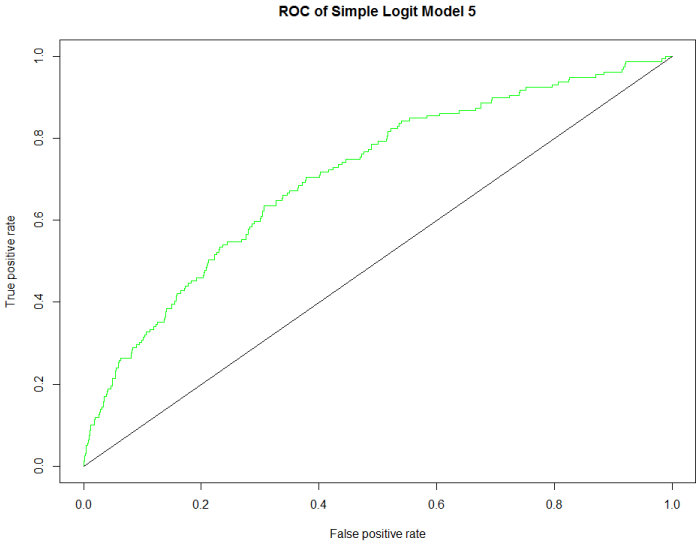

Figure 6.1. ROC Curve of Simple Model 5 Built with Logistic Regression.

The variable importance of the variables used in the model is evaluated, it is seen from Table 6.12 that conversion rate of member, the average of last two view actions and

total amount spent factors have highest importance value as specified in the output of summary table of Simple Model 5. The importance score is scaled between 0 and 100.

Table 6.12. Variable importance of Simple Model 5 built with logistic regression.

Independent Variables	**Overall**
Conversion rate of member	6.520
Average of last two views	5.551
Total amount spent	3.908
Average views per day	2.824
Mean view frequency	2.821
Channel (web mobile)	2.760
Time elapsed since the last visit	2.646
Month (September)	2.178
Channel (mobile)	2.139
Average of last two purchases	2.099
Channel (web)	2.018
Time elapsed since the last purchase	1.766
Average purchases per day	1.460
Month (January)	1.422
Month (June)	1.402
Month (October)	1.342
Month (May)	1.232
Month (February)	1.08
Month (November)	1.079
Month (July)	1.043
Month (August)	0.460
Month (December)	0.314
Month (March)	0.190

6.3.2. Boosting Method Results

When all statistical methods are benchmarked, the other best approach is selected as boosting.

We assessed the predictive performance of boosting method by visualizing its model in Figure 6.2 with AUC method. The area under the ROC curve of Simple Model 5 built with boosting method equals to 0.73. In the figure shown below, the points above the diagonal show good classification results of our model.

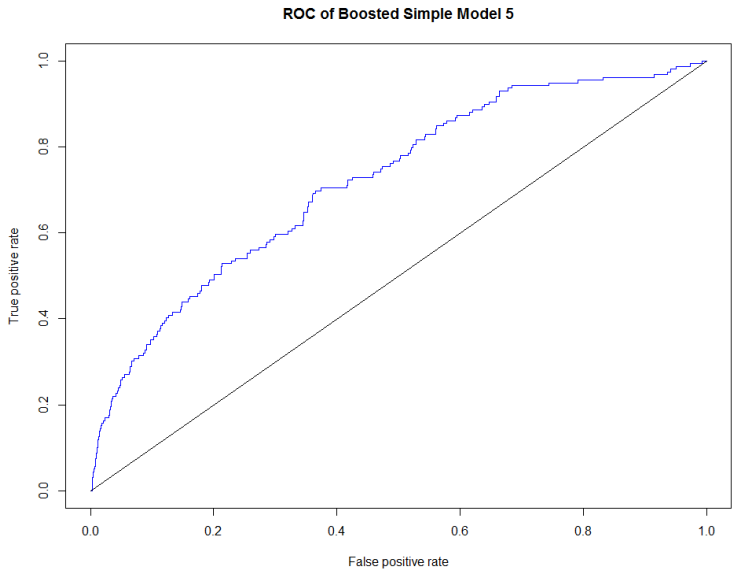

Figure 6.2. ROC Curve of Simple Model 5 by Using Boosting Method.

The variable importance of Simple Model 5 built with boosting method is assessed, month, the average of last two views and time elapsed since the last purchase factors are found as the most important variables, shown in Table 6.13 and Figure 6.3. Like logistic regression method, average time of last two view factor is seen as a significant factor affecting the purchasing decision.

Table 6.13. Relative influences of factors used in boosted Simple Model 5.

Independent Variables	**Relative Influence**
Month	23.978
Average of last two views	17.431
Time elapsed since the last purchase	15.271
Conversion rate of member	9.203
Total amount spent	8.149
Average purchases per day	7.058
Average of last four purchases	5.481
Time elapsed since the last visit	4.450
Average views per day	3.809
Mean view frequency	3.481
Channel	1.688

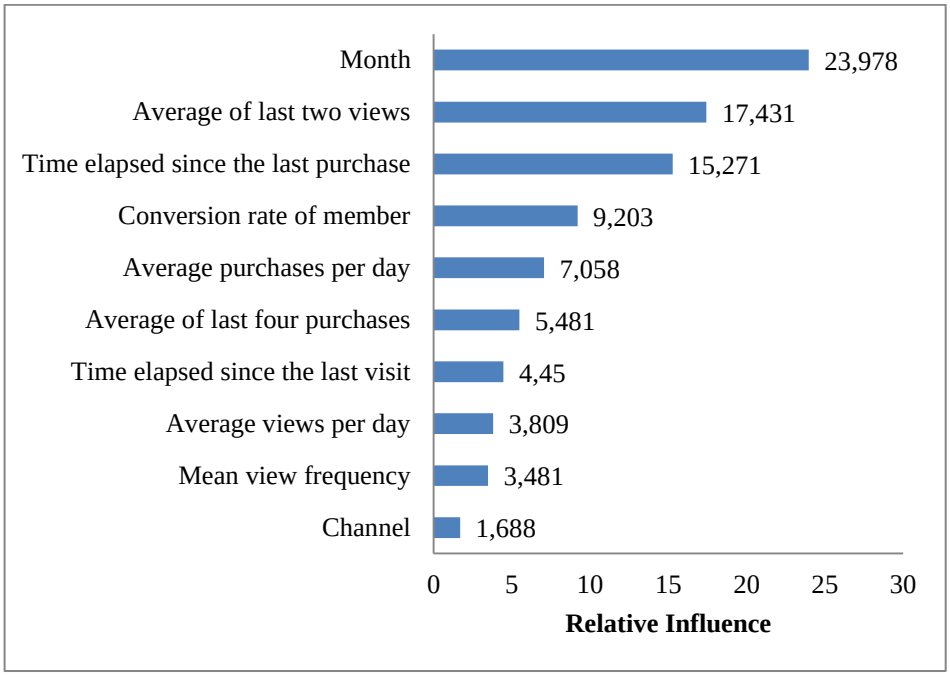

Figure 6.3. Relative Importance Plot of Boosted Simple Model 5.

In order to visualize the marginal effects of boosted Simple Model 5 predictors, the following plots in Figure 6.4 are drawn for each variable with the principle of holding other covariates constant. In these partial dependence plots, the predicted probabilities for each covariate by integrating out all other factors can be interpretable.

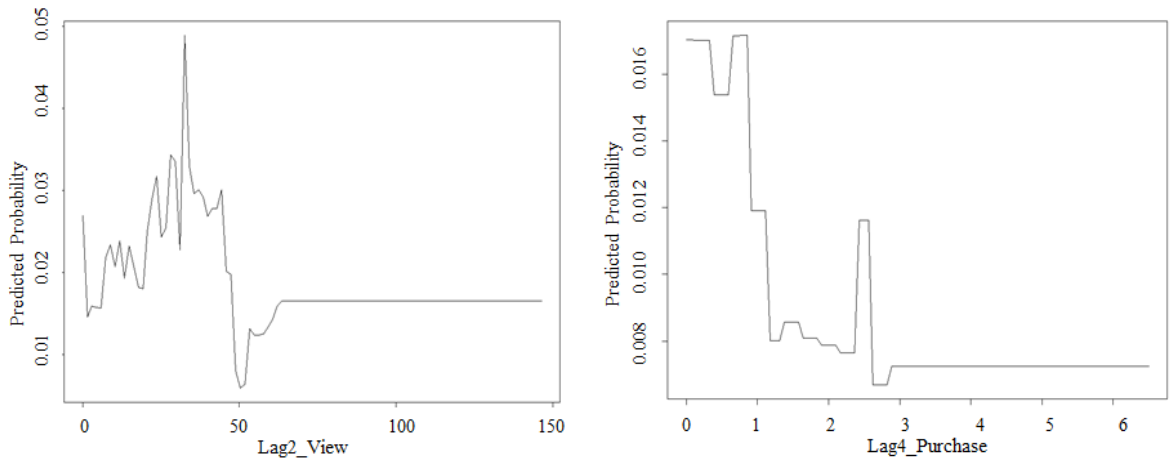

Figure 6.4. Marginal Plots of Predictors in Boosted Simple Model 5.

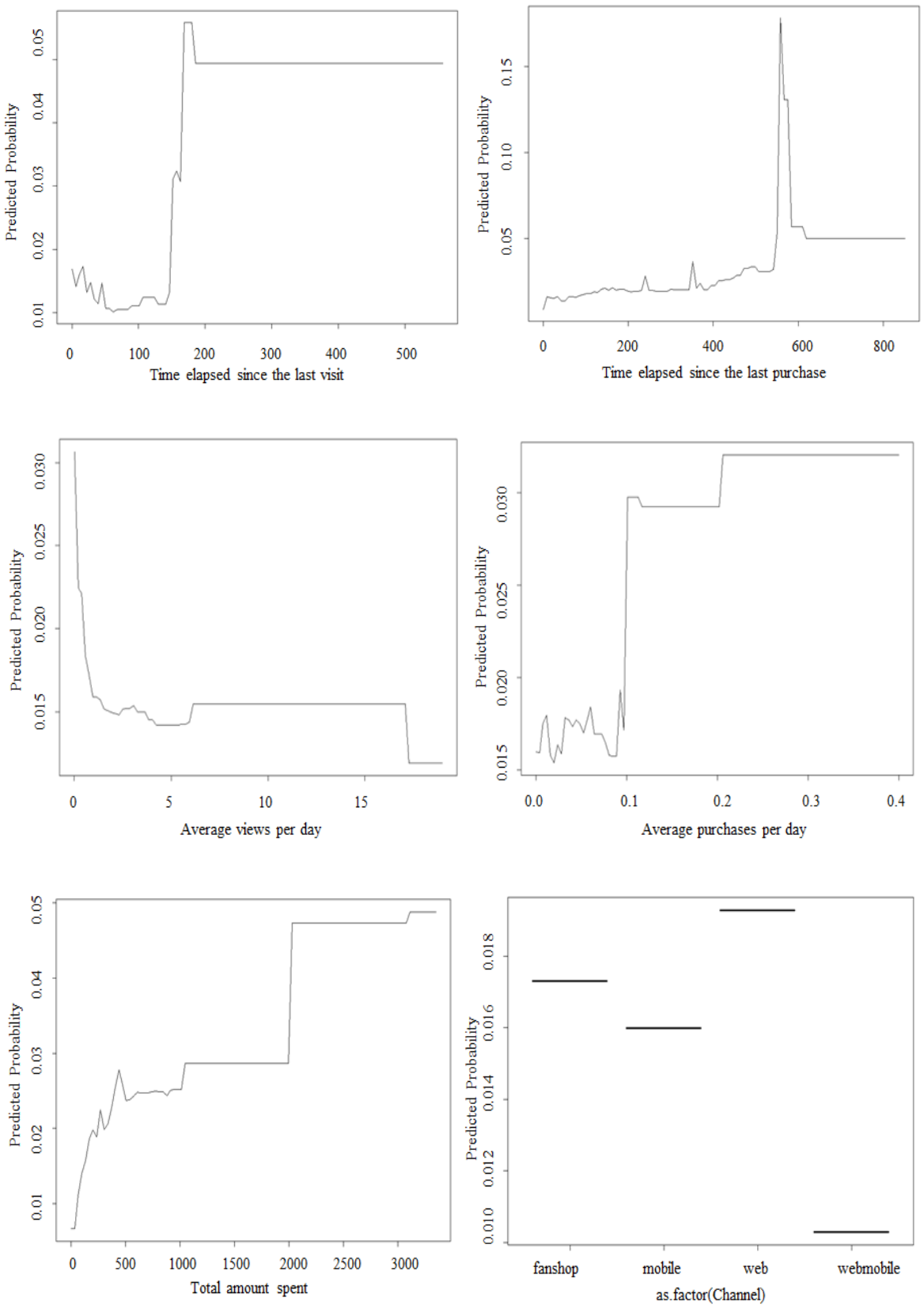

Figure 6.4. Marginal Plots of Predictors in Boosted Simple Model 5.(cont.)

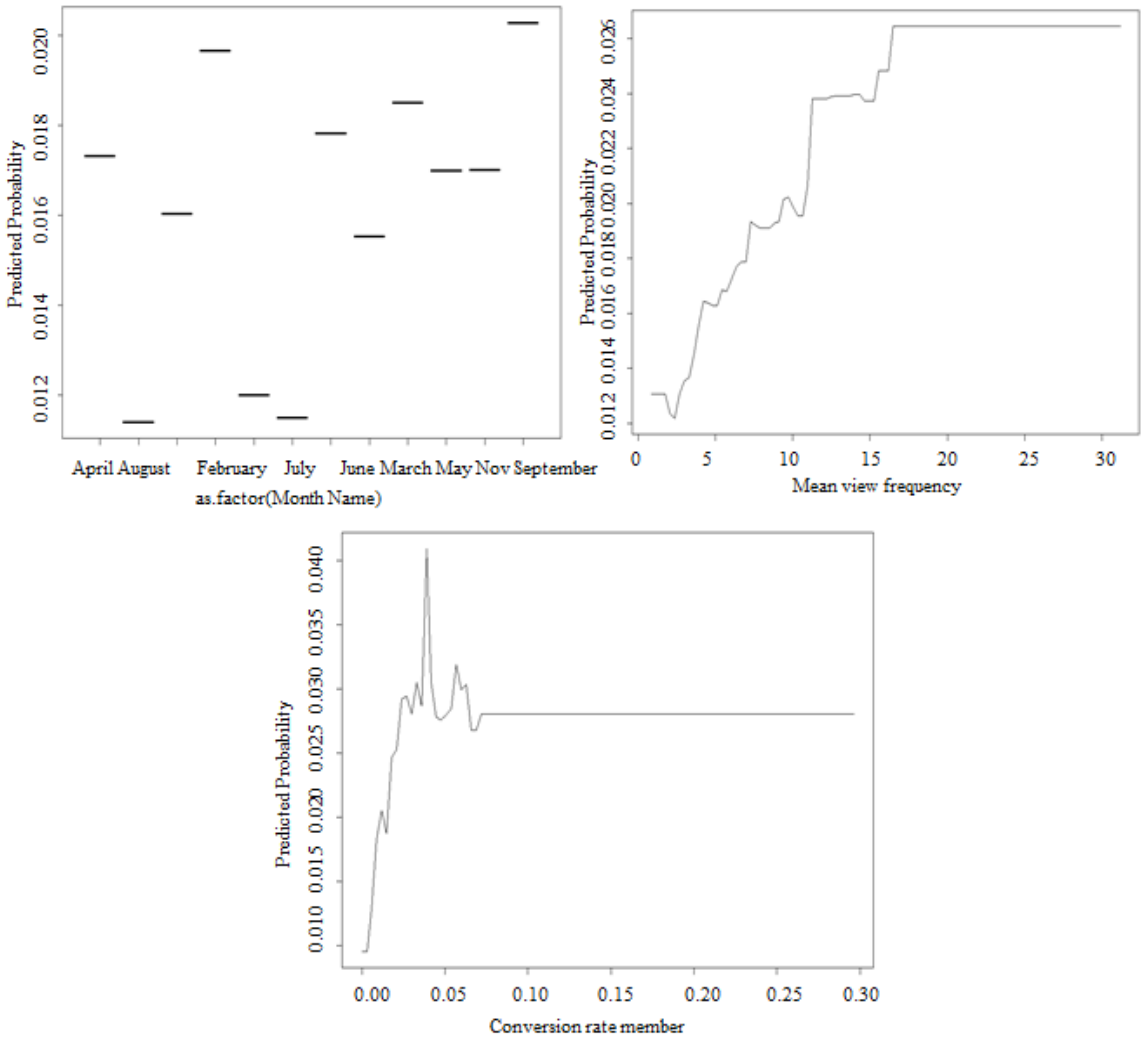

Figure 6.4. Marginal Plots of Predictors in Boosted Simple Model 5.(cont.)

According to Figure 6.4, the predicted probabilities of the factors which are the mean of last two view actions, time elapsed since the last purchase, time elapsed since the last visit, total amount spent and conversion rate of member are higher than that of other plots. That is, these variables are more important since their impact on the output is larger. Supportively, four of these factors except from time elapsed the since last visit are seen as the most important variable in Figure 6.3. If their predicted probabilities are compared between each other, it is seen that time elapsed since the last purchase has highest estimated likelihood.

In addition, when relative importance plots are evaluated separately, the predicted probability of buying increases with the change of the factors which are time elapsed since the last visit, time elapsed since the last purchase, average purchases per day, total amount spent, mean view frequency, conversion rate of member and total amount spent. The month September and channel web affect the probability in a positive manner. Besides, the average of last two view actions factor results in an increase up to a certain point and after a threshold the predicted probability decreases and stabilizes. However, the mean of last four purchase actions and average views per day have a negative influence on the estimated likelihood of purchasing. Similar to logistic regression outputs, total amount spent, time elapsed since the last visit, average of last two views, mean view frequency, conversion rate of member and month September boost the buying likelihood; while the mean of last four purchases and average views per day cause a decrease in the purchasing probability.

Furthermore, we compared all possible alternatives of models with cross validation method, whose results were interpreted in the above parts. In addition to the advantage of measuring our models performed the set of data that has not seen already with cross validation method, we found best parameters of boosting method. Therefore, we applied ten fold cross validation method with five replications to our models. According to cross validation results seen in Table 6.14, the best parameters of boosting method are selected in Table 6.15. The parameter combinations resulting the highest AUC values are selected as appropriate.

By using best parameters, models are built with the all training set and performed with test set.

Table 6.14. Cross validation AUC results of boosting method.

Parameters			Average AUC									
Depth	Shrinkage	Number of Trees	Model 1	Simple Model 1	Model 2	Simple Model 2	Model 3	Simple Model 3	Model 4	Simple Model 4	Model 5	Simple Model 5
1	0,001	300	0,6795	0,6796	0,6782	0,6793	0,7025	0,7019	0,7024	0,7019	0,6757	0,6985
3	0,001	300	0,7002	0,7028	0,6972	0,6978	0,7237	0,7241	0,7246	0,7241	0,6757	0,7191
5	0,001	300	0,7073	0,7116	0,7028	0,7079	0,7286	0,7305	0,7284	0,7305	0,6757	0,7275
1	0,005	300	0,6977	0,6961	0,6971	0,6964	0,7247	0,7242	0,7250	0,7242	0,6757	0,7174
3	0,005	300	0,7110	0,7140	0,7073	0,7093	0,7310	0,7315	0,7309	0,7315	0,6757	0,7293
5	0,005	300	0,7187	0,7219	0,7139	0,7192	0,7320	0,7345	0,7326	0,7345	0,6757	0,7333
1	0,01	300	0,7016	0,7034	0,7011	0,7037	0,7297	0,7295	0,7293	0,7295	0,6757	0,7272
3	0,01	300	0,7219	0,7247	0,7203	0,7232	0,7355	0,7368	0,7350	0,7368	0,6757	0,7357
5	0,01	300	0,7267	0,7285	0,7252	0,7285	0,7372	0,7395	0,7370	0,7395	0,6757	0,7384
1	0,05	300	0,7232	0,7256	0,7239	0,7275	0,7363	0,7389	0,7359	0,7389	0,6757	0,7370
3	0,05	300	0,7292	0,7328	0,7323	0,7349	0,7385	0,7436	0,7403	0,7436	0,6757	0,7395
5	0,05	300	0,7281	0,7296	0,7321	0,7330	0,7393	0,7411	0,7375	0,7411	0,6757	0,7384
1	0,001	500	0,6888	0,6865	0,6884	0,6878	0,7086	0,7092	0,7097	0,7092	0,6757	0,7114
3	0,001	500	0,7026	0,7044	0,6993	0,6992	0,7265	0,7270	0,7267	0,7270	0,6757	0,7218
5	0,001	500	0,7094	0,7149	0,7040	0,7093	0,7296	0,7320	0,7295	0,7320	0,6757	0,7287
1	0,005	500	0,7006	0,7008	0,7000	0,7009	0,7294	0,7289	0,7289	0,7289	0,6757	0,7252
3	0,005	500	0,7188	0,7222	0,7178	0,7197	0,7337	0,7353	0,7341	0,7353	0,6757	0,7342
5	0,005	500	0,7243	0,7277	0,7229	0,7272	0,7360	0,7382	0,7357	0,7382	0,6757	0,7378
1	0,01	500	0,7088	0,7137	0,7090	0,7136	0,7319	0,7333	0,7316	0,7333	0,6757	0,7320
3	0,01	500	0,7267	0,7300	0,7282	0,7312	0,7389	0,7411	0,7386	0,7411	0,6757	0,7394
5	0,01	500	0,7304	0,7329	0,7324	0,7348	0,7409	0,7433	0,7409	0,7433	0,6757	0,7412
1	0,05	500	0,7261	0,7279	0,7290	0,7300	0,7380	0,7399	0,7373	0,7399	0,6757	0,7373
3	0,05	500	0,7292	0,7279	0,7312	0,7345	0,7364	0,7406	0,7365	0,7406	0,6757	0,7376
5	0,05	500	0,7240	0,7238	0,7270	0,7280	0,7322	0,7370	0,7311	0,7370	0,6757	0,7307
1	0,001	1000	0,6953	0,6932	0,6943	0,6937	0,7214	0,7211	0,7225	0,7211	0,6757	0,7153
3	0,001	1000	0,7055	0,7086	0,7035	0,7046	0,7294	0,7301	0,7293	0,7301	0,6757	0,7260
5	0,001	1000	0,7146	0,7189	0,7088	0,7142	0,7312	0,7332	0,7314	0,7332	0,6757	0,7312
1	0,005	1000	0,7090	0,7136	0,7090	0,7133	0,7317	0,7334	0,7316	0,7334	0,6757	0,7323
3	0,005	1000	0,7278	0,7306	0,7281	0,7310	0,7390	0,7415	0,7388	0,7415	0,6757	0,7399
5	0,005	1000	0,7306	0,7326	0,7322	0,7356	0,7410	0,7440	0,7407	0,7440	0,6757	0,7414
1	0,01	1000	0,7196	0,7240	0,7210	0,7253	0,7352	0,7376	0,7353	0,7376	0,6757	0,7363
3	0,01	1000	0,7316	0,7336	0,7339	0,7366	0,7420	0,7448	0,7419	0,7448	0,6757	0,7416
5	0,01	1000	0,7324	0,7333	0,7358	0,7378	0,7420	0,7449	0,7428	0,7449	0,6757	0,7421
1	0,05	1000	0,7258	0,7292	0,7292	0,7311	0,7365	0,7387	0,7363	0,7387	0,6757	0,7363
3	0,05	1000	0,7217	0,7225	0,7253	0,7282	0,7282	0,7326	0,7296	0,7326	0,6757	0,7307
5	0,05	1000	0,7159	0,7109	0,7197	0,7203	0,7221	0,7263	0,7189	0,7263	0,6757	0,7222
BEST			0,7324	0,7336	0,7358	0,7378	0,7420	0,7449	0,7428	0,7449	0,6757	0,7421

Table 6.15. Best parameters of boosting method for different models.

Parameters			Best AUC									
Depth	Shrinkage	Number of Trees	Model 1	Simple Model 1	Model 2	Simple Model 2	Model 3	Simple Model 3	Model 4	Simple Model 4	Model 5	Simple Model 5
3	0,01	1000		0,7336			0,7420					
5	0,01	1000	0,7324		0,7358	0,7378		0,7449	0,7428	0,7449	0,6757	0,7421

6.3.3. Computation Time of Logistic and Boosting with respect to Data Size

In Table 6.16, the computation time of building a model with a training data set depending on the data size can be comparable for logistic and boosting methods. It is showed that training a model takes more time when a boosting method is used. Naturally, the more data size used to build a model, the more computation time is required to have result. Nevertheless, the processing time of logistic regression does not change after the threshold of %60 data size.

Table 6.16. Computation times of logistic regression and boosting by data size.

The Percentage of Data Size	Computation Time of Logistic Regression (sec)	Computation Time of Boosting (sec)
% 20	0.09	1.24
% 40	0.15	2.25
% 60	0.36	3.42
% 80	0.35	4.51
% 100	0.36	5.70

Moreover, the increase in the computation time of boosting method is more than that of logistic regression, seen in Figure 6.6. That indicates our data set can be modelled more appropriately with logistic regression method.

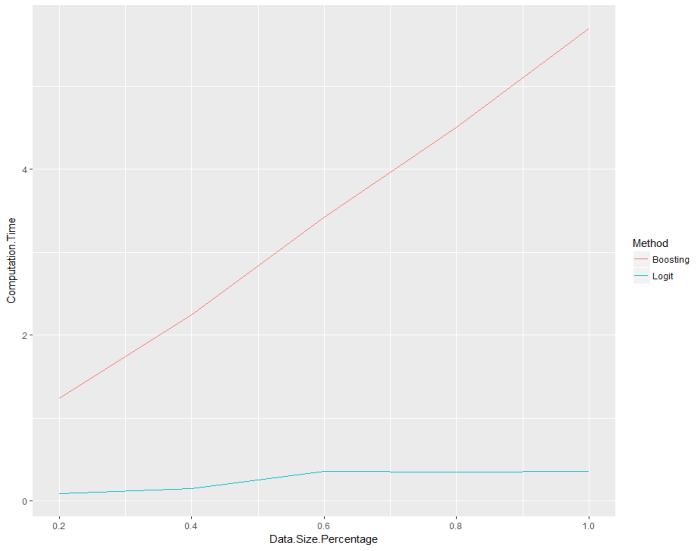

Figure 6.5. The Plot of Computation Time of Logistic and Boosting versus Data Size.

6.3.4. Computation Time of Boosting Method according to Parameters

In this part, we investigated how the parameters of boosting affect its computation time. When the effect of the number of trees on the computation time of boosting model is examined in Table 6.17, it is found that necessary computation time to establish a boosted model increases with the number of trees. Figure 6.7 shows the rise of processing time with respect to the number of trees.

Table 6.17. Computation times of boosting according to the number of trees.

Number of Trees	Computation Time (in seconds)
300	6.37
500	10.27
1000	19.84

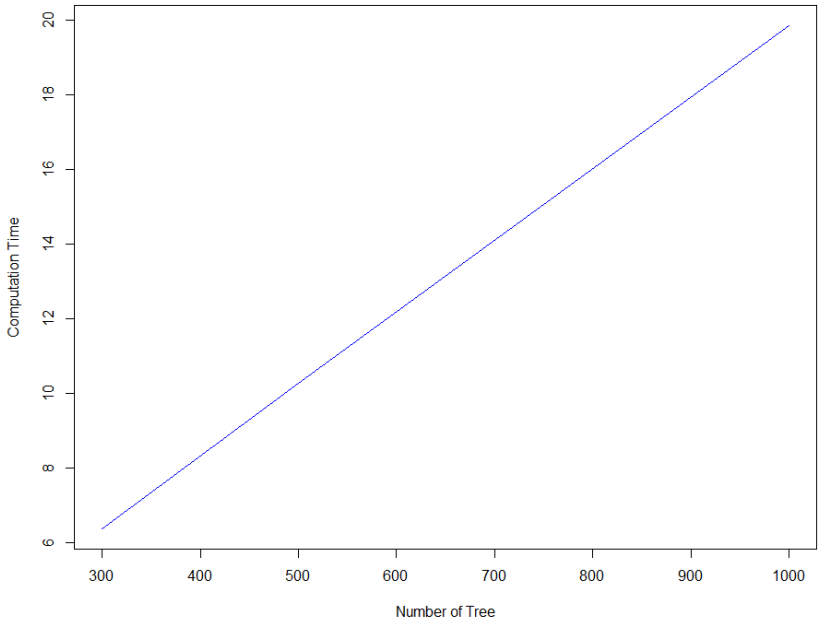

Figure 6.6. The Plot of Boosting Method Computation Time wrt the Number of Trees.

It can be interpretable from Table 6.18 and Figure 6.8 that if the shrinkage parameter, which controls the learning rate of boosting, is selected as 0.01, training a boosted model takes more time. The processing time of building a model decreases barely and shows an increase and decline with the change of shrinkage parameter.

Table 6.18. Computation times of boosting according to the shrinkage parameter.

Shrinkage	**Computation Time (in seconds)**
0.001	19.69
0.005	19.58
0.010	20.25
0.050	19.26

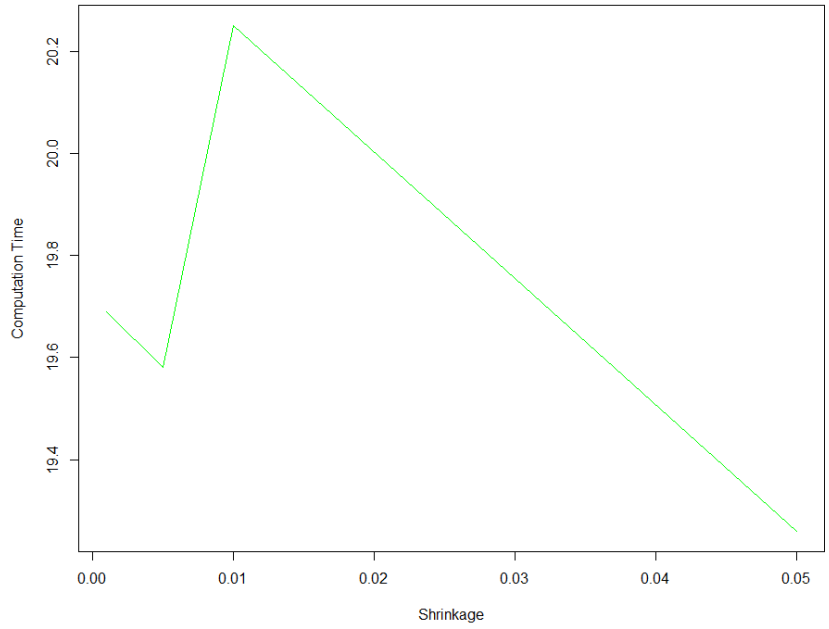

Figure 6.7. The Plot of Boosting Method Computation Time wrt Shrinkage.

When the depth parameter increases, the processing time to build a boosted model requires more time as well, seen in Table 6.19 and Figure 6.9.

Table 6.19. Computation times of boosting according to the depth parameter.

Depth	Computation Time (in seconds)
1	9.08
3	14.53
5	19.31

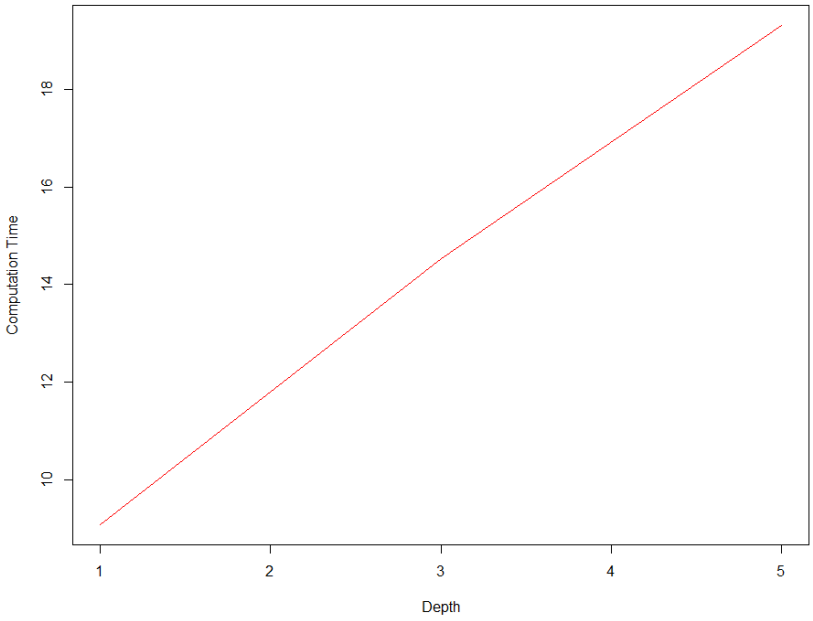

Figure 6.8. The Plot of Boosting Method Computation Time wrt Depth.

7. CONCLUSION

The digital environment has altered the way the customers and marketers interact, enabled new challenges by facilitating to access the information. The convenience of just a few clicks away information availability has brought about a rapid evolvement of customers' preferences and behaviors. The differentiation of customer journey has triggered recently the change in the strategies and the occurrence of dynamic market place. Therefore, the focus of marketers has shifted to follow the customers' footsteps in order to get to know them more closely and to meet their needs better.

The opportunities enabled by technology make possible to track customer transactions with an abundance of data. However, modelling the dynamics of customers' behavior using the clickstream data which tracks sequential transactions is an emerging but yet poorly studied field. Also, customers' behavior in today's world requires a stronger analytical effort to predict their propensity to view, click and buy. If and only if the customer behavior can be predicted successfully, the companies will be able to target their customers according to their tendencies, offer relevant contents to attract them, serve better by offering the products at the price they want. Therefore, identifying key attributes affecting customer decision making process through the analysis of click-stream data come into prominence.

In this research, we aimed to model customer purchasing behavior using clickstream data. We started to our study by examining historical pathways of consumers to define our metrics for modelling. We investigated how their past visits affect their purchasing decision, when customers are more prone to visit the website via which channel, what time of day they tend to buy more, which channels are used to end up the purchasing activity. According to the results of descriptive data analyses, we found that purchasing activities are influenced by add to cart activities and campaigns positively. Consumers prefer to use web to complete their purchasing activities, yet; mobile shopping is gaining importance day by day. Moreover, we determined that consumers are more inclined to visit the website at nights with mobile phones, while they often tend to turn their visits into an order over the web at midday. After detailed analyses of data, we defined our dynamics which are

factors related to the past actions of members, channel and seasonality effects, customer behavior frequencies and sales determinants month, channel, day and season based, also drivers associated with the conversion rate.

These predictors were used to build our various models with their different combinations. Since a sequence of choices made by customers is binary decision about whether to buy or not the product, we used classification methods. In order to find the better statistical approach to estimate the purchasing attitude of customers in the near future, we applied three classification methods, which are logistic regression, random forest and boosting. We examined all possible alternatives of our models and optimized our parameters with the help of cross validation method. Then, we benchmarked these approached with respect to their predictive performance measured with AUC method and requiring computation time.

After building our models, we indicated that logistic regression and boosting methods outperforms to all other methods in terms of predicting consumers' tendency to buying. The findings showed that logistic regression prediction accuracy equals to 0.71; whereas the AUC value of boosting method number is 0.73. Also, according to logistic regression method, conversion rate of member, average of last two view actions and total amount spent are most important three factors influencing the purchasing probabilities, while month, average of last two views and time elapsed since the last purchase are identified most essential three drivers of boosting. It is found that with respect to two method outputs, average of last two views, conversion rate of member, total amount spent, month September, mean view frequency and time elapsed since the last visit affect purchasing probability in a positive manner, whereas average views per day and average of last four purchases have a negative effect on buying probability. As to the comparison of the processing time of logistic regression and boosting methods, it was pointed out that logistic regression takes less time to train a model.

In conclusion, the objective of this thesis is to understand customers' rationale behind their choices and to estimate their steps beforehand with a reasonable model. There are still many areas that need to be investigated in future studies. It would be of interest to examine the effects of demographics, website design characteristics and level of Internet knowledge etc. on customer purchasing behavior. Also, product and brand choices, the impacts of promotions and campaigns may also be investigated. In addition, further

extensions of our proposed model could also capture both the influences of past behaviors and their two-way interactions. The same model would be built with a larger data set allowing to segment customers according to their similarities; in thus way the prediction performance may be developed.

REFERENCES

Artun, Ö. and D., Levin, 2015, *How to use predictive analytics to understand which consumers are most likely to buy*, Hubspot, https://blog.hubspot.com/agency/predictive-analytics-buy, accessed at December 2017.

Baydogan, M. G. and G., Runger, 2014, "Learning a symbolic representation for multivariate time series classification", Data Min Knowl Disc, Springer, 29, pp. 400–422.

Bednarowska, Z. and B., Jedruszek, 2012, *PMR: Nearly 70% of young people buy online*, www.research-pmr.com, pp.1-6, accessed at December 2017.

Bertsimas, D., n.d., *The Analytics Edge*, https://www.edx.org/course/analytics-edge-mitx-15-071x-3, accessed at December 2016.

Bucklin, R. E., J. M., Lattin, A., Ansari, S., Gupta, D., Bell, E., Coupey, J. D. C., Little, C., Mela, A., Montgomery and J., Steckel, 2002, "Choice and the Internet: From Clickstream to Research Stream", *Marketing Letters*, 13 (3), pp. 245–258.

Bucklin, R. E. and C., Sismerio, 2009, "Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing", Science Direct, *Journal of Interactive Marketing*, 23, pp. 35-48.

Burns, R., Logistic Regression, n.d., https://www.staff.ncl.ac.uk/mike.cox/III/spss10.pdf, accessed at December 2017.

Chaffey, 2017, *Ecommerce Conversion Rates*, Smart Insights, https://www.smartinsights.com/ecommerce/ecommerce-analytics/ecommerce-conversion-rates/, accessed at December 2017.

Chen, L. and Q., Su, 2013, "Discovering User's Interest at E-Commerce Site Using Clickstream Data", *IEEE*, 978-1-4673-4843-0/13, pp. 1-6.

Chiang, K. P. and R. R., Dholakia, 2003, "Factors Driving Customer Intention to Shop Online: An Empirical Investigation", *Journal of Consumer Psychology,* 13(1&2), pp. 177-183.

Close, A. G. and M., Kukar-Kinney, 2010, "Beyond Buying: Motivations behind Consumers' Online Shopping Cart Use", *Journal of Business Research,* 63, pp. 986-992.

Degeratu, A. M., A., Rangaswamy and J., Wu, 2000, "Consumer choice behavior in online and traditional supermarkets: The effects of brand name, price, and other search attributes", *International Journal of Research in Marketing,* 17, pp. 55–78.

eMarketer, 2014, *Worldwide Ecommerce Sales to Increase Nearly 20% in 2014*, http://www.emarketer.com/Article/Worldwide-Ecommerce-Sales-Increase-Nearly-20-2014/1011039, accessed at October 2016.

eMarketer, 2016, *Worldwide Retail Ecommerce Will Reach $1.915 Trillion This Year*, https://www.emarketer.com/Article/Worldwide-Retail-Ecommerce-Sales-Will-Reach-1915-Trillion-This-Year/1014369, accessed at October 2016.

eMarketer, 2017, *Consumer Behavior Roundup*, https://www.emarketer.com/public_media/docs/eMarketer_Consumer_Behavior_Roundup_2.pdf, accessed at December 2017.

Fan, M., Z., Sheng, L., Hao and Y., Tan, 2012, "A Hidden Markov Model for Conversion Rate Dynamics in Online Retail", *Thirty Third International Conference on Information Systems,* Economics and Value of IS, pp. 1-12.

Fawcett, T., 2006, "An Introduction to ROC Analysis", Science Direct, *Pattern Recognition Letters,* 27, pp. 861–874

Hastie, T., R., Tibshirani and J., Friedman, 2008, *The Elements of Statistical Learning*, Second edition, Springer, Stanford, California.

Hormann, W. and M. G. Guler, 2015, Unpublished Lecture Notes, IE-508 Statistical Inference Lecture Notes, Boğaziçi University.

Hou, J., 2015, "Online Stock Trading: Do Demographics, Internet Usage, and Attitudes Matter?", *International Journal of Business and Social Science,* 6 (2), pp. 8-11.

Huang, J. and C. X., Ling, 2005, "Using AUC and Accuracy in Evaluating Learning Algorithms", *IEEE Transactions on Knowledge and Data Engineering*, 17 (3), pp. 299-306.

Iwanaga, J., N., Nishimura, N., Sukegawa and Y., Takano, 2016, "Estimating Product-Choice Probabilities from Recency and Frequency of Page Views", *Knowledge-Based Systems*, Elsevier, 99, pp. 157-167.

James, G., D., Witten, T., Hastie and R., Tibshirani, 2013, *An Introduction to Statistical Learning with Applications in R*, Springer, New York.

Johnson, E. J., W. W., Moe, P. S., Fader, S., Bellman and G. L., Lohse, 2002, "On the Depth and Dynamics of Online Search Behavior", pp. 1-30.

Karimi, S., K. N., Papamichail and C. P., Holland, 2015, "The effect of prior knowledge and decision-making style on the online purchase decision-making process: A typology of consumer shopping behavior", *Decision Support Systems*, Elsevier, 77, pp.137–147.

Kurmiawan, S., 2000, "Modeling Online Retailer Customer Preference and Stickiness: A Mediated Structural Equation Model", *AIS Electronic Library*, PACIS 2000 Proceedings, pp. 238-252.

Lariviere, B. and D., Van den Poel, 2005, "Predicting Customer Retention and Profitability by Using Random Forests and Regression Forests Techniques", *Expert Systems with Applications,* 29, pp. 472-484.

Lee, S., S., Lee and Y., Park, 2007, "A Prediction Model for Success of Services in E-commerce Using Decision Tree: E-customer's Attitude towards Online Service", *Expert Systems with Applications,* 33, pp. 572-581.

Levin, A. M., I. P., Levin and J. A., Weller, 2005, "A Multi-Attribute Analysis of Preferences for Online-Offline Shopping: Differences across Products, Consumers, and Shopping Stages", *Journal of Electronic Commerce Research*, 6 (4), pp. 281-290.

Lim, Y.J., A., Osman, S. N., Salahuddin, A. R., Romle and S., Abdullah, 2016, "Factors Influencing Online Shopping Behavior: The Mediating Role of Purchase Intention", *Procedia Economics and Finance*, 35, p. 402.

Moe, W. W., 2003, "Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream", *Journal of Consumer Psychology*, 13 (1-2), pp. 29-39.

Moe, W. W. and P. S., Fader, 2004, "Dynamic Conversion Behavior at E-Commerce Sites", *Management Science*, 50 (3), pp. 326-335.

Moe, W. W., 2006, "An Empirical Two-Stage Choice Model with Varying Decision Rules Applied to Internet Clickstream Data", *Journal of Marketing Research*, 43 (4), pp. 680-692.

Nielsen, 2014, *E-Commerce: Evolution or Revolution in the Fast-Moving Consumer Goods World?*, http://www.nielsen.com/us/en/insights/reports/2014/e-commerce-evolution-or-revolution-in-the-fast-moving-consumer-goods-world.html, accessed at June 2016.

Nielsen, 2016, *Global Connected Commerce*, http://www.nielsen.com/us/en/insights/reports/2016/global-connected-commerce.html, accessed at June 2016.

Olbrich, R. and C., Holsing, 2011, "Modeling Consumer Purchasing Behavior in Social Shopping Communities with Clickstream Data", *International Journal of Electronic Commerce*, 16 (2), pp. 15-40.

Parise, S., B., Iyer and D., Vesset, 2012, *Four Strategies to Capture and Create Value from Big Data,* Ivey Business Journal, http://iveybusinessjournal.com/publication/four-strategies-to-capture-and-create-value-from-big-data/, accessed at September 2016.

Punj, G., 2012, "Income effects on relative importance of two online purchase goals: Saving time versus saving money?", *Journal of Business Research,* Elsevier, 65, pp. 634–640.

Rewatkar, S., 2014, "Factors Influencing Consumer Buying Behavior: A Review (with reference to Online Shopping)", *International Journal of Science and Research,* 3 (5), pp. 1347-1349.

Rodgers, S. and M. A., Harris, 2003, "Gender and E-Commerce: An Exploratory Study", *Journal of Advertising Research,* 43 (3), pp. 322-330.

Sato, S. and Y., Asahi, 2012, "The Model of Purchasing and Visiting Behavior of Customers in an E-commerce Site for Consumers", DOI: 10.7763/IPEDR.2012.V52.15, pp.72-76.

Senecal, S., P. J., Kalczynski and J., Nantel, 2005, "Consumers' Decision-making Process and Their Online Shopping Behavior: a Clickstream Analysis", *Journal of Business Search,* Science Direct, 58, pp. 1599-1608.

Shankar, V., A. K., Smith and A., Rangaswamy, 2003, "Customer satisfaction and loyalty in online and offline environments", *International Journal of Research in Marketing,* Elsevier, 20, pp. 153–175.

Statista, 2017, *E-commerce Market Report 2017,* https://www.statista.com/outlook/243/100/ecommerce/worldwide, accessed at December 2017.

Sultan, M. U. and M. D. N., Uddin, 2011, *Customers' Attitude towards Online Shopping*, Master of Business Administration, Gotland University.

Swinyard, W. R. and S. M., Smith, 2003, "Why People (Don't) Shop Online: A Lifestyle Study of the Internet Consumer", *Psychology & Marketing,* Wiley Periodicals, 20 (7), pp. 567–597.

TechTarget, 2016, *E-commerce*, http://searchcio.techtarget.com/definition/e-commerce, accessed at October 2016.

Toronto University, n.d., Stepwise Logistic Regression, http://www.utstat.toronto.edu/~brunner/oldclass/appliedf11/handouts/2101f11StepwiseLogisticR.pdf, accessed at October 2016.

Tulsyan, S., 2017, *How does Clickstream Analytics help improve an e-commerce portal's performance?*, http://blog.algoscale.com/clickstream-analytics-help-improve-e-commerce-portals-performance/, accessed at December 2017.

Turan, T., 2011, *Factors Affecting Online Shopping Behavior of Turkish Consumers*, Master of Arts, Bogazici University.

Van Wezel, M., and R. Potharst, 2007, "Improved customer choice predictions using ensemble methods", *European Journal of Operational Research*, Science Direct, 181, pp. 436-452.

Wang, N., D., Liu and J., Cheng, 2008, "Study on the Influencing Factors of Online Shopping", *Proceedings of the 11th Joint Conference on Information Sciences*, Atlantis Press.

Yeh, L., E. M. Y., Wang and S. L., Huang, 2007, A Study of Emotional and Rational Purchasing Behavior for Online Shopping, In: D. Schuler, *OCSC'07 Proceedings of the 2nd international conference on Online communities and social computing*, Heidelberg, Springer, Berlin, pp. 222-227.

Zhou, L., L., Dai and D., Zhang, 2007, "Online Shopping Acceptance Model - A Critical Survey of Consumer Factors in Online Shopping", *Journal of Electronic Commerce Research*, 8 (1), pp. 41-62.